Author's response to revisions of acp-2019-1214 (Gliss et al. AeroCom study)

Dear Nikolaos Mihalopoulos,

We would like to thank the reviewers for their constructive and helpful comments which we address below. We would also like to thank you for granting the required extensions, which enabled us to address all reviewer comments in great detail.

We followed most of the reviewers' suggestions, which resulted in major revisions and restructuring of the document. As a result of the revisions, we slightly modified the title of the manuscript. We changed the title from:

Multi-model evaluation of aerosol optical properties in the AeroCom phase III Control experiment, using ground and space-based columnar observations from AERONET, MODIS, AATSR and a merged satellite product as well as surface in situ observations from GAW sites

To:

AeroCom phase III multi-model evaluation of the aerosol lifecycle and optical properties using ground and space based remote sensing as well as surface in situ observations

We are convinced that we were able to address all comments and we believe that the revised manuscript is now adequate for publication in ACP.

Below we answer the comments of both reviewers. Before that, we summarise major changes that have been made in the revised manuscript.

Please also note that we did not submit a *diff* file as requested by ACP, as we believe that such a file would not help in this case, given the considerable changes applied with most paragraphs rewritten. We hope that you and the reviewers see this in the same way. Please let us know if you want us to provide such a diff file regardless.

Kind regards, The authors

Changes in data used

Before addressing the comments from the reviewers below, here we first summarize all changes and updates that have been applied to the model and observational data used in the paper (including bug fixes in the analysis that affect the results).

1. Changes in observation data used

- EBAS dataset was updated (this affects the GAW in situ results, however, differences between both versions were investigated and had minor impact on our results).
- AERONET was updated (also with minor impacts on results).
- MODIS was updated now using combined DT and DB products. This was a recommendation from R. Levy through private communication. It resulted only in minor changes when evaluating the models with the updated MODIS satellite dataset, with slight improvements of NMB of ca. +1% and R of ca. 0.03 - 0.05 (when colocating in 5x5 degree resolution, based on monthly statistics).
- All satellite AOD products: due to increased uncertainties in satellite products, only measurements showing AOD>0.02 were considered. Changes in global NMB results (models vs. satellites) are minor (less than 1% difference in NMB).
- AATSR4.3-SU: due to larger uncertainties in satellite retrievals of AE, only AE values were used where 0.05 < AOD < 1.5. This decision is based on recommendations by P. North through private communication.

2. Changes in model data used

- ECHAM-HAM:

- Deposition rate diagnostics were corrected (with impacts on lifetime estimates).
- Resubmitted speciated optical depths at ambient conditions, as required by AeroCom (before, species optical depths represented dry aerosol).
- **ECHAM-SALSA:** resubmitted speciated ODs at ambient conditions, as required by AeroCom (before, species optical depths represented dry aerosol).
- ECMWF-IFS: was updated to more recent version 46r1 (before 45r1 was used), now with a complete set of required variables (fine / coarse AOD, surface dry scat. and abs. coeffs.). Differences can be seen in the online web visualization of the results: https://aerocom-evaluation.met.no/overall.php?project=aerocom&exp=PIII-optics201_9-P
- **SPRINTARS:** was resubmitted due to an error in the calculation of the Angstrom exponent (AE). This mostly affected results in AE (large improvement) and other investigated variables to a minor degree.

- GISS-OMA: in the initial submission version, the fSST (i.e., no nudging applied) run was used as a temporary solution since CTRL was erroneous before submission in Dec 2019. The latter was corrected and resubmitted and is now used instead of fSST. Changes in bias between both versions are around 5-15% but signs of biases are mostly preserved (only surf. scattering seems to be shifted towards more positive bias). The results for both versions are also available online (see link above).
- **CAM5-ATRAS:** was updated since biomass burning emissions were incorrect (they were counted twice).
- **OsloCTM:** Updated dry diagnostics for EBAS evaluation, with improvement in dry absorption and dry scattering bias.
- GFDL-AM4: Updated results of coarse AOD. Accidentally, in the previous version of the manuscript, coarse AOD was computed as the sum of seasalt and dust optical depth, rather than the difference between total AOD and fine mode AOD (this mistake was due to an old setting from a preliminary analysis configuration file). It has been corrected in the revised version of the manuscript. Normalized mean bias vs. AERONET SDA changed from -24% (when using AOD>1um=SS+DUST) to -65% (when using AOD>1um=AOD-AOD<1um).
- AeroCom ensemble MEAN and MEDIAN fields: were updated accordingly and now include all 14 models (note that INCA is missing fine and coarse AOD and surface dry scat. and abs. data, so this is not included in the corresponding MEAN and MEDIAN fields).

Relevant changes in the analysis code

A bug was fixed in the analysis software pyaerocom, which was related to merging of overlapping time-series data. Overlapping time-series data can happen in the EBAS data files used, due to possible resubmissions of the (same) data from the data providers to the EBAS database. This bug only affects the results from the surface in situ comparisons of dry scattering and absorption data at GAW sites, accessed through the EBAS database. Changes in model NMBs and bias at GAW sites changes by ca 0% - +2% for scattering and +3% - +8% for absorption.

All these updates were incorporated and all affected results and figures/tables were recomputed. We remark that none of these changes had major impacts on the results, with respect to our discussion and interpretation.

General remarks from the authors based on comments from both reviewers

We would like to thank the reviewers for their thoughtful comments and very good suggestions. Both reviewers found major flaws in the structure of the paper and in the way the results were discussed and interpreted. Both reviewers thus, suggested major improvements related to the structure and interpretation of the results. We agree with these suggestions and as a result, we have reorganized the structure of the paper substantially and also the discussion and interpretation of our results. We are convinced that this will help the reader to understand our results and their implications in a wider context of the associated literature.

The applied changes directly or indirectly affect / clarify most of the individual comments from the reviewers below. Therefore, we summarise the major updates here and will refer to them when needed in our individual answers to the reviewers below. We will mark our responses and relevant changes in red color.

Major changes in the manuscript between initial submission (ACPD version) and the revised manuscript:

- **Appendices:** We removed the Appendix sections completely and incorporated the content either in the main manuscript, or in the new supplemental material 2. We also removed unnecessary/unused Figures from the Appendix (i.e., scatter plots from sensitivity studies, site bias heatmaps for each model and GAW site for surface scattering and absorption data). Interpretation of model biases at individual sites is certainly of interest but beyond the scope of this overview paper focussing on the global modelling of the aerosol.
- In this context, Table A1 (which contained some more details about models) was removed and relevant information was included in the updated model overview table (Table 2 in the revised manuscript).
- Section 2: was renamed from "Methods" to "Data and Methods". Most important changes in subsections:
 - **Observations used (Sect. 2.1):** individual paragraphs of AERONET, GAW in situ data and used satellite datasets have been revised and improved. An introduction paragraph for the used satellite data was added.
 - Model introduction (Sect. 2.2): We improved the introduction of the individual models and relevant information is now summarised in the revised model overview table 2 in the manuscript. Section 2 now also includes a short introduction of the content of the supplementary material 1 (questionnaire about model assumptions related to optical properties

and aerosol lifecycle). Detailed introductions of individual models have been separated from the discussion of their results (i.e., they were formerly included in the discussion Section 4) and have been moved into the new supplement 2. Details related to the computation of ensemble MEAN and MEDIAN fields have been moved into the new Sect. 2.3.1.

- Sect. 2.3 "Data processing and statistics" (formerly "Data analysis"): As mentioned above, this section now includes the details related to the ensemble model (Sect. 2.3.1). Also the discussion of the applied STP correction for the in situ data was moved from the introduction section of these data into a new subsection 2.3.2 "Model STP correction for comparison with GAW in situ data".
- Results from spatiotemporal representativity studies (formerly Sect.
 2.4): the results from several sensitivity studies were before included in Sect. 2.4 and were misplaced there. They were thus moved to the end of the (revised) Section 4 which discusses results from the model comparison with observations (details related to changes in Sect. 4 are provided below).
- Results from satellite evaluation (formerly Sect. 2.5): We removed the discussion of results from the satellite vs. AERONET intercomparison in (e.g. former Sect. 2.5) into the new supplement 2. This was done because this discussion distracted from the main purpose of the paper, which is model evaluation. Relevant satellite biases compared to AERONET are mentioned in the introduction of the satellite datasets used (in Section 2.1.3f) and are included in Table 1. Associated uncertainties of satellite retrievals that may impact the interpretation of the results from the model assessments are discussed where needed, in the new result and discussion section 4.
- Presentation and discussion of results (Sect. 3&4): The presentation and discussion of the results was presented before in "Sect. 3 Results" and "Sect. 4 Discussion of results from individual models". As indicated by both reviewers this organisation of the content was not optimal. Thus, large efforts have been made to improve this, resulting in major restructuring of these sections. In the revised manuscript, results and discussion are organised in 2 sections:
 - Sect. 3 Results Model diversity: this section focuses on presentation and discussion of inter-model diversity related to aerosol lifecycle and optical properties, on a global scale (this section does not include any comparison with observations). This includes much of the content of the previous Sect. 3 (i.e. modelled diversities in emissions, lifetimes, burdens, AODs, MECs) but in addition now also includes a comparison with results

from AeroCom phase I (e.g., Textor et al., 2006, Kinne et al., 2006).

- Sect. 4 Results Optical properties evaluation: this section presents and discusses the results from the optical properties evaluation. As suggested by reviewer 1, the discussion of the results is now categorised by the individual optical parameters that were evaluated, rather than by the individual models (as was done before).
- Figures and tables: as a result of the above discussed updates and in order to improve the interpretability of our results, most figures and tables have been updated. As a result, we drastically reduced the number of Figures in the main manuscript from before 12 to 7 in the revised manuscript. Here we list the most important changes:
 - We removed results from the AeroCom ensemble MEAN from all affected figures (former Figs. 3-7 and 10-12), since we focus our discussion on the MEDIAN results in the revised manuscript.
 - Removed former Figure 2. The figure showed satellite biases and correlations compared to AERONET and was related to the former Sect.
 2.5 which has been moved in the supplement (see also comment above).
 - Figures related to aerosol lifecycle and optical properties: Former 0 Figures 3-7 (blue colored heatmaps of speciated emissions, lifetimes, burdens, MECs and ODs) have been merged into one table (Table 3 in the revised manuscript). Also, MAC of BC, OA and DU was added to that table for the models that provided the required diagnostics. Emissions of DMS and SO2 were removed and, instead, the total source strength of SO4 and OA is provided now which is derived based on the reported total deposition rates. The color coding was changed to illustrate the deviation from the ensemble median. In addition, results from associated AeroCom phase I (AP1) studies (Kinne et al., Textor et al., 2006) were added and are discussed in the new Section 3. A better comparison with older AeroCom studies was requested by both reviewers and we agree that this adds substantially to the quality of the paper by putting the new results into historical context. The outcomes of this intercomparison are discussed accordingly in the new Sect. 3.
 - New Figure 2: In that context, to better investigate intro-model diversity, a new Figure was created that connects diversities (and deviations from the AP3 ensemble median) in emissions, lifetimes, burdens, MECs and resulting optical depths for each species.
 - **New Fig. 3:** Another new Figure illustrates the simulated total AODs for each model as a stacked bar-chart of its component optical depths. It

further includes the median composition from Kinne et al., 2006 and global estimates from AERONET and the merged satellite AOD product, as well as information related to which models provided clear-sky or all-sky optics. This Figure helps the reader to visualize the main changes in AOD composition since AP1.

- Figure 4 global maps of annual averages and diversities from the ensemble model (formerly Fig. 8): We applied a discrete colour mapping (before a continuous mapping was used) as suggested by reviewer 2. We also removed displaying of the corresponding observation sites (and the associate info in the legends) as these are provided in Fig. 1 and were out of context in this figure which focuses on the intra-model diversity and not on the comparison with observations.
- Figure 5 bias maps of ensemble model compared to some observations used (formerly Fig. 9): this Figure now shows 8 maps of biases of ENS-MED compared to AOD (AERONET, merged satellite), AODf and AODc (AERONET), AE (AERONET, AATSR-SU) and in situ scattering and absorption. Before only AOD and surf. scattering was shown in a single map. This way it is easier to interpret and separate the results.
- New Table 4 Model biases and correlation coefficients compared to the various observations used (formerly Figs.10&11): the 2 figures were merged into one table and also here a discrete color mapping is applied now for NMB and Pearson R. It also includes results from ensemble fields of the 25th (Q1) and 75th (Q3) percentile, in order to illustrate the intra-model variability in biases.
- Figure 6 summary of results from comparison with observations (formerly Fig. 12): added AOD results from merged satellite product (before only AERONET was shown). Also, the color coding was changed: before the colours indicated the correlation coefficients, now they indicate the individual models. Also, the ensemble median results and corresponding interquartile range are now plotted as boxplots.
- New Figure 7: this new figure shows model biases in the Angstrom Exponent (AE) compared to AERONET in different AE bins (e.g. 0-0.5, 0.5-1, ..., 2.5-3). This is done to visualize differences in model bias between fine and coarse-mode dominated aerosol measurements.
- Table 5 results from spatiotemporal representativity studies (formerly Table 3): The table shows differences in bias and correlation retrieved for the individual tests performed (e.g., results from low vs high resolution experiments). Formerly, only relative differences in biases and

correlations were visualised, now the actual biases and retrieved correlations are reported for each test case and resolution. In addition, for the temporal representativity tests, differences in bias and correlation are presented for model / obs comparison that (1) use monthly averages computed with 25% coverage constraint for temporal resampling (e.g., at least 7 daily values per month) as done in the paper, compared to (2) results that do not require any coverage constraint (i.e., a single daily value is enough to retrieve a monthly "mean"). These results are discussed in the representativity section which was moved from Sect. 2.4 into Sect. 4.6 in the revised manuscript.

- Former table A1: Was removed as discussed above.
- Former table A2 Sensitivity of model / satellite comparisons to the choice of resolution: This table was recomputed and was moved from the Appendix into Sect. 3 in the new supplement 2. The associated text was revised. Also, the horizontal resolution of the individual models was added.
- To summarise, this is the new order of tables and figures in the revised manuscript:
 - Table 1: Observations and variables.
 - Table 2: Model overview.
 - Table 3 (coloured): Per-species emissions, lifetimes, burdens, ODs, MECs and MACs for each model as well as AP3 and AP1 ensemble median and diversities (formerly Figs. 3-7).
 - Table 4 (coloured): Results from model evaluations (biases and correlation coefficient) compared to the various observation records (formerly Figs. 10 & 11).
 - Table 5: results from spatiotemporal sensitivity analysis.
 - Figure 1: Maps of annual averages from selected ground and space based observations.
 - Figure 2: Diversity in aerosol lifecycle parameters.
 - Figure 3: AOD composition (stacked bar chart)
 - Figure 4: Maps of annual averages from ensemble median model and diversities.
 - Figure 5: Maps of model biases compared to selected observations (same observations as used in Fig. 1).
 - Figure 6: Summary of model biases compared to selected observations (box plot, models in different colors).
 - Figure 7: Model biases in AE as a function of the considered AE range.

We believe that the new structure and the updated / new visualization will help the reader to better understand and link the results.

As mentioned above, please note that, as a result of the changes, we slightly changed the title of the paper from:

Multi-model evaluation of aerosol optical properties in the AeroCom phase III Control experiment, using ground and space based columnar observations from AERONET, MODIS, AATSR and a merged satellite product as well as surface in situ observations from GAW sites

To:

AeroCom phase III multi-model evaluation of the aerosol lifecycle and optical properties using ground and space based remote sensing as well as surface in situ observations

Below we answer the individual comments by both reviewers. Where appropriate, we refer to the substantial changes that were summarised in this section. We note that in most cases it is not possible to provide explicitly the applied "**Changes to the manuscript**" due to the major changes summarised here. We hope that the reviewers and the editor understand that.

Author's responses to the comments from 2 reviewers

Comments from Reviewer #1

Received and published: 15 April 2020 Review report of acp-2019-2014 manuscript

The submitted work deals with the intercomparison of models contributing to the AEROCOM initiative as well as to the evaluation of key simulated aerosol optical properties against corresponding measurements provided by ground-based networks and satellite sensors. It is clear that the topic fits well to the scientific purposes of ACP.

Nevertheless, after reading carefully the text I have the feeling that it looks more as a technical report rather than a scientific paper. My concern is that there is a "disparity" between the number of figures (including also Appendices) and the discussion (interpretation) of the outcomes. Moreover, it is needed a reconstruction of the structure in order to facilitate the reader to understand the tools, the methods and the findings.

We have emphasized the scientific dimension of the paper by applying the changes summarised in the general remark above. We are convinced that the updated / new visualisations will help the reader to better understand and link the results in a deeper scientific context.

Summarizing, the submitted paper can be published after taken into account the comments listed below.

1. The abstract is too long providing a lot of numbers. It is better to reduce it, highlighting the major findings of your work without stating in detail the metrics obtained from the evaluation/intercomparison analysis.

The abstract was revised as a result of the better presentation and interpretation of our results. It has also been shortened.

2. Section 2: I cannot understand why you have to discuss your results here. It is more straightforward to move them in a subsection of the relevant part of the manuscript (i.e., Results). Also, consider renaming Methods to Observations and models (or Data).

We agree with the reviewer. This issue has been resolved in the revised version of the manuscript, as a result of the major changes summarised above.

3. Section 2.2: Introduce here all the models used in your analysis.

This issue has been resolved in the revised version of the manuscript, as a result of the major changes summarised above.

4. Section 2.2 must be improved. Please consider rewriting both paragraphs.

We agree that the former model introduction section was not sufficient. As discussed above, the section was revised and information relevant to the paper was added to the model overview table. Furthermore, references and short summaries to additional in-depth information about the models are provided in the new Sect. 2.2.

5. I would suggest changing the title in Section 2.4. Please move this part to Results. Also, in this section (as well as in many parts of text) the interpretation is poor containing just statements from the metrics.

As stated above, the former Section 2.4 (about representativity) is now section 4.6 in the results, however, we kept the title of this section "Representativity of the results" as it describes well what this section is about.

We improved the description, presentation and interpretation of the results from these studies. In addition, as mentioned above, considerable emphasis was placed on improving the discussion and interpretation of our results throughout the paper. In this context, we refer particularly to the revised discussion sections 3 & 4 and the revised conclusion section.

6. Why is useful for your analysis the evaluation of the satellite products since their reliability has been assessed in depth in previous relevant studies? As mentioned above, the discussion of results from the satellite evaluation vs AERONET was moved to supplement 2 and where available, references were added. However, we note that in the case of AATSR SU v4.3 dataset there is not really any published literature available, so we found it important to document at least relevant relative biases between satellites and AERONET. We believe that having these metrics (bias and correlation) vs AERONET available in the paper is helpful for the reader when assessing the results from the model evaluation.

7. In general, it is missing the inter-comparison (connection) of your results with those reported in Kinne et al. (2006).

We agree with the reviewer. This issue has been resolved in the revised version of the manuscript, as a result of the major changes summarised above. Comparisons with results from the AeroCom Phase I studies by Kinne et al., 2006 and Textor et al., 2006 are discussed mostly in the new Sect. 3 and also in Sect. 4.

8. Line 132: Provide wavelengths for AERONET AOD and Angstrom.

Comparison wavelengths for each observation dataset and variable are now provided in a dedicated column in Table 1 in the revised version of the manuscript. Discussion of the measurement wavelengths has been expanded in the text.

9. Line 173: Why the AAE is universally constant and not aerosol-type dependent?

We are aware that AAE is certainly not universally constant and is aerosol type dependent. Unfortunately, in the measurements we don't necessarily have spectral absorption information or time varying information about aerosol type. We chose a constant value of 1 for AAE as we believe it is a justified assumption and significantly increases the number of sites which we can use here. We have performed an analysis and determined that the error for the adjustment assuming a constant AAE tends to be relatively small. We understand that this was not discussed well enough and added the following text in Sect. 2.1.2 in the revised manuscript:

Changes to the manuscript:

"For the in situ AC data used in this study, most of the measurements are performed at wavelengths other than 550 nm (see sect. 1 in supplement 2). These were converted to 550 nm assuming an absorption Ångström exponent (AAE) of 1 (i.e., a $1/\lambda$ dependence, e.g., Bond and Bergstrom, 2006). This is a fairly typical assumption when the spectral absorption is not measured. For about 50 % of the sites, absorption was measured at ~530 nm meaning that even if the true AAE had a value of 2, the wavelength-adjusted AC value would only be underestimated by ca 4%. For another 25% of the sites, absorption was measured at ~670 nm. For these sites the impact of an incorrect AAE value is larger (ca 26% overestimation for an actual AAE of 2 and ca 6% for AAE=1.25). The remaining 25% of sites typically utilized wavelengths between these two values. Schmeisser et al. (2017) suggest that, across a spatially and environmentally diverse set of sites measuring spectral in situ absorption (many included here), that the AAE is typically between 1 and 1.5." 10. Lines 177-179: Please be more specific on how the corrections of PAMB and TAMB are applied.

We have now included the assumed standard temperature and pressure values and tried to clarify the text about this adjustment. We moved the discussion into a new subsection 2.3.2 in the revised manuscript.

11.Lines 184-189: Provide a short description and interpretation of the obtained findings, both for scattering and absorption coefficients.

This refers to the bias heatmaps for each EBAS site shown in the appendix. Even though they are interesting themselves, we decided to remove these figures as the investigation of results at individual sites is beyond the scope of this comprehensive global study. (One of the co-authors is working on this in a separate paper). The figures and lines 184-189 have now been removed.

12. Please put more effort on explaining the results for the absorption coefficient.

We agree with the reviewer that the absorption related discussions could see some improvements. This issue has been resolved in the revised version of the manuscript by (1) discussing and connecting better BC lifecycle related parameters and MACs in Sect. 3 and (2) by largely extending the discussion of the results from surface absorption coefficient comparisons in Sect. 4.5 in the revised manuscript. The results are now also discussed in the context of previous studies.

13. Figure 2: Could you please explain how the discrimination of fine and coarse AOD has been done for MODIS? Why for MODIS-Terra there are results for fine/coarse AODs and not for MODIS-Aqua?

This is not relevant anymore for the revised version of the manuscript since Figure 2 is removed in the revised version and also because the fine / coarse data from MODIS was not further used in the paper (i.e. only total AOD is used from MODIS, cf. also answer to point 2 above).

14. Section 4: See my comment 3. Present the results based on the considered parameters instead of separately for each model.

We agree with the reviewer and followed this suggestion as summarised above.

15. Lines 560-561: Clarify better this sentence.

These lines were part of the former Section 4.4 (discussion of results from ECHAM-SALSA), which has been completely removed, as a result of the major restructuring of the paper (see prev. point and information above).

Comments from reviewer #2

Received and published: 16 April 2020

General remarks:The present manuscript presents the results of the annual evaluation (for the year 2010) of the aerosol optical properties of 14 global aerosol models participating in the AeroCom Phase III Control Experiment. The observational products used in this exercise include in-situ observations (AERONET and GAW) and satellite retrievals (AATR-SU, MODIS and MERGED-FMI). It is obvious the significant effort that the authors are doing for summarising all the results. **However, the large amount of results related to the assessment of the observations and the ensemble and individual model evaluation results makes it difficult to follow all the discussions. While the results of the study are interesting to be published, their presentation and discussion are not yet sufficient enough to be published at Atmospheric Chemistry and Physics in the current form.** Consider being publishing after addressing revisions which are explained below.

We agree with the reviewer. As discussed in the summary above ("General remark from the authors based on comments from both reviewers") the paper has seen major restructuring and improvement in its discussions. We believe that these changes bring more clarity and focus into the paper.

General comments: As a general comment, I would like to emphasise the effort of the authors for synthesizing all the information in this manuscript. **However, a large number of figures, tables and supplementary material can introduce some confusion to the reader.**

We agree that the presentation of the results (and its order) was confusing in the initial submission and we believe that these issues have been resolved in the revised manuscript, as a result of the major updates summarised above.

You refer Tables and Figures that at the same time refer other Tables and Figures from the Appendix (see, for example, Table 3).

This should be resolved due to the major restructuring of the paper (see comments above) and since there is no Appendix anymore. We are convinced that the content is now clearer.

Also, it is hard to can get clear conclusions of the comparison because of the **mixture of different models and variables.** I mean the use of some models for the analysis of the representativity (Section 2.4) are not considered in the AeroCom ensemble (see Table A1)

We agree with the reviewer and refer to the major changes discussed above. We believe that the presentation and discussion of our results are much easier to follow while providing much deeper scientific interpretation at the same time. Regarding the sensitivity analyses: we remark now clearly in Sect. 4.6 where models were used for representativity studies that are not used throughout the rest of the paper. We also discuss associated implications and assess uncertainties.

or some models that are providing AOD and others, AOD clear-sky for the comparison.

We substantially improved the information about which models are providing clear sky and all sky optics (see e.g., model info Table 2 in the paper or new Figure 3) and also highlighted models that only provided all sky in relevant figures. We also discuss differences in AOD between models that submitted clear sky optics and models that submitted all-sky optics.

As it is shown in Table A1, there are some models (as INCA and ECMWF-IFS) are not considered in the AeroCom ensemble. Is there any advantage to keeping some models outside the AeroCom ensemble?

We reconsidered the way the ensemble is calculated and now includes all models considered in the paper. In the initial submission, ECMWF-IFS and INCA had not submitted all diagnostics, i.e. they had not provided fine and coarse mode optical depth diagnostics or dry scattering and absorption coefficient data. Therefore they were also not included in the ensemble fields for AOD and AE, since we wanted consistency between the different variables in terms of which models contributed to the ensemble. This was changed and now all models are included in the ensemble. Please note in this context that the model version of ECMWF-IFS was updated and has now all diagnostics available. INCA is the only model that did not provide fine / coarse AOD and dry surface scattering and absorption and hence, the corresponding MEAN and MEDIAN fields do not include INCA. Differences are discussed in the new paragraph 2.1.3 that discusses the details of the ensemble composition and computation.

Changes to the manuscript:

The following sentence was added in Sect. 2.1.3 in the revised manuscript:

"[...]. Please also note that the ensemble total AOD includes results from INCA which are not included in AODf and AODc (seeTab. 2). This results in a slightly smaller total AOD in the

ensemble when inferred from AODf+AODc (which does not include INCA) compared to the computed AOD field (which includes INCA)."

Is there any reason for selecting the year 2010 for the comparison? Maybe, the observation's availability or the emissions considered in the modelling simulations? You should include some words in the manuscript.

The year 2010 for AeroCom Phase III CTRL was chosen by the AeroCom consortium for better comparability with older AeroCom studies (e.g. phase II) and also because many more measurements became available between 2000 (which was used in AP1) and 2010. We added the following text in Sect. 2.2 in the revised manuscript.

Changes to the manuscript:

"The year 2010 was chosen as a reference year by the AeroCom consortium and is used throughout many phase II and III experiments for inter-comparability of different experiments and model generations. The AeroCom phase I simulations (e.g., Dentener et al., 2006; Kinne et al., 2006; Schulz et al., 2006;Textor et al., 2006) used the year 2000 as a reference year. One of the main reasons to update the reference year from 2000 to 2010 was that many more observations became available between 2000 and 2010 and also to account for changes in the present day climate, for instance, due to changing emissions and composition (e.g., Klimont et al., 2013; Aas et al., 2019; Mortier et al., 2020a)."

It is value the effort that the authors include for the assessment and representativity of the different observational datasets used in the comparison. As it is indicated in Table 1, authors are considering different time-frequency for the various databases. I understand that all the observational datasets are converted to monthly averages for the comparison with the monthly averages of the model. Could you include information about the delivered output frequency of the model?

We do not think that it will add useful content for the reader to provide the originally delivered frequencies, as we consistently resample to monthly resolution. It may even cause confusion, as to which original frequencies were provided in the diagnostics before resampling to monthly. Thus, we have chosen to not provide this information in the paper. For interested users, the information is readily available in the AeroCom database and also in the "Information" tab in the online visualisation of the results, which is referred to many times in the paper

(https://aerocom-evaluation.met.no/infos.php?project=aerocom&exp=PIII-optics2019-P).

Changes to the manuscript:

We added the following sentence in the model description Sect. 2.2 in the revised manuscript:

"Details on the AeroCom phase III experiments can be found on the AeroCom wiki page (AeroCom wiki, 2020). The wiki also includes information on how to access the model data

from the different AeroCom phases and experiments, which is stored in the AeroCom database."

Moreover, you mention that you are computing the AeroCom mean and median at 2x3 and use the raw resolution of the rest of the models for the AERONET and GAW comparison. Meanwhile, you regrid all the models to 5x5 for the satellite comparison that also you compare with AERONET. Are there the results of AeroCom mean/median consistent in both comparisons?

The choice of 2x3 for the ensemble model is due to the fact that the lowest provided model version is 2x3. We have investigated the sensitivity of that choice by comparing with a median computed at 1x1 resolution (i.e., lower resolution models were interpolated to 1x1) and the differences were marginal. The results of this comparison are online available here:

https://aerocom-evaluation.met.no/overall.php?project=aerocom&exp=PIII-optics2019-e ns

Regarding the 5x5 choice for the model vs satellite comparisons: this was a compromise between spatial and temporal resolution, i.e. by spatially averaging 25 grid-points into one grid-point we increased the temporal sampling coverage such that the 25% temporal coverage constraint is met, which we require to resample from daily to monthly (which is rarely given in the original 1x1 and also not in 2x3). By regridding to 5x5 degrees instead, before temporal averaging, we found that sufficient temporal sampling coverage was given at most locations. We remark that differences in the results from the satellite / model evaluations were investigated already in the initial version of the paper and were summarised in Table A2 therein. This Table A2 was moved into the new supplement 2 (Table 3 therein) and implications are discussed in the corresponding Section 3 and are also discussed briefly in the main manuscript Sect. 4.5. We emphasize that the discussion of our results in Sect. 3 & 4 aims to address the increased uncertainties associated with the satellite data in a better way than it did in the initial submission.

Also, the observational aerosol products that you are considering are only available during the daytime, can you quantify this uncertainty? Figures A3-A5 should support this discussion because you are comparing the impact of considering 3hourly (FigureA3) and hourly (Figure A4) basis vs monthly, which is 14% and 8% respectively. However, it is difficult to understand the impact when you are comparing different models and different variables.

This is indeed a very important point and we extended the discussion in the revised sections 2.1.1 (AERONET introduction) and 4.5 (representativity of results) accordingly, following the suggestion to incorporate our findings from the high resolution tests:

Changes to the manuscript:

We added the following paragraph in Sect. 2.1.1 in the revised manuscript:

"The sun photometer measurements only occur during daylight and cloud free conditions. Thus, the level 2 daily averages used here represent daytime averages rather than 24h averages (as provided by the models). Because of the requirements for sunlight and no clouds, the diurnal coverage at each site shows a more or less pronounced seasonal cycle depending on the latitude (e.g., only mid-day measurements at high latitudes in winter) and the seasonal prevalence of clouds in some regions. This is a clear limitation when comparing with 24h monthly means output from the models (as done in this study). However, these representativity issues were found to have minor impact for the model assessment methods used in this study (details are discussed in Sect. 4.5)."

We added the following paragraph in Sect. 4.5 in the revised manuscript:

"One further uncertainty related to the representativity of the results is that AERONET only measures during the daytime, while the models computed 24h averages (as indicated in Sect. 2.1.1). This will cause shifts in the intrinsic weighting applied when computing the network averaged statistics used throughout this paper (e.g., wintertime measurements at high latitudes are restricted to noon-time if they occur at all). In addition, it could introduce systematic errors at locations that show a persistent and pronounced diurnal profile. In this context, note that the GAW in situ observations are not affected by this as they measure continuously, night and day regardless of cloud conditions. The latter is reflected in the very similar results in Test 1 (i.e., hourly vs monthly comparison of ACdry). Since the results of test 2 (AERONET 3hourly vs monthly) show very good agreement as well, we believe that uncertainties associated with diurnal variations of AOD are likely small compared to the large uncertainties associated with the correct modelling of the AOD, reflected by the considerable biases (and their diversity) found here among the models. Furthermore, AOD represents the whole atmospheric column and, thus, should be less sensitive to diurnal variations than the near surface measurements. A detailed investigation of associated impacts of diurnal variability is desirable but beyond the scope of this paper. Also in that context, it would be interesting to investigate the extent to which global climate models need to be able to reproduce amplitudes in diurnal variability of certain tracers and physical processes and which phenomena can be sufficiently parameterised in lower temporal resolution."

From the satellite comparison, you are considering for different datasets MODIS-Aqua, MODIS-Terra, AATSR-SU and MERGE-FMI, is there is a recommendation that you can provide in the manuscript about the most reliable for model evaluation purposes?

We believe that the merged dataset is the most complete one and in particular for long-term studies extending before 2000 it should be used. But it is not the only high-quality dataset, and also the individual datasets (which are self-consistent) are available. The analysis of the individual datasets in Sogacheva, et al., 2020 shows possible significant regional biases between them, but also shows similar regional temporal patterns. Knowing the availability and the performance of different products, which is discussed in Sogacheva, et al., 2020, a user can select a dataset covering the period under study for model evaluation or trend analysis. We intend to continue work on the merged dataset to consolidate and extend it further. However, as we do not go into depth regarding the assessment of individual satellite datasets we are hesitant to give a recommendation in the paper.

To help the reader, I would move the results of the individual models (Sect. 4) to a Supplement.

As described above, we have restructured our results substantially and focus now in Sect. 4 on the individual parameters rather than the individual models. Relevant findings from the individual models from the former Sect. 4 were included in the revised Sections 3 & 4.

I would keep the most important findings related to the multi-model comparison in the main discussion. Considering that you are including different models, it would be expected to find more discussion about possible improvements to have into consideration for the model community as aerosol emissions, size distribution, hygro-scopicity or aerosol optical properties used.

Connecting with the previous comment and discussion above, we have largely improved the discussion of individual models based on our findings. However, given the large diversity in our results (see new Sect. 3), it is difficult to give general recommendations in the paper and possible improvements are likely very model specific. However, by combining the results of the aerosol lifecycle and diversity analysis (which includes comparisons of major aspects, such as significantly decreased BC burden, large shift in relative contributions of dust and sea salt with sea salt dominating the natural AOD) with the results from the optical properties evaluation (e.g. many models likely simulate too fine particles for coarse dominated aerosol) we are able to provide indications of possible areas for model improvements. However in this global overview paper it is difficult (and not the purpose) to diagnose individual aspects in the detail necessary to make recommendations on the individual model level. Nonetheless, we provide clues and indications of major flaws of individual models and give recommendations about which aspects should be investigated in more detail.

In Section 5, it is where you introduce the results considering AOD clear-sky from model experiments. Is there any difference in the comparison with satellites between those models that delivered AOD or AOD clear-sky?

In general, it was recommended to provide clear-sky (CS) diagnostics over all-sky (AS) for AOD and other columnar remote sensing variables, since the measurements should represent clear-sky conditions (also the satellites). The new Figure 3 (stacked bar chart of speciated AOD) indicates that intra-model AOD variability is mostly linked with large differences in speciation and individual component ODs rather than the choice of CS vs AS and reasons for the diversity is rather to be searched in the modelled mass and mass-to-optics conversion than the treatment of CS vs AS optics. Thus, we kept the focus on the discussion of the related aspects determining the reported AODs for each

model rather than focusing on CS / AS aspects. Where appropriate, we remark to differences associated with CS / AS treatment in Sect. 3 & 4 (e.g. for SPRINTARS). However, a detailed investigation is beyond the scope of this paper.

Finally, please, revise the references to figures, tables and sections there is a mixture of formats.

Thank you for observing this, we have revised all references and tried to make reference to figures/tables/sections consistent.

Minor comments:

Page 2 Line 16: Capital letters in Aqua and Terra, i.e., MODIS-Aqua and MODIS-Terra. This has been resolved in the revised version of the manuscript.

Page 2 Line 21: Capital letters in Terra, i.e., MODIS-Terra. This has been resolved in the revised version of the manuscript.

Page 3 Lines 51-55: Add a reference as Boucher, O., Randall, D., Artaxo, P., Brether-ton, C., Feingold, G., Forster, P., ... & Rasch, P. (2013). Clouds and aerosols. In Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (pp.571-657). Cambridge University Press. The reference has been added accordingly.

Page 3 Lines 58-59: Add a reference. This refers to the following sentence:

"Both natural and anthropogenic emissions are highly uncertain due to lack of measurements and information or documentation flow."

The sentence has been removed in the revised introduction section.

Page 3 Lines 64-65: Add a reference related to the quantification of DMS.

This refers to the following sentence:

"Marine dimethyl-sulfide (DMS) and volcanic emissions are responsible for approximately a third of the global anthropogenic sulphur budget."

The sentence has been removed in the revised introduction section.

Page 4 Line 83: Introduce the GAW acronym. This has been resolved in the revised version of the manuscript. Page 4 Line 102: What are the advances between the set of models used in Kinne et al.(2006) and the ones considered in the present study? Is there any common/different feature between both studies? Is there any major improvement in the optical properties calculation from the modelling side?

Advances compared to Kinne et al., 2006 have been collected in an additional question G9 in the optics questionnaire which is included as supplement 1. Main changes since AeroCom Phase I (also in the optics calculations) are now included in the text in Sect. 3 & 4, mostly based on the reported literature values in associated AP1 papers.

Page 4 Line 103: Remove "aerosol optical depth", it is already introduced.

This has been resolved in the revised version of the manuscript.

Page 5 Line 121: Introduce MEC and OD.

MEC and OD are now introduced at the end of the revised introduction section. In addition the new section 2.2.1 "Model diagnostics" introduces them in the scientific context.

Page 5 Line 137: In Figure 1 in the AATS-SU AE map, it is observed fine aerosols (high AE values, > 1) in Antarctica, could you add a comment on it?

The aerosol retrievals show highest accuracy over ocean and darker surfaces, with higher uncertainty over bright desert surfaces, or for measurements at large solar zenith angles (e.g. over Antarctica). We have added a comment to this effect in the manuscript. Please see the revised paragraph in Sect. 2.1.5 "AATSR SU v4.3 data" for comments related to uncertainties.

Page 6 Section 2.1.1: Consider to mention that AERONET SDA products (AOD coarse and AOD fine) are provided at 500nm. Also, you should include the number of sites used for the comparison in Table 1 and do reference to the location in Figure 1, for example.

More information about AERONET wavelengths are now provided in Table 1 and in the AERONET introduction section 2.1.1 in the revised manuscript. Number of stations are listed for each network/measurement platform - labeled '#st.' as described in the table caption. We have added letters to each pane in figure 1 and the individual panes are referenced when a particular measurement/network is discussed.

Page 6 Line 163: Introduce STP.

We have introduced STP including the assumed standard temperature and pressure values and tried to clarify the text about this adjustment (see also answer to comments from reviewer 1 above).

Pages 6-7 Section 2.2.2: You should include the number of the final selection of sites used for the comparison shown in Table 1, instead of mentioning the ones excluded and do reference of the location in Figure 1, for example.

We believe the reviewer refers to section 2.1.2, which introduces the surface in situ data. This section has been revised and includes now also a statement on the total number of sites considered for absorption and scattering. Please also note Section 1 in supplement which provides tables containing detailed information about each GAW site used in the climatological time-series.

Changes to the manuscript:

In addition to some further changes in Section 2.1.2 we added the following sentence:

"After applying the RH constraint, removing urban sites from consideration, and resampling to monthly climatology, data from 39 sites with scattering data and from 39 sites with absorption data (not necessarily the same sites as for scattering) were available for model assessment (see Table 1)."

Page 8 Lines 216: What Level are you considering in the study? I suppose that it is Level 3, but it is better to mention again here.

Yes, we used level 3 data and we clarified this in the revised paragraph 2.1.5 (before 2.1.4), by not mentioning level 2 anymore and stating the following.

Changes to the manuscript:

In addition to some further changes in Section 2.1.5 we added the following sentence:

"This study uses the level 3 output, which is provided at daily and monthly 1x1 resolution, intended for climate model comparison."

Page 8 Section 2.1.5: It should be mentioned that MODIS and AATS products are considered inside this MERGED-FMI.

This has been clarified by adding the following sentence in Sect. 2.1.6 (before 2.1.5).

Changes to the manuscript:

"It should be noted that MODIS and AATSR products are considered inside this MERGED-FMI data-set."

Page 8 Line 234: Table is in capital letters.

This has been resolved in the revised version of the manuscript.

Page 9 Line 255: Indicates that this is Appendix C

This has been resolved in the revised version of the manuscript. The content from Appendix C was now included in the new Section 2.3 "Data processing and statistics" as part of the major restructuring summarised above (i.e., the revised manuscript does not have any appendices anymore).

Page 9 Line 265: Specify the reference to cf. 1.

This has been resolved in the revised version of the manuscript.

Page 10 Lines 296-298: In Section 2.4, for the representativity of the results you are combining different models and observational datasets for concluding that "the overall difference is of the order of 10% and 0.2 for NMB and correlation, respectively". This is supposed shown in Table 3, but the numbers are not coincident.

We believe that there was a misunderstanding, perhaps the reviewer used an older pdf version of the manuscript that had this wrong reference to Table 3 (which was corrected between initial submission and the ACPD version, which refers to the correct table A2). Nonetheless, we remark that Table A2 is now included in the supplement 2 and the discussion of these results in Section 4.5 in the revised manuscript has also been updated.

Page 11 Section 2.5: Because you mention the bias to Europe and North America because of the density of sites. Is there any regional results in the comparison satellite vs AERONET that you can consider to include?

As discussed above we have removed section 2.5 (comparison of satellites with AERONET) and put some of the key findings in the individual satellite sections. Since most of the analysis is based on global results and most of our key findings based on the comparison with the ground based observations, we did not include any additional results of the satellite assessment.

Page 12 Line 341: Add Figure 3.

We are not sure what the reviewer means with this comment, however, this should be resolved as a result of the major restructuring applied.

Page 13 Line 370: Missing ".". This was fixed (line 372 in original submission). Page 13 Line 379: Extra space.

This was fixed (line 382 in original submission).

Table 2: Add a reference to the supplementary material (the excel table).

We added a sentence in the caption of table 2: More details about the models can be found in the supplementary material 1&2.

Figure 2: Replace the continuous colour palettes for a new one with categories as in Figure 1. It is easier for the reader to associate each colour to the corresponding category.

Figure 2 has been removed in the revised manuscript (see comments above for details).

Figures 4-8: Add the corresponding legend associated with the colours. Another possibility is to keep the numbers and use the colour scale to indicate what models are above/under the AeroCom median/mean which is the reference.

This has been resolved in the revised version of the manuscript, see summary above related to updates in these Figures. The Figures were merged into the new Table 3 in the revised manuscript and the colour coding has been updated indicating deviations from the median.

Figures 10-12: Replace the continuous colour palettes for a new one with categories as in Figure 9. It is easier for the reader to associate each colour to the corresponding category.

This has been resolved in the revised version of the manuscript.

Figures A1 and A2: The sites in the heatmap are organised by alphabetical order, to can distinguish a pattern, maybe it would be better to create clusters per continent, latitude, longitude. Replace the continuous colour palettes for a new one with categories as in Figure 9. It is easier for the reader to associate each colour to the corresponding category.

As discussed above, these Figures were removed in the revised manuscript, as a discussion of results at individual sites is beyond the scope of this global study.

Multi-model AeroCom phase III multi-model evaluation of the aerosol lifecycle and optical properties in the AeroCom phase III Control experiment, using ground and space based columnar observations from AERONET, MODIS, AATSR and a merged satellite product remote sensing as well as surface in-situ in situ observationsfrom GAW sites

Jonas Gliß¹, Augustin Mortier¹, Michael Schulz¹, Elisabeth Andrews², Yves Balkanski³, Susanne E. Bauer^{20,19}, Anna M. K. Benedictow¹, Huisheng Bian^{4, 5}, Ramiro Checa-Garcia³, Mian Chin⁵, Paul Ginoux⁶, Jan J. Griesfeller¹, Andreas Heckel⁷, Zak Kipling⁹, Alf Kirkevåg¹, Harri Kokkola¹⁰, Paolo Laj¹¹, Philippe Le Sager¹², Marianne Tronstad Lund¹⁵, Cathrine Lund Myhre¹³, Hitoshi Matsui¹⁴, Gunnar Myhre¹⁵, David Neubauer¹⁶, Twan van Noije¹², Peter North⁷, Dirk J. L. Olivié¹, Samuel Rémy²¹, Larisa Sogacheva¹⁷, Toshihiko Takemura¹⁸, Kostas Tsigaridis^{19,20}, and Svetlana G. Tsyro¹

¹Norwegian Meteorological Institute, Oslo, Norway

²Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA

³Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Gif sur Yvette Cedex, France

⁴Maryland Univ. Baltimore County (UMBC), Baltimore, MD, USA

⁵NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

⁶NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

⁷Dept. of Geography, Swansea University, Swansea, UK

⁹European Centre for Medium-Range Weather Forecasts, Reading, UK

¹⁰Atmospheric Research Centre of Eastern Finland, Finnish Meteorological Institute, Kuopio, Finland

¹¹Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, Institute for Geosciences and Environmental Research (IGE), Grenoble, France

¹²Royal Netherlands Meteorological Institute, De Bilt, the Netherlands

¹³NILU -Norwegian Institute for Air Research, Kjeller, Norway

¹⁴Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

¹⁵CICERO Center for International Climate and Environmental Research, Oslo, Norway

¹⁶Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

¹⁷Finnish Meteorological institute, Climate Research Program, Helsinki, Finland

¹⁸Research Institute for Applied Mechanics, Kyushu University, 6-1 Kasuga-koen, Kasuga, Fukuoka, Japan

¹⁹Center for Climate Systems Research, Columbia University, New York, USA

²⁰NASA Goddard Institute for Space Studies, New York, USA

²¹HYGEOS, Lille, France

Correspondence: Jonas Gliß (jonasg@met.no)

Abstract.

Within the framework of the AeroCom (Aerosol Comparisons between Observations and Models) initiative, the present day state of the art modelling of aerosol optical properties has been assessed using simulated data representative for the year 2010,

is assessed from 14 global aerosol-models participating in the Phase III Control experiment . The model versions are close or

- 5 equal to those used for phase III control experiment (AP3). The models are similar to CMIP6and AerChemMIP and inform also on bias in state of the art ESMs. Modelled column optical depths (total, fine and coarse mode AOD) and Exponents (AE) were compared both with ground based observations from the Aerosol Robotic Network (AERONET, version 3)as well as space based observations from AATSR-SU instruments. In addition, the modelled AODs were compared with MODIS (Aqua and Terra) data and a satellite AOD data-set (MERGED-FMI) merged from 12 different individual AOD products. Furthermore,
- 10 for the first time, the modelled near surface scattering (under dry conditions) and absorption coefficients were evaluated against measurements made at low relative humidity at surface in-situ GAW sites.

Statistics are based mainly on normalised mean biases and Pearson correlation coefficients from colocated model and observation data in monthly resolution. Hence, the results are mostly representative for the regions covered by each of the observation networks. Model biases established against satellite data yield insights into remote continental areas and oceans,

- 15 where ground-based networks lack site coverage. The satellite data themselves are evaluated against AERONET observations, to test our aggregation and re-gridding routines, suggesting relative AOD biases of -5%, -6%, +9% and +18% for AATSR-SU, MERGED-FMI / AerChemMIP Earth System Models (ESMs) and provide a robust multi-model ensemble. Inter-model spread of aerosol species lifetimes and emissions appears to be similar to that of mass extinction coefficients (MECs), MODIS-aqua and MODIS-terra, respectively, with high correlations exceeding 0.8. Biases of fine and coarse AOD and AE in AATSR are
- 20 found to be +suggesting that aerosol optical depth (AOD) uncertainties are associated with a broad spectrum of parameterised aerosol processes.

Total AOD is approximately the same as in AeroCom phase I (AP1) simulations. However, we find a 50% decrease in the OD of black carbon (BC), attributable to a combination of decreased emissions, lifetimes and BC MEC. Relative contributions from sea salt (SS) and dust (DU) have shifted from approximately equal in AP1 to SS contributing about 2%, -16% and +14.7%

25 respectively, at AERONET sites, with correlations of the order of 0.8./3 of the natural aerosol optical depth (OD) in AP3. This shift is linked with a decrease in DU mass burden, a lower DU MEC, and a slight decrease in DU lifetime, suggesting coarser DU particle sizes in AP3 compared to AP1.

The AeroCom MEDIAN Relative to observations, the AP3 ensemble median and most of the participating models underestimate the all aerosol optical properties investigated, relative to remote sensing observations. AERONET AOD is underestimated

- 30 by that is, total AOD as well as fine and coarse AOD (AOD_f, AOD_c), Ångström exponent (AE), dry surface scattering (SC_{dry}) and absorption (AC_{dry}) coefficients. Compared to AERONET, the models underestimate total AOD by circa $21\% \pm 17\% 20\%$ (as inferred from the ensemble median and interquartile range). Against satellite data, the model ensemble AOD biases range from -38% (MODIS-terra) to -17-37% (MODIS-Terra) to -16% (MERGED-FMI). Correlation coefficients of model AODs with AERONET, MERGED-FMI and AATSR-SU are high (0.8 - 0.9) and slightly lower against the two MODIS data-sets (0.6,
- a multi-satellite AOD product), which we explain by differences between individual satellites and AERONET measurements themselves. Correlation coefficients (R) between model and observation AOD records are generally high (R > 0.75), suggesting that the models are capable of capturing spatio-temporal variations in AOD. We find a much larger underestimate in coarse AOD_c (~-45% -± 0.8). Investigation of fine and coarse AODs from the MEDIAN model reveals biases of -1025%) than in fine

 AOD_f (~-15% \pm 20%25%) with slightly increased inter-model spread compared to total AOD. These results indicate problems

40 in the modelling of DU and SS. The AOD_c bias is likely due to missing DU over continental land-masses (particularly over the US, SE-Asia and -41S-America), while marine AERONET sites and the AATSR SU satellite data suggest more moderate oceanic biases in AOD_c.

 $\frac{\text{Column AEs are underestimated by about 10\% \pm 29\% against AERONET and -13\% and -24\% against AATSR-SU, respectively.}{\text{The differences in bias against AERONET and AATSR-SU are in agreement with the established satellite bias against AERONET16\%}}$

- 45 For situations where measurements show AE > 2, models underestimate AERONET AE by circa 35%. In contrast, all models (but one) exhibit large overestimates in AE when coarse aerosol dominates (bias ca +140% if observed AE < 0.5). Simulated AE does not span the observed AE variability. These results indicate that most of the AOD bias is due to missing coarse AOD in the regions covered by these observations models overestimate particle size (or underestimate the fine mode fraction) for fine dominated aerosol and underestimate size (or overestimate the fine mode fraction) for coarse dominated aerosol. This
- 50 must have implications for lifetime, water uptake, scattering enhancement and the aerosol radiative effect, which we can not quantify at this moment.

Underestimates are also found when comparing the models against the surface GAW observations, showing AeroCom MEDIAN-Comparison against GAW in situ data results in mean bias and inter-model variation of -44 variations of -35% \pm 22% and -3225% and -20% \pm 34% for scattering and absorption coefficients 18% for SC_{drv} and AC_{drv}, respectively. Dry scattering

- shows higher underestimation than AOD at ambient relative humidity and is in agreement. The larger underestimate of SC_{drx} than AC_{dry} suggests the models will simulate an aerosol single scattering albedo that is too low. The larger underestimate of SC_{dry} than ambient air AOD is consistent with recent findings that suggest that models tend to models overestimate scattering enhancement due to hygroscopic growth. Broadly The broadly consistent negative bias in AOD and scattering suggests a underestimate of aerosol radiative effects in current global
- 60 aerosol models.

The large diversity in the surface absorption results suggests differences in the model treatment of light absorption by black earbon (BC), dust (DU) and to a minor degree, organic aerosol (OA). Considerable diversity is found among the models <u>Considerable inter-model diversity</u> in the simulated near surface absorption coefficients, particularly in regions associated with dust (e.g. Sahara, Tibet), biomass burning (e.g. Amazonia, Central Australia) and biogenic emissions (e.g. Amazonia). Regions associated with high anthropogenic BC emissions such as China and India exhibit comparatively good agreement for

all models.

65

Evaluation of modelled column AEs shows an underestimation of $9\% \pm 24\%$ against AERONET and -21% against AATSR-SU. This suggests that overall, models tend to overestimate particle size, with implications for lifetime and radiative transfer calculations. optical properties is often found in regions that are, unfortunately, not or only sparsely covered by ground

70 based observations. This includes, for instance, the Sahara desert, Amazonia, central Australia and the South Pacific. This highlights the need for a better site coverage in the observations, which would enable us to better assess the models, but also the performance of satellite products in these regions.

An investigation of modelled emissions, burdens and lifetimes, mass-extinction-coefficients (MECs) and optical depths (ODs) for each species and model reveals considerable diversity in most of these parameters. These are discussed in detail

75 for each model individually. Inter-model spread of acrosol species lifetime appears to be similar to that of mass extinction coefficients, suggesting that AOD uncertainties are still associated to a broad spectrum of parameterised aerosol processes. Using fine mode AOD as a proxy for present day aerosol forcing estimates, our results suggest that models underestimate aerosol forcing by circa -15%, however, with a considerably large interquartile range suggesting a spread between -35% and +10%.

80 1 Introduction

The global aerosol remains one of the largest uncertainties for the projection of future Earth's climate, in particular because of its impact on the radiation balance of the atmosphere (IPCC (2014))(IPCC, 2014). Aerosol particles interact with radiation through scattering and absorption, thus directly altering the atmosphere's radiation budget (aerosol-radiation interactions, or ARI). Moreover, they serve as cloud condensation nuclei (CCN) and can thus , among other things, influence further climate

- 85 relevant components such as clouds and their optical properties (e.g., cloud droplet number concentrations, cloud optical depth) and lifetime as well as cloud coverage and precipitation patterns (aerosol-cloud interactions, or ACI) (IPCC (2014)) (IPCC, 2014). Since 2002, the "Aerosol Comparisons between Observation and Models" (AeroCom) project has attempted to federate global aerosol modelling groups to provide state-of-the art multi-model evaluation and, thus, to provide updated understanding of aerosol forcing uncertainties and best estimates. Multi-model ensemble results have often been shown to be
- 90 more robust than individual model simulations, outperforming them when compared with observations. This paper attempts to provide a new reference, including multi-model ensemble median fields to inform further model development phases.

Aerosol optical properties such as the aerosol scattering and absorption coefficients, the aerosol optical depth (AOD) and the Ångström exponent (AE) are important components of aerosol direct forcing calculations, as they determine how aerosols interact with incoming and outgoing long and shortwave radiation. A special case is aerosol absorption, because it is capable

95 of changing the sign of aerosol forcing. Improved insight about aerosol optical properties, including their spatial and temporal distributions, would be very helpful to better constrain the aerosol radiation interactions. The evaluation of these parameters is thus the focus of this paper.

A challenging part of modelling the global aerosol is its comparatively high variability in space and time (e.g., Boucher et al., 2013), as compared to well-mixed greenhouse gases such as carbon dioxide and methane. The radiative impact aerosols exert de-

- 100 pends on the amount and the properties of the aerosol. Emissions, secondary formation of aerosol and lifetime combined lead to different amounts of aerosol in transport models. The lifetime of aerosol particles in the atmosphere is of the order of one week and is, to first order, dependent on their size. Particles in the accumulation mode (particle diameter between $0.3-1 \mu m$) show the longest residence times due to less effective atmospheric sink processes. The sources of aerosol are complicated since not all aerosol particles are directly emitted. Instead, particles can also be formed in the atmosphere (secondary aerosol) which is
- 105 dealt with in various degrees of complexity in models. In addition, atmospheric aerosol particles undergo continuous alteration

(e.g.Tsigaridis et al. (2014)). Both natural and anthropogenic emissions are highly uncertain due to lack of measurements and information or documentation flow, growth, mixing) due to micro-physical processes that occur on lengths and timescales that cannot be resolved by global models, such as nucleation, coagulation, gas-to-particle conversion or cloud processing.

Natural aerosols constitute a large part of the atmospheric aerosol, being composed of sulphur and organic components,

- 110 They are dominated by sea salt (SS) and dust (DU) which make up more than 80% of the total aerosol mass. Natural aerosol precursors include volcanic and biogenic sulphur (SO₄), volatile organic compounds (BVOCs) as well as sea salt and dust. Emissions of sea BC and OA from wildfires. Sea salt and dust emissions are strongly dependent on local meteorology and surface properties and, thus, require sophisticated parameterisations in global models with comparatively coarse resolution. In models, these emissions are usually computed based on simulated winds and constitute a major source of uncertainty
- (e.g. Carslaw et al. (2013)). Marine-These parameterisations are sensitive to simulated near-surface winds, soil properties (in case of dust) and model resolution (e.g., Guelle et al., 2001; Laurent et al., 2008). Major sources of natural SO₄ aerosol are marine emissions of dimethyl-sulfide (DMS) and volcanic emissions are responsible for approximately a third of the global anthropogenic sulphur budget. Both eruptive and passively degassing volcanic sulphur emissions are highly uncertain, with estimates ranging between 1 50 Tg (e.g. Andres and Kasgnoc (1998), Halmer et al. (2002), Textor et al. (2004), Carn et al. (2017)
 120).

In addition, atmospheric aerosol particles undergo continuous alteration (e.g. growth, mixing) due to micro-physical processes that occur on lengths and timescales that cannot be resolved by global models, such as nucleation or gas-to-particle conversion. SO₂ emissions (e.g., Seinfeld and Pandis, 2016). Uncertainties in natural aerosol emissions constitute a major source of uncertainty for estimates of the radiative impact of aerosols on the climate system (e.g., Carslaw et al., 2013), mainly because of non-linearities in the aerosol-cloud interactions and in the resultant cloud albedo effect (Twomey, 1977).

The chemical and physical properties of aerosol particles determine how they interact with radiation. They are highly dependent on the aerosol type and state of mixing. Aerosol optical properties such as the aerosol scattering and absorption coefficients, the aerosol optical depth (AOD) and the Ångström exponent (AE) are closely linked with aerosol forcing estimates as they determine how aerosols interact with incoming and outgoing long and shortwave radiation. A key parameter that

125

- 130 determines the efficiency of scattering and absorption of radiation is the complex refractive index (n + iκ), which depends on aerosol type (chemical composition) and mixing. It is accounted for in models in different ways (e.g. volume mixing, Maxwell Garnett, core-shell, e.g. Klingmüller et al. (2014)) Major absorbing species are black carbon, followed by dust and, to a certain degree, organic aerosols (e.g., Samset et al., 2018, and references therein). Also anthropogenic dust may exert forcing on the climate system (e.g., Sokolik and Toon, 1996). The absorptive properties of dust aerosol , for instance, are dependent.
- 135 dent on the mineralogy and size of the dust particles, resulting in some dust types being more absorptive than others (e.g., Lafon et al. (2006)), which absorbing than others (e.g., Lafon et al., 2006). This has direct implications for forcing estimates (e.g., Claquin et al. (1998)).

Scattering and absorption coefficients are derived from these extinction efficiencies and depend on particle size distribution and wavelength. In general, water uptake will enhance the light extinction efficiency. This is mostly relevant for scattering, since

140 absorptive aerosols such as dust and black carbon are generally considered to be hydrophobic (which can, to a minor degree be

violated in aged aerosol due to mixing, e.g. Cappa et al. (2012)). For instance, between 0% and 40% relative humidity (RH) (a range that is often considered "dry" for the purposes of GAW in-situ measurements), the light scattering can be enhanced (up to about 20% due to hygroscopic growth for some types of aerosol (e.g. Zieger et al. (2013)). This is important when comparing models with in-situ observations, since the latter are often performed at low humidity (RH<40%) but not at absolutely dry

145 conditions (GAW Report 227 (2016)). Some models tend to overestimate the scattering enhancement factor at low RH (and high RH) and hence, the amount of light scattering (?)(e.g., Claquin et al., 1998). Several measured parameters can be used to evaluate model simulations of aerosol optical properties.

AOD is the vertically integrated light extinction (absorption + scattering) due to an atmospheric column of aerosoland is a function of wavelength. AAOD. AAOD (the absorption aerosol optical depth) is the corresponding equivalent for the

absorptive power of an aerosol column and tends to be small compared relative to AOD (ca .-5-105-10% of AOD). Both AOD (mostly dominated by scattering) and AAOD (absorption) are of particular relevance for aerosol forcing assessments (e.g. Bond et al. (2013)). Major absorbing species are black carbon (BC), followed by dust (DU) and, to a certain degree, organic aerosols (OA) (e.g. Samset et al. (2018) and references therein)(e.g., Bond et al., 2013). Remote sensing of these parameters by sun photometers, for instance, within the Aerosol Robotic Network (AERONET Holben et al., 1998), or via satellite borne instruments have provided an enormous observational database to compare with model simulations.

Simulating the AOD (and AAOD) in a global model is a challenging task as it requires many prerequisites to be correct, not only the assumptions on optics (e.g. shape and refractive index, atmospheric radiative transfer), but also the emissions, transport, ageing, sources and sinks of all aerosol species , which determine the aerosol composition in space and time. Therefore, it is useful to also investigate other related optical parameters that can help to assess model performance. The AE, for

- 160 instance, describes the wavelength dependency of acrosol extinction and is The AE describes the wavelength dependence of the light extinction due to acrosol and can be measured via remote sensing using AOD estimates at different wavelengths. AE depends on the acrosol species (and state of mixing), due to differences in the refractive indices and size domains (e.g., Seinfeld and Pandis, 2016). It is a qualitative indicator of acrosol size since it is inversely related to the size of the acrosol size (i.e.larger particles exhibit less spectral dependence of scattering, resulting in smaller value of the AE
- 165). It can thus, provide a qualitative assessment of modelled particle size, smaller AE suggests larger particles). However, for mid-visible wavelengths (e.g. Schuster et al. (2006)). For instance, an underestimation of AE suggests an overestimation of the particle size. Like AE, around $0.5 \mu m$, as used in this paper), the spectral variability of light extinction flattens for particle sizes exceeding the incident wavelength. This can create considerable noise in the AE versus size relationship, especially for multi-modal aerosol size distributions, as discussed in detail by Schuster et al. (2006). Global AE values, which
- 170 combine data from regions dominated by different aerosol types, have potential to further complicate the interpretation of model simulated AE in comparison with observations. Nonetheless, the comparison of modelled AE with observations can still provide qualitative insights into the modelled size distributions.

Model and observational estimates of fine and coarse mode AOD can also give insights into the particle size domains, which can help establish provide another view onto the light extinction in both size regimes. This is because these parameters also

175 depend on the actual amount (mass) of aerosol available in each mode. The coarse mode is dominated by the natural aerosols

(sea salt and dust). Hence, individual assessment of extinction due to fine and coarse particle regimes can provide insights into differences between natural and anthropogenic aerosols(since the major natural constituents, dust and sea salt, dominate the coarse mode AOD). It should be noted that the split between fine and coarse mode is not straightforward in models (for example, some size bins may span the size cut) or for remote sensing instruments which rely on complex retrieval algorithms.

- 180 The comparison to surface in situ measurements of scattering and absorption coefficients offers a valuable performance check of the models, independent of remote sensing. One factor that impacts both remote sensing and in situ measurements is water uptake by hygroscopic aerosols. In general, water uptake will enhance the light extinction efficiency (e.g., Kiehl and Briegleb, 1993) . This is mostly relevant for scattering, since absorbing aerosols such as dust and black carbon typically become slightly hygroscopic as they age, due to mixing with soluble components (e.g., Cappa et al., 2012). Even at low relative humidity
- 185 (RH < 40%, a range that is often considered "dry" for the purposes of Global Atmosphere Watch (GAW) in situ measurements GAW Rep aerosol light scattering can be enhanced by up to 20% due to hygroscopic growth (e.g., Zieger et al., 2013). Recent work showed that some models tend to overestimate the scattering enhancement factor at low RH (and high RH) and hence, overestimate the light scattering coefficients at relatively dry conditions (Latimer and Martin, 2019; Burgos et al., 2020).
- Kinne et al. (2006) provided a first analysis of modelled column aerosol optical properties of 14 aerosol models <u>participating</u> in the initial AeroCom <u>phase 1 (AP1)</u> experiments. They found that, on a global scale, <u>aerosol optical depth (AOD) AOD values</u> from different models compared well to each other and generally well to global annual averages <u>involving trusted ground based</u> references from <u>AERONET</u> (model biases of the order of -20% to +10% to -2010%). However, they also found considerable diversity in the aerosol speciation among the models, mainly related to differences in transport and water uptake. The diversity for carbonaceous aerosol remained small as similar approaches were adopted in all models. They concluded that this diversity
- 195 in component contribution adds added (via differences to in aerosol size and absorption), to uncertainties for to uncertainties in associated aerosol direct radiative effects. Textor et al. (2006) used the same model data as Kinne et al. (2006) and focused on the diversities in the modelling of the global aerosol, by establishing differences between modelled parameters related to the aerosol lifecycle, such as emissions, lifetime and column mass burden of individual aerosol species. One important result from Textor et al. (2006) is that the model variability of global aerosol emissions is highest for dust and sea salt, which is
- attributed to the fact that these emissions were computed online in most models, while the agreement in the emissions of the other species (OA, SO_4, BC) were due to the usage of similar emission inventories. Since then, in the framework of AeroCom, several studies have investigated different details and aspects of the global aerosol modelling, focusing on individual aerosol species and forcing uncertainty. However, it became clear that a common base or control experiment was again needed to compare the current aerosol models contributing to assessments such as the Coupled Model Intercomparison Project Phase
- 205 6 (CMIP6, Eyring et al., 2016) or the upcoming report of the Intergovernmental Panel on Climate Change (IPCC), against updated measurements of aerosol optical properties and to assess aerosol life cycle differences. This study aims to provide this basic assessment and will also facilitate interpretation of other recent AeroCom phase III experiments.

This study thus investigates modelled aerosol optical properties of simulated by the most recent models participating in the AeroCom phase III 2019 control experiment (in the following denoted CTRL,) (AeroCom wiki, 2020, in the following denoted AP3-CTRI

210 on a global scale. Making It makes use of the increasing amount of observational data which have become available during the

past decade, we are able to two decades. We extend the assessment of modelled optical properties beyond what was originally presented in Kinne et al. (2006). Here, we use observations of ground and space based by Kinne et al. (2006) and use ground and space-based observations of the above introduced columnar variables of total, fine and coarse AOD and AE as well as and, for the first time, surface in-situ in situ measurements of scattering and absorption coefficients, primarily from surface observa-

215 tories contributing to Global Atmospheric Watch (GAW), obtained from the World Data Centre for Aerosols (GAW-WDCA) archive.

This paper is structured as follows. The next section <u>Section</u> 2 introduces the observations (OBS), variables (VAR) and models (MOD) observation platforms, parameters and models used, followed by a discussion of the analysis details for the model evaluation and a short section discussing the representativity of the results. The section ends with a brief discussion

- 220 of results from a validation study investigating the performance of the satellites used against ground based AERONET data. Section ?? starts with an overview of (e.g., statistical metrics, re-gridding and co-location). The results are split into two sections. Section 3 provides an inter-model overview of the diversity in globally averaged emissions, burdens, lifetimes, mass-extinction-coefficients (MECslifetimes, burdens as well as mass-extinction and mass-absorption coefficients (MECs, MACs) and optical depths (ODs) for each model and aerosol species,⁻¹. This is followed by a discussion of the results from
- 225 the AeroCom MEDIAN model and regional model diversity in the optical parameters considered. The section ends with a discussion of diversity of simulated aerosol optical properties (AOD, AE, scattering and absorption coefficients) in the context of the species specific aerosol parameters (e.g., lifetime, burden, etc.) from each model. Section 4 presents and discusses the results from the comparison of modelled optical properties with the different observation records used. These results are presented in the form of performance charts of retrieved normalised biases and correlation coefficients for each OBS/VAR/MOD
- 230 combination. This is followed by a dedicated section??, which discusses the results for each model individually in order to identify strengths and weaknesses of each model in comparison with the observations and the other models. The paper ends with our conclusions from this comprehensive inter-comparison study. observational data sets. The observational assessment section ends with a short discussion of the representativity of the results.

2 Methods

235 2 Data and Methods

In this section, we first describe the ground and space based observation networks / platforms and variables that are used in this study (Section 2.1). Section 2.2 introduces the 14 global models used in this paper. Finally, Section 2.3 contains relevant information related to the data analysis (e.g., computation of model ensemble, co-location methods and metrics used for the model assessment).

¹Note that throughout this paper AOD denotes total "aerosol optical depth", while OD denotes "optical depth" of individual species (e.g., OD_{SO4})

240 2.1 Observationsand variables

Several ground and space-based observations have been utilised in order to establish perform a comprehensive evaluation at all scales . Table 1summarises all variables and observation networks that have been used. They (Tab. 1). These are introduced in more detail below.

Fig. 1 shows yearly average mean values of the observed AERONET AODs and column extinction Angstrom exponents.
245 Dust dominated regions such as Northern Africa and Southwest Asia are clearly visible both in the coarse AOD and the AE, but also in the total AOD, indicating its importance for the global AOD signal due to dust. The displayed satellite fields of AOD (MERGED-FMI) and AE (ATSR-SU) are particularly useful in remote regions and over the oceans where ground based measurements are less common, and, thus, add substantially to the global picture when assessing models. For example, satellites capture the nearly constant ocean AOD background of around 0.1 (mostly arising from sea salt) which is not really measured by
250 the land dominated, ground-based observation networks. The AE from AATSR-SU for instance, shows a latitudinal southwards

- decreasing gradient in remote ocean regions, indicating coarse(r)particle sizes, likely due to cleaner and, thus, more sea salt dominated regions. Transatlantic dust transport results in an increased particle size west of the Sahara (c. g. Kim et al. (2014)) as is captured by AATSR-SU. Finally, as can be seen in the lowermost panel of Fig. 1, in-situ sites from GAW show the highest density in Europe, followed by North America, while other regions are poorly represented. The differences in the individual
- 255 paragraphs below. Figure 1 shows maps of the annual mean values of the spatial coverage for each observation data-set are important to keep in mind when interpreting the results presented in Sect. ?? (especially Figs. ?? and ??).

The following subsections introduce briefly each of the variables considered (from some of the observation data-sets usedplatforms used). It is discussed below in Sect. 2.1.7. Note that the wavelengths in Table 1 reflect the wavelengths used for comparison with the models, however, the original measurement wavelengths may be different as noted below.

260 2.1.1 AERONET

The Aerosol Robotic Network (AERONET, Holben et al. (1998)) is a (AERONET Holben et al., 1998) is a well established, ground-based , well established remote sensing network based on sun photometer measurements of columnar optical properties. The network comprises several hundred measurement sites around the globe (see Fig. 1).

a.c.d.e for the 2010 sites). In this paper, cloud screened and quality assured daily aggregates of AERONET AODs, AOD_f, AOD_c and AE from the version 3 (Level 2) Sun and SDA products (e.g. O'Neill et al. (2003), Giles et al. (2019)
 (e.g., O'Neill et al., 2003; Giles et al., 2019)) have been used. No further quality control measures have been applied due to the already high quality of the data. Only site locations below 1000 m altitude were considered in this analysis.

For the analysis, the spectral AOD values. The sun photometers measure AOD at multiple wavelengths. For comparison with the model output (which is provided at 550 nm), the measurements at 500 nm and 440 nm were used to derive an the total

AOD at 550 nm, using the provided AE – data to make the wavelength adjustment (the 500 nm channel was preferred over the 440 nm channel). Similarly, the AOD_f and AOD_c data provided at 500 nm via the AERONET spectral deconvolution algorithm (SDA) product, were shifted to 550 nm using the AE data. The SDA product (O'Neill et al., 2003) computes AOD_f and AOD_c

in an optical sense, based on the spectral curvature of the retrieved AODs in several wavelength channels and assuming bimodal aerosol size distributions. Thus, as pointed out by O'Neill et al. (2003) it does not correspond to a strict size cut at a

- 275 certain radius, such as the R=0.6 μ m established in the AERONET Inversion product (Dubovik and King, 2000). Compared to the Inversion product, the SDA product used here tends to overestimate the coarse contribution (O'Neill et al., 2003) which suggests that, on average, the effective cut applied in the SDA product is closer to the strict threshold of R=0.5 μ m required from the models within the AP3-CTRL experiment (see Sect. 2.2 for details). The implications of this difference are discussed in Section 4. It should also be noted that the AE provided by AERONET is calculated from a multi-wavelength fit to the four
- 280 AERONET measurement wavelengths, rather than from selected wavelength pairs.

Data from the short term DRAGON campaigns (Holben et al. (2018)) was (Holben et al., 2018) were excluded in order to avoid putting giving too much weight on the associated regions (that show to the associated campaign regions (with high density of measurement sites) with respect to the in the computation of network averaged statistical parameters used in this study. No further site selection has been performed, since potential spatial representativity issues associated with some AERONET sites

285 were found to be of minor relevance for this study (Sect. ??, Fig. ??). Fig. 1 shows the sites used for all variables, where colors indicate the 2010 mean values at each location. Table 1 includes relevant information about the data-set4.5).

The sun photometer measurements only occur during daylight and cloud free conditions. Thus, the level 2 daily averages used here represent daytime averages rather than 24h averages (as provided by the models). Because of the requirements for sunlight and no clouds, the diurnal coverage at each site shows a more or less pronounced seasonal cycle depending on the

290 latitude (e.g., only mid-day measurements at high latitudes in winter) and the seasonal prevalence of clouds in some regions. This is a clear limitation when comparing with 24h monthly means output from the models (as done in this study). However, these representativity issues were found to have minor impact for the model assessment methods used in this study (details are discussed in Sect. 4.5).

2.1.2 Surface in-situ in situ data

295 Surface in-situ

Surface in situ measurements of the aerosol light scattering (SC) and absorption coefficients – (AC) were accessed through the GAW-WDCA database EBAS (http://ebas.nilu.no/). As with AERONET, only sites with elevations below 1000 m were considered. Annual mean values of scattering and absorption are shown in Figure 1g,h. The in situ site density is highest in Europe, followed by North America, while other regions are poorly represented. The EBAS database also includes various

- 300 observations of atmospheric chemical composition and physical parameters, although those were not used here. For both scattering and absorption variables, only level 2 data from the EBAS database were used (i.e., quality controlled, hourly averaged, reported at STP)standard temperature and pressure (STP); $T_{std} = 273.15$ K, $P_{std} = 1013.25$ hPa). All data in EBAS have version control, and a detailed description of the quality assurance and quality control procedures for GAW aerosol in-situ in situ data are available in ?Laj et al. (2020). Additionally, for this study, data was only considered if it was associated with
- 305 the EBAS categories *aerosol* or *pm10*. The *aerosol* category indicates the aerosol was sampled using a whole air inlet, while

pm10 indicates the aerosol was sampled after a 10μ m aerodynamic diameter size cut. It was assumed whole air and pm10 would provide the better comparison with model simulations than measurements with smaller cut size (e.g., pm2.5 or pm1).

Invalid measurements were removed based on values in the flag columns provided in the data files. Furthermore, outliers were identified and removed using value ranges of $\{-10, 1000\}$ Mm⁻¹ and $\{-1, 100\}$ Mm⁻¹ for scattering and absorption coefficients, respectively. The outliers were removed in the original 1h time resolution before averaging to monthly resolution

for comparison with the monthly model data. For most of the absorption data,

For the in situ AC data used in this study, most of the measurements are performed at wavelengths other than 550 nm (see Sect. 1 in supplement 2). These were converted to 550 nm assuming an absorption Angstrom Angström exponent (AAE) of 1 (e. g. Bond and Bergstrom (2006)). For the scattering coefficients (i.e., a $1/\lambda$ dependence, e.g., Bond and Bergstrom, 2006).

- 315 This is a fairly typical assumption when the spectral absorption is not measured. For about 50% of the sites, absorption was measured at ~530 nm meaning that even if the true AAE had a value of 2, the wavelength-adjusted AC value would only be underestimated by ca 4%. For another 25% of the sites, absorption was measured at ~670 nm. For these sites the impact of an incorrect AAE value is larger (ca 26% overestimation for an actual AAE of 2 and ca 6% for AAE=1.25). The remaining 25% of sites typically utilized wavelengths between these two values. Schmeisser et al. (2017) suggest that, across a spatially
- 320 and environmentally diverse set of sites measuring spectral in situ absorption (many included here), that the AAE is typically between 1 and 1.5.

The majority of in situ scattering sites used here included a measurement at 550 nm (see Tab. 2 in supplement 2), so for these data no wavelength adjustment was necessary. The remaining few sites measuring around 520 nm were shifted to 550 nm assuming a scattering AE (SAE) of 1 (we note that this is rather at the lower end of typically measured SAEs, see Andrews et al., 2019)

325 . However, we assess the uncertainties similar to those discussed above for AC, indeed, the change in model bias as compared to an assumed SAE=1.5 was found to be <0.5%. As mentioned previously, the in situ measurements are, ideally, made at low RH (RH ≤ 40%) but are not absolutely dry (i.e., RH = 0%). Control of sample relative humidity is not always perfect so, depending on site and conditions, the measurement RH could exceed 40%. Because the model data with which the in situ scattering data will be compared is reported at RH = 0%, only measurements at RH ≤ 40% were considered . For the model
330 evaluation, the 2010 monthly model data was converted to STP using the following formula:

$$X_{\rm STP} = X_{\rm AMB} \times \left(\frac{p_{\rm STP}}{p_{\rm AMB}}\right) \cdot \left(\frac{T_{\rm AMB}}{T_{\rm STP}}\right)$$

310

335

to minimize discrepancies due to potential scattering enhancement at higher RH values. While maintaining the measurement RH < 40% is typically assumed to minimize the confounding effect of water on aerosol properties (GAW Report 227, 2016), Zieger et al. (2013) suggest that there may be noticeable scattering enhancement even at RH = 40% for some types of aerosol (see their Figure 5b).

where p_{STP} and T_{STP} are standard IUPAC standard pressure and temperature, and p_{AMB} and T_{AMB} are air pressure and temperature at the corresponding site location. The correction was performed on a monthly basis using the station altitude to estimate the pressure and monthly near surface (2m)temperature from ERA5While observations from other platforms and

networks relied solely on 2010 data for the model assessment (see Tab. 1), many in situ sites began measurements after 2010 so

- a slightly different approach was taken in order to maximise the number of sites with monthly aggregated data. For any given in situ site, all data available between 2005 2015 was used to compare with the 2010 model output. The climatology for each in situ site was computed requiring at least 30 valid daily values for each of the climatological months over the 10 year period. Prior to that, daily values were computed from the hourly data applying a minimum 25% coverage constraint (i.e., at least 6 valid hourly values per day). It should be noted that the in situ data is collected continuously day and night regardless of cloud
- 345 conditions and, thus, daily data will represent the full diurnal cycle in most cases. As can be seen in column "Cov" in Tables 1 & 2 of the supplementary material 2, for most of the in situ sites, the 25% coverage constraint for the resampling from hourly to daily was typically met. Note that about half of all available hourly SC measurements in the 2005 - 2015 period were not considered here, either because the measured RH exceeded 40% or because RH data were missing in the data files.

A few urban in situ sites were removed from consideration for the model analysis, as these sites are likely not representative on spatial scales of a typical model grid. For scattering coefficients got excluded the sites excluded are: Granada; Phoenix; National Capitol - Central, Washington D.C; and for absorption coefficients: Granada; Leipzig Mitte; Ústí n.L.-mesto. After applying the RH constraint, removing urban sites from consideration, and resampling to monthly climatology, data from 39 sites with scattering data and from 39 sites with absorption data (not necessarily the same sites as for scattering) were available for model assessment (see Table 1).

355

The biases of each model for individual in-situ sites are shown in Appendix Figs. ?? and ?? for scattering and absorption, respectively.Due to the limited number of stations, and in order to increase temporal sampling coverage of the monthly aggregates used, a 2005-2015 climatologywas used to compare with the 2010 model output (unlike for the other observationswhich solely used 2010 data , see Tab. 1) Tables 1 & 2 in the supplementary material 2 provide detailed information about each of the

360 absorption and scattering sites used. This includes the original measurement wavelengths as well as temporal coverage for the computation of the climatology. The elimatology for each site was computed requiring at least 30 valid daily values over the 10 year period, for each of the months. Prior to that, daily values were computed from

2.1.3 Satellite data sets - Introduction

365 In addition to the hourly data applying a minimum 25% coverage constraint (i.e.at least 6 valid hourly values per day)ground-based observations, data from four different satellite data sets (MODIS Aqua & Terra, AATSR SU v4.3 and a merged AOD satellite data set) were used to evaluate optical properties from the AP3 models. The four satellite data sets are introduced below.

Even though the satellite observations usually come with larger uncertainties and may exhibit potential biases against ground-based column observations (e.g., Gupta et al., 2018), we believe that it is a valuable addition to not only evaluate

370 models at ground sites but also incorporate satellite records for an assessment of model performance. The main advantage of satellite data is the spatial coverage relative to ground-based measurements. Satellites provide more coverage over land masses than AERONET and in addition, they are the primary observational tool for column optical properties over oceans. Because of AERONET's reliability and data quality, it is generally accepted as the gold standard for column AOD measurements.

Therefore, all four satellites used in this paper, were evaluated against AERONET data, in order to establish relative biases and

- 375 correlation coefficients. Details related to this satellite assessment are discussed in supplement 2 and are briefly mentioned in the introduction sections for each individual satellite below. The results from this satellite assessment are also available online (see Mortier et al., 2020a), allowing an interactive exploration of the data and results (down to the station level) and include many evaluation metrics (e.g., various biases, correlation coefficients, root mean square error (RMSE)). These comparisons of the individual satellites against AERONET provide context for the differences in the model assessments discussed below in
- 380 Section 4. It should be noted, however, that the retrieved biases for each satellite data set provide insights into the performance of each satellite product at AERONET sites, which are land dominated. Satellites often have different retrieval algorithms over land and ocean (e.g., Levy et al., 2013), and the aerosol retrieval tends to be more reliable over dark surfaces, such as the oceans, than over bright surfaces, such as deserts (e.g., Hsu et al., 2004).

2.1.4 MODIS data

385 Daily gridded level 3 AOD data data from the Moderate Resolution Imaging Spectroradiometer (MODIS) has have been used from both satellite platforms (Terra and Aqua) for evaluation of the models. The merged land and ocean global product (named Aerosol_Optical_Depth_Land_Ocean_Mean AOD_550_Dark_Target_Deep_Blue_Combined_Mean in the product files) of the recent collection 6.1 was used. This is an updated and improved version of collection 6 (e.g. Levy et al. (2013), Sayer et al. (2014)(e.g., Levy et al., 2013; Sayer et al., 2014). For changes between both data-setsdata sets, see Hubanks (2017).

390 2.1.5 AATSR SU v4.3 data

Details about the MODIS data sets used are provided in Table 1. Compared to AERONET, both Aqua and Terra exhibit positive AOD biases, suggesting an overestimation of ca +9% and +17%, respectively, at AERONET sites and for the year 2010 (for details see supplement 2). The larger overestimate for Terra is in agreement with the findings from Hsu et al. (2004). The AATSR-

395 2.1.5 AATSR SU v4.3 data

The AATSR SU v4.3 SU data-set data set provides gridded AOD and associated parameters from the AATSR Advanced Along Track Scanning Radiometer (AATSR) instrument series, developed by Swansea University (SU) under the ESA Aerosol Climate Change Initiative (CCI). The AATSR instrument on ENVISAT covers the period 2002-2012-2002 - 2012 and in this study, data from 2010 is are used. The instrument's conical scan provides two near simultaneous views of the surface, at solar reflective wavelengths from 555-nm to 1.6- μ m.

400

Over land, the algorithm uses the dual-view capability of the instrument to allow estimation without a priori assumptions on surface spectral reflectance (North (2002), Bevan et al. (2012))(North, 2002; Bevan et al., 2012). Over ocean, the algorithm uses a simple model of ocean surface reflectance including wind-speed and pigment dependency at both nadir and along-track
view angles. The retrieval directly finds an optimal estimate of both the AOD at 550 nm, and size, parameterised as relative

- 405 proportion of fine and coarse mode aerosol. The local composition of fine and coarse mode is adopted from the MACv1 aerosol climatology (Kinne et al. (2013))(Kinne et al., 2013). The local coarse composition is defined by fraction of non-spherical dust and large spherical particles typical of sea salt aerosol, while fine mode is defined by relative fractions of weak and strong absorbing aerosol. A full description of these component models is given in (de Leeuw et al. (2015))de Leeuw et al. (2015). Further aerosol properties including AE and AAOD ((calculated between 550 and 856 nm) and absorption aerosol optical depth
- (AAOD, not used in this study) are determined from the retrieved AOD and composition. Aerosol is properties are retrieved over all snow-free and cloud-free surfaces. The most recent version SU AATSR AATSR SU V4.3 (North and Heckel (2017))
 (North and Heckel, 2017) advances on previous versions by improved surface modelling and shows reduced positive bias over bright surfaces. The output at L2 is total column AOD at 550 nm, at 10 km resolution, and associated aerosol properties. Retrieval uncertainty and comparison with sun photometer observations show highest accuracy retrieval over ocean and darker
- 415 surfaces, with higher uncertainty over bright desert surfaces, and land surface at southern latitudes (Popp et al. (2016)). The surfaces (e.g., desert, snow) and for large zenith angles (Popp et al., 2016). This study uses the level 3 outputis re-gridded to, which is provided at daily and monthly 1x1 degree 1° x 1° resolution, intended for climate model comparison.

In this study, AE as well as Specifically, AATSR SU values for AE and total, fine and coarse AODs are used. Results (normalised biases and correlation coefficients) from an inter-comparison with AERONETmeasurements is shown in Fig. ??

420 (discussed in more detail in Sect. ??) The AE calculation is only performed for 0.05 < AOD < 1.5 due to increased retrieval uncertainty of AE at low and high AODs.

In comparison with AERONET, the AATSR data exhibits an AOD bias of ~-4%, suggesting a slight underestimation of AOD at AERONET sites, in contrast to the two MODIS products used (see Tab. 1). To our knowledge, this AATSR product (SU V4.3) has not been evaluated against AERONET in the literature. Thus, these results comprise an important finding of

425 this study. Biases of AOD_f , AOD_c and AE against AERONET were found to be + 1.6%, - 14.7% and + 14.3%, respectively (see web visualisation, Mortier et al., 2020a).

Initial comparisons within the Aerosol CCI project suggest that the fine mode fraction of total AOD may be overestimated over the ocean, with consequently some high bias in AE. The AE provided by AATSR is estimated for the range 550-870 nm, and some difference may be expected also with AERONET derived AE using a different wavelength range (e.g., Schuster et al., 2006)

2.1.6 Merged satellite AOD data

430

435

The MERGED-FMI data-set (1995-2017)data set, developed by the Finnish Meteorological Institute, includes gridded L3 level 3 monthly AOD products merged from 12 available satellite products (Sogacheva et al. (2019)) (Sogacheva et al., 2020). It should be noted that MODIS and AATSR products are considered inside this MERGED-FMI data set. It is available for the period 1995-2017, however, here only 2010 data are used.

Compared to AERONET measurements from 2010, this merged satellite product has shown excellent performance with the highest correlation (R=0.89) among the four satellites used and only a slight underestimation of AOD (bias of -5.4%) at

AERONET sites (see supplement 2 and Mortier et al., 2020a). The merging method is based on the results of the evaluation of the individual satellite AOD products against AERONET. Those results were utilised to infer a regional ranking which was

- 440 then used to calculate a weighted AOD mean. Because it is combined from the individual products of different spatial and temporal resolution, the AOD merged product is characterised by the best possible coverage, compared with other individual satellite products. The AOD merged product is at least as capable of representing monthly means as the individual products (Sogacheva et al. (2019))(Sogacheva et al., 2020). Standard pixel-level uncertainties for the merged AOD product were estimated as the root mean squared sum of the deviations between that product and other eight merged AOD products calculated
- 445 with different merging approaches applied for different aerosol types (Sogacheva et al. (2019)) (Sogacheva et al., 2020).

2.1.7 Global distribution of optical properties investigated

The previous sections introduced the individual ground and space-based observation records and optical properties variables that will be used in this paper for the model assessment. Figure 1 provides an overview of the global distribution of these optical properties. The displayed global maps show annual mean values of all variables considered, both for the ground-based networks
and for a selection of the satellite observations. Fig. 1a,c,d shows yearly average mean values of the observed AERONET AODs (total, coarse and fine, respectively). Column Ångström exponents from AERONET are shown in Fig. 1e. Dust dominated regions such as Northern Africa and Southwest Asia are clearly visible both in the coarse AOD and the AE, but also in the total AOD, indicating the importance of dust for the global AOD signal. The satellite observations of AOD (MERGED-FMI) and AE (ATSR-SU) (Fig. 1b,f) are particularly useful in remote regions and over the oceans where ground-based measurements are less common. Thus, they add substantially to the global picture when assessing models. For example, satellites capture the nearly constant AOD background of around 0.1 over the ocean (mostly arising from sea salt) which cannot be obtained from the land dominated, ground-based observation networks. The AE from AATSR-SU shows a latitudinal southwards decreasing

gradient in remote ocean regions, indicating dominance of coarse(r) particle size distributions, which is likely due to cleaner and, thus, more sea salt dominated regions. Transatlantic dust transport results in an increased particle size west of the Sahara
(e.g., Kim et al., 2014) as is captured by AATSR-SU. Finally, it is difficult to observe global patterns in the in situ scattering and absorption data due to the limited spatial coverage of the measurements, as can be seen in the lowermost panels (Fig. 1g,h). The differences in the spatial coverage for each observation data set will be important to keep in mind when interpreting the results presented in Sect. 4.

2.2 Models

- 465 This study uses output from 13–14 models that are participating in the AeroCom 2019 control experiment(, denoted in the following as CTRL). For this AP3-CTRL experiment. Details on the AeroCom phase III experiments can be found on the AeroCom wiki page (AeroCom wiki, 2020). The wiki also includes information on how to access the model data from the different AeroCom phases and experiments, which is stored in the AeroCom database. Note, that the database location and information about it might change in the future, the intention is however to keep updated information available via the website
- 470 https://aerocom.met.no (last access: 14.09.2020). Table 2 provides an overview of the models used in this paper. For the

<u>AP3-CTRL</u> experiment, modellers were asked to submit simulations of at least the years 2010 and 1850, with 2010 meteorology and prescribed (observed) sea-surface temperature and sea ice concentrations. <u>Modellers were asked to use CMIP6</u> emission inventories , and using emission inventories from CMIP6 (Eyring et al., 2016), when possible. <u>Details concerning the</u> anthropogenic and biomass burning emissions are given in (CEDS, Hoesly et al., 2018) and (BB4CMIP, van Marle et al., 2017)

- 475 . In this paper, only the 2010 model output is used. The year 2010 was chosen as a reference year by the AeroCom consortium and is used throughout many phase II and III experiments for inter-comparability of different experiments and model generations. The AeroCom phase I simulations (e.g., Dentener et al., 2006; Kinne et al., 2006; Schulz et al., 2006; Textor et al., 2006) used the year 2000 as a reference year. One of the main reasons to update the reference year from 2000 to 2010 was that many more observations became available between 2000 and 2010 and also to account for changes in the present day climate, for instance,
- 480 due to changing emissions and composition (e.g., Klimont et al., 2013; Aas et al., 2019; Mortier et al., 2020b).

Detailed information about the models and on emissions, humidity growth and particularly their treatment of aerosol optics has been collected from the modelling groups through a questionnaire. The tabulated responses are provided in the AeroCom optics questionnaire (supplementary material). An overview of all models is provided in table 2. More details about supplementary material 1. The first table (Spreadsheet "Table: General questions") contains general information that applies

- 485 to the total aerosol, such as mixing assumptions, treatment of clear-sky optics and water uptake parameterisations. The second table (Spreadsheet "Table: Species specific") contains aerosol species specific information such as the complex refractive index at 550 nm, humidity growth factors, particle density, as well as details regarding the emission data sets used. Further information related to OA emissions and secondary formation is provided for most models in a third spreadsheet ("Table: OA details"). In addition, Sect. 4 of supplement 2 provides further information for each of the modelsis provided in the corresponding discussion
- 490 section ??., mostly complementary to Table 2. The

2.2.1 Model diagnostics

Requested diagnostics fields for AP3-CTRL are available online (see, AeroCom diagnostics sheet, 2020). In addition, variables for dry (at RH = 0%) extinction (EC_{dry}) and absorption (AC_{dry}) coefficients were requested (at model surface level) from the modelling groups participating in this study. These are needed for the comparison with the GAW surface in situ observations (Sect. 2.1.2). Note that in a few cases, some diagnostic fields used in this study could not be provided by some of the modelling groups.

To obtain model values that were comparable with observations, additional processing was required for some variables. The 500 \underline{AOD}_{c} fields were not directly submitted but were computed as the difference: AOD - . For the comparison with the surface in-situ data, modellers were asked to provide fields of extinction (EC_{dry}) and absorption (AC_{dry}) at dry conditions (RH=0%). The dry scattering (SC_{dry})was \underline{AOD}_{f} . The AE fields were computed from the provided AOD at 440 nm and 870 nm² via $\underline{AE} = -\ln(\underline{AOD}_{440}/\underline{AOD}_{870}) / \ln(440/870)$. Dry scattering coefficients (SC_{dry}), for the comparison with the surface in situ

²For GISS-OMA, 550 nm and 870 nm AODs were used for AE calculation as 440 nm AOD data was missing.

data were computed via $SC_{dry dry} = EC_{dry dry}$ - $AC_{dry dry}$. Some of the models that provided these data - submitted dry

- 505 EC, but ambient AC , in which case that combination was used, (indicated in Table 2). For these models, dry scattering was derived in the same way $SC_{drx} = EC_{drx} - AC_{amb}$ consistent with the idea that absorbing aerosol tend tends to be hydrophobic. Note that for some models, not all required fields were available, which is indicated by gaps in the resulting heat-map plots shown in Figs. ??-?? and Figs. ??, ??). The latter may be violated to some degree for models that include internally mixed BC modes with hydrophilic species, such as SO_4 . However, an investigation of differences between dry and ambient absorption
- 510 coefficients revealed that the overall impact on the results is minor, both for models with internally mixed BC modes and for models with externally mixed modes.

Some of the models reported the columnar optical properties based on clear-sky (CS) assumptions, while others assumed all-sky (AS) conditions to compute hygroscopic growth and extinction efficiencies. These choices are indicated in Table ?? and differences can be inferred from Fig. ??. 2 and details related to the computation of CS optics can be found in supplement 1.

515 2.2.2 Computation of AeroCom ensemble mean and median

525

535

The monthly AeroCom ensemble mean and median fields were computed in $2^{\circ} \times 3^{\circ}$ resolution (Tab. 2). Model fields were all re-gridded to this resolution before the ensemble mean and median was computed. Only those models were considered that had submitted all required optical property variables used in this study (Tab. 1). Those used for constructing the ensemble model are indicated in Table **??**. In this paper, the output from the median model is used (denoted MEDIAN below)if not

520 otherwise indicated. The following modelled global average values have been retrieved of species specific model parameters to be compared in section 3 to assess lifecycle aspects of model diversity:

In addition local diversity δ fields were computed for each variable and are defined as follows:

- 1. Diversity ensemble mean: δ_1 =std. Emissions and formation of aerosol species (in units of Tg / yr). The secondary aerosol formation of SO₄, NO₃ and OA by chemical reactions in the atmosphere is difficult to diagnose. Thus, they are diagnosed here from total deposition output.
- 2. Lifetimes of major aerosol species (in units of days), computed from column burden and provided wet+dry deposition rates. The lifetimes can give insights into the efficiency of removal processes in the models.
- 3. Global mass burdens (in units of Tg) for each species. These values enable comparisons amongst the models in terms of aerosol amount present on average.
- Modelled speciated optical depths (ODs) at 550 dev.nm. This unitless quantity provides another way of looking at contributions from different species to total AOD based on their optical properties rather than their burden.
 - 5. Modelled mass extinction coefficients (MECs) at 550 nm for each species (in units of m²/mean (not so good g), calculated by dividing the species optical depth with the corresponding species mass burden (e.g., OD_{DU} /meaningful in case of outliers) LOAD_{DU}). The MEC determines the conversion of aerosol mass to light extinction, and can provide insights into the variability of modelled size distributions or hygroscopicity.
 - 17

- 6. Diversity ensemble median: δ_{2} = IOR Additionally, modelled mass absorption coefficients (MACs) at 550 nm for light absorbing species (BC, DU, OC) are presented. These are calculated by dividing the species absorption optical depth (AAOD) with the corresponding species mass burden (e.g., AAOD_{BC} /median (outlier resistant definition) where IOR denotes the interguartile range (i.e.the difference between the 3rd and 1st quartile). LOAD_{BC}).
- We note again that detailed introductions for each model are provided in supplement 1 and in Section 4 of supplement 2, in 540 addition to the summary Table 2.

2.3 Data analysisprocessing and statistics

The analysis of the data was performed using the pyaerocom software (Appendix ??). Most of the analysis in this study was performed with the software pyaerocom (Github: https://github.com/metno/pyaerocom, Website: https://pyaerocom.met.no/.

- 545 last access: 14.09.2020), pyaerocom is an open source Python software project that is being developed and maintained for the AeroCom initiative, at the Norwegian Meteorological Institute. It provides tools for harmonisation and co-location of model and observation data, and dedicated algorithms for the assessment of model performance at all scales. Evaluation results from different AeroCom experiments are uploaded to a dedicated website that allows exploration of the model and observation data and evaluation metrics. The website includes interactive visualisations of performance charts (e.g., biases, correlation
- 550 coefficients), scatter plots, bias maps and individual station and regional timeseries data, for all models and observation variables, as well as bar charts summarising regional statistics. All results from the optical properties evaluation discussed in this paper are available via a web interface (see Mortier et al., 2020c).

The ground and space based observations are colocated co-located with the model simulations by matching with the closest model grid-point in the originally provided model resolution.

- 555
 - In the case of ground-based observations (AERONET and GAW in-situin situ), the model grid-point grid-point closest to each measurement site is used. For the satellite observations, both the model data and the (gridded) satellite product are regridded to a resolution of $5^{\circ} \times 5^{\circ}$ and the closest model grid-point grid-point to each satellite pixel is used. The choice of this rather coarse resolution is a compromise, mostly serving the purpose of increasing the temporal representativity (i.e., more data points per grid cell) in order to meet the time resampling constraints (defined below). For the comparison of satellite AODs
- with models, a minimum AOD of 0.01 was required, due to the increased uncertainties related to satellite AOD retrievals at 560 low column burdens. The low AODs were filtered in the original resolution of the level 3 gridded satellite products, prior to the co-location with the models.

Since many model fields were only available in monthly resolution, the colocation co-location of the data with the observations (and the computation of the statistical parameters used to compare the models) was performed in monthly resolution.

Any model data provided in higher temporal resolution was averaged to obtain monthly mean values, prior to the analysis. For 565 the higher resolution observations (see Table 1), the computation of monthly means was done using a hierarchical resampling scheme, requiring at least $\frac{25}{25}$ coverage. Practically, this means that the daily AERONET data was were resampled to monthly, requiring at least 7 daily values in each month. For the hourly in-situ in situ data, first a daily mean was computed (requiring at least 6 valid hourly values) and from these daily means, monthly means were computed requiring at least 7 daily

- 570 values. Data that did not match these coverage constraints were invalidated.
 - Throughout this paper, the discussion of the results will use two statistical parameters to assess the model performance. The : the normalised mean bias (NMB) is defined as NMB = $\frac{\sum_{i}^{N} (m_i o_i)}{\sum_{i}^{N} o_i}$ where m_i and o_i are the model and observational mean, respectively, and the Pearson correlation coefficient (R)to assess the model performance. These parameters were computed for each variable / obs. network / model combination and are presented in the form of heat-map performance charts (Figs. ??,
- 575 **??**). More evaluation metrics, such as normalised RMSE or fractional gross error are available online in the web visualisation (Mortier et al., 2020c), but are not further considered within this paper.

The next section Section 4.5 presents several sensitivity studies that were performed in order to investigate the spatiotemporal representativity of this analysis strategy, which is based on network-averaged, monthly aggregates, as. This was done because representativity (or lack thereof) comprises a major source of uncertainty (e.g. Schutgens et al. (2016), Schutgens et al. (2017)

580 , Sayer and Knobelspiesse (2019))(e.g., Schutgens et al., 2016, 2017; Sayer and Knobelspiesse, 2019). The focus here was to assess how such potential representation errors affect the biases and correlation coefficients used in this paper to assess the model performance and comparison with other models.

2.4 Representativity of the results

590

As described in the previous Section 2.3, monthly aggregates of the models and observations were colocated in space and time. 585 The resulting point cloud of monthly mean values from all sampling coordinates (sites-

2.3.1 AeroCom ensemble mean and median

For all variables investigated in this paper, the monthly AeroCom ensemble mean (ENS-MEAN) and median (ENS-MED) fields were computed and have been made available in the AeroCom database, for future reference. This was done in order to enable an assessment of the AP3 model ensemble, which we consider to represent the most likely modelling output of the state of the art aerosol model versions participating in the AP3-CTRL exercise.

The ensemble fields were computed in a latitude/aggregated satellite pixels) was then used to compute the biases and correlation coefficients. These are then used to assess the performance of individual models and longitude resolution of $2^{\circ} \times 3^{\circ}$, which corresponds to the lowest available model resolution (i.e., of models EC-Earth and TM5, see Table 2). Model fields

- 595 were all re-gridded to this resolution before the ensemble mean , discussed in the following sections (Figs. ?? and ??). The comparison of the (often)temporally incomplete observational records (that are sampled at distinct locations) can introduce considerable representation errors both on spatial and on temporal scales (see e.g. Schutgens et al. (2016), Schutgens et al. (2017) , Wang et al. (2018), Sayer and Knobelspiesse (2019) and references therein). These errors can affect established biases between model and observations but also other performance measures such as and median were computed. In this paper, only the output
- 600 from the median model is used. Note that results from the mean model are not further discussed below but are available online (see Mortier et al., 2020c). In addition to the median (50th percentile), also the 25th (Q1) and 75th (Q3) percentiles

were computed and evaluated against the observations like any other model. This was done to enable an assessment of model diversity in the retrieved biases and correlation coefficients. We consider this to be the major source of uncertainty for this study. Therefore, several sensitivity studies have been performed in order to investigate how potential spatio-temporal

- 605 representation errors affect the global monthly statistical parameters used in this study. Temporal representation uncertainties were investigated 1. for in-situ absorption coefficients using hourly TM5 data from the AeroCom INSITU experiment evaluated against GAW measurements (Fig. ??) and 2. for columnar AOD using 3-hourly data from ECMWF-IFS, evaluated against AERONET AODs (Fig. ??). In addition, spatial representativity errors were investigated by colocating the ensemble mean AOD field both with observations from all AERONET sites (available in 2010)and with a selection of sites that are considered
- 610 representative on spatial scales covered by a typical model grid cell. The latter was selected based on Wang et al. (2018) using only sites that show an absolute spatial representation error smaller than 10% and

In addition, local diversity fields were computed for each variable by dividing the the interquartile range (IQR = Q1 – Q3) by the ensemble median: $\delta_{IQR} = IQR$ / median, which corresponds to the result of this comparison is shown in Fig.?? central 50% of the models as a measure of diversity (this is different than Kinne et al., 2006, who use the central 2/3). Note that the IQR

615 is not necessarily symmetrical with respect to the median. In order to enable a better comparison with the AP1 results from Textor et al. (2006) and Kinne et al. (2006), a second set of diversity fields were computed as follows: $\delta_{std} = \sigma /$ (ensemble mean), where σ is the standard deviation.

Note that the ensemble AE fields were computed from the individual models' AE fields. In case of the ensemble median, this will give slightly different results compared to a computation of a median based on median 440 and 870 AOD fields. This

620 is because the median computation is done in AE space and not in AOD space.

Please also note that the ensemble total AOD includes results from INCA which are not included in AOD_f and AOD_c (see Tab. 2). This results in a slightly smaller total AOD in the ensemble when inferred from $AOD_f + AOD_c$ (which does not include INCA) compared to the computed AOD field (which includes INCA).

2.3.2 Model STP correction for comparison with GAW in situ data

625 One further model processing note: since the GAW in situ measurements are reported at STP conditions (Section 2.1.2), the 2010 monthly model data were converted to STP using the following formula:

$$X_{\rm STP} = X_{\rm amb} \times \left(\frac{P_{\rm std}}{P_{\rm amb}}\right) \cdot \left(\frac{T_{\rm amb}}{T_{\rm std}}\right) \tag{1}$$

X_{STP} and X_{amb} are the model value of absorption (or scattering) at STP and ambient conditions, respectively. P_{amb} and T_{amb} are the ambient air pressure and temperature at the corresponding site location. The correction factor was estimated on a
monthly basis, where P_{amb} was estimated based on the station altitude (using the barometric formula and assuming a standard atmosphere and T_{amb} was estimated using monthly near surface (2m) temperature data from ERA5. This correction may introduce some statistic error mostly due to natural fluctuations in the pressure and possible uncertainties in the ERA5 temperature data. However, we assess this additional uncertainty to be small for the annual average statistics discussed below.

3 Results and discussion - Model diversity of aerosol lifecycle and optical properties

- 635 The focus of this section is to establish a global picture and to try to understand model diversity in relevant parameters related to the aerosol lifecycle (i.e., global emissions, lifetimes and burdens) as well as the simulated aerosol optical properties (i.e., speciated MECs, MACs and ODs). The results of these 3 sensitivity studies are summarised in Tab. 5) and show that the overall differences are of the order of 10% and 0.2 for NMB and correlation, respectively. For the in-situ absorption inter-comparison, the results in monthly resolution show better performance in nearly all statistical parameters, compared to hourly (Fig. ??).
- 640 goal is to develop an understanding of how, based on the models, processes and parameterisations link emissions to optical properties. A comparison of modelled optical properties with the various observation records is presented in the following Section 4.

From these results, we conclude that differences in these network averaged statistics, arising from spatio-temporal representation errors, are small compared to the diversity in the results found among the different models participating in this study (shown

- 645 in Figs. ?? and ??). Most of the discussion in this section focuses on the model ensemble median and associated diversities (δ_{IQR}) . Section 3.1 focuses on diversity in the treatment of the different aerosol species in the models, starting with an overview of simulated global aerosol emissions, lifetimes and mass burdens (Sect. 3.1.1), followed by a discussion of simulated ODs, MECs and MACs for each species (Sect. 3.0.1). Section 3.1 provides and discusses the global distribution of the simulated aerosol optical properties and their spatial diversity.
- 650 Based on these findings, and due to the fact that some model data was only available in monthly resolution, it was therefore decided that all model and observation comparisons in this study would be performed in monthly resolution. This was done because we believe that it will make the inter-model results more consistent and hence, more suitable for *inter*-comparison, since they carry similar representation errors (which are introduced by the incompletely sampled observational records).

3.1 Lifecycle and optical properties for each aerosol species

- 655 The small differences in bias and correlation that we find in our sensitivity tests (Figs.??, ??, ??) are important results that indicate that the magnitude of spatio-temporal representation uncertainties (in statistical parameters derived from annual averages over whole networks) is of the order of $\pm 10\%$. For non-geostationary satellites, the absolute temporal representation errors are likely larger due to the low sampling coverage, combined with cloud contamination in certain regions Table 3 provides an overview of global annual mean values of emissions, lifetimes, burdens, ODs, MECs and, where available, MACs.
- for each aerosol species (i.e., BC, DU, NO₃, OA, SO₄ and SS) and for each model. Gaps in the table indicate where models did not provide a requested variable. Also included are the median (MED) and diversity estimates (δ_{IQR}, δ_{std}) for each species and variable. Note that these are computed directly from the values provided in Table 3, not using the ensemble median fields. For comparison, median and δ_{std} from the AeroCom phase 1 (e. g. South Pacific). A detailed investigation of these uncertainties is beyond the scope of this work. Nonetheless, a further simple sensitivity study was performed aiming to investigate, how our choice of resolution in the satelliteAP1) simulations are provided as well. The colours in the table provide an indication of the sign and bias of the individual model values relative to the AP3 median.

Figure 2 provides a different view of the data provided in Table 3, by illustrating how the diversity of the individual parameters contributes to the resulting model ensemble diversity in species OD, similar to illustrations used earlier in Schulz et al. (2006) (their Figure 8) and Myhre et al. (2013) (their Figure 14). This visualisation makes it easier to link the diversity in speciated

670 ODs with the uncertainty in modelling the processes controlling the OD of each species.

3.1.1 Aerosol lifecycle: from emissions to mass burdens

As explained above, global aerosol emission and formation (in Table 3) were estimated either using the provided emission fields as for primary aerosols BC, DU, SS and POA, or using the equivalent total emissions as for SO_4 , OA and NO_3 based on total deposition. For simplicity we also call the equivalent total emissions, which include secondary formation from precursors,

675 "emissions" in this section. Note, that only major aerosol species are included in our study; aerosol precursor species that are provided by some few models (e.g., NO_x, NH₄ or VOCs) are not analysed.

Emissions are highest for sea salt (4980 Tg /model comparison (i. e. based on $5^{\circ} \times 5^{\circ}$ resolution and monthly averages) would affect the results (NMB and R), as compared to an analysis that is performed in daily resolution and using the highest available horizontal resolution for each model yr), followed by dust (1440 Tg /satellite_yr), SO₄ (143 Tg / yr), OA (116 Tg / yr,

- 680 of which ca 75 Tg / yr are due to primary emissions), NO_3 (33 Tg / yr), and BC (10 Tg / yr). Compared to AP1, the median emissions have decreased for all species except organic aerosols. For prescribed anthropogenic emissions, the differences between AP1 and AP3 may partly be due to differences in the emissions inventories. AP1 used inventories for the year 2000 whereas, here, the 2010 emissions are used (for details see supplement 1, section S6). Differences are likely also due to changes in the modelling setups and emission parameterisations.
- 685 Changes in parameterisations of online calculated natural DU and SS emissions are an explanation for their decreased emissions, 20% and 13%, respectively, compared to AP1. DU diversity has increased slightly relative to AP1, while SS diversity has decreased, however, with a standard deviation of circa 150%, it is still very large. As in AP1, the reasons for diversity in DU and SS emissions can be found in a range of parameters: surface winds, regions available to act as a source (semi-arid and arid areas for DU, sea-ice free ocean for SS), power functions used in the wind-emission relationship, aerosol size and
- 690 other factors. As an example, different size cutoffs are applied in the models when computing the source strength (see Sect. 2.2). For instance, EMEP includes dust particles with sizes up to $10 \,\mu$ m, TM5 and EC-Earth consider sizes up to $16 \,\mu$ m, while ECMWF-IFS considers sizes up to $20 \,\mu$ m. While the higher size cut explains higher emissions for the IFS model, it does not explain why the TM5 dust emissions are lower than those in the EMEP model.

The emission strengths of dust and sea salt reflect the surface wind distribution, which exhibits a larger tail in the distribution at higher resolution and in free-running atmospheric models. Meteorological nudging that was required for AP3-CTRL leads to lower emissions (e.g., Timmreck and Schulz, 2004). Most of the models in the AP1 simulations had implemented free-running atmospheric models, but operated at lower resolution, which should cancel out to a certain degree and make AP1 and AP3 similar when it comes to effective surface wind distribution. Better documented wind distributions could help explain emission differences. For instance, SPRINTARS (one of the highest resolution models, see Tab. 2) exhibits a negative departure from the

700 median in SS emissions, but an above average DU source (ca 1900 Tg / yr). The latter is comparable to that of OsloCTM3 and

EMEP, which both use reanalysis winds at different resolutions. Also noteworthy are considerable differences in SS emissions between the two ECHAM models (ECHAM-SALSA emits ca 30% less SS but 18% more dust than ECHAM-HAM) even though these two models use the same emission parameterisation (see Sect. 4 in supplement 2), the same meteorology for nudging and have the same resolution (see Tab. 1 for an overview of the satellites used2). This was done for each model that

- 705 provided daily (or higher resolution) data and for the variables AOD, , and AE. The results are shown in Table ??. In most cases, the higher resolution data results in slightly less negative biases and differences can be up to +10% in NMB (e.g. AE SPRINTARS vs. AATSR-SU) . However, in most cases the differences are marginal and are well below 5%. indicates that nudging and higher resolution in AP3 are not the sole explanation for the AP3 decrease in the dust and sea salt emission strengths against AP1 and that inconsistencies remain.
- Finally, we want to stress that the uncertainties established here and discussed above are not to be misinterpreted with corresponding uncertainties over sub-domains or at specific locations and times, which can be significantly larger as shown in the various literature referred to above. Considerable diversity is also observed for OA emissions (64%), which is a result of multiple organic aerosol sources, represented differently by the models (supplement 1). Uncertainties are associated with the primary organic particle emissions (POA, diagnosed in only four models) and biogenic and anthropogenic secondary
- 715 organic aerosol formation (SOA), DMS derived MSA, as well as biomass burning sources. As can be seen in supplement 1, there are also considerable differences among the models related to the conversion of organic carbon (OC) from the different sources to total organic mass. For instance, some models use a constant factor for all types of OC "emissions" (most commonly 1.4, though Tsigaridis et al., 2014, had suggested this value is too low) while others use different conversion factors for fossil fuel and biomass burning sources (ranging between 1.25-2.6). Conversion factors of 1.14 are reported for the NorESM model for monoterpene and isoprene as well as 8.0 for MSA (which is formed in the atmosphere via oxidation
- of DMS). Moreover, models show considerable differences in OA related emission inventories used. All these differences combined explain the high diversity associated with OA "emissions", which deserves further attention.

3.2 Evaluation of satellite products at AERONET stations

All satellite data-sets were evaluated against the ground based AERONET data in order to establish an estimate of the relative differences (biases, correlation coefficients) between the different data-sets when comparing with the models. The evaluation of the gridded satellite level 3 products was performed in the same manner as the evaluation of the models (see previous Sect. 2.3The decrease of SO₄ "emissions" compared to AP1 can not be explained by a change in anthropogenic SO₂ emissions between 2000 and 2010. Although Klimont et al. (2013) showed a decrease, the updated CEDS inventory (Hoesly et al., 2018) shows an increase of SO₂ emissions and was used in AP3. The increased variability in sulphate "emissions" may be due to

730 considerable differences in the treatment of natural sulphur sources. The anthropogenic emissions are prescribed by CEDS and should be more consistent among the models, although loss of SO₂ and the chemical formation of SO₄ certainly contribute to "emission" variability. Estimates of volcanic sulfur emissions range between 1 – 50 Tg / yr (SO₂, e.g., Andres and Kasgnoc, 1998; Halmer et al. Note that ECMWF-IFS did not consider volcanic emissions, and EMEP only considered major European sources (i.e., degassing from Etna and the Aeolian Islands, and the 2010 Eyjafjallajökull eruption in Iceland), which explains their comparatively

- 735 low SO₄ emissions. GEOS, despite including volcanic emissions, also shows comparatively low SO₄ emissions (ca 95 Tg / yr). This could be due to a too inefficient conversion of SO₂ (and DMS) to SO₄ in GEOS. In terms of BC emissions, models agree well which is not surprising, since most models used the CMIP6 BC emission inventories (see supplement 1). Note that for this analysis the satellite data was used in the original 1° × 1° resolution.ECLIPSE BC, SO_x, NO_x and NH₃ emissions, used by EMEP, are somewhat lower compared to CMIP6. Emissions of NO₃ show a remarkable high diversity of 286% (see Fig.
- 2) with values ranging from 5.4 Tg / yr (TM5) up to 128 Tg / yr (GEOS), which is on the same order of magnitude as SO_4 and OA. Natural sources of NO_x (soil, lightning) and formation of secondary NO_3 with ammonium, dust and sea salt provide several degrees of freedom for model formulation. NO_3 has only been implemented in some models in recent years and was not considered in the AP1 simulations.
- The results of this analysis lifetimes (computed from burden and total deposition) are shown in Figure ?? and reveal generally high correlation with AERONET measurements (R > 0.80). In terms of NMB, AATSR-SU and the MERGED-FMI product show slight underestimations (NMB ≈ -5%) while MODIS Aqua and Terra yield slightly overestimated AODs of approximately +9% and +18the second panel in Table 3. Associated diversities are illustrated in Figure 2. OA has the longest lifetime with 6 days, followed by BC (5.5 d), SO₄ (4.9 d), NO₃ (3.9 d), DU (3.7 d) and SS (0.56 d). The largest differences compared to AP1 are found for BC which shows a decrease in lifetime of ca 15%, and in SO₄ and SS, showing increased

lifetimes of ca 20% and 37%, respectively.

We remark that this analysis is biased by the uneven distribution of AERONET sites (highest density in Europe and North America, Fig. 1) and that problematic regions in the satellite retrievals. In addition, the latter two species show a notable increase in lifetime diversity compared to AP1. In the case of sulphate, the increased variability is in agreement with the changes in

- 755 emissions discussed above (i.e., it may reflect an increase in the natural fraction). This is consistent with the increase in SO₄ lifetime compared to AP1, since DMS derived and volcanic emissions are often released into the free troposphere, where the residence time is larger. For sea salt, the increased lifetime relative to AP1 could indicate a shift towards smaller particle sizes, but could also be due to differences in assumptions about water uptake. These changes in SS lifetime and lifetime diversity will impact the conversion to optical properties, as shall be seen below. The decreased BC lifetime may be due to changes in the
- The treatment of BC in the models. For instance, in AP1 most models assumed external mixing (see Tab. 2 in Textor et al., 2006) while many models in AP3 treat BC as an internal mixture (e.g. Sahara)may not be well represented in this comparison., with hygroscopic SO_4 , see supplement 1). This may impact the effective hygroscopicity of aged BC and thus, the wet scavenging efficiency. Earlier studies also showed that BC in older models was likely transported too efficiently to the upper troposphere, with a too long lifetime as a consequence (Samset et al., 2014). The dust lifetime is slightly decreased compared to AP1 and,
- 765 with ca 56% the associated inter-model diversity is slightly increased. The fact that the DU lifetime diversity is larger than the diversity in DU emissions and burden indicates differences in the models regarding dust size assumptions. For instance, ECMWF-IFS shows the lowest lifetimes both for dust and sea salt, which is subject of an ongoing development³. In the case of sea salt, the short lifetime for ECMWF-IFS is related to the emission scheme used (based on Grythe et al. (2014)), resulting

³Personal communication with Z. Kipling and S. Rémy

in too coarse SS particles. In the case of dust, the scheme used by ECMWF-IFS is based on Nabat et al. (2012) and tends

to produce too much dust. In addition, it is possible that the DU emission size distribution (which is based on Kok, 2011) is coarser than in the other models (which is also reflected by a below average DU MEC).

In case of the AATSR-SU data, the retrieval includes a conservative cloud mask utilising thermal channels in additional to optical, The modelled atmospheric mass burdens are shown in the third row of Table 3. They are essentially a result of their "emissions", and lifetimes, discussed in the previous paragraphs. Consistently, the largest burdens are found for dust, followed

775 "emissions", and lifetimes, discussed in the previous paragraphs. Consistently, the largest burdens are found for dust, followed by SS, OA, and thereby avoids retrieval near cloud edges. Evaluation under aerosol CCI of six data-sets showed AATSR and SeaWifs exhibited the lowest bias (with SeaWifs) with respect to ocean and coastal sun photometers (Popp et al. (2016)).

4 Results

SO₄, while burdens for NO₃ and BC are small.

- The section the results from the model evaluation are presented, starting with an overview of annual averaged emissions, burdens, lifetimes, MECs and ODs for each aerosol species and model, followed by a discussion of the results from the ensemble model and regional model diversity. Finally, the results of the optical property evaluation are presented. This is followed by Compared to AP1, a discussion section for the results from the individual models. notable decrease of ca 40% in the BC burden is found, which is in agreement with decreased emissions and lifetimes discussed above. However, the
- 785 associated variability of the simulated BC burdens (ca 50%) is comparatively large. Since the BC emissions are relatively harmonised among the models, this variability is likely due to (ageing/mixing induced) differences in the BC removal efficiencies, particularly in strong source regions such as China and India (e.g., Riemer et al., 2009; Matsui et al., 2018). The sea salt burden is increased by ca 36% relative to AP1. This can be explained by the increased lifetime compared to AP1, suggesting a shift towards smaller particle sizes for sea salt. The observed high diversities in sea salt emissions (54%) and
- lifetime (92%) have a compensating effect on the variability in the associated burden, which is only 38% (see Fig. 2f). This indicates discrepancies in the assumptions about the associated size distributions of this predominately coarse aerosol (see e.g., ECMWF-IFS vs. NorESM in Tab. 3). However, not all models show such a "compensation effect" of emissions and lifetime for SS. For instance, both SPRINTARS and EMEP exhibit below median SS emissions and lifetimes, resulting in the lowest SS burdens for these models. The lower SS emissions of EMEP are due to the fact that only SS particles below 10 μm are
- 795 <u>simulated by the model.</u>

3.1 Modelled emissions, burdens, lifetimes, MECs and ODs

Figures ??, ??, and ?? show the global annual average of emissions, lifetimes, burdens and MECs, for each aerosol species and for each model, respectively. The colors in the performance charts are applied row-wise in order to highlight differences between the models. Also included in each plot are mean, median and diversity (IQR)for each species. Note that The dust burden is decreased by an 20% compared to AP1, which can be explained by the lower AP3 emissions and lifetimes. The

800 burden is decreased by ca 20% compared to AP1, which can be explained by the lower AP3 emissions and lifetimes. The

associated diversity in dust burden is approximately the same as for the emissions and lifetimes (i.e., unlike for SS, for dust no "compensating" effect of emissions and lifetimes can be seen, see Fig. 2).

The sulphate burden is only slightly decreased relative to AP1, a result of the decrease in emissions, which is almost counterbalanced by the increased lifetime. In terms of diversity, however, inter-model differences in SO_4 "emissions" and

805 lifetimes have an enhancing effect on the associated SO_4 burden diversity (72%). Interestingly, models that have below average SO_4 emissions also tend to have below average SO_4 lifetimes and vice versa (in contrast to sea salt, where a compensating effect was observed).

The OA burden is slightly increased compared to AP1, however, the variability is comparable. Because the OA lifetime decreases slightly between AP3 and AP1, changes in the burden are due to differences in emissions. This is difficult to tease

810 out as organic aerosol treatment and the latter are computed directly from the provided table values for each species and model, and not using the ensemble mean and median fields (which were used for the inter-comparison with the measurements shown in Figs. ?? and ??). In addition, Figure ?? shows corresponding averages for the individual optical depths (ODs) of each aerosol species inclusion of different sources is very different than it was in AP1.

NO₃ shows the highest diversity in burden (>300%), with values ranging from 0.08 Tg (OsloCTM3) to 0.93 Tg (GEOS).

- 815 This is likely associated with the wide range in the corresponding emissions, indicating disagreement in the formation of nitrate aerosol. According to the AeroCom phase III nitrate experiment, the majority of NO₃ formed in the atmosphere is associated with atmospheric dust and sea salt in the coarse mode (Bian et al., 2017). Differences in the association of NO₃ with coarse particles and, thus, nitrate particle size can explain the large diversity in NO₃ lifetime (ranging from 2.5 to 10.4 days). For instance, TM5 and EC-Earth show the longest NO₃ lifetimes, which is likely due to the fact that nitrate is described only by
- 820 its total mass and assumed to be present only in the soluble accumulation mode (van Noije et al., 2014). The comparatively small nitrate burden of OsloCTM3 (0.08 Tg) is because the reported NO_3 diagnostics only include fine nitrate and coarse NO_3 particles are included in the sea salt diagnostics, however, with almost no impact on the burden of sea salt. A careful budget analysis for nitrate would need more information on its chemical formation and particle size distribution and most importantly, more consistency among the models in the reported nitrate diagnostics.

825 3.0.1 Diversity in optical properties: speciated MECs, MACs and ODs

Global annual average MEC values for each species are provided in the 5th part of Table 3. They represent here the link between *dry* aerosol mass and its size distribution and the resulting *ambient* air total light extinction (i.e. BC, DU, OA,) and SS; and their sum) as well as OD, reported clear-sky (CS) and, absorption + scattering of the water containing particles) associated with a given species. Since the MEC values here were computed via OD_{i.amb}/or all-sky AOD (where provided).

830 In the following we briefly discuss the main results from this global perspective of Burden_{i,dry} (*i* denoting an aerosol species) they represent the whole atmospheric column and they include the effects of water uptake (while the species specific burden values represent just the dry aerosol component). Because the MEC (and MAC) values reported here will include the water contribution to the species OD they will be larger for hygroscopic aerosol such as sea salt or sulfate than the corresponding values for dry conditions (e.g., Table 5 in Hand and Malm, 2007). This is partly balanced by smaller specific

- 835 extinction for larger particles. Notably, the aerosol life-cycles. The focus in this section will be on the discussion of the ensemble median results and the corresponding model-spread in percent, derived via the provided MEDIAN and δ_{IQR} . Results of individual models are discussed in Sect. ??. model-derived MECs for dust shown in Table 3, are fairly consistent with measurement-based dry mass scattering efficiencies for dust (Hand and Malm, 2007). This consistency is reassuring because dust is typically considered to be hydrophobic in models, meaning there shouldn't be a large discrepancy between MEC for
- 840 dry and ambient conditions. Note that the split of total AOD into speciated ODs is not trivial for internally mixed aerosols. The general recommendation for such diagnostics is to split proportional to the dry volume fractions of the species. The latter may result in too much water uptake associated with hydrophobic particle fractions. This can have implications for the computed BC MECs as discussed below.

The median emissions (Fig. ??) are highest for sea salt (5090 Tg/yr), followed by dust (1430 Tgspecies specific MECs found

in this AP3 analysis are mostly similar to those reported for AP1. The largest difference is in the DU MEC (ca 20% decrease), suggesting that the AP3 models tend to simulate larger dust particles compared to AP1. This is consistent with the observed slight reduction in DU lifetime for AP3. AP1 models likely simulated too fine (or a too large fine fraction) dust particles as suggested by comparisons with AE by Huneeus et al. (2011). As shall be seen in Sect. 4, the AP3 models considered here still tend to overestimate AE in dust dominated regions. This combination of results suggests that the simulation of dust aerosol size has been improved since AP1, however, dust particles are likely still too fine.

MECs of sea salt and sulphate are comparable with AP1, however, both show a decreased inter-model variability. While

860

- one could conclude that this may suggest better agreement in the modelled size distributions of SS and SO_4 , the dramatic increase in the variability of their lifetimes suggests differently. The better agreement in MEC may also be linked with similar assumptions in microphysical properties in AP3 (e.g., refractive index or density, see supplement 1) or assumptions about (and
- 855 impacts of) hygroscopic growth of these hydrophilic species (e.g., for SS most of the light extinction is linked to high water uptake). However, from the broad diagnostic overview provided here, it is difficult to understand what drives this behaviour.

 NO_3 shows the highest MEC variability of all species, though, again, only 9 models consider this species. However, compared to the spread in its burden and emissions, the NO₃ MEC diversity is "small" (see Fig. 2) and is similar to the corresponding lifetime diversity (<100%). TM5 and EC-Earth exhibit the largest NO_3 MECs because in these models the particles are associated with the optically more efficient accumulation mode. Other models (such as GEOS) appear to have their NO_3 more tied with DU and SS and exhibit smaller NO_3 MECs.

The BC MEC values exhibit a diversity of ca 20%, smaller than in AP1, however several models were excluded from the ensemble calculations. The MAC values (shown in the 6th part of Table 3) represent the absorptive fraction of MEC (computed as AAOD_{i,anub}/yr) Burden_{i,dry}) and should fulfill by definition: MAC_i < MEC_i. However, for some models (ECHAM-HAM,
 sulphur species (192 TgECHAM-SALSA, EMEP and INCA), BC MAC values are up to 3 times larger than their BC MEC values. TM5 and EC-Earth did not submit AAOD_{BC} but AAOD_{tot} and BC MAC was estimated for those two models via MAC_{BC}* = AAOD_{tot} /yr), OA (78 Tg Burden_{BC} (assuming BC was the dominant absorber). This resulted in estimated MAC values exceeding 12 m²/yr, primary) g, even in regions where the weak absorbers DU and OA do not add absorption. BC MECs for TM5 and BC (10 TgEC-Earth are around 6 m²/yr). Models agree well in their BC emissions, which is expected since most

- 870 models used the CMIP6 BC emission inventories (see AeroCom optics questionnaire (supplementary material)). g (comparable to the ECHAM MEC values), meaning the BC MACs for these two models also exceed their MECs. In the case of EMEP, the MAC_i > MEC_i discrepancy results from an inconsistency between the choice of the prescribed MAC literature value and the ways MECs are computed in the model (for details, see Sect. 4 in supplement 2). In the case of INCA, the apparent inconsistency between BC MEC and MAC is related to BC absorption enhancement for external mixing which is based
- 875 on Wang et al. (2016) and improves the agreement with BC MAC observations. However, the BC absorption enhancement effects are only considered for the computation of $AAOD_{BC}$, not OD_{BC} . For the other affected models (EC-Earth, TM5, ECHAM-HAM, ECHAM-SALSA) discrepancies between BC MAC and MEC likely arise from nonlinear internal mixing rules that may not be properly accounted for when computing the component optical depths based on the species volume fraction, as recommended by the AeroCom protocol.⁴
- 880 These inconsistencies with calculating aerosol optical properties will primarily impact the BC OD and MEC estimates, because (a) aerosol absorption contributes a large fraction of the column extinction making up BC OD and (b) the OD due to absorbing aerosol is small (around 0.002) relative to the other (mostly scattering) species. The highest diversity is found for OA (64%) followed by sea salt (51%) and DMS (42%)and dust (32%). These differences are not surprising, since the emissions of these species are typically computed online (fully or partly) in inconsistent BC MAC, MEC and OD values for
- 885 the affected models are indicated in brackets in Table 2 and were excluded from the computation of the corresponding AP3 ensemble median and diversity estimates (and accordingly, also from Fig. 2). These inconsistencies may have already been an issue for some models in Kinne et al. (2006) where similar recommendations related to the computation of component ODs were given by AeroCom.

As can be seen in Table 3, BC is by far the most efficient absorbing species, suggesting a BC MAC value of around $8.5m^2/g - 1000$

890 almost 2 orders of magnitude more efficient than dust or OA at 550 nm. This value is slightly larger than MAC values suggested for fresh BC based on extensive surveys of fresh BC MAC values reported in the models and hence, are highly dependent on the individual parameterisations applied (see AeroCom optics questionnaire (supplementary material)).

The lifetimes shown in Fig. ?? were computed using the provided burdens (Fig. ??) and total deposition for each variable (not shown). BC lifetime is around 5.5 days and, in contrast to the BC emissions, shows a rather high diversity of 42% between the models. The modelled lifetimes show the largest diversity with literature, for instance, 7.5±1.2 m²/g (at 550 nm) recommended by Bond and Bergstrom (2006) or 8.0±0.7 m²/g from Liu et al. (2020). Bond and Bergstrom (2006) note that for aged BC, the MAC may be enhanced by 35 – 80%, with the enhancement due to coatings as well as changes in morphology. They suggest BC MAC values in the range 9 – 12 m²/g for ambient BC. Measurements across Europe indicate MAC values ranging between

⁴For the two ECHAM models, OD_{BC} and $AAOD_{BC}$ are diagnosed as follows: OD_{BC} is computed from the BC volume fraction (dV_{BC}) , relative to other abundant species (i.e., $OD_{BC} = OD_{tot} * dV_{BC}$), while for the computation of $AAOD_{BC}$, dV_{BC} is weighted by the respective imaginary refr. indices of all species in the mixture. For instance, if BC is the only absorber, then $AAOD_{tot}=AAOD_{BC}$. $AAOD_{tot}$, however, is computed in the model via $OD_{tot} * (1-SSA)$, where SSA is the single-scattering-albedo of the mixture. Then, if 1-dV_{BC} > SSA it follows that $AAOD_{BC} > OD_{BC}$ (i.e., in cases where the SSA of the mixture is smaller than the scattering volume fraction). A correction suggested by H. Kokkola (through pers. communication) would be to compute the scattering component as $SCOD_{tot}=OD_{tot}*SSA$ and then computing $SCOD_{BC}$ accordingly by weighting its volume fraction with the real part of the refr. indices (of all species in the mixture), then $OD_{BC}=SCOD_{BC}+AAOD_{BC}$.

2.7 days (GEOS) up to around 4.3 – 22.7 m²/g (at 637 nm) as summarised by Zanatta et al. (2016). They propose a value of

900 $10\frac{\text{days} (\text{TM5 and EC-Earth})}{\text{and OA have lifetimes of around 5-} m^2/g (at 637 nm) to be representative for a mixed boundary layer at European background sites, which would translate to 11.6 m²/g at 550 nm assuming AAE=1.$

The range in species dependent MACs in Table 3 is ca 50% after excluding the models mentioned above. Given the harmonised BC emissions used, this diversity indicates differences in the BC treatment in the models. The BC MAC values derived from the AP3 models reflect uncertainties related to assumptions about optical properties of aged BC (e.g., absorption enhancement)

- but could also indicate uncertainties related to the size and density (which varies between 1 2.3 g/m³ among the AP3 models) and / or assumptions about the refractive index m = n + ik of BC (see supplement 1). We suspect though, that the largest differences are due to assumptions about BC aging (e.g., coatings) and 6 days, respectively (and diversities of around 30%). The ensemble median lifetimes of dust and sea salt are around 0.6 and 3.7 days, respectively. However, the individual models tend to show show high variability in these (globally dominant) species with diversities of around 100% and 52% for the
- 910 residence times of dust and sea saltprocesses affecting its size distribution since models using nearly the same m exhibit large differences in MAC. For instance, GEOS, OsloCTM3 and SPRINTARS all use $m \approx 1.75 \pm 0.45i$ at 550 nm but show MACs of 7.8, 13.0 and 3.1 m²/g, respectively.

The modelled atmospheric burdens for each species are shown in Fig. ??). They mostly reflect the corresponding diversities that could be associated with their main sources (emissions) and sinks (deposition). Dust and sea salt burdens The variations in

915 MAC may also impact BC lifetime - coatings may result in lifetime variations by changing the hygroscopicity of BC - however, we find no clear relationship between BC lifetime and MAC.

A comparison of MECs and MACs for DU and OA suggests that ca 5% and ca 2% of their total extinction is due to absorption at 550 nm, respectively. DU MAC and MEC diversities are similar. However, we find a slightly larger increase in DU MAC diversity compared to DU MEC, when considering a consistent model ensemble. This is likely linked to a larger

- 920 disagreement in the DU imaginary refractive index (*ik*) compared to its real part (see supplement 1). It reflects uncertainties related to assumptions about dust absorptive properties, which depend on mineralogy and size, for instance, show considerable variability among the models, with median values of 15 ± 8 Tg and 9 ± 3.4 Tg, respectively. The highest relative hematite and soot content (Kandler et al., 2007). MAC diversity for OA also exhibits an apparent increase compared to MEC and shows considerable variability in *ik*, based on a small model ensemble.
- 925 The following paragraphs focus on the discussion of speciated optical depths, that is, (1) how they result from the above discussed parameters, (2) how they contribute to total AOD and (3) how they compare with the AP1 data.

SS makes the largest contribution to total AOD with a median value of 0.044, followed by SO_4 (0.035), OA (0.022), DU (0.021) and to a lesser degree, NO₃ (0.005) and BC (0.002) (see Tab. 3). The largest diversity is found for optical depth of OA, followed by SO_4 , SS, NO₃(among the 8 models accounting for it) and burdens range between 0.08 Tg (OsloCTM3)and

930 0.93 Tg (GEOS)The modelled BC burdens also exhibit a considerable spread of around 65% with a median value of 0.16 Tg (Fig. ??). Since the BC emissions are relatively harmonised among the models, the variability in the BC burden is likely due to (ageing / mixing induced) differences in the BC deposition efficiencies, particularly in strong source regions such as China and India (e, DU and BC. Figure 2 illustrates how OD diversity is linked to diversities in emissions, lifetime and MEC. The

diversity in SS ODs is almost twice as large as SS diversities in burden and MEC. This reflects that models with a high SS mass

- 935 burden also tend to have an above average MEC (and vice versa), possibly, simulating smaller and thus, optically more efficient SS particle sizes (e.g., NorESM2). However, the wide spread in SS parameters is likely also linked to varying contributions of water due to different parameterisations of SS hygroscopicity (see discussion in Burgos et al., 2020). The role of water adds another level of complexity to the relationship between aerosol lifecycle and optical properties for hygroscopic species such as SS.
- Similar to sea salt, an "amplifying" combination of burden and MEC diversities is found for OA. As can be seen in Figure 2d, this results in a prominent departure from the median for all modelled OA ODs, with none of the models being close to the ensemble median. Sulphate ODs exhibit a similar spread as the sulphate burdens, while the range in MEC is smaller. This suggests that models tend to agree better in their simulated SO₄ sizes (also supported by comparatively low diversity in lifetime) and optical properties. It further suggests that the disagreement in OD is primarily related to uncertainties in the SO₄
 "emissions". g. Riemer et al. (2009), Matsui et al. (2018)).
- MECs are shown The diversity in BC and dust OD is smaller than the associated burden and MEC diversities for these two species (see Fig. 2a), indicating a compensating effect on OD variability. We can partly explain this. For instance, in Fig. ?? and were calculated NorESM2 a low DU burden (a result of low emissions and lifetimes) is compensated by a large MEC, resulting in an close-to average dust OD. This appears to be contradictory, as one would expect that a shorter DU lifetime would
- 950 reflect larger particles and thus a small MEC. However, NorESM2 assumes some hygroscopicity for dust (i.e., DU hygroscopic growth factor $\kappa = 0.069$, see supplement 1), which may lead to an efficient wet-removal pathway also for small DU particles. In addition, MEC values are not only size dependent. However, variations in the dust refractive index shows good agreement in its real part among the models (see supplement 1).
- The speciated ODs are also shown in Figure 3 as a stacked bar-chart. The plot also includes median values from AP1 from 855 Kinne et al. (2006) as well as global estimates of the total AOD from AERONET and the merged satellite product. The latter were scaled to represent global averages, using a scaling factor that was computed from ENS-MED, by dividing the OD (Fig. ??) for each species by the corresponding burden (Fig. ??). The two ECHAM models (indicated with a star) were not considered to compute mean, median and IQR, since the OD_i values were diagnosed at dry conditions, while all other models reported ambient speciated ODs. This explains the comparatively low MECs for these two models. Diversities of MECs are of the same
- 960 order of magnitude as for the burdens, indicating different treatment related to the underlying assumptions that determine the extinction efficiencies for each species. The largest diversities in MEC are found for BC, DU and (global average of ENS-MED with its average when co-located with AERONET and the merged satellite, respectively. On average, the global total AOD from models has not substantially changed from AP1 to AP3, although it is lower than the AOD from both observational data sets (details will be discussed below in Sect. 4). Most notable compared to AP1 is a shift in the natural contribution from SS and
- 965 DU, respectively. AP3 models show a shift towards more sea salt, with SS making up 2/3 of the natural OD. Interestingly, this shift is likely not due to the changes in the emitted mass of these species (since SS emissions have decreased more than DU since AP1), but likely originate in changes in the simulated size distributions, with changes in DU and SS lifetimes and MECs as discussed above. Figure 3 also shows that the BC OD is decreased by a factor of 2 compared to AP1. This marks a

substantial change in this important species and can be explained by the combination of decreased BC emissions and lifetime

970 (and thus, burden) and MECs. This manifests in substantial model underestimates of surface absorption coefficients, as shall be seen below in Section 4.

Based on 8 models, water makes up between approximately 40 50%) - 65% of the total AOD. The water OD is correlated with SS OD (R=0.63; and R=0.80 when SPRINTARS is excluded). SPRINTARS appears to be an outlier, its SS OD is only 0.013, related to issues described above. Also the fraction of water (water OD / total OD) to fraction of sea salt (SS OD / total

975 OD) exhibits a strong correlation of R=0.81 (SPRINTARS excluded). Textor et al. (2006) notes that water in AP1 makes up about 1/2 of the aerosol mass and also observed a relationship between water and SS mass.

Ultimately, the model spread in the burdens, combined with the diversity in MECs results in a considerable large diversity in the speciated ODs shown in Fig. ??. For most species, the model spread exceeds Also indicated on Figure 3 is whether models provided clear-sky (CS) or all-sky (AS) AOD. Five models report only AS AOD, six models report only CS AOD and

- 980 three models report both. The ensemble median AOD values (CS AOD=0.132 and AS AOD=0.128) are slightly higher than the median AOD value of approximately 0.125 reported in Kinne et al. (2006) (see Fig. 1 therein). AOD diagnostics used by Kinne et al. (2006) were based on undocumented clear-sky/all-sky assumptions so it is difficult to explain the slight increase in AOD. The comparison with remote sensing observations discussed in the following Section 4 is performed using CS AOD if available. For a given model, CS AOD is smaller than AS AOD (where both are available). This is expected, since CS AOD
- 985 is not as much affected by hygroscopic growth, while AS AOD reflects conditions which include supersaturated environments needed for cloud formation. This is also reflected in the larger diversity in AS AODs compared to CS AODs (see Tab. 3) as models utilize different RH ceilings for AS conditions. For the three models that report both AS and CS AOD, the largest difference is found for SPRINTARS where the CS AOD is almost 50% (i. e. SS, , smaller compared to the AS AOD, while the results from NorESM and ECHAM-HAM suggest circa 10% lower AOD under clear sky conditions. This could indicate that
- 990 SPRINTARS exhibits higher global cloud coverage or increased impacts of hygroscopic growth. Its water AOD, however, is close to the median.

EMEP and ECMWF-IFS reported AS AOD; surprisingly they are among the models with the lowest AOD. In the case of ECMWF-IFS this seems to be mostly due to too small OA and SO_4 ODs. SO_4 is likely too low due to missing volcanic degassing (see supplement 1). Low OA OD could be related to size and/or assumptions about hygroscopicity, as both MECs

and lifetimes are comparatively low. In the case of EMEP, on the other hand, the low AOD is a combination of too little SS, OA) while, interestingly, SO_4 and DU which is, to a certain degree, compensated for by its large OA OD. The latter is a result of comparatively strong "emissions" (POA+SOA). EMEP also simulates the largest NO_3 OD (due to contribution from both fine and coarse nitrate).

Finally, it is interesting to note that, despite the spread in OD values for different species, the simulated total AOD (both

1000 CS and AS, where provided) indicate indicates much better agreement between among the models (11% in within 8% for CS AOD and 17% in AS). Note that for these estimates of diversity, the two ECHAM models were excluded, since they reported dry speciated ODs. With values of around 0.13, the total AODs (CS and AS) and agree well with the values found in Kinne et al. (2006)(see Fig. 1 therein)24% for AS) – a characteristic that has not changed since Kinne et al. (2006).

3.1 Modelled annual global distributions of optical properties and their diversity

- 1005 Figure 4 shows global maps from the ensemble median (Sect. 2.3.1) for each variable (left) and corresponding diversities (right). Also included are annual average values of the corresponding ground observationsat the individual site locations. The legends in each plot provide average values of The diversity maps provide insights into the regional model-spread. They are useful to identify regions where models tend to disagree. This may also help to explain differences between models and observations. The AOD and AE diversity maps from this study can be compared with the diversity maps for these properties
- 1010 presented in Kinne et al. (2006) (their Figure 4). Kinne et al. (2006) report the largest diversity in AOD at high latitudes and over central Asia which is similar to what we see here. The AE diversity map in Kinne et al. (2006) is also similar the highest AE diversity is observed over northern Africa and the southern ocean. They note that these are regions dominated by dust and sea salt respectively, suggesting that assumptions about aerosol size for these species are important to properly simulate AE. The high diversity in the ambient column AE in the South Pacific, is likely due to both differences in assumptions about the
- 1015 initial SS size distributions and in the simulated water uptake of SS.

Figure 4 also presents maps for AOD_f and AOD_c. On a global scale AOD_f dominates the total AOD and the median values indicate a fine-mode fraction of ca 60% (note that the total AOD field includes one more model than AOD_f and AOD_c, thus here AOD≠AOD_f+AOD_c, see Sect. 2.3.1). AOD_f is highest over regions with strong anthropogenic sources (e.g., China, India, eastern US, etc.) but also in regions associated with biogenic emissions and / or biomass burning (e.g., Amazonia, central Africa). The diversity in AOD_f is highest in remote regions (high latitudes and high altitudes such as the Andes and the ensemble model (global and at obs.sites) and the observations (at obs.sites).

- The latter provide an indication of potential biases between model and observations and how representative the observation locations are with respect to the whole globe.AODsHimalayas). This could be due to differences in aging and removal processes affecting long range transport, or differences in local sources (e.g., few models include oceanic POA emissions). However, it
- 1025 could also be affected by "inter-model noise", for instance, show an annual average value of 0.21 at the AERONET site locations, while the ensemble model shows a corresponding value of 0.16, indicating an absolute AODbias of about 0.05 (or 24%)because these regions typically show smaller burdens, or because of differences in the methods that models use to diagnose AOD_f and AOD_c (e.g., some models use the dry radius for the split, others the ambient radius , see supplement 1, and discussion below in Sect. 4). Another region with low AOD_f and high diversity is the Bay of Bengal and the Indian Ocean,
- 1030 likely linked to differences in the modelling of the mass outflow from the heavily polluted Indo-Gangetic Plain, in combination with complex meteorology prevalent in this region (e.g., Pan et al., 2015).

The plotted diversity maps provide insights into the regional model-spread. These may be useful, for instance, to identify regionswhere models tend to disagree which ultimately may help to explain differences observed when comparing the models with observations (which may be performed in different regions due Outside of high latitudes, AOD_c is most diverse over

1035 China, southern Africa, the eastern US and western South America. AOD_c exhibits the opposite pattern as AE, which is expected as AE is inversely related to size. However, the patterns in diversity for AE and AOD_c are different. Outside of high latitudes, the AE diversity is highest in regions dominated by natural aerosol, while the highest AOD_c diversity occurs

primarily over anthropogenic source regions. Both AE and AOD_c show a high diversity over the Sahara, which is consistent with the considerable variability related to the DU lifecycle and optical properties, discussed in the previous section. For

- 1040 remote oceanic regions, the coarse AOD shows much less diversity than AE. This indicates large diversity related to SS size assumptions and modelling. For instance, an increased variability in AOD_f is found in the South Pacific, where AE also shows high inter-model variability but AOD_c diversity is rather small. This again highlights the need for the models to lack of spatial coverage). reevaluate fine and coarse SS, due to the increased extinction efficiency at finer particle sizes, even though most of the SS mass resides in the coarse mode. China is also one of the regions showing considerable diversity in AOD_c . This could
- 1045 be related to dust storms, which regularly affect China (e.g., Sun et al., 2001). The fact that models tend to agree in AE in this region indicates similar assumptions in the modelled dust size. However, this needs further investigation due to the complex relationship between AE and size (Schuster et al., 2006), particularly for regions such as China, which are affected both by fine (anthropogenic) and coarse (natural) aerosol (also indicated by the AE \approx 1.25 prevalent over China).

The overall highest diversity , for instance, is found for the simulated surface in situ aerosol absorption coefficientsand is particularly prominent, in particular in Amazonia, a region of substantial regular biomass burning events ((e.g., Rissler et al., 2006)

particularly prominent, in particular in Amazonia, a region of substantial regular biomass burning events ((e.g., Rissler et al., 2006), which were peaking in early September in 2010) and also new particle formation (NPF) events from biogenic emissions(see Mortier et al., 2006).
 Reasons for these differences may be a combination of the different treatments of SOA formation (and absorptive properties of OA), or potential differences in the emission altitudes (see AeroCom optics questionnaire (supplementary material) of smoke

plumes (see supplement 1 for details). The diversity in in-situ simulated in situ surface absorption is also high in the South

1055 Pacific / Antarctica and Australia, which is also Australia, another region affected by regular biomass burning events as well as dust emissions. Interestingly, models tend to agree in major source regions such as China and India(, simulating low diversity in surf. absorption). aerosol absorption at the surface.

The dust dominated Sahara region also shows considerable diversity in <u>simulated</u> surface absorption but little diversity in <u>simulated</u> surface scattering. This is an indication of We explain this by noting the considerable differences in the treatment of

- 1060 dust absorption optical properties. The increased assumed imaginary indices for dust at 550 nm (see supplement 1). The high diversity in AE in this region also suggests differences in dust size distribution, which may, to a certain degree, be the simulated dust size distributions and is linked with the increased diversity seen in , which reflects the diversity between the models found AOD_c. These results reflect the diversity among models for dust emissions, burdens, lifetimes and MECs (Figs. ?? and ??, ??, ??) . Explaining these dust related differences in detail is beyond the scope of this work, and needs further investigationand
- 1065 lifetimes, as well as MECs and MACs (Tab. 3) as suggested already by Huneeus et al. (2011).

Another notable region is the (comparatively clean) Several clean regions exhibit high diversity in one or more variables. The South Pacific and Antarctica which shows a belt exhibit a zone of high diversity in simulated surface absorption (but not scattering) and AE and considerable diversity in (over land) in addition to the diversity in AE, AOD_f discussed above. In addition, models tend to show high variability in AOD_c over land in this remote area. This behaviour may arise from a

1070 combination of differences in sources, lifetimes and points to considerable differences in long range transport of the aerosol (e.g. dust shows > 50% diversity in lifetime, see also Li et al. (2008)). It may also be due to differences in the absorption

optical properties of OA (due to organic Ocean emissions), combined with potential differences in sea ice retreat. Most likely, it is a combination of all these effects.

Furthermore, elevated (e.g., dust exhibits > 50% diversity in lifetime, see also Li et al., 2008), or treatment of organic emissions

- 1075 from the ocean. Elevated and / or mountainous desert regions such as the Southern Peruvian and Northern Chilean Andes , Tibet and Tibet also show high diversity in . These regions however, associated with generally low AODs and thus such differences may not AOD_c. Unfortunately, most ground-based observations provide little or no coverage in these remote regions, where the models exhibit high diversity. These model results therefore lend support to the idea of expanding measurements in undersampled locations, in order to better evaluate models.
- 1080 While presenting model variability in terms of percentages sheds some light into differences into the modelling of the aerosol in remote and clean regions (with low aerosol loading), it gives equal weight to variability independent of the abundance of aerosol. However, the diversity in pristine areas is small in terms of absolute value and is thus unlikely to have a significant impact on the global radiation budget.

Unfortunately, most ground based observations (used in the following Sect. ?? to evaluate the individual models) provide 1085 little or no coverage in these remote regions, where the models show high diversity

4 Results and discussion - Optical properties evaluation

1090

This section presents and discusses the results from the optical properties evaluation, shown in Table 4 and Figures 5 & 6. Most of the discussion is based on the results from the ensemble model (ENS-MED). A detailed assessment of individual models on a regional or seasonal scale is beyond the scope of this paper. However, where there are clear outliers among the individual models we note model assumptions and attempt interpretation. Note that detailed results for each model and

observation data set are provided online via two web interfaces: a) the new interactive visualization (see Mortier et al., 2020c), including regional statistics and b) the old AeroCom web interface (see Schulz et al., 2020), obtained with a different analysis tool AeroCom (IDL based), which also allows interactive viewing of earlier AeroCom results.

Figure 5 shows global maps of annual mean biases retrieved when evaluating the ensemble AODs against the merged satellite product as well as biases established against AERONET AODs and the surface in-situ scattering measurements. The legend provides the network biases and correlation coefficients for each data-setco-locating the ensemble median model (ENS-MED) against some of the various observational parameters and datasets (see Sect. 2.3 for methods). Note that not all satellite datasets used later are shown in this figure.

South-east Asia appears to be a region where modelled AOD is low (by about -40%) both compared to MERGED-FMI 1100 and to AERONET. It can also be clearly seen that the underestimated scattering (by 44% over all GAW stations) is mostly representative for Europe and The individual panes in the bias maps indicate that, in general, models simulate lower values for all parameters considered, even though there are some regions / locations where models overestimate the observations. It

should be noted that the differences in NMB for AOD_c and AE between AERONET and AATSR-SU for the models primarily reflect the respective biases found in the satellite assessment (i.e., ca -15% for AOD_c and +15% for AE, see Sect. 2.1.5). For

1105 instance, the US, where the site density is highest. These regions also show underestimated AERONET AODs, but only by about -14% (as can be seen in web visualisation, see Appendix ??).

Furthermore, models tend to underestimate scattering and AOD at the few available polar sites. This is also the case for surface absorption (e. g. Barrow, Alert, Tiksi and Neumeyer in Figs. ??, ??). AE results from AATSR-SU suggest that the ENS-MED model underestimates AE by 22% (Tab. 4). Figure 5 suggests that most of this bias is due to underestimations of AE

- 1110 over the oceans. This would suggest that the AP3 models tend to overestimate the ambient particle size in marine environments. However, models tend to yield rather diverse results at some of these stations, showing over and underestimations (e.g. absorption at Barrow, scattering at Neumeyer) the comparison with AERONET suggests that AE in AATSR-SU is overestimated by ca 14%, which may be linked with the different wavelength regimes used in AATSR and AERONET to retrieve AE (see Sects. 2.1.1 and 2.1.5, see also Schuster et al., 2006). The extent to which this bias translates to the oceans would need to
- 1115 be investigated to make a clear statement about whether current models are capable of simulating AE in marine environments. A detailed investigation of this is beyond the scope of this paper but desirable, given the important role of sea salt also for cloud formation, lifetime and optical properties in clean and remote marine environments, and the impact on assessing the indirect aerosol effect (e.g., Fossum et al., 2020, and references therein).

4.1 Results from optical properties evaluation

- 1120 Figures ?? and ?? show Table 4 shows performance matrices of the normalised mean bias (NMB, top) and the Pearson correlation coefficient , respectively(R, bottom). These are displayed for each model, variable and observation data-set used. The data set used, respectively. They represent averages over all site locations, for each observation platform. The evaluation results from the AeroCom ensemble mean and median (median and first and third quantile fields (ENS-MED, Q1, Q3, Sect. 2.3.1) are plotted displayed in the rightmost column, suggesting that overall, models moderately underestimate the selected optical
- 1125 properties, both when evaluated against ground and space-based remote sensing and in situ observations. In terms of bias (NMB) the mean model shows slightly better performance compared to the observations with up to +10% improvement (e.g. surface scattering and). In terms of correlation (Fig. ??) both median and mean show similar results. Relative biases between the different satellite AOD products mostly resemble the biases found when evaluating the satellites against AERONET (Fig. ??). However, columns. Note that for the AOD comparison with the MERGED-FMI product, 3 different results for NMB and
- 1130 R are provided, 1. representing the whole globe (denoted MERGED-FMI), 2. only over land (MERGED-FMI-LAND) and 3. only over ocean (MERGED-FMI-OCN). The land / ocean filtering was done using gridded masks provided by the Task Force on Hemispheric Transport of Air Pollution (TF HTAP), which are constrained to a latitude range from 60°S to 60°N.

While most models moderately underestimate the selected optical properties compared to the ground based observations, the satellites can show significantly different results as can be seen, for instance, in observations, they show surprisingly good

1135 agreement in terms of correlation (e.g., ENS-MED R is between 0.72 and 0.88, see lower pane in Tab. 4). This suggests that spatial and temporal variations are mostly captured by the models, despite underestimating the absolute magnitude of the investigated optical parameters. However, for some variables individual models perform quite poorly (e.g., AOD_c by

ECHAM-HAM and NorESM2). Compared to the from NorESM2 vs.AATSR-SU and AERONET, respectively. This is because the satellites generally show higher spatial coverage and are thus, also sensitive to the oceans (Fig.1).

1140 This demonstrates the usefulness of incorporating satellite data, even though these may carry larger uncertainties and representativity errors (Sect. ??). For instance, compared to AODs from the two MODIS instruments, models show the largest negative biases, which mostly reflects the results from the satellite evaluation (Sect. ??, Fig. ??, i.e. positive biases of +9% and +18% for Aqua and Terra against AERONET).-

The individual satellite AOD data sets, the models show high correlation (≥ 0.78 for ENS-MED) and differences in NMB 1145 for and AE between AERONET and AATSR-SU for the models mostly reflect the respective biases found in the satellite assessment (mostly reflect biases of the satellites established against AERONET (see Tab. 1, i.e. ca. -15% for and +15% for AE, the largest AOD underestimate of ca. -35% is found against the MODIS satellites).

The comparison with the surface in-situ data shows considerably large negative biases (and the lowest correlations) of -44%TM5 and -32%, for dry scattering and absorption, respectively at the GAW site locations (Fig. 1). In case of scattering, a

1150 small fraction (but likely not more than 20%) of these biases may be due to the fact that models reported at RH=0% and the observations are being performed at RH between 0% - 40%.

Correlation coefficients (Fig. ??)are generally high for the median model (> 0.6) but can be as low as 0.12 for individual model assessments.

5 Discussion of results from individual models

1155 In this section, the results shown in Figs. ??EC-Earth appear to be the best performing individual models when all optical variable comparisons are considered - ?? and Figs. ?? and ?? are discussed for each model individually. This includes a small introduction into each of the models.

4.1 CAM5-ATRAS

The Community Atmosphere Model version 5 (CAM5) with the Aerosol Two-dimensional bin module for foRmation and

1160 Aging Simulation (ATRAS)(Matsui (2017); Matsui and Mahowald (2017)) calculates the following atmospheric aerosol and chemistry processes: emissions, gas-phase chemistry, new particle formation, condensation of sulphate, nitrate, and organic aerosols, coagulation, cloud activation, aqueous-phase chemistry, dry and wet deposition, and aerosol-radiation and aerosol-cloud interactions. Aerosol particles from 1 nm to 10000 nm in dry diameter are represented with a two-dimensional sectional representation with 12 size bins and 8 BC mixing state bins. Meteorological nudging was used for temperature and wind

1165 fields in the free troposphere (they exhibit mostly low biases (<800 hPa)using the MERRA2 data.

The sources and burden of OA exceed the model ensemble by 90% and 50%, respectively (Figs.??, ??) $\pm 10\%$) and correlations close to the ensemble model for most parameters. The similarity in their results is not surprising, given the similarity of their model setups (see Sect. 4 in supplement 2). SPRINTARS is the model that most consistently underestimates observations, which is not surprising as it was the model that consistently had the lowest burdens and component ODs in

- 1170 the inter-model comparison, particularly for DU and SS (see Tab. 3). This is likely because the ATRAS model considers OA formation from semi-volatile and intermediate volatility organic compounds in addition to anthropogenic and biogenic VOCs based on the volatility basis set approach (Matsui et al. (2014a), Matsui et al. (2014b), Matsui (2017)). ECHAM-HAM, ECHAM-SALSA and NorESM2 are the models exhibiting the least correlation with the observational data sets, particularly for AOD_c and SC_{dry}, and over the oceans (Tab. 4). This could be due to their assumptions related to SS, discussed above.
- 1175 Indeed, there is a tendency for lower correlation in SC_{dry} at coastal GAW sites (see Mortier et al., 2020c). The burden is lower than the model ensemble, consistent with Matsui and Mahowald (2017). The burdens of BC, , SS, and DU and the lifetimes of all aerosol species in CAM5-ATRAS are similar to those in the model ensemble (Figs. ??-??). BC MEC islarger than the model ensemble by 40% (Fig. ??) likely because the ATRAS model calculates the enhancement of absorption by BC aging processes explicitly by resolving the BC mixing state with 8 bins (pure BC, BC-free, comparatively long SS lifetimes for these
- 1180 three models could result in larger particles (e.g., due to more swelling and less wet deposition) which will impact the ambient SS MECs and ODs (which are largest for these models, see Fig. 3). Also, these three models exhibit the highest H₂O AODs (among the models who submitted this diagnostic), which likely results from enhanced impacts of SS hygroscopic growth. Additionally, the importance of SS parameterisation is demonstrated by comparisons with Tegen et al. (2019) who find higher correlations and a good agreement with observations of size distributions for ECHAM-HAM, using a different SS emission
- 1185 parameterisation scheme than is used by ECHAM-HAM in this study. Tegen et al. (2019) also show a positive bias in AE in ECHAM-HAM compared to AERONET, suggesting the model simulates more fine particles than are observed, while the version of ECHAM-HAM used here shows a negative bias in AE compared to AERONET (and 6 internally-mixed BC bins). The value of BC MEC in this study $(9.5m^2g^{-1})$ is close to that in Matsui et al. (2018) $(10m^2g^{-1}AATSR-SU)$, suggesting too many coarse particles are being simulated. These results suggest that the versions of ECHAM-HAM, ECHAM-SALSA
- 1190 and NorESM2 used here overestimate the sea salt size, either due to the parameterisation (e.g., online emission and dry size distribution) or hygroscopicity or some combination of both. This could also explain the lower correlation in these models, which is particularly apparent over the oceans as can be seen from the different AOD results over land and ocean in Fig. 4 (i.e., from the MERGED-FMI product): RH over the oceans is, on average, likely more smoothly distributed in space and time than the actual SS emissions (which strongly depend on near surface wind speeds) which could have a smoothing effect on the spatio-temporal variability of the SS AOD signal, manifesting in a lower correlation compared to observations (which have
- 1195 the spatio-temporal variability of the SS AOD signal, manifesting in a lower correlation compared to observations (which have less swelling).

The biases of aerosol optical properties in CAM5-ATRAS are similar to those in the model ensemble (Fig. ??). Model simulations generally agree well with the observations for AOD (MERGED-FMI and AATSR-SU), coarse mode AOD (ATSR-SU), fine mode AOD(AERONET and AATSR-SU), and AE (AERONET) (Fig. ??) with correlation coefficients exceeding 0.6 (Fig.

1200 **??**). Simulated AOD is underestimated by 21% compared with AERONET AODFigure 6 presents another way of looking at the NMB. As with the top pane of Table 4, it is clear that the models have a tendency to underestimate the observations, with largest underestimates (>25%) for AOD_c (against AERONET) and by 33 - 57% compared with MODIS AOD, which is consistent with Matsui and Mahowald (2017). Scattering and absorption coefficients are also underestimated by 24 - 40% compared with the GAW observations. SC_{drv} (at GAW sites).

4.1 AOD, AOD_f and AOD_c

1210

Two configurations of the atmospheric composition model TM5 (Tracer Model 5) are included in this study (van Noije et al. (2014)): a standalone version of TM5, and an atmosphere-only version of This section presents and discusses the results for total AOD as well as AOD_f and AOD_c . The latter two diagnostics can provide insights into the differences associated with the modelling of anthropogenic aerosol (more fine dominated) and natural aerosol (dominated by dust and sea salt).

- As mentioned above, models typically underestimate observed total AOD regardless of measurement platform (i.e, by ca. 20% against AERONET and 16 37% against the various satellite products, see Tab. 4). However, the CMIP6 climate model EC-Earth3-AerChem (?). The standalone model is driven by meteorological and surface fields from the ERA-Interim reanalysis (Dee et al. (2011)), whereas in the climate model there is online interaction between TM5 and the atmospheric
- 1215 general circulation model, which is based on model cycle 36r4 of ECMWF's Integrated Forecasting System (IFS). The sets of meteorological and surface variables that drive TM5 are similar in both configurations. In the EC-Earth simulations analyzed in this study, sea surface temperatures and sea ice concentrations were prescribed using AMIP forcing fields provided for CMIP6; in addition, vorticity, divergence and surface pressure fields were nudged to ERA-Interim, using a Newtonian relaxation scheme with a time constant of 8h and 15 min in the whole atmosphere. TM5 uses the aerosol scheme M7 (Vignati et al. (2004)).
- 1220 which represents sulphate, black carbon, organic aerosols, sea salt and mineral dust with seven lognormal size distributions or modes. Aerosol components are assumed to be internally mixed inside the modes. The formation of secondary organic aerosols in the atmosphere is described following Bergman et al. (2020)). Ammonium-nitrate and methane sulphonic acid (MSA) are described by their total mass, and assumed to be present only in the soluble accumulation mode (see van Noije et al. (2014) for more details). TM5 has an interactive tropospheric chemistry scheme (Williams et al. (2017)) model AOD biases established
- 1225 against the four satellites mostly resemble the biases of each satellite found when compared with AERONET (see Tab. 2). Thus, we concentrate the discussion of satellite AOD results on the MERGED-FMI data set, which includes both MODIS and AATSR and shows good performance at AERONET sites (NMB=-5.5%, R=0.89). This is the reason why individual results from AATSR-SU and MODIS (Terra and Aqua) are not shown in Figure 6.
- The consistent AOD underestimate in the models means that they are either simulating too little mass loading in the column and/or underestimating the column optical extinction efficiency (e.g., related to assumptions about size, MEC and/or composition). Figure 6 suggests that most of the AOD bias is due to missing (or optically too inefficient) coarse aerosol, which also describes the aqueous-phase oxidation of dissolved sulphur dioxide in clouds. When calculating the dust source, TM5 does not include particles with dry diameter larger than 16 µm. This may explain why the mean emitted dust mass is smaller than in other models . Differences in 10 m wind speeds generally reduce the dust emissions from the main source regions in
- 1235 EC-Earth compared to TM5 (Fig. ??), leading to proportionally lower dust burdens. Sea salt emissions, on the other hand, which depend on 10 m wind speeds and sea surface temperatures, are very similar in the two models exhibits the largest diversity (in terms of bias) among the models and shows lower correlation than AOD and AOD_f (Tab. 4). The mean OA lifetime in However, as pointed out above, about half of the models computed AOD_f using the dry particle radius (see also

supplement 1). These models likely attribute some extinction to the fine mode that should be attributed to AOD_c if hygroscopic

swelling was accounted for. 1240

- The impact of this uncertainty was investigated using 7 models that submitted diagnostics of fine and total SS OD. Four of these, CAM5-ATRAS, GEOS, GISS-OMA and SPRINTARS computed AOD f based on dry radius, while the other three, EC-Earth, TM5 and ECMWF-IFS used the ambient particle radius. The first 4 models suggest a SS fine-mode fraction of circa 26% while the latter 3 models suggest 15%. Thus, circa 11% of the SS OD is erroneously attributed to AOD $_{f}$ when the dry radius is
- used to split fine and coarse mode. On a global scale, ca 33% of the total AOD is 9% longer than in TM5, and in both models are 1245 longer than in the other models. This may be in part due to the use of interactive chemistry in TM5 (and EC-Earth), which may lead to a depletion of oxidants over regions with high biogenic VOC emissions, thereby increasing their lifetime (?). The aerosol optical properties in TM5 are calculated based on Mie theory, where the mixing rules of Bruggeman and Maxwell-Garnett are applied as approximations of the refractive index of the internally mixed modes. The contributions of the individual aerosol
- 1250 components are estimated by distributing the resulting total ambient extinction of each mode over the individual dry aerosol components, using volume weighting. In this way the extinction due to due to SS (Tab. 3). Hence, circa 3% too much of the presence of water is associated with the other aerosol components. This will enhance the species AOD and MEC values for TM5 and EC-Earth compared to models in which the water contribution is not included, such as ECHAM-HAM-total AOD is attributed to AOD_f if the dry radius is used, assuming that SS is the dominant species affected by this error. This is a fair assumption as other hydrophilic species typically reside in the accumulation mode, and ECHAM-SALSA (Fig. ??). dust is
- 1255

assumed to be hydrophobic in most models.

Compared to the observations (Figs. ?? and ??), both TM5 and EC-Earth show similar performance and are generally in good agreement with observations in terms of bias (NMB), outperforming the ensemble values in all comparisons. Particularly, AOD and show good performance with biases smaller than 10% and high correlation (R < 0.79), with being slightly overestimated

- and AODbeing slightly underestimated. The latter is due to a slightly underestimated, both against AERONET and AATSR-SU, 1260 which is also reflected in the slightly positive AE bias. Comparison of the diagnosed dry scattering with surface in-situ measurements (at RH<40%) results in biases of -15%. The corresponding comparison of dry absorption, indicates a slightly better performance in TM5 (-2%) than in EC-Earth (-7%), which may be Thus, for the affected models, AOD_c is likely slightly shifted towards less negative biases by ca 3%, while AOD_f would show larger underestimations, accordingly. The diversity in
- AOD_c could be in parts due to the fact that the dust burden in TM5 is about 35% larger than in EC-Earth (and corresponding 1265 MECs are similar). The latter would also explain why biases are less negative (by about +10%) in TM5-different methodologies used in the models to determine the size threshold, however, this is not the sole explanation, since the AOD_f exhibits less diversity. Also, the diversity in AOD_c is consistent with the large diversity found for dust and sea salt aerosol among the models (see Sect. 3) and highlights the large uncertainties associated with the modelling of these natural aerosols (see Sect.
- 1270 3). Attempts to address these uncertainties are also reflected in the substantial changes in AP3 compared to EC-EarthAP1 (e.g., less and optically more inefficient dust and more sea salt in AP3). Figure 5b does not allow a clear statement related to over or underestimates of dust and sea salt. For instance, in several remote ocean regions ENS-MED exhibits positive biases (e.g., South Pacific), whereas in other regions slightly negative biases or good agreement are found. This suggests that,

overall, models manage to capture the overall magnitude of the sea salt contribution. Most of the coarse AOD bias seems to

1275 be associated with continental land-masses (e.g., SE-Asia, Arabian Peninsula, Fig. 5d), which is also indicated by the lower underestimate when comparing AOD_c against AATSR, instead of AERONET (Tab. 4).

4.2 ECHAM-HAM

We stress that the fact that models tend to match the magnitude of the AOD in sea salt dominated regions does not necessarily reflect that the processes leading to these AODs are represented correctly. In this context we refer once more to the large

diversity (and compensating effects) associated with SS emissions (computed online), burdens, lifetime and MECs and the substantial changes since AP1 (see Sect. 3 for details). This is further supported by the typically decreased correlations found when comparing models with satellite datasets (which "see" both oceans and continental land masses, see, for example, results from MERGED-FMI, LAND versus OCN in Tab. 4). For instance, ENS-MED exhibits comparatively low correlation in the SW and S-Pacific compared to the satellites (see online results, Mortier et al., 2020c). However, these regions are affected by high cloud coverage throughout the year, and thus the lower correlation may also be due to representation errors in the monthly satellite aggregates used.

The global aerosol-climate model ECHAM6.3-HAM2.3 (ECHAM-HAM in the following) is part of the fully coupled aerosol chemistry climate model ECHAM-HAMMOZ (Tegen et al. (2019), Schultz et al. (2018)). Aerosol microphysical processes in ECHAM-HAM are described with the modal M7 aerosol model (Vignati et al. (2004)) in contrast to ECHAM-SALSA

- 1290 which employs the sectional aerosol scheme SALSA (Kokkola et al. (2018)). The aerosol representation in ECHAM-HAM has been evaluated in Tegen et al. (2019) but using different aerosol emissions (different inventories for anthropogenic and biomass burning emissions as well as a different sea salt emission parameterisation). For the CTRL experiment the sea salt emission parameterisation from Guelle et al. (2001) was chosen, firstly because the one proposed by Long et al. (2011) and Sofiev et al. (2011) resulted in an underestimation of the sea saltconcentrations (Tegen et al. (2019)) and secondly, to be
- 1295 consistent with the CTRL setup of ECHAM-SALSA (Sect.??.However, this comes at the price of larger sea salt particles (on average), resulting in a slightly decreased correlation against AERONET compared to Tegen et al. (2019). The latter, however, may to a certain degree also be affected by different representation errors as Tegen et al. (2019) use 6-hourly data to colocate in time, while this study relies on monthly means (Sect. ??, particularly Tab.5).
- AOD over land is lower than in AERONET or MODIS observations (Fig. ??)which may be due to several reasons, for 1300 instance because is missing, too low emissions of OA or a misrepresentation of SOA (the OA burden in ECHAM-HAM is lower than in most other models, see Fig. ?? and Tegen et al. (2019))and INCA are the only models which slightly overestimate AOD, when compared with two satellite products (MERGED-FMI and AATSR).

is overestimated over ocean and dusty regions which is indicated by the stronger overestimation compared to AATSR-SU (dominated by ocean) than to AERONET (more representative of land). The coarse mode AOD on the other hand is underestimated

1305

5 over land (too low compared to AERONET, Fig. ??)but overestimated over the subtropical ocean (as can be seen in web visualisation of the results), leading to almost no bias compared to AATSR-SU. Except for regions dominated by dust aerosol AE is biased low. The overestimation of AE in dust dominated regions combined with the overestimation of fine mode

AOD and the longer lifetime of dust particles compared to other models (Fig. ??)indicates a too small size of dust particles. The underestimation of AE compared to AERONET and AATSR is surprising since fine mode AOD is overestimated. The

1310 overestimates by ECHAM-SALSA and INCA have some of the largest contributions of SS and sulphate to AOD (Fig. ??). The aerosol size distribution of ECHAM-HAM agrees reasonably well with observations (Tegen et al. (2019)) and Tegen et al. (2019) find a positive bias of AE compared to AERONET. This could be related to the different sea salt emission parameterisation applied in CTRL or may be affected by temporal sampling errors (Schutgens et al. (2016), Sayer and Knobelspiesse (2019)).

4.2 ECHAM-SALSA

- 1315 3). Both models exhibit overestimates of AOD over the ocean as can be seen in comparison with MERGED-FMI-OCN (see Tab. 4) suggesting an overestimated contribution of sea salt to AOD. In the case of ECHAM-SALSA, this is further confirmed by the overestimated AOD_c (by ca 24%) compared to the AATSR data set (in addition, ECHAM-SALSA is one of the models using dry radius to compute AOD_f, thus likely overestimating the fine-mode fraction diagnosed through AOD_c and AOD_f). However, note again that AOD_c from AATSR is underestimated by ca 15% at AERONET sites (Tab. 1). Thus, it is difficult to 1320 draw clear conclusions about the magnitude of the model bias over marine environments.
 - SALSA is AOD underestimates are mostly associated with SE-Asia and Amazonia, Siberia and high latitudes in general (see Fig. 5a,b). The latter may be associated with insufficient transport towards the poles (e.g., Stohl, 2006) or associated phenomena affecting the radiative properties of the sectional aerosol microphysics module within the ECHAM-HAMMOZ aerosol-chemistry-climate model (Kokkola et al., 2018) alongside the modal aerosol module M7 (Tegen et al., 2019). The implementation
- 1325 of SALSA to ECHAM-HAMMOZ and its evaluation against satellite retrievals, ground based remote sensing retrievals, and in situ observations has been described by Kokkola et al. (2018). One change in these model simulations compared to those in Kokkola et al. (2018) are, in addition to using anthropogenic emissions required for AEROCOM III simulations, is using sea salt emission parameterisation of Guelle et al. (2001) for the reasons described in the previous section ??.
- As the atmospheric model is the same in ECHAM-HAM and ECHAM-SALSA, results between the two model configurations are quite similar. An overall view of the performance of SALSA is that the values fall within the spectrum of model ensemble values except for the burdens of BC and SU for which SALSA predicts highest values of all models (Fig.??). The BC lifetime is highest among all models (9.6 days, Fig. ??) which explains the high burden. On the other hand, reasons for the high burden are not obvious and , since corresponding emissions and lifetimes are comparable with the other models. It may hence be related to the oxidation efficiency of sulphate from its precursors (DMS, SO2). Arctic atmosphere (such as arctic haze, e.g., Tunved et al., 2013)
- perhaps linked with an insufficient attribution of the extensive wildfires in Russia in the summer of 2010 (e.g., Mielonen et al., 2012)
 . The negative bias in S-America (i.e., AERONET AOD~0.5 while ENS-MED~0.3, see web results, Mortier et al., 2020c) mostly
 arises from an underestimate in the biomass burning season in 2010. However, compared to the AP1 simulations (which simulated a too early
 the AP3 models match the timing of the biomass burning season well (which peaked in Aug.-Sept. in 2010).
- When comparing the total simulated (clear sky) AOD of SALSA to the observations (Fig.??), values are biased low compared
 1340 to AERONET as well as MODIS Aqua and Terra. The latter is likely due to Explaining the AOD underestimate in SE-Asia is not trivial. SE-Asia has substantial anthropogenic emissions which are fairly harmonised among the positive biases found for the

MODIS instruments (Fig. ??)especially also because a positive AOD bias is found against the other two satellites (AATSR and MERGED-FMI) <u>AP3 models (cf. Sect. 3)</u>. This indicates, that SALSA underestimates AOD over most of the land area while overestimating AOD over the oceans. Exceptions for the underestimation are Australia and North Africa where SALSA exhibit

- 1345 high values for the total AOD. This is likely This can also be seen by the comparatively low diversity in surface SC_{dry} and AOD_f in Figure 3.1. The region is also strongly affected by a pronounced seasonality including an intensive biomass burning season, dust inflow (e.g., transported from the Arabian Peninsula), monsoon seasons (i.e., seasonality in wet deposition), and other factors impacting the abundance and properties of aerosol (e.g., fog and hygroscopicity), making it difficult for models to simulate regional and seasonal aerosol loadings (e.g., Pan et al., 2015). However, as can be seen in Figure 5d most of the AOD
- 1350 bias in this region is due to missing (or optically too inefficient) coarse particles, which is also supported by the overestimated AE in that region (e.g., Fig. 5e). Thus, the bias could be related to insufficient dust transport or too coarse (and optically less efficient) dust particles. This hypothesis is consistent with the generally low burden, MEC and OD associated with dust in AP3 discussed above (Sect. 3). A detailed investigation of the AP3 models in this important region is desirable but beyond the scope of this paper, due to the contribution of dust to the AOD and is also reflected in the coarse mode AOD. Compared
- 1355 to AATSR-SU, the coarse mode AOD of SALSA is significantly overestimated with a normalized bias of +24%, while the AERONET comparison indicates good agreement over land in (NMB=-3%). On the other hand, over regions affected by dust, coarse mode AOD is overestimated in SALSA. For example, AERONET sites north of Africa exhibit simulated values higher than those measured. While the apparent high overestimation against AATSR-SU may be, to a certain degree, due to low biased AATSR-SU data (Fig. ??), these results indicate that possible overestimates in are likely due to ocean regions.
- 1360 Regions with high dust loads also exhibit overestimation of coarse mode AOD. These is in agreement with the findings of Kokkola et al. (2018) who find large positive biases in over the oceans, in addition to dusty regions. This is expected to be due to high simulated relative humidity in ECHAM over the oceans or too high hygroscopicity for SS aerosol. It is noteworthy that although coarse mode AOD is overestimated over regions where AOD is dominated by sea salt and dust, their emissions are not higher in SALSA (Fig. ??) and it is likely that the simulated size distribution of SALSA is such that SS and DU particles influence radiation effectively, complexity of the prevailing processes.

4.2 ECMWF-IFS

As part of the Copernicus Atmosphere Monitoring Service (CAMS; It is also interesting to compare the relative differences in model simulations of AOD for the different observation platforms. NorESM2, for instance, is the only model that seems to exhibit a weaker performance (in terms of bias) simulating AERONET AOD than simulating any of the satellite AODs. The

- 1370 large underestimate over land could be linked with its comparatively low SO₄ and dust optical depths (see Tab. 3), resulting in substantial AOD underestimates compared to AERONET. The seemingly better performance when compared to the satellite products is due to an overestimate of the SS contribution (particularly in the southern oceans), ECMWF runs a version of the IFS model that includes prognostic aerosol and tropospheric chemistry schemes to produce global forecasts of atmospheric composition. The underlying meteorological model is essentially identical to that used for operational medium-range weather
- 1375 forecasting and is documented at , but at a lower resolution of 40 km to offset the cost of the extra schemes. The results

presented here are from a "cycling forecast" configuration, that is, a forecast with free-running aerosols and chemical species (no assimilation of atmospheric composition), with meteorology reinitialised at 00 UTC each day from operational ECMWF analyses.

The aerosol component is described in Rémy et al. (2019) and based on the earlier work of Morcrette et al. (2009). This is
an externally-mixed hybrid bin/bulk scheme, consisting of three size bins each for desert dust (up to 20μm dry radius) and which compensates to some degree for the underestimate over land (see also Fig. 3). The latter is likely due to the SS size assumptions discussed above, resulting in a shift towards finer, and optically too efficient sea salt aerosol. The overestimated sea salt (up to 20μm radius at 80% relative humidity), and bulk tracers for organic matter, black carbon and sulphate aerosol. For organic matter and black carbon, there are separate hydrophobic and hydrophilic tracers, with a fixed *ageing* timescale for conversion of the former to the latter. There is also an precursor tracer driving the sulphate production via a latitude-optical

- depth manifests particularly in the comparison with the AOD_c data from AATSR (NMB=44%) and the 6% overestimate compared to MERGED-FMI-OCN. As discussed above, the lower correlation over the oceans suggests too much hygroscopic swelling, likely resulting in a smoothing of the spatio-temporal variability of marine aerosol.
- In contrast, EMEP and GEOS exhibit larger underestimates in AOD when compared with the satellites. In the case of EMEP, 1390 this is likely due to too little SS mass, for the reasons discussed above, and temperature-dependent conversion timescale. There is no separate DMS tracer, and no primary sulphate aerosol emission, but all sulphate and precursor emissions are treated as (resulting in a seemingly large contribution of in Fig. ??). The tropospheric chemistry scheme is described in Flemming et al. (2015), but in the version described here this is not directly coupled to the aerosol scheme. it can be seen particularly in the increased underestimate in AOD_c. In the case of GEOS, the larger AOD underestimate against the satellites 1395 could be due to too coarse and optically too inefficient sea salt particles (comparatively low SS MEC, see Tab. 3), in agreement

with Bian et al. (2019).

Compared to the other AP3 models, the total sea salt emissions and burden are very large, as can be see in Figures ?? and ??. Emissions are three times larger than the ensemble mean, but due to a short lifetime (see Figure ??) the burden is only three times larger. However, the sea salt contribution to AOD remains similar to other models because the large size distribution

1400 reduces the extinction per unit mass. These are known issues with the emission scheme in this version of the model (based on Grythe et al. (2014)), and the subject of ongoing development.

The model also has one of the smallest sulphate burdens, which appears to be the result of both relatively low total sulphur emissions and a short lifetime (In contrast to the large underestimate of AOD_c , models tend to agree better in the AOD_f , showing less underestimation and higher correlation (Tab. 4). To some degree, this may be linked with more harmonised

1405 emissions of the associated anthropogenic aerosol species. However, as was shown in Section 3, models also show considerable diversity in fine aerosol species (e.g., OA, SO₄, see Fig. ??). Organic aerosol emissions are higher than most models, although the burden and lifetime are similar to other models. This is likely due to the fact that there is no secondary organic precursor scheme, and secondary organic production is included instead as if it were a primary emission.

Although correlation coefficients for AOD(Figure ??) for this model exhibit relatively high values, there is a significant low 1410 bias against all the 2), a result of differences associated with secondary formation but also due to differences in MEC. Hence, it is interesting to see that models tend to agree better in AOD_f when compared with observations. In addition, the associated particle sizes in the accumulation mode are optically more efficient. Therefore, the AODdata-sets (satellite and AERONET, Fig. ??). This is likely related to the relatively short lifetimes of many species compared to other models, which can be seen in Figure ??. There is also a low bias against both AERONET and AATSR AE, suggesting that particles are on average too large; this means the metal best have and the metal best in sector by the deset best in sector by the deset best in sector by the deset best in the metal best in

1415 this may well be dueat least in partto the unusually high sea salt burden in the model noted above.

4.2 EMEP MSC-W

The EMEP MSCW model is a chemical transport model, designed for policy related applications to combat acid deposition, eutrophication and health adverse air pollution (Simpson et al. (2012)). It calculates the mass concentrations of all main anthropogenic and natural aerosols, contributing to the health related indicators and . The results presented in the paper were obtained in a

- 1420 model run at $0.5 \times 0.5^{\circ}$ grid, driven by 3-hourly ECMWF-IFS meteorology and using ECLIPSE6b emissions (ECLIPSE6a for shipping)_f may be influenced also by dust or sea salt, dependent on the region even though their mass primarily resides in the coarse regime. For instance, as discussed above, approximately 15% of the sea salt OD resides in the fine mode. The dust fine-mode fraction was found to be circa 30% and was retrieved based on dust fine mode ODs submitted by the same models used above to estimate the SS fine-mode fraction (i.e., both for the year of 2010. The model includes aerosols
- 1425 with diameters up to 10 µm and calculates the mass concentrations aerosols in fine and coarse mode. Then, the extinction and absorption coefficients are calculated for the individual aerosol components using mass extinction/absorption coefficients and accounting for aerosol hygroscopic growth (aerosol effective radii, growth factors and specific extinction efficiencies are tabulated) (Schulz et al. (2012)).

The calculated all-sky AODis -10% lower compared to globally averaged annual AOD from AERONET (correlation 0.76).

- 1430 Comparison with satellite AODshows suggests underestimations between 34%-51%, and the relative differences here mostly reproduce the biases observed between the satellites (Fig. ??). These results indicate that EMEP underestimates AOD more over the oceans than over land. Evaluation results against those observations for different world regions are inconclusive in terms of model bias (inferred from web visualisation of the results, Appendix ??). Furthermore, fine AODis overestimated by 20% compared with AERONET data and slightly (by only 11%) underestimated compared to AATSR-SU, whereas coarse
- 1435 AOD is considerably underestimated (by 68 and 70% respectively). Consistently with that CAM5-ATRAS, GEOS, GISS-OMA, SPRINTARS, EC-Earth, TM5, ECMWF-IFS). Compared to AERONET, ENS-MED overestimates AOD_f in several regions including the Mediterranean, the AE is somewhat overestimated (by 36% and 44%)US, Australia and the Arabian Peninsula (see Fig. 5c). The latter could be an indication of too fine dust size distributions contributing to the AOD_f . Overall, these overestimates of AOD_f in some regions tend to have a compensating effect on the underestimate in AOD_c , resulting in a
- 1440 seemingly better model performance for total AOD (see Fig. 5). This could be due, in part, indicating a disproportion between the contributions to AOD from the fine and coarse aerosols. This suggests that either the EMEP model calculates too few coarse particles or the applied MECs are too low (which may be the case for dust, Fig. ??). One of the possible reasons for that to uncertainties related to aerosol size (e.g., too much fine dust, which is optically more efficient and results in an overestimate of AOD_f, while missing in the coarse mode). However, it could also be simply due to too much fine anthropogenic aerosol

- such as SO_4 and too little coarse aerosols. GFDL-AM4, for instance, exhibits considerable overestimates of AOD_f compared to both AERONET and AATSR, while strongly underestimating AOD_c , resulting in a quite good agreement in total AOD. All three parameters (AOD, AOD_f and AOD_c) exhibit high correlation. Compared to the ensemble median, this model shows above average MECs and ODs for both OA and SO_4 , which could explain the overestimated AOD_f . DU and SS, on the contrary, are close to the ensemble median, which would explain the underestimate of AOD_c . Other models that overestimate
- 1450 AOD_f include ECHAM-HAM, OsloCTM3 (both against AERONET and AATSR) and EMEP (only against AERONET). It is difficult to identify a common characteristic amongst these models that might explain the overestimate, but in the case of ECHAM-HAM and OsloCTM3 it may be due to comparatively large sulphate burdens and ODs which could be linked with above average DMS emissions (not shown). For EMEP, on the other hand, it is more likely due to its relatively large contributions from OA, NO3 and NH4 (see Fig. 3), since SO₄ appears to be underrepresented (see Tab. 3). Another possible
- reason for EMEPs overestimation of AOD_f is that fine sea salt and dust particles are assumed to have diameters smaller than 2.5-µm µm, so that the extinction due to sea salt and dust aerosols with diameters between 1-21-2.5-µm µm contributes to the (overestimated) fine mode rather than the (underestimated) -coarse mode. These results suggest a complex interplay among various model assumptions related to composition and size. (Mortier et al., 2020b) also find poorer model performances (larger biases and higher inter-model variability) in long term-trends of the AOD_c, as compared to measurements of total AOD and AOD_f.

Aerosol specific ODs (Fig. ??) of and OA are somewhat larger than the corresponding ensemble median values

4.2 Column Angstrom Exponent (AE)

Models are fairly consistent in their underestimate of AE (see Tab. 4). This suggests that they are (1) either simulating larger particles than are observed or (2) overestimating the fine-mode fraction. EMEP shows the largest overestimate in AE, both

- 1465 compared to AERONET and AATSR-SU. This is in agreement with the relatively large loads for those components (Fig. ??) and may be due to the fact that the model calculates both fine ammonium nitrate and coarse on sea salt and dust. Also, the OA burdens include both primary sources as well as anthropogenic and biogenic secondary OA. For the other aerosols, EMEP calculated aerosol loads and AODs are somewhat smaller than the mean / median values. The resulting MECs are in general agreement with the ensemble, with the exception of , which is one of the largest (probably due to too effective hygroscopic growth). The latter, however, is compensated by the comparatively low burden (emissions from ECLIPSE6b used by the EMEP
- 1470 growth). The latter, however, is compensated by the comparatively low burden (emissions from ECLIPSE6b used b model are smaller than from CMIP6).

The small discrepancy between Total AOD and the sum of the aerosol specific AODs is because the modelled BC AOD is only due to anthropogenic emissions (and does not include forest fires) and DU AOD is only due to windblown dust (while some fugitive anthropogenic dust is also included in the total AOD).

1475 Absorption coefficient is diagnosed from BC and dust mass concentrations, using mass-absorption coefficients. Compared to the climatological GAW observations (at RH<40%), likely related to the lower MEC and burden for SS and DU relative to the ensemble model (Tab. 3) and the 2010 dry (RH=0%) modelled absorption coefficients are biased low (by 40%) cut-off for fine SS and DU at 2.5 μm. INCA, in contrast, exhibits the largest underestimate of AE for both AERONET and AATSR-SU which is likely linked with the comparatively long dust and sea salt lifetimes, and the correlation is 0.66, which is a fair result given

- 1480 the crude simulation approach. The dry scattering coefficient is underestimated by 47% on average (R = 0.74).
 - corresponding high burdens (Tab. 3). The two ECHAM models exhibit similar species optical depths to those simulated by INCA (Fig. 3), so other factors (e.g., assumptions about coarse mode size distributions) may also play a role.

4.3 **GEOS**

As discussed above, the comparison with AATSR-SU (Fig. 5f) suggests that models underestimate AE over the oceans,
however, there is a large uncertainty due to difficulties in validating satellite products over the oceans. In this context, the apparent large overestimate of AE over Australia compared to AATSR (Fig. 5f) is likely associated with retrieval errors in the satellite product. The few AERONET sites in Australia suggest that ENS-MED underestimates AE (Fig. 5e).

GEOS is a global Earth system model, containing components for atmospheric circulation and composition, ocean circulation and biogeochemistry, land surface processes, and data assimilation (Rienecker et al. (2008)). The version of GEOS Earth

- 1490 System Model (with a GOCART aerosol module) used for this study is Icarus-3_3_p2. The simulations run at a spatial resolution of 1.0° x 1.0° latitude and 72 vertical levels from surface up to 0.01hPa (85km) with "replay" mode, denoting simulations driven by the reanalysis meteorological fields from the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA2). This is done to assure that weather and climate patterns are accurately represented for the simulated time. The GOCART module includes major aerosol types of black carbon (BC), organic carbon (OC), brown
- 1495 carbon (BRC), sulphate, nitrate, ammonium, dust, and sea salt (Chin et al. (2002), Colarco et al. (2010), Bian et al. (2019)
). The emissions of dust , sea salt, DMS, and biogenic VOCs are model calculated time-varying fields. All other aerosol emissions used in this study follow the instruction of the AeroCom Phase III History experiment. The major updates on this GOCART version include newly implemented nitrate and ammonium (Bian et al. (2017)), anthropogenic and biomass burning SOAs,as well as separate treatment of optical properties for brown carbon (from biomass burning source) and organic carbon
- 1500 (from all other sources Figure 5e shows that models tend to overestimate AE in dusty regions (e.g., Sahara, Arabian Peninsula), suggesting that they tend to simulate too small dust particles. In the case of the Arabian Peninsula this is consistent with the overestimated AOD_f and the underestimated AOD_c (Fig. 5c,d).

The emissions of aerosols and their precursors used in this GEOS study are similar to those of the ensemble median except OA, which is closer to the ensemble mean (Fig.??). The simulated atmospheric burdens are within 30% of the ensemble median

- 1505 with the exception of dust and nitrate (Fig. ??). The higher dust burden given by GEOS can be explained by its long lifetime (with 9.7 days the longest among the models, Fig. ??). However, the higher nitrate burden cannot simply be explained with its lifetime (Fig. ??). According to the AeroCom Phase III nitrate experiment, the majority of nitrate formed in the atmosphere is associated with atmospheric However, several regions exhibit a larger underestimate in AOD_c than AOD_f in ENS-MED, while also underestimating AE (e.g., Europe). As can be seen in Table 4, five out of the 14 models (ECHAM-HAM, ECMWF-IFS,
- 1510 GISS-OMA, OsloCTM3, SPRINTARS) show such apparent inconsistencies among AE, AOD_f and AOD_c in comparison with AERONET. To some degree, these seemingly inconsistent results for these three parameters may be affected by uncertainties related to the separation of AOD_f and AOD_c both in the models (as discussed above) and the observations (as discussed for

AERONET in Sect. 2.1.1). GISS-OMA, OsloCTM3 and SPRINTARS are among the models which use the dry particle radius to split AOD into AOD_f and AOD_c . In addition, OsloCTM3 simulates dust and sea salt in coarse mode (Bian et al. (2017)). A

1515 careful budget analysis for nitrate would need more information in its chemistry formation and particle size distribution, which is beyond the scope of this paper.

In general, aerosol optical fields (i.e. AE, AOD, Sc. coeff., and Abs. coeff.) simulated by GEOS agree well with various ground station and satellite observations (Fig. ??) and show comparatively high correlation (mostly-8 size bins, and one of these bins contributes both to the R<0.5µm and R>0.7, Fig. ??). The established biases are close to the MEDIAN results, except

- 1520 for the , which is shows overestimations both against AERONETand AATSR-SU (Fig. ??). This is consistent with the high dust and sea salt burdens (Fig. ??) as discussed above 0.5μm regime but is accounted for in the AOD_f diagnostic computed for AeroCom. In the case of ENS-MED, the results may be affected by the averaging choices related to AE and AOD_f (see Sect. 2.3.1). However, even though these uncertainties are not negligible they cannot explain the large differences observed in the biases. ECHAM-HAM, for instance, shows an AE underestimation of 23% compared to AERONET, while at the same time overestimating AOD_f by ca 11% and underestimating AOD_c by ca 57%.
 - 4.3 GFDL-AM4

The Geophysical Fluid Dynamics Laboratory Atmospheric Model version 4 has cubed-sphere topology with 96 × 96 grid boxes per cube face (approximately 100 km grid size) and 33 levels in the vertical, contains an aerosol bulk model that generates mass concentration of aerosol fields (sulphate, carbonaceous aerosols, sea salt and dust) from emissions and a "light"
1530 chemistry mechanism designed to support the aerosol model but with prescribed ozone and radicals (Zhao et al., 2018a). The model is driven by time-varying boundary conditions, and natural and anthropogenic forcings developed in support of CMIP6 (Eyring et al., 2016), except for ship emission of (BC ship emission is included), which has unintentionally not been included. The dust is emitted from constant sources with their erodibility expressed This analysis clearly shows the difficulty of interpreting model AE biases averaged over whole regions and on an annual basis. As Schuster et al. (2006) point out, the
1535 relationship between AE and the aerosol size distribution is complex and allows only qualitative statements. Here, we found

- that models tend to underestimate the global median AE, which suggests that overall they tend to simulate an excess of fine particles or relatively too little coarse aerosol. In order to investigate this in more detail, an additional sensitivity study was performed in which the model biases were analysed as a function of surrounding topography (Ginoux et al., 2001). The sea salt emissions are based on Mårtensson et al. (2003) and Monahan et al. (1986) for fine and coarse mode particles, respectively.
- 1540 Aerosols are externally mixed except for black carbon, which is internally mixed with sulphate. The optical properties of the mixture are calculated by volume weighting of their refractive indices using a Mie code. In the present configuration, the model is run with observed sea surface temperatures (SSTs) and sea-ice distribution (Taylor et al., 2000). In addition, the wind components are nudged, with a 6-hour relaxation time, towards the NCEP-NCAR re-analysis provided on a T62 Gaussian grid with 192 longitude equally spaced and 94 latitude unequally spaced grid points (Kalnay et al., 1996). This resolution is
- 1545 lower than in GFDL-AM4, which may create a low bias of aerosol emission depending on surface winds. In Fig. ??, aerosol emission from GFDL-AM4 are within 25% of the ensemble mean, except for and , which are the lowest among all models

essentially because ship emissions are missing in the simulations. The lower emissions of sulphur compounds does not translate in low atmospheric burden (Fig. ??) as their lifetime is among the highest between the models (Fig. ??).either because of weak oxidation or deposition, the observed AE. The results of this study are shown in Figure 7 where all models were co-located with

- 1550 AERONET AE observations and segregated by different AE intervals. For coarse mode dominated aerosol with AE<1 (i.e., Bins 1&2 in Fig. 7) the AP3 models tend to overestimate AE, with INCA being the only exception. For instance, ENS-MED overestimates AE substantially (ca +140%) if only AE measurements below values of 0.5 are considered. On the other hand, the other aerosols have a shorter lifetime than other models (Fig. ??) while their burdens are well within 25% the AP3 mean values (Fig. ??). The opposite bias between sulphur compounds and the other aerosols suggest an issue with oxidation of rather than
- wet or drydeposition. In Figure ?? the MEC values are within the diversity of the AP3 models except for sea salt which is lower 1555 by a third. This may be because of the cap on hygroscopic growth at 97% relative humidity or the emission parameterisation, as for more fine mode dominated aerosol (i.e., AE>2, Bins 5&6), all models tend to show larger underestimates in AE compared to the general results presented above that consider the full range (non-binned) AE measurements. For instance, ENS-MED shows a bias of ca -60% if only AE>2 measurements are considered. These results suggest that simulated fine-dominated aerosol is larger than it should be (based on the observations) while coarse dominated aerosol is smaller than it should be. The

latter is consistent with the above observation that AE is substantially overestimated in dusty regions.

1560

4.3 Dry surface scattering coefficient (SC_{dry})

This and the following section present and discuss the results from the comparison with surface "dry" (RH<40%) in situ measurements at GAW sites (see Fig. 1g,h for site locations). Compared to the previously discussed columnar variables, 1565 these comparisons are more sensitive to modelling uncertainties associated with the vertical dimension (such as transport and mixing). On the other hand, these comparisons of "dry" variables are "closer" to the simulated aerosol, as they minimise (but do not eliminate) the impact of water uptake on the optical properties.

As with loading through the column, most models also underestimate loading at the scheme of Mårtensson et al. (2003) generates much less sea salt sub-micron particles than Monahan et al. (1986). An alternative explanation is that dry deposition velocity is too strong. The GFDL-AM4 AODs from individual species surface as indicated by the primarily blue shading for 1570 surface scattering in Table 4. The model median bias is -35% for dry scattering, however, the individual model results show considerable diversity, similar to those for AOD_c (Fig. ??)are within the AP3 model diversity except BC, which has the highest value most likely due to the treatment of its internal mixing with sulphate. This high bias will convert into high bias of fine mode AOD, as it appears in Figure ?? where the positive biases of fine mode AOD compare to AERONET and AATSR-SU

- 1575 are the largest among all models. Other normalized biases are relatively weak compared to other models (Figure ??). AOD bias is slightly negative against AERONET and the different satellites. The differences in these biases mostly represent the biases found for the different satellites at AERONET stations (Fig6). Note that these "global" surface scattering values are primarily representative of Europe and the US, where the in situ site density is highest. ??). However, it is important to note, that this model version reported all-sky AOD However, as can be seen in Figure 5a, g the magnitude of underestimation appears
- 1580 to be much larger at many GAW sites in Europe and the US, relative to the AERONET results (which also appear to show

spatially smoother variation in the biases). For comparison, over Europe and the US, ENS-MED exhibits an AOD bias of -13% compared to AERONET (see web results, Mortier et al., 2020c).

It is difficult to observe any spatial tendencies due to the sparsity of sites, although it appears that models tend to overestimate scattering in the eastern US, a region that also shows fairly good performance in AOD, both compared to AERONET and

1585 MERGED-FMI (see Fig. 5). In addition, models tend to underestimate SC_{dry} at the few polar sites, a pattern that can also be observed in the AOD data, suggesting problems related to poleward mass transport.

The discrepancy in biases between surface and column loading could be related to the simulated vertical profiles or could indicate that too much light extinction is attributed to water uptake. The latter would be consistent with a recent study by Burgos et al. (2020) who find that current climate models tend to overestimate the scattering enhancement due to hygroscopic

- 1590 growth. NorESM2, for instance, shows a large underestimate of SC_{dry} at GAW sites (ca 62%). It also shows a large underestimate of ambient column AOD over land (compared to AERONET ca -46%), which suggests that the model is missing aerosol mass, particularly more fine mode dominated species such as SO_4 and OA (cf., Tab. 3). The larger underestimate of SC_{dry} compared to AOD could indicate that water uptake may compensate ambient AOD to some degree for the missing fine mass. A similar behaviour can be observed for the two ECHAM models. ECHAM-HAM, ECHAM-SALSA and NorESM2, are among the
- 1595 models with the largest contribution of water to AOD (Tab. 4) and they are the models that show the largest underestimate and lowest correlation in SC_{dry} . SPRINTARS is another model showing a large negative bias, however it exhibits a slightly better correlation with the observations compared to ECHAM-HAM, while most other models report AOD at clear-sky, which would likely shift the biases towards increased underestimation of AOD (ECHAM-SALSA and NorESM2. SPRINTARS' underestimation of SC_{dry} is also more comparable with the large underestimations found in the other variables suggesting that
- 1600 this model is missing mass rather than over or under emphasizing water availability or uptake. Overall, the correlation between modelled and observed SC_{drx} is generally a little lower than the columnar variables. This is not surprising since the surface measurements do not represent a whole atmospheric column and are thus, more sensitive to associated uncertainties in the vertical transport (e.g.Sect. ??, see also AeroCom optics questionnaire (supplementary material)) . Overall, optical properties are well correlated with observations with coefficients greater than 0.74 except for the scattering and
- 1605 absorption coefficients provided by the surface in-situ data with values at 0.49 and 0.57, respectively (Fig.??).Concerning the Angstrom exponent, one set of value (AERONET) gives poor correlation (0.52) while another (ATSR-SU) provides reasonable correlation (0.74), mixing in the boundary layer, convection and associated changes in lifetime and long range transport) which determine the near surface aerosol mixture, size distributions and associated scattering.

4.4 GISS-OMA

1610 ECMWF-IFS and GISS-OMA is the short name of the GISS ModelE Earth system model (Kelley et al. (2020)), coupled with the One-Moment Aerosol scheme (OMA; Bauer et al. (2020)). In OMA, all aerosols are externally mixed and tracked by their total mass only, except for sea salt and dust where 2 and 5 size-resolved sections are used, respectively. OMA tracks sulphate, nitrate, ammonium, carbonaceous aerosols (black and organic carbon), dust (up to $16 \mu m$) and sea salt (up to $4 \mu m$) are the only two models that show a slight overestimation of SC_{dry} (12% and 18%, respectively), and both exhibit fairly good correlation
1615 (R~0.60). Based on the other results for these models (see Tabs 3&4), no clear explanations can be provided for these results and it would require more detailed investigations of the temporal and spatial distributions (particularly in the vertical), predominant aerosol types and mixtures, and the resulting size distributions.

Relevant to this work, a random maximum cloud overlap is calculated in the column, which is then used to define a totally eloudy or totally cloud-free state in radiation, using a pseudo-random number generation. This is described in Hansen et al.

- (1983). For all-sky AOD calculations 100% relative humidity is used, while for clear-sky we use ambient. This applies to the whole atmospheric column, as dictated by the random maximum cloud overlap calculation. In GISS-OMA there is no calculation from AE. Instead, we calculate it from the AOD calculations in radiation, which are probably underestimating AOD at 870nm by about 10%. Issues with simulated aerosol size likely also play a role in the ability of models to simulate surface scattering. Spectral scattering measurements are available for most surface sites so AE for the surface scattering at two size
- cuts (PM1 and PM10, Andrews et al., 2019) which could provide a further constraint on the aerosol size evaluation.

The results from the evaluation of optical properties shown in Figs. ?? and ?? show a comparatively good agreement with However, note again that this inter comparison bears some uncertainties, for instance, the model data corresponds to RH=0% while the measurements correspond to RH<40%, which is typically considered "dry". Measurements by Zieger et al. (2013)

- 1630 suggest that hygroscopic growth at low RH could lead to an enhancement in scattering of up to 20% for the so-called "dry" aerosol measured at RH=40%. On average, the observations in terms of bias and correlation. The simulated CS AOD shows a bias of -26% against AERONET, which is slightly lower than the ensemble median. In comparison with the satellites, biases of -14% and -19% are found against the MERGED-FMI data-set and AATSR-SU. Similar to the other models, and as explained above, the comparison with MODIS AODs indicates larger negative biases (and slightly decreased correlation)
- 1635 as these satellites show the overall highest AODs (Fig. ??) . Considering these relative biases established for the satellites at AERONET sites, AE, and show similar results when compared with AERONET and AATSR-SU, with biases of the order of -20 to -40% for all three variables. measurements considered here were performed at ~24% RH, thus the impact of light scattering enhancement in the observations should be well below 20% (although that will be site / season specific).

A possible explanation for these underestimated AODs could be that burdens of and sea salt are comparatively low (Fig. 1640 ??), which is also reflected by the fact that both and appear to be underestimated, both against AERONET and AATSR-SU. In case of sea salt, however, the comparatively low burden is likely due to low emissions (Fig. ??) and may, to a certain degree, be compensated by a relatively high SS MEC (+44%compared to median, Fig. ??). A comparatively low burden for nitrate (-33%) is compensated by the largest MEC (ca. +166%). The increased dust emissions, together with an increased lifetime yield a comparatively high burden (Fig. ??) and Another uncertainty is the use of a climatology utilizing data available between

1645 2005-2015. Comparison with climatological values was done to increase the number of sites considered as many sites began measuring after 2010 (see Figure 3 in Laj et al., 2020). However, we investigated the differences and found that SC_{drx} is larger for most models when comparing them with only 2010 surface measurements (not shown). For instance, ENS-MED shows an underestimation of -42% when compared to 2010 observations, as opposed to -35% when compared to the 2005-2015 climatology as was done here. The same comparison was done for AC_{dry}, indicating minimal differences in the AC_{dry} bias

- 1650 (ENS-MED bias of -21.3% for 2010 and -20.3% using the climatology). The 2010 results are also available online and can be compared with the fact that the corresponding DU OD is close the climatology based results (see EBAS-CLIM, EBAS-2010 in Mortier et al . One factor that likely impacts the climatological results is that (1) the median (Figsite density increased in more recent years (e.g., Laj et al., 2020), particularly in Europe and the US, and (2) these regions are associated with negative trends in scattering and absorption (Collaud Coen et al., 2020; Mortier et al., 2020b). This could shift the weight in the climatology to more recent
- 1655 measurements, or to regions affected by larger changes between 2005-2015. While it would be useful to further explore the impact of climatology versus exact temporal matching, such effort is beyond the scope of this paper. Fortunately, our choice of using the surface data climatology for model evaluation does not appear to impact our results substantially, primarily because of the large differences among the models and their tendency to underestimate these variables substantially (see Tab. 4). ??) may be due to the low dust MEC (??). In case of BC,

1660 4.4 Dry surface absorption coefficient (AC_{dry})

Models tend to underestimate the surface aerosol absorption coefficients. The results of ENS-MED suggest an average underestimation of ca 20%. Interestingly, compared to SC_{dry} , the low burden (likely arising from short lifetime) is compensated by the highest BC MEC among the models AC_{dry} observations are better correlated with the model simulations (R~0.75) and also less underestimated. This could be linked with the fact that AC_{dry} depends less on hygroscopic growth and that absorption is

- 1665 mostly associated with a single species (BC), which reduces impacts of mixing on size distributions and optical properties. This is particularly the case for anthropogenic source regions like Europe and the US, which is consistent with the low inter-model diversity in AC_{drx} in Europe and the US, compared to SC_{drx} (see Fig. 4j,l). Note however, that the AC_{drx} results primarily represent Europe (and less the US compared to SC_{drx}) as can be seen in Figure 5g,h.
- Compared to the in-situ measurements, GISS-OMA shows good agreement (NMB=1%) and comparatively high correlation
 with surface scattering, and fairly good performance also for the surface absorption coefficients (NMB=-24%) At the few arctic sites, models tend to significantly underestimate AC_{drv} (apart from Barrow, where good agreement is found) as suggested by the ENS-MED results shown in Figure 5g,h. This could be due to insufficient mass transport into the Arctic or could also be linked with arctic winter time phenomena resulting in increased domestic BC and OC emissions, combined with trapped air-masses due to typically strong winter time inversions, with substantial impacts on arctic climate change (see e.g., Sand et al., 2013, and references the substantial impacts).
- 1675 . The Russian wildfires in 2010 (e.g., Mielonen et al., 2012) may also have played a role in this. Another pattern is an apparent overestimate of AC_{drx} at the few northern Scandinavian sites, as opposed to the underestimations in most of Europe. Whether this behaviour is linked with 2010 wildfires and / or insufficient transport and local emissions, with comparatively low correlation (R=0.52). should be investigated in more detail. Two sites in Asia (both on the Korean peninsula) also show overestimated AC_{dry}, while exhibiting underestimated SC_{dry} (see Fig. 5g). Reasons for that should be examined in detail
- 1680 but may be linked with overestimated impacts of local anthropogenic emissions and / or transport. The same applies for the Cape Point site in South Africa.

4.5 **INCA**

With a bias of -55% SPRINTARS exhibits the largest underestimate of AC_{drv} . This is consistent with the results found for all other extensive variables investigated for this model. SPRINTARS also exhibits the lowest BC MAC value (which was also the case in AP1,

1685 This would most certainly contribute to the underestimate in AC_{drv} as the BC burden for SPRINTARS is comparable to those for the other models. This further suggests that the assumptions describing the conversion from BC mass to absorption may play a role and are likely linked with the simulated size distributions (which depend on aging and mixing) and with the lifetime of BC. Unfortunately, SPRINTARS did not provide the diagnostics to derive BC lifetime. However, as all other species show below average lifetimes in SPRINTARS, it is not unlikely that BC does so too. In summary, the short lifetimes and large

1690 underestimations in loading point to too efficient removal processes in SPRINTARS.

The INCA (INteraction with Chemisty and Aerosols) and ORCHIDEE land surface modules has been coupled to LMDZ dynamical core to conform the LMDZORINCA model. It has been run with forced sea-surface temperatures, sea-ice concentrations and with nudged monthly wind-fields from ERA-Interim. The comparisons with the climatological simulations without nudged-winds shows slightly larger emissions of those aerosols driven interactively by wind at the surface Balkanski et al. (2004), Schulz et al. (2009)

- 1695 . The aerosol modelling in INCA relies on a modal approach to represent the size distribution of DU, SS, BC, NO3Five models overestimate surface absorption (EC-Earth, TM5, SO2 and OA with a combination of accumulation and coarse log-normal modes (both soluble and insoluble). Since these runs use a simplified chemistry scheme, DMS emissions are prescribed and not interactively calculated, and the secondary organic aerosols are not simulated. Hence the organic aerosols are underestimated by this model (low burden in Fig. ??). The current version is modelling BC as internally mixed with sulphate (Wang et al. (2016)
- 1700), where the refractive index is estimated using the Garnet-Maxwell method. This results in an increased and more accurate BC absorption. On the other hand, the dust refractive index is deduced from dedicated experiments Biagio et al. (2017, 2019)showing a marked impact on the longwave part of the spectrum. This results in a less absorbing dust acrosol. BC emissions are derived from inventories and are equally partitioned between surface and altitudeECMWF-IFS, GFDL-AM4 and OsloCTM3). With a bias of +58%, OsloCTM3 shows the largest overestimate of AC_{dry}. This is related to the implementation of new
- absorption parameterisations in the model, resulting in a fairly high BC MAC of $13 \text{ m}^2/\text{g}$, in combination with a strong vertical gradient of BC (and OA) in the lower atmosphere (not shown). The considerable difference in OsloCTM3's ability to simulate AC_{dry} (overestimated) and SC_{dry} (underestimated) are likely linked with the fact that SO₄, one of the main contributors to scattering, does not show such a strong vertical gradient⁵.

The emissions of dust and sea salt have values close to the ensemble mean. With LMDZORINCA the global emitted mineral

- 1710 dust is 1560 Tg/yr (Fig. ??) is within the interval proposed by Kok et al. (2017) for fine and coarse modes. The simulations are based on a coarse insoluble mode (MMD= $2.5 \mu m$ -As with ECMWF-IFS's overestimate of SC_{dry}, no clear explanation can be provided for the positive bias in near surface AC_{dry} for this model. However, the fact that both variables are overestimated in this model suggests that the results are linked with differences in the vertical transport in ECMWF-IFS compared to the other models. This is also supported by the observation that all species column ODs for ECMWF-IFS are average or below-average
- 1715 (see Tab. 3). GFDL-AM4, TM5 and σ). Meanwhile, an improved version with 4 modes (?) shows that including larger particles implies significant higher emissions, although burdens do not increase as substantially as emissions due to the small lifetime

⁵Pers. communication with G. Myhre.

of larger particles (?). Sea salt the emissions amount to 4030 Tg/yr and include accumulation and coarse soluble modes (the super-coarse mode is calculated but not included in this estimation). OA emissions (48.3 Tg/vr) are underestimated compared to other models (ensemble mean 98.2 Tg/yr) because SOA formation is not accounted for. This also explains the comparatively

1720

low burden of OA (0.79 Tg, Fig. ??). All lifetimes are close to the ensemble central values but for sea salt which has a lifetime of 3.1 days.

Our values of MEC are close to the ensemble mean. For those species modelled by a single mode (like dust) we expect less spatial variation of MEC compared to other models with several modes. Regarding optical properties, the AE is highly underestimated both against AERONET and AATSR-SU (ca. -65%EC-Earth did not provide the diagnostics to compute BC

1725 MACs. However, for TM5 and EC-Earth, indications of fairly high BC MACs were found above (see discussion in Sect. 3.0.1). Thisis due to a smaller dynamical response for wavelength in the visible with respect to, together with their above average BC mass burdens, explains their higher AC_{drv} values compared to most of the other models. The total AOD indicates a slight overestimation compared to the multi-model central values, which is likely due to the overestimations of and dust contributions to optical depth, which may partially be compensated by the expected lower values of OA optical depths (Fig. ??).

1730 4.5 NorESM2

However, as for OsloCTM3, vertical transport may also play a role and would need to be investigated to fully explain the results.

The atmosphere module in NorESM2 (NorESM2-MM, see Seland et al. (2020)), CAM6-Nor (Olivić et al. (2020)), is an updated version of CAM5.3-Oslo, for which optical properties have been described and validated by Kirkevåg et al. (2018).

- 1735 Seen in conjunction with these studies, the results presented here can be interpreted as follows. The dust burden is the lowest (5.7 Tg)among the AP3 models, and also low compared to the burden in the un-nudged NorESM2-LM simulation (9.9 Tg), and in CAM5.3-Oslo with fSST and nudged meteorology for year 2000 (16.3 TgKoch et al. (2009) found that AP1 models tend to underestimate column AAOD (compared to AERONET). However, they overestimated near surface BC mass concentrations over Europe (by ca 120%) and the US (by ca 20%) while underestimating in Asia (by ca 60%). Their comparison with
- aircraft measurements over the Americas suggests that aloft, models overestimate BC in the tropics and at mid-latitudes by a 1740 factor of ca 8, while underestimating high latitude BC concentrations at altitude. Based on these results (i.e., underestimated AAOD, overestimated BC mass), they conclude that BC mass to optics conversion is likely underestimated, and they suggest investigation of possible improvements to the associated optical parameters (e.g., BC refractive index, particle size, coating). The lifetime of dust is 1.9 days and is about the same in all these simulations. This is consistently also the lowest among the AP3
- models. The large drop in burden from CAM5.3-Oslo and mostly underestimated AAODs described in Koch et al. (2009) are 1745 consistent with the underestimated surface AC_{dry} found here, but not directly comparable since AAOD represents the whole column and is also sensitive to potential water uptake of aged BC. However, recall that, compared to AP1, the un-nudged NorESM2 is to a large degree a result of tuned dust emissions, while the change between the un-nudged (1870 Tg/yr) and the nudged (1090 Tg/yr) NorESM2 simulations with fSST is consistent with the considerably lower U10 (especially over land)
- 1750 and dust emissions in nudged vs. free meteorology. While NorESM2 sea salt emissions are among the lowest for simulated BC

mass is significantly lower in AP3, the burden is mid-range, and with the highest MEC (4.1(see Sect. 3). A comparison of our results for BC MACs and refractive indices with the AP1 values (summarised in Table 1 in Koch et al., 2009) suggests a slight increase in BC MAC from ca 7.4 m²/g), this model has the highest sea salt AOD values, which is reflected in the positive coarse mode bias against AATSR satellite observations (Fig ??). The relatively high MEC is likely due to SS particle sizes

- 1755 which are shifted towards the more optically efficient accumulation mode, compared to other AP3 models. Sea salt MEC was even higher in CAM5.3-Oslo (5.0AP1) to 8.5 m²/g), but a change in assumed RH (from all-sky to clear-sky) for hygroscopic growth brought about a ca. 19% reduction. The excessive sea salt AOD is a result of tuning of the CMIP6 control simulation for (AP3). Also, some of the AP3 models simulate increased absorption by assuming a BC imaginary refractive index of around ik = 0.75i (e.g., CAM5-ATRAS, ECHAM models, NorESM2with respect to radiative balance at TOA. Compared with
- 1760 AERONET (mainly continental stations) AOD is underestimated, particularly by fine mode particles. One possible reason may be that nitrate aerosols and anthropogenic SOA are not taken into account in the model. Despite missing anthropogenic SOA, our OA burden is still among the highest compared to the other models. Due to the overestimated extinction by sea salt, AE is more underestimated compared to satellite (ocean areas dominate) than to AERONET (mainly continental stations), but the over-all AE bias is close to the AeroCom TM5, GISS-OMA) while most of the AP1 models used values around
- *ik* = 0.45 which will increase their MACs for a given BC size distribution. The extent to which this translates into over or underestimations of surface absorption or BC mass concentrations and AAOD should be investigated in detail, particularly also due to the importance of the vertical distribution of absorbing aerosol for forcing estimates (e.g., Samset et al., 2013). Based on AeroCom phase II (AP2) experiments, Samset et al. (2014) find that a BC lifetime of less than 5 days is required to reproduce observations of BC in remote regions. Samset et al. (2013) emphasize the importance of correctly treating the
- 1770 <u>BC vertical uplift and associated long range transport in order to properly assess the impacts on forcing. Considering these findings, the AP3 mean. The large underestimate in surface scattering and absorption compared to EBAS is consistent with the underestimated AOD over the continents, but as for the majority of the models, the negative bias here is stronger than for the vertically integrated AOD values (compared to AERONET). The high negative bias in surface absorption is consistent with the low dust burden, resulting from the low emissions and short lifetimes, compared to the other models (Figs. ??-??) BC lifetime</u>
- 1775 of 5.5 days is still too long, but is an improvement compared to the lifetime in AP1 (6.5 days). To summarise, it appears that the AP3 models tend to simulate less BC mass but have worked on improving their parameterisations that determine the conversion of BC mass into optical properties.

4.5 OsloCTM3

The OsloCTM3 is a global, offline CTM driven by 3-hourly meteorological data from To conclude, as with SC_{drx}, additional
 model diagnostics and observations are needed to clearly diagnose impacts of vertical transport and mixing near the surface.
 Relating surface measurements to vertical profiles is also key (e.g., Leaitch et al. (2019)).

4.5 **Representativity of the results**

As described in Section 2.3, monthly aggregates of the models and observations were co-located in space and time. The resulting monthly mean values from all sampling coordinates (sites / aggregated satellite pixels) were then used to compute

- 1785 the biases (NMB) and correlation coefficients (R). Based on these metrics, the European Centre for Medium Range Weather Forecast (ECMWF) Integrated Forecast System (IFS)model, and is an updated version of the OsloCTM2 used in previous AeroCom phases (Søvde et al. (2012), Lund et al. (2018)). The model is run in a 2.25°x2.25° horizontal resolution, with 60 vertical levels (the uppermost centered at 0.1 hPa) using the Community Emission Data System (CEDS) performance of individual models was assessed in the previous sections. The comparison of the (Hoesly et al. (2018), van Marle et al. (2017)
- 1790) emission inventory. The treatment of transport and scavenging, as well as individual aerosol modules, is described in detail in Lund et al. (2018) and references therein. In OsloCTM3, the absorption properties have been updated, with BC mass absorption coefficient (MAC) following formula in Zanatta et al. (2016) and a weak absorption implemented for OA (Lund et al. (2018)).OsloCTM3 has a BC MAC value of 12 m²/gand BC MEC is among the highest between the models (Fig.??). The implementation of stronger absorption contributes to the high positive bias (+73%) in surface absorption compared
- to the surface in-situ observations and in contrast to the other models, which tend to underestimate surface absorption at the in-situ locations (Fig. ??) often) temporally incomplete observational records (that are sampled at distinct locations) can introduce considerable representation errors both on spatial and on temporal scales (see e.g., Schutgens et al., 2016, 2017; Wang et al., 2018).
 These errors can affect established biases between model and observations but also other performance measures such as correlation coefficients. We consider this to be the major source of uncertainty for the network averaged statistics (NMB, R)
- 1800 used in this study for the model assessment and inter-comparison (e.g. The burden of nitrate is low, and sulphate high compared to the other models, whereas all other aerosol species in OsloCTM3 are close to model mean values. An evaluation of the burdens and AOD simulated by the OsloCTM3 for year, Tab. 4). Therefore, several sensitivity studies have been performed in order to investigate how potential spatio-temporal representation errors affect the global monthly statistical parameters used in this study.

1805 The following three sensitivity studies were performed in order to investigate temporal and spatial representation uncertainties:

(Temporal): Using GAW in situ AC_{drx} measurements to investigate impacts of temporal representativity errors at the surface. Note that only 2010 CEDS emissions against in-situ and remote sensing observations is provided by Lund et al. (2018). The optical properties for aerosols emitted from biomass burning assume internally mixed aerosol and thus, the reported AOD from BC and OA includes only fossil fuel and biofuel emissions (Fig. ??). This results in lower AOD from OA for OsloCTM3 compared to the other models. The combined BC+OA contribution to AOD amounts to 0.0086. Only all-sky (AS) AOD is provided from OsloCTM3 (Tab. ?? for models that provided CS). This is done because a reliable sub-grid scale parameterisation for RH is unavailable, in order to avoid the AOD used in the radiative transfer calculations to be biased low or high. Compared with the observations, AOD is slightly underestimated, both at AERONET sites (-6%) and the satellite comparisons suggest slightly higher underestimations. The low bias (ca. -20% observation data was used for this sensitivity study (instead of the 2005-2015 climatology of in situ measurements)for AE is consistent between ground and satellite retrievals and is also reflected in the low bias for coarse and high bias for

fine AOD (Fig. ??). In contrast to surface absorption,... The in situ data are co-located in the provided hourly resolution with hourly TM5 data from the surface seattering is biased low compared to observations, which would result in a strong low bias in single scattering albedo. Correlation with the observationsis generally among the higher ones compared to the other models (Fig. ??) AeroCom INSITU experiment. Network averaged statistics (NMB, R) are computed from this hourly co-located data and are compared with results obtained using our analysis strategy (based on monthly averages, see Sect. 2.3). Note that this TM5 model version is slightly different from the AP3-CTRL TM5 model run used in this paper. This is because for AP3-CTRL, no high resolution data was available for this sensitivity study. However, the choice of the model should have limited impact on the results of this study, as most temporal representativity issues arise from incomplete observation records.

- 2. (Temporal): Same as 1., but using AOD from AERONET in order to investigate impacts of temporal representativity errors on columnar variables. For this comparison, the AERONET all-points product is used (instead of the daily product) and is compared with 3-hourly output from the ECMWF-IFS model (see Sect. 4 in supplement 2).
- 1830 3. (Spatial): Spatial representation uncertainties were investigated by using a selection of AERONET sites that is considered representative on spatial scales of a typical model grid. The selection was done based on Wang et al. (2018), using only sites that show an absolute spatial representation error smaller than 10%. This subset was co-located with the AP3 ensemble median (ENS-MED) and results were compared with our results (without AERONET site selection).

The results of these three sensitivity studies are summarised in Table 5. In general, the retrieved differences in NMB and R are small (well below 10% change in NMB and up to ca 0.2 change in R). For instance, for experiments 1 and 2 the difference in NMB is only 0.2% and 1%, respectively and the correlation coefficients are slightly improved in the monthly resolution (which is not surprising, as the coarser resolution will lead to smoother results). Also shown in Table 5 are monthly NMBs and Rs retrieved without applying the 25% time sampling coverage constraint for the observations. For these results, the departures from the high-resolution data set are increased, which illustrates the importance of these resampling constraints.

1840 4.6 SPRINTARS

SPRINTARS (Takemura et al. (2005, 2009)), coupled with a coupled atmosphere-ocean general circulation model (MIROC, Tatebe et al. (2019)), is used in this study although there is also a version coupled with a global cloud resolving model, NICAM For experiment 3, the difference in NMB is ca 5%, however, the total number of considered AERONET sites is reduced from 250 to 50, so this difference could also arise from regional shifts in the considered site coverage (e.g., the 50 sites from Wang et al., 2018, co

- 1845 . Overall, compared to the magnitude and inter-model diversity of biases and correlation coefficients observed in our results (e.g., Sato et al. (2016)). The calculated dust and sea salt emissions with nudged wind field by meteorological reanalysis data are smaller than those without nudging since the emission amounts strongly depend on the wind speed near the surface (see also Sect. ??), which are proportional to 3rd and 3.41th powers, respectively. The 6-hourly reanalysis data cannot represent the gust of wind. The difference between the case with and without nudging is larger in finer horizontal resolution. SPRINTARS
- 1850 has one of the finest resolutions among the participating models in this study. SPRINTARS estimates the global and annual

total emissions of dust and sea salt to be 1390 Tg/yr and 3390 Tg/yr, respectively (Fig.??) with the horizontal resolution of T85 (approx.1.4°×1.4°). Both the lifetime of sea salt and dust are short compared to the other models (Fig. ??), and in case of dust this may be attributed to strong wet deposition over the outflow regions. This, combined with the low emissions, explains the low burdens of these natural species (Fig. ?? which is consistent with the high underestimation of the (Fig. ??). On the

1855 other hand, Tab. 4), we consider these spatio-temporal uncertainties to be acceptable and they do not affect our results (or their interpretation) substantially.

One further uncertainty related to the representativity of the results is that AERONET only measures during daytime, while the models computed 24h averages (as indicated in Sect. 2.1.1). This will cause shifts in the intrinsic weighting applied when computing the network averaged statistics used throughout this paper (e.g., wintertime measurements at high latitudes are

- 1860 restricted to noon-time if they occur at all). In addition, it could introduce systematic errors at locations that show a persistent and pronounced diurnal profile. In this context, note that the GAW in situ observations are not affected by this as they measure continuously, night and day regardless of cloud conditions. The latter is reflected in the calculated AE by SPRINTARS is underestimated, which would rather suggest an overestimation of particle size. However, for this model, this could be attributed to an inappropriate computation of standard deviations of log-normal size distributions of and OA, when calculating optical
- 1865 properties (based on the Mie theory). An internal investigation has confirmed that the diagnosed AE calculated from prognostic mass mixing ratio of each aerosol component is around 1.5 over the industrialized and biomass burning regions, with the appropriate standard deviations of very similar results in Test 1 (i.e., hourly vs monthly comparison of AC_{dry}). Since the results of test 2 (AERONET 3hourly vs monthly) show very good agreement as well, we believe that uncertainties associated with diurnal variations of AOD are likely small compared to the size distributions. This revision (which is not shown in large
- 1870 uncertainties associated with the correct modelling of the AOD, reflected by the considerable biases (and their diversity) found here among the models. Furthermore, AOD represents the whole atmospheric column and, thus, should be less sensitive to diurnal variations than the near surface measurements. A detailed investigation of associated impacts of diurnal variability is desirable but beyond the scope of this article) results in a better AOD performance, with an global annual mean of 0.112, as opposed to 0.072 found in this study (Fig. ??). paper. Also in that context, it would be interesting to investigate the extent
- 1875 to which global climate models need to be able to reproduce amplitudes in diurnal variability of certain tracers and physical processes and which phenomena can be sufficiently parameterised in lower temporal resolution.

Overall, the underestimated dust and sea salt sources result in an underestimation and low correlation in all optical properties that have been investigated in this study (Figs.?? and ??). Consistently, the largest negative biases are found in the evaluation of the coarse AOD, both for AERONET and AATSR (Fig.??) For non-geostationary satellites, the absolute temporal representation errors are likely larger due to the low sampling coverage, combined with cloud contamination in certain regions (e.g., the South Pacific). A detailed investigation of these uncertainties is beyond the scope of this work. Nonetheless, a further simple sensitivity study was performed that investigates how our choice of resolution for the satellite / model inter-comparisons affects the retrieved metrics (NMB and R) presented in Tab. 4. Similar to the three experiments above, this was done by comparing co-location results based on the original satellite products (i.e., 1° × 1°, daily aggregates) with results based on

57

our analysis strategy (i.e., $5^{\circ} \times 5^{\circ}$, monthly aggregates), using models that also provided daily diagnostics for the optical properties investigated. The results of this investigation are presented in Section 3 in supplement 2 (see Tab. 3 therein). They include results for the variables AOD, AOD_f, AOD_c and AE. In most cases, co-location in the higher spatio-temporal resolution results in positive shifts in the respective model NMB and the associated differences can be up to +13% (e.g., AE SPRINTARS

1890 vs. AATSR-SU). However, in most cases the differences are marginal and are well below 5%. Correlation coefficients are generally higher in the lower resolution data, for the same reasons mentioned above (i.e., due to the smoothing effect, intrinsic in the data aggregation).

Finally, we want to stress that the results from these sensitivity studies give insights into uncertainties in evaluation metrics (i.e., bias, correlation) based on monthly means, and averaged over many observation locations around the globe. We emphasise

1895 that representation uncertainties may be significantly larger over sub-domains or at specific locations and times, as shown in the various literature referred to above.

5 Conclusions

In this study a comprehensive inter-comparison of 14 models from the Phase III AeroCom Control AeroCom phase III (AP3) control experiment has been performed. The focus was on the assessment of the modelled column integrated Inter-model

- 1900 diversity of key parameters associated with the aerosol lifecycle and optical properties was investigated for the major aerosol species. These results were compared with results from the AeroCom phase I (AP1) experiments to identify significant differences and improvements in the modelling of the global aerosol. In addition, the models were compared to aerosol observations made at mid visible wavelengths. For this comparison, remote sensing observations of the columnar aerosol optical properties AOD, , AOD₆, and AE, as well as from AERONET and several satellite products were used. Furthermore,
- 1905 for the first time, surface (dry) scattering the models were compared with near surface in situ dry scattering (SC_{dry}) and absorption coefficients . The columnar data was compared to ground based observations from AERONET as well as to several space based observations. In addition to the model evaluation, the performance of the satellite products - in the resolution as aggregated and used for this study - was investigated by comparison with AERONET observations. This was done in order to establish potential relative biases when evaluating the models using satellite observations (Fig. ??) . From this analysis,
- 1910 AATSR-SU and MERGED-FMI showed slight underestimations of AOD (ca. -5%) (AC_{drx}) at GAW sites (mostly located in Europe and the US). Finally, the spatial and temporal representativity of the results was evaluated.

The results suggest that overall, models tend to underestimate all optical properties investigated. Comparison of the modelled AODs with AERONET (mostly land based) and MODIS Aqua and Terra showed overestimations of about 10% and a merged satellite AOD product (better global coverage) show mostly consistent results and suggest that AP3 models underestimate the

1915 AOD by approximately -20±20%, respectively at AERONET sites. AE from AATSR-SU was found to be biased high by about 15% against AERONET, while was found to be underestimated by about 15%. from AATSR-SU showed good agreement with AERONET.All satellite products showed high correlation against AERONET (based on computed ensemble median and IQR). A large fraction of the AOD bias is due to an underestimation of the AOD due to coarse particles – the ensemble median results 1920 13%, however note that these are relative biases (i.e., the average fine mode fraction at AERONET sites is ca. 70%). However, some uncertainties and inconsistencies remain in the way models diagnose AOD_c and AOD_f , and whether the modelled size cuts are comparable with the observed size cuts.

The results of the model evaluation against all ground based observations are summarised in Fig. 6. It shows results of the AeroCom MEDIAN and MEAN (triangles) and corresponding uncertainties estimated from the results of the individual models

- 1925 (plotted as circles). The AE is underestimated by about -9% and shows considerable spread between the models. This suggests that, on average, the simulated particle size is overestimated. This may imply a too short aerosol lifetime or too large fraction of coarse particles present in the models. It may also impact the atmospheric radiation budget due to shifts in the wavelength dependency of aerosol scattering. While the underestimated AE suggests too coarse particles in the models, the analysis of the reveals an underestimation by -40%, with a considerable large underestimation of AODe is associated with uncertainties in the
- 1930 modelling of the natural aerosols sea salt and dust. These two species exhibit wide diversity in their online computed emission strengths and lifetimes, suggesting differences in the simulated size distributions among models, which has clear implications for the conversion to optical extinction. Models tend to agree slightly better in their dust and sea salt MECs, compared to the diversity in lifetimes and emissions, which is likely due to the fact that coarser particle sizes exhibit less spectral variability at mid visible wavelengths than fine particles.
- 1935 Notably, while the total AOD is comparable to that found during the AP1 evaluations, the relative contribution of dust and sea salt to total AOD has changed substantially since AP1 (where both species contributed approximately equally to AOD). In the AP3 simulations, sea salt is the dominant natural species, contributing approximately 2/3 of the combined SS+DU contribution to AOD. While the emissions of both species decrease in AP3, this shift in their relative contribution is mostly due to lifetime changes. In the case of sea salt the emissions decrease is largely compensated for by an increase in its lifetime, which suggests
- 1940 that smaller sea salt particles are being simulated (or removal pathways are less efficient). On the other hand, the dust lifetime is slightly decreased compared to AP1, resulting in a ca 25% lower global dust burden.

Interestingly, the comparison of simulated AE with measurements of the AE from AERONET suggests that models (still) simulate too fine dust aerosol or overestimate the fine mode fraction of coarse dominated aerosol. This would likely translate into an overestimation of the dust MEC which is, however, reduced compared to AP1 and deserves further attention.

- 1945 Unfortunately, a clear assessment of the modelled sea salt size distributions based on the AE is difficult within the scope of this paper. This is mostly due to a lack of ground based measurements over the oceans and large uncertainties in the satellite AE data, but also because of the impact of water uptake on the simulated size distributions. The longer sea salt lifetime compared to AP1 suggests smaller dry sea salt particles in AP3 models, which has implications for water uptake and associated light scattering enhancement and possibly also for cloud formation and properties (e.g., cloud lifetime and albedo)
- 1950 in clean marine environments. Indications for and implications of smaller sea salt particles are reflected in results from a few models showing above average sea salt lifetimes, burdens, MECs and water contribution to AOD, as well as overestimates (and decreased correlation) of total AOD and AOD_c, particularly in comparison with the satellites over the oceans. An overestimated contribution of water uptake to the light extinction is also supported by a considerable underestimate and inter-model spread

The average AOD bias amounts to -20% and shows highest consistency (lowest spread) between the models. The AOD
 bias primarily appears to arise from the low, since shows a smaller bias (-10%, i.e. smallest underestimation) against the respective observations, with a similar spread as for AOD. Compared to Kinne et al. (2006), our AOD bias indicates a slightly larger underestimation in the more recent model versions (in the near surface "dry" scattering coefficients with a bias of approximately -35±25% found when comparing the ensemble median with GAW in situ aerosol scattering measurements.

Similar to AP1, the contribution of the other aerosol species (SO_4 , OA and BC) to total AOD shows considerable variability

- 1960 among the models, while the simulated total AODs are more consistent. The AP3 relative to earlier AeroCom phases). This may partly be attributed SO₄ lifecycle and optical parameters are mostly comparable with AP1. The source strength of OA (POA + secondary formation) shows an increase of ca 20% relative to AP1 which translates into a similar increase in its mass burden and the contribution of OA to AOD. However, the contribution of OA to AOD exhibits large variability among the models, which is mostly due to the fact that in this study, 10 out of 14 models reported clear-sky (CS) AODs (see Tab. ?? and
- 1965 AeroCom optics questionnaire (supplementary material)), while the AOD diagnostics used by Kinne et al. (2006) were likely based on more models reporting AOD under all-sky (AS) assumptions. This hypothesis is underpinned by a +20% increase in NMB in NorESM2, when using AS instead of CS (results available here, leftmost simulation:). The recent findings from the trends analysis by ? indicate, that observed model biases remain mostly constant over time and are not particular for the investigated year 2010. models that tend to simulate an above average OA burden also show an above average MEC and vice
- 1970 versa. The BC emissions are harmonised, however, lifecycle processes related to BC exhibit considerable diversity, resulting in a decrease in the BC mass burden by almost a factor of 2 compared to AP1, mostly due to a decrease in BC lifetime from 6.5 days in AP1 to 5.5 days in AP3. At the same time BC MAC is slightly increased with a median of 8.5 m²/g in AP3 but is still lower than recommended literature values. The results indicate improvement in the BC size assumptions and impacts of aging (e.g., hygroscopicity due to internal mixing, or absorption enhancement effects) but are also likely related to shifts in the
- 1975 assumed BC refractive index towards more absorbing aerosol (in some models). These changes from AP1 to AP3 reflect the considerable effort that went into improving models treatment of BC. However, considering the finding of Samset et al. (2014) the BC lifetime in AP3 remains too long.

Surface dry scattering and absorption coefficients are underestimated by about -40% and -30%, respectively (The lower BC burden is also reflected in considerable underestimations (bias ca $-20\pm18\%$ based on ensemble and IQR) of the near surface

- 1980 dry absorption coefficients at GAW sites, Fig. 1). Both variables show considerable spread between the models, similar to the results for . In case of the dry scattering comparison, the large negative bias may be attributed to differences in the However, there is considerable diversity among the individual 39 GAW sites considered (most of which are located in Europe), indicating impacts of transport and vertical mixing. The SC_{dry} results show larger underestimates (-35%) and more inter-model diversity than the AC_{dry} comparisons. These results reflect uncertainties related to mixing and water uptake for scattering
- 1985 enhancement due to hygroscopic growth. Models used here reported scattering at RH=0%, while in the observations scattering is measured at RH that can be somewhere between 0% and 40%. Thus, on average, the measurements should show larger scattering due to hygroscopic growthand since the ambient AOD bias is less negative (-20%) are consistent with recent

findings by Burgos et al. (2020) indicating that current climate models overestimate the light scattering enhancement associated with aerosol water uptake.

- 1990 A detailed investigation of model biases compared to AERONET AE in different observed AE regimes suggests that models overestimate size (or underestimate the fine mode fraction) in fine mode dominated regimes, while coarse mode dominated regimes indicate the opposite, that is, models simulate not coarse enough particles (or overestimate the contribution of fine aerosol to extinction). Even though the AERONET AE measurements are mostly land based, this could further indicate that not only the hydrophobic dust, but also sea salt particles are too small. This could explain a possible overestimation of the light
- 1995 scattering enhancement, which impacts fine particles more than coarse particles (Zieger et al., 2013). However, the models overestimate the scattering enhancement factor due to hygroscopic growth, as found by ? (Fig. we note that such a hypothesis would need to be investigated in more detail for the AP3 models to make clear statements related to the sea salt size and water uptake.

The newly introduced NO₃ aerosol component in 9 of the 14 models exhibits very large variability in all lifecycle and

optical parameters investigated and it was shown that this is mostly linked to differences in the assumed size distributions but, to some degree, also due to inconsistencies in the associated diagnostics submitted to AeroCom. However, the impact of these uncertainties on total AOD are small as NO₃ contributes a small fraction to the total AOD (<5therein). From a qualitative perspective, a potential overestimation of the scattering enhancement factor in the models agrees well with our finding that models underestimate (ambient) AOD less than dry scattering (by about a factor of 2). % based on the median results of this
 study).

Altogether it is noteworthy that most models underestimate consistently several of the different extensive aerosol optical properties (AOD, fine and coarse mode AOD, scattering and absorption coefficients), both derived from in-situ and remote sensing sensors. This suggests that aerosol loads might be underestimated in the models for the year 2010. Such underestimates are partly compensated by different aerosol optical models and The quality of the constructed AeroCom median ensemble

- 2010 model is a very solid reference for the parameters investigated. In terms of correlation, only in a few instances do individual models outperform it; however no single model is better than the ensemble model on all parameters. This paper and its associated supplements, the AeroCom database and web interfaces are available as reference for further investigations, in particular for developing recommendations for global aerosol modelling. Model diagnostics are comprehensive but clearly not enough to understand all aspects on how models simulate the path from emissions to optical properties. The documented
- 2015 inconsistencies in aerosol life cycle and mass to absorption / scattering coefficients are hopefully an encouragement for modellers to further investigate their individual parameterisations and diagnostics.

The consistency of model performance against different AOD observational datasets speaks in favor of the quality of the observations. The performance variation against different AOD datasets provides at the same time a rough error estimate or robustness of our evaluation method. Our sensitivity tests with respect to spatio-temporal resolution and site selection indicate

2020 that the results are robust. Surface in situ scattering and sun photometer / satellite derived AOD data have been used for the first time in a consistent way for evaluating a multi-model ensemble revealing an underestimate of aerosol AOD, for instance, higher mass exctinction coefficients, though smaller for anthropogenic, fine mode aerosol. Using fine mode AOD as a proxy for

present day aerosol forcing estimates, our results suggest that models underestimate aerosol forcing by circa -15%. However, the associated inter-model spread (quantified by the IQR) is between -35% and +10% and summarizes the large disparity between the individual models, stressing the need for further research.

In future studies the biases found in this study should be investigated, for instance, by incorporating different additional aspects into the analysis – such as model resolution (particularly vertical), regional and seasonal variations, profile extinction data (to investigate "where" the mass is located) and or column water content (to assess hygroscopic growth). Delving into the details of assumed size distributions, particularly for natural aerosol could also help resolve some of the discrepancies

2030 reported on here. In addition, a comparison with surface-mass concentration measurements at the surface and aloft could provide valuable insights related to the question , of whether the models are missing mass or whether assumptions about optical properties are causing the underestimated scattering coefficients and optical depth. Such an analysis would certainly benefit also from a better global coverage of surface measurement sites, since the analysis performed in this study is mostly representative for Europe and the US, where the density of GAW sites is highest (Fig. 1).

2035 6 GAW site evaluation biases

2025

Figs. ?? and ?? show the established biases (NMB) of near surface scattering and absorption coefficients, at all GAW sites (Sect. ??) used in this paper.

6 Sensitivity studies related to spatiotemporal representativity results

- As introduced in Sect. ?? and summarised in Tab. 5, several tests have been performed in order to investigate the spatiotemporal
 representativity and associated uncertainties. The results of tests related to temporal representativity errors are shown in Figs.
 ??, ??, the former being an analysis of monthly vs. 3hourly AOD data vs. AERONET and the latter being an analysis of hourly vs. monthly using surface in-situ absorption data. Both tests do not indicate that the magnitude of these uncertainties in the network-averaged annual statistics exceed 10% in NMB or 0.15 in correlation. Particularly, the results from the in-situ test differ by only 2.4% in NMB which may be attributed to the fact that these data generally shows more continuous sampling
 coverage throughout the 24h of each day as these techniques do not rely on the availability of sunlight.
 - An investigation of spatial representativity errors was done for AERONET AODs, by choosing a subset of sites considered representative based on Wang et al. (2018). The result is shown in Fig. **??** and also does not show substantial differences in light of the diversity found in between the models (Figs. **??** and **??**).

6 Pyaerocom and web visualisation

2050 Most of the analysis in this study was performed with Pyaerocom (Github: , Website:). It is an open source python software project and is being developed and maintained at the Norwegian Meteorological Institute, focussing on model evaluation for aerosol models and the AeroCom initative.

A dedicated website is associated to this study and allows to explore the data from many angles and includes interactive visualisations of performance charts, seatter plots, bias maps and individual station timeseries data, for all models and observation variables, as well as barcharts summarising regional statistics. All results from the optical properties evaluation discussed in this paper are available online at: (last access: 20.12.2019)

2055

. The pre-industrial aerosol state (although available in the AeroCom AP3 control experiment) has not been incorporated yet, but should be included to provide more insight into aerosol forcing estimates and link to the CMIP6 model ensemble results on historical climate evolution.

2060 *Code and data availability.* Most of the data analysis was performed using the open source software pyaerocom (version 0.10.0, release upcoming). All additional analysis scripts are stored in a private GitHub repository and can be provided upon request. All data used in this study is stored on servers of the Norwegian Meteorological Institute and can be provided upon request. The GAW data used is accessible through the EBAS database.

Author contributions. JGl, MS and AM designed the study. JGl did most of the analysis and wrote most of the manuscript, together with substantial contributions from MS, EA. AM, JGr and MS contributed the interactive websites associated to this study. EA, AH, PL, PN, CM, JGr and LS helped analysing and interpreting the observational data used. AB and JGr helped organising the AeroCom database and AeroCom infrastructure needed. YB, SB, HB, RCG, MC, PG, ZK, AK, HK, PS, ML, HM, GM, DN, TN, DO, SR, TT, KT and ST provided model data and model interpretations. All co-authors gave feedback to the manuscript.

Competing interests. No competing interests.

- Acknowledgements. Data providers from all observation networks and satellite products and all people involved in the production of these products are highly acknowledged for sharing and submitting their data and for support. JG-JGI wishes to thank O. Landgren and H. Svennevik for providing data and code used in this study, and S. Kinne for helpful comments via personal communication. TT was supported by the supercomputer system of the National Institute for Environmental Studies, Japan, and JSPSKAKENHI Grant Number Japan Society for the Promotion of Science (JSPS) KAKENHI (JP19H05669), and Environment Research and Technology Development Fund
 (JPMEERF20202F01) of the Environmental Restoration and Conservation Agency, Japan. TvN and PLS acknowledge funding from the
- European Union's Horizon 2020 research and innovation programme project CRESCENDO under grant agreement No 641816. DN ac-knowledges funding from the European Union's Horizon 2020 research and innovation programme project FORCeS under grant agreement No 821205. SEB and KT acknowledge funding from NASA's Atmospheric Composition Modeling and Analysis Program (ACMAP), con-tract number NNX15AE36G. PG acknowledges partial funding from NASA's Earth Surface Mineral Dust Source Investigation (EMIT), program number NNH12ZDA0060-EVI4. They also thank Jingbo Yu for running the GISS model. Resources supporting this work were

provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space

Flight Center. HK acknowledges the Academy of Finland Projects 317390 and 308292. The ECHAM-HAMMOZ model is developed by a consortium composed of ETH Zurich, Max Planck Institut für Meteorologie, Forschungszentrum Jülich, University of Oxford, the Finnish Meteorological Institute and the Leibniz Institute for Tropospheric Research, and managed by the Center for Climate Systems Modeling

- 2085 (C2SM) at ETH Zurich. PN and AH are supported under the ESA Aerosol Climate Change Initiative (Aerosol CCI). HM acknowledges funding from the Ministry of Education, Culture, Sports, Science, and Technology and the Japan Society for the Promotion of Science (MEXT/JSPS) KAKENHI Grant Number JP17H04709. AK., DO, JGI, AK, DO, and MS acknowledge support from the Research Council of Norway funded projects EVA (229771), INES (270061), and KeyClim (295046), and the Horizon 2020 project CRESCENDO (Coordinated Research in Earth Systems and Climate: Experiments, Knowledge, Dissemination and Outreach, no. 641816). High performance
- 2090 computing and storage resources were provided by the Norwegian infrastructure for computational science (through projects NN2345K, NN9560K, NS2345K, and NS9560K). GM has received support from the project SUPER (no. 250573), funded through the Research Council of Norway. HB acknowledges funding from NASA's Atmospheric Composition Modeling and Analysis Program (ACMAP), contract number NNX17AG31G. MC acknowledges support by the NASA Earth Science programs. We thank two anonymous reviewers for careful reading and constructive suggestions.

2095 References

2105

- Aas, W., Mortier, A., Bowersox, V., Cherian, R., Faluvegi, G., Fagerli, H., Hand, J., Klimont, Z., Galy-Lacaux, C., Lehmann, C. M. B., Myhre, C. L., Myhre, G., Olivié, D., Sato, K., Quaas, J., Rao, P. S. P., Schulz, M., Shindell, D., Skeie, R. B., Stein, A., Takemura, T., Tsyro, S., Vet, R., and Xu, X.: Global and regional trends of atmospheric sulfur, Scientific Reports, 9, 953, https://doi.org/10.1038/s41598-018-37304-0, https://doi.org/10.1038/s41598-018-37304-0, 2019.
- 2100 AeroCom diagnostics sheet: AeroCom diagnostics sheet, https://docs.google.com/spreadsheets/d/1NiHLVTDsBo0JEBSnnDECNI2ojUnCV1xuy2PFrsRJV edit?usp=sharing, last access: 2020-08-19, 2020.

AeroCom wiki: AeroCom Phase III experiments description, https://wiki.met.no/aerocom/phase3-experiments, last access: 2020-08-22, 2020.

Andres, R. J. and Kasgnoc, A. D.: A time-averaged inventory of subaerial volcanic sulfur emissions, Journal of Geophysical Research Atmospheres, https://doi.org/10.1029/98JD02091, 1998.

- Andrews, E., Sheridan, P. J., Ogren, J. A., Hageman, D., Jefferson, A., Wendell, J., Alástuey, A., Alados-Arboledas, L., Bergin, M., Ealo, M., Hallar, A. G., Hoffer, A., Kalapov, I., Keywood, M., Kim, J., Kim, S.-W., Kolonjari, F., Labuschagne, C., Lin, N.-H., Macdonald, A., Mayol-Bracero, O. L., McCubbin, I. B., Pandolfi, M., Reisen, F., Sharma, S., Sherman, J. P., Sorribas, M., and Sun, J.: Overview of the NOAA/ESRL Federated Aerosol Network, B. Am. Meteorol. Soc., 100, 123–135, https://doi.org/10.1175/BAMS-D-17-0175.1, 2019.
- 2110 Balkanski, Y., Schulz, M., Claquin, T., Moulin, C., and Ginoux, P.: Global Emissions of Mineral Aerosol: Formulation and Validation using Satellite Imagery, in: Advances in Global Change Research, pp. 239–267, Springer Netherlands, https://doi.org/10.1007/978-1-4020-2167-1_6, https://doi.org/10.1007%2F978-1-4020-2167-1_6, 2004.
 - Bauer, S. E., Tsigaridis, K., Faluvegi, G., Kelley, M., Lo, K. K., Miller, R. L., Nazarenko, L., Schmidt, G. A., and Wu, J.: Historical (1850-2014) aerosol evolution and role on climate forcing using the GISS ModelE2.1 contribution to CMIP6, Journal of Advances in Modeling
- 2115 Earth Systems, n/a, e2019MS001 978, https://doi.org/10.1029/2019MS001978, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/ 2019MS001978, 2020.
 - Bergman, T., Kerminen, V.-M., Korhonen, H., Lehtinen, K. J., Makkonen, R., Arola, A., Mielonen, T., Romakkaniemi, S., Kulmala, M., and Kokkola, H.: Evaluation of the sectional aerosol microphysics module SALSA implementation in ECHAM5-HAM aerosol-climate model, Geoscientific Model Development, 5, 845–868, https://doi.org/10.5194/gmd-5-845-2012, https://www.geosci-model-dev.net/5/845/2012/,
- 2120 2012.

2125

- Bergman, T., Makkonen, R., Schrödner, R., Swietlicki, E., Phillips, V., Le Sager, P., and van Noije, T.: Description and evaluation of a secondary organic aerosol and new particle formation scheme within TM5-MP v1.1, in preparation, 2020.
- Bevan, S. L., North, P. R. J., Los, S. O., and Grey, W. M. F.: A global dataset of atmospheric aerosol optical depth and surface reflectance from AATSR, Remote Sensing of Environment, 116, 199–210, https://doi.org/https://doi.org/10.1016/j.rse.2011.05.024, http:
- Biagio, C. D., Formenti, P., Balkanski, Y., Caponi, L., Cazaunau, M., Pangui, E., Journet, E., Nowak, S., Caquineau, S., Andreae, M. O., Kandler, K., Saeed, T., Piketh, S., Seibert, D., Williams, E., and Doussin, J.-F.: Global scale variability of the mineral dust long-wave refractive index: a new dataset of in situ measurements for climate modeling and remote sensing, Atmospheric Chemistry and Physics, 17, 1901–1929, https://doi.org/10.5194/acp-17-1901-2017, https://doi.org/10.519/acp-17-1901-2017, 2017.

//www.sciencedirect.com/science/article/pii/S0034425711002239, 2012.

2130 Biagio, C. D., Formenti, P., Balkanski, Y., Caponi, L., Cazaunau, M., Pangui, E., Journet, E., Nowak, S., Andreae, M. O., Kandler, K., Saeed, T., Piketh, S., Seibert, D., Williams, E., and Doussin, J.-F.: Complex refractive indices and single scattering albedo of global dust aerosols in the shortwave spectrum and relationship to iron content and size, https://doi.org/10.5194/acp-2019-145, https://doi.org/10.5194/acp-2019-145, 2019.

- Bian, H., Chin, M., Hauglustaine, D. A., Schulz, M., Myhre, G., Bauer, S. E., Lund, M. T., Karydis, V. A., Kucsera, T. L., Pan, X., Pozzer,
 A., Skeie, R. B., Steenrod, S. D., Sudo, K., Tsigaridis, K., Tsimpidi, A. P., and Tsyro, S. G.: Investigation of global particulate nitrate from
 - the AeroCom phase III experiment, Atmospheric Chemistry and Physics, 17, 12911–12940, https://doi.org/10.5194/acp-17-12911-2017, https://www.atmos-chem-phys.net/17/12911/2017/, 2017.
 - Bian, H., Froyd, K., Murphy, D. M., Dibb, J., Darmenov, A., Chin, M., Colarco, P. R., da Silva, A., Kucsera, T. L., Schill, G., Yu, H., Bui, P., Dollner, M., Weinzierl, B., and Smirnov, A.: Observationally constrained analysis of sea salt aerosol in the marine atmosphere,
- 2140 Atmospheric Chemistry and Physics, 19, 10773–10785, https://doi.org/10.5194/acp-19-10773-2019, https://www.atmos-chem-phys.net/ 19/10773/2019/, 2019.
 - Bond, T. C. and Bergstrom, R. W.: Light Absorption by Carbonaceous Particles: An Investigative Review, Aerosol Science and Technology, 40, 27–67, https://doi.org/10.1080/02786820500421521, https://doi.org/10.1080/02786820500421521, 2006.
 - Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., Kinne,
- 2145 S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser, J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S.: Bounding the role of black carbon in the climate system: A scientific assessment, Journal of Geophysical Research: Atmospheres, 118, 5380–5552, https://doi.org/10.1002/jgrd.50171, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrd.50171, 2013.
- 2150 Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., and Zhang, X.: Clouds and Aerosols, book section 7, p. 571–658, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.016, www.climatechange2013.org, 2013.
 - Burgos, M. A., Andrews, E. J., Titos, G., Benedetti, A., Bian, H., Buchard, V., Curci, G., Kirkevåg, A., Kokkola, H., Laakso, A., Lund,
- 2155 M. T., Matsui, H., Myhre, G., Randles, C., Schulz, M., van Noije, T., Zhang, K., Alados-Arboledas, L., Baltensperger, U., Jefferson, A., Sherman, J., Sun, J., Weingartner, E., and Zieger, P.: A global model-measurement evaluation of particle light scattering coefficients at elevated relative humidity, Atmospheric Chemistry and Physics Discussions, 2020, 1–30, https://doi.org/10.5194/acp-2019-1190, https: //www.atmos-chem-phys-discuss.net/acp-2019-1190/, 2020.
 - Cappa, C. D., Onasch, T. B., Massoli, P., Worsnop, D. R., Bates, T. S., Cross, E. S., Davidovits, P., Hakala, J., Hayden, K. L., Jobson, B. T.,
- Kolesar, K. R., Lack, D. A., Lerner, B. M., Li, S.-M., Mellon, D., Nuaaman, I., Olfert, J. S., Petäjä, T., Quinn, P. K., Song, C., Subramanian, R., Williams, E. J., and Zaveri, R. A.: Radiative Absorption Enhancements Due to the Mixing State of Atmospheric Black Carbon, Science, 337, 1078–1081, https://doi.org/10.1126/science.1223447, https://science.sciencemag.org/content/337/6098/1078, 2012.
 - Carn, S. A., Fioletov, V. E., McLinden, C. A., Li, C., and Krotkov, N. A.: A decade of global volcanic SO2 emissions measured from space, Scientific Reports, 7, 44 095, https://doi.org/10.1038/srep44095, https://doi.org/10.1038/srep44095, 2017.
- 2165 Carslaw, K. S., Lee, L. A., Reddington, C. L., Pringle, K. J., Rap, A., Forster, P. M., Mann, G. W., Spracklen, D. V., Woodhouse, M. T., Regayre, L. A., and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing, Nature, 503, 67–71, https://doi.org/10.1038/nature12674, https://doi.org/10.1038/nature12674, 2013.
 - Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., Martin, R. V., Logan, J. A., Higurashi, A., and Nakajima, T.: Tropospheric Aerosol Optical Thickness from the GOCART Model and Comparisons with Satellite and Sun Photometer Measurements,

- 2170 Journal of the Atmospheric Sciences, 59, 461–483, https://doi.org/10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2, https://doi. org/10.1175/1520-0469(2002)059%3C0461:TAOTFT%3E2.0.COhttp://0.0.0.2, 2002.
 - Claquin, T., Schulz, M., Balkanski, Y., and Boucher, O.: Uncertainties in assessing radiative forcing by mineral dust, Tellus B, 50, 491–505, https://doi.org/10.1034/j.1600-0889.1998.t01-2-00007.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0889.1998.t01-2-00007.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0889.t01-2-00007.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0889.t01-2-00007.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.t01-2-00007.x, https://onlinelibrary.wiley.com/d
- 2175 Colarco, P., da Silva, A., Chin, M., and Diehl, T.: Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012820, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JD012820, 2010.
 - Collaud Coen, M., Andrews, E., Alastuey, A., Arsov, T. P., Backman, J., Brem, B. T., Bukowiecki, N., Couret, C., Eleftheriadis, K., Flentje, H., Fiebig, M., Gysel-Beer, M., Hand, J. L., Hoffer, A., Hooda, R., Hueglin, C., Joubert, W., Keywood, M., Kim, J. E., Kim, S.-W.,
- 2180 Labuschagne, C., Lin, N.-H., Lin, Y., Lund Myhre, C., Luoma, K., Lyamani, H., Marinoni, A., Mayol-Bracero, O. L., Mihalopoulos, N., Pandolfi, M., Prats, N., Prenni, A. J., Putaud, J.-P., Ries, L., Reisen, F., Sellegri, K., Sharma, S., Sheridan, P., Sherman, J. P., Sun, J., Titos, G., Torres, E., Tuch, T., Weller, R., Wiedensohler, A., Zieger, P., and Laj, P.: Multidecadal trend analysis of in situ aerosol radiative properties around the world, Atmospheric Chemistry and Physics, 20, 8867–8908, https://doi.org/10.5194/acp-20-8867-2020, https://acp.copernicus.org/articles/20/8867/2020/, 2020.
- 2185 de Leeuw, G., Holzer-Popp, T., Bevan, S., Davies, W. H., Descloitres, J., Grainger, R. G., Griesfeller, J., Heckel, A., Kinne, S., Klüser, L., Kolmonen, P., Litvinov, P., Martynenko, D., North, P., Ovigneur, B., Pascal, N., Poulsen, C., Ramon, D., Schulz, M., Siddans, R., Sogacheva, L., Tanré, D., Thomas, G. E., Virtanen, T. H., von Hoyningen Huene, W., Vountas, M., and Pinnock, S.: Evaluation of seven European aerosol optical depth retrieval algorithms for climate analysis, Remote Sensing of Environment, 162, 295–315, https://doi.org/10.1016/j.rse.2013.04.023, http://www.sciencedirect.com/science/article/pii/S0034425713003507, 2015.
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.
- Dentener, F., Kinne, S., Bond, T., Boucher, O., Cofala, J., Generoso, S., Ginoux, P., Gong, S., Hoelzemann, J. J., Ito, A., Marelli, L., Penner, J. E., Putaud, J.-P., Textor, C., Schulz, M., van der Werf, G. R., and Wilson, J.: Emissions of primary aerosol and precursor gases in the years 2000 and 1750 prescribed data-sets for AeroCom, Atmospheric Chemistry and Physics, 6, 4321–4344, https://doi.org/10.5194/acp-6-4321-2006, https://acp.copernicus.org/articles/6/4321/2006/, 2006.
- 2200 Dubovik, O. and King, M. D.: A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements, Journal of Geophysical Research: Atmospheres, 105, 20 673–20 696, https://doi.org/10.1029/2000JD900282, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD900282, 2000.
 - ERA5: Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), https://cds.climate.copernicus.eu/cdsapp#!/home, accessed: 2019-12-27.
- 2205 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, https://www.geosci-model-dev.net/9/1937/2016/, 2016.

- Flemming, J., Huijnen, V., Arteta, J., Bechtold, P., Beljaars, A., Blechschmidt, A.-M., Diamantakis, M., Engelen, R. J., Gaudel, A., Inness, A., Jones, L., Josse, B., Katragkou, E., Marecal, V., Peuch, V.-H., Richter, A., Schultz, M. G., Stein, O., and Tsikerdekis, A.: Tropospheric
- 2210 chemistry in the Integrated Forecasting System of ECMWF, Geoscientific Model Development, 8, 975–1003, https://doi.org/10.5194/gmd-8-975-2015, https://www.geosci-model-dev.net/8/975/2015/, 2015.
 - Fossum, K. N., Ovadnevaite, J., Ceburnis, D., Preißler, J., Snider, J. R., Huang, R.-J., Zuend, A., and O'Dowd, C.: Sea-spray regulates sulfate cloud droplet activation over oceans, npj Climate and Atmospheric Science, 3, 14, https://doi.org/10.1038/s41612-020-0116-2, https://doi.org/10.1038/s41612-020-0116-2, 2020.
- 2215 GAW Report 227: WMO/GAW Aerosol Measurement Procedures, Guidelines and Recommendations, https://library.wmo.int/doc_num.php? explnum_id=3073, accessed: 2020-09-14, 2016.
 - Giles, D. M., Sinyuk, A., Sorokin, M. G., Schafer, J. S., Smirnov, A., Slutsker, I., Eck, T. F., Holben, B. N., Lewis, J. R., Campbell, J. R., Welton, E. J., Korkin, S. V., and Lyapustin, A. I.: Advancements in the Aerosol Robotic Network (AERONET) Version 3 database – automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol
- 2220 optical depth (AOD) measurements, Atmospheric Measurement Techniques, 12, 169–209, https://doi.org/10.5194/amt-12-169-2019, https://www.atmos-meas-tech.net/12/169/2019/, 2019.
 - Ginoux, P., Chin, M., Tegen, I., Prospero, J. M., Holben, B., Dubovik, O., and Lin, S.-J.: Sources and distributions of dust aerosols simulated with the GOCART model, Journal of Geophysical Research: Atmospheres, 106, 20255–20273, https://doi.org/10.1029/2000JD000053, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD000053, 2001.
- 2225 Gliß, J.: Geonum python library, https://geonum.readthedocs.io/en/latest, last access: 2020-08-20, 2017.
 - Grythe, H., Ström, J., Krejci, R., Quinn, P., and Stohl, A.: A review of sea-spray aerosol source functions using a large global set of sea salt aerosol concentration measurements, Atmospheric Chemistry and Physics, 14, 1277–1297, https://doi.org/10.5194/acp-14-1277-2014, https://acp.copernicus.org/articles/14/1277/2014/, 2014.

Guelle, W., Schulz, M., Balkanski, Y., and Dentener, F.: Influence of the source formulation on modeling the atmospheric global distribution

- 2230 of sea salt aerosol, Journal of Geophysical Research: Atmospheres, 106, 27 509–27 524, https://doi.org/10.1029/2001JD900249, https: //agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD900249, 2001.
 - Gupta, P., Remer, L. A., Levy, R. C., and Mattoo, S.: Validation of MODIS 3 km land aerosol optical depth from NASA's EOS Terra and Aqua missions, Atmospheric Measurement Techniques, 11, 3145–3159, https://doi.org/10.5194/amt-11-3145-2018, https://amt.copernicus.org/ articles/11/3145/2018/, 2018.
- Halmer, M. M., Schmincke, H. U., and Graf, H. F.: The annual volcanic gas input into the atmosphere, in particular into the stratosphere: A global data set for the past 100 years, Journal of Volcanology and Geothermal Research, https://doi.org/10.1016/S0377-0273(01)00318-3, 2002.
 - Hand, J. L. and Malm, W. C.: Review of aerosol mass scattering efficiencies from ground-based measurements since 1990, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2007JD008484, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/
- 2240 2007JD008484, 2007.
 - Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J.-I., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geoscientific Model Development, 11, 369–408, https://doi.org/10.5194/gmd-11-369-2018, https://www.geosci-model-dev.net/11/369/2018/, 2018.

- 2245 Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization, Remote Sensing of Environment, 66, 1–16, https://doi.org/https://doi.org/10.1016/S0034-4257(98)00031-5, http://www.sciencedirect.com/science/article/pii/S0034425798000315, 1998.
- Holben, B. N., Kim, J., Sano, I., Mukai, S., Eck, T. F., Giles, D. M., Schafer, J. S., Sinyuk, A., Slutsker, I., Smirnov, A., Sorokin, M., Anderson,
 B. E., Che, H., Choi, M., Crawford, J. H., Ferrare, R. A., Garay, M. J., Jeong, U., Kim, M., Kim, W., Knox, N., Li, Z., Lim, H. S., Liu, Y.,
 Maring, H., Nakata, M., Pickering, K. E., Piketh, S., Redemann, J., Reid, J. S., Salinas, S., Seo, S., Tan, F., Tripathi, S. N., Toon, O. B.,
 and Xiao, Q.: An overview of mesoscale aerosol processes, comparisons, and validation studies from DRAGON networks, Atmospheric Chemistry and Physics, 18, 655–671, https://doi.org/10.5194/acp-18-655-2018, https://www.atmos-chem-phys.net/18/655/2018/, 2018.
- Hsu, N. C., Si-Chee Tsay, King, M. D., and Herman, J. R.: Aerosol properties over bright-reflecting source regions, IEEE Transactions on
 Geoscience and Remote Sensing, 42, 557–569, https://doi.org/10.1109/TGRS.2004.824067, 2004.
 - Hubanks, P. A.: Collection 6.1Change Summary DocumentMODIS Atmosphere Level-3 Algorithmand Global Products, Tech. rep., https://atmosphere-imager.gsfc.nasa.gov/sites/default/files/ModAtmo/L3_C61_Changes_v2.pdf, 2017.
 - Huneeus, N., Schulz, M., Balkanski, Y., Griesfeller, J., Prospero, J., Kinne, S., Bauer, S., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Fillmore, D., Ghan, S., Ginoux, P., Grini, A., Horowitz, L., Koch, D., Krol, M. C., Landing, W., Liu, X., Mahowald,
- 2260 N., Miller, R., Morcrette, J.-J., Myhre, G., Penner, J., Perlwitz, J., Stier, P., Takemura, T., and Zender, C. S.: Global dust model intercomparison in AeroCom phase I, Atmospheric Chemistry and Physics, 11, 7781–7816, https://doi.org/10.5194/acp-11-7781-2011, https://www.atmos-chem-phys.net/11/7781/2011, 2011.

IPCC: Climate Change 2014, https://doi.org/10.1017/CBO9781107415324, 2014.

- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., and others: The
 NCEP/NCAR 40-year reanalysis project, Bulletin of the American Meteorological Society, 77, 437–472, 1996.
- Kandler, K., Benker, N., Bundke, U., Cuevas, E., Ebert, M., Knippertz, P., Rodríguez, S., Schütz, L., and Weinbruch, S.: Chemical composition and complex refractive index of Saharan Mineral Dust at Izaña, Tenerife (Spain) derived by electron microscopy, Atmospheric Environment, 41, 8058 – 8074, https://doi.org/https://doi.org/10.1016/j.atmosenv.2007.06.047, http://www.sciencedirect.com/science/article/ pii/S1352231007006061, 2007.
- 2270 Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., Ackerman, A. S., Aleinov, I., Bauer, M., Bleck, R., Canuto, V., Cesana, G., Cheng, Y., Clune, T. L., Cook, B. I., Cruz, C. A., Del Genio, A. D., Elsaesser, G. S., Faluvegi, G., Kiang, N. Y., Kim, D., Lacis, A. A., Leboissetier, A., LeGrande, A. N., Lo, K. K., Marshall, J., Matthews, E. E., McDermid, S., Mezuman, K., Miller, R. L., Murray, L. T., Oinas, V., Orbe, C., Pérez García-Pando, C., Perlwitz, J. P., Puma, M. J., Rind, D., Romanou, A., Shindell, D. T., Sun, S., Tausnev, N., Tsigaridis, K., Tselioudis, G., Weng, E., Wu, J., and Yao, M.-S.: GISS-E2.1: Configurations and
- 2275 Climatology, Journal of Advances in Modeling Earth Systems, n/a, e2019MS002025, https://doi.org/10.1029/2019MS002025, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002025, 2020.
 - Kiehl, J. T. and Briegleb, B. P.: The Relative Roles of Sulfate Aerosols and Greenhouse Gases in Climate Forcing, Science, 260, 311–314, https://doi.org/10.1126/science.260.5106.311, https://science.sciencemag.org/content/260/5106/311, 1993.
- Kim, D., Chin, M., Yu, H., Diehl, T., Tan, Q., Kahn, R. A., Tsigaridis, K., Bauer, S. E., Takemura, T., Pozzoli, L., Bellouin, N.,
 Schulz, M., Peyridieu, S., Chédin, A., and Koffi, B.: Sources, sinks, and transatlantic transport of North African dust aerosol: A multimodel analysis and comparison with remote sensing data, Journal of Geophysical Research: Atmospheres, 119, 6259–6277, https://doi.org/10.1002/2013JD021099, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD021099, 2014.

- Kinne, S., Schulz, M., Textor, C., Guibert, S., Balkanski, Y., Bauer, S. E., Berntsen, T., Berglen, T. F., Boucher, O., Chin, M., Collins, W., Dentener, F., Diehl, T., Easter, R., Feichter, J., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Herzog, M., Horowitz,
- 2285 L., Isaksen, I., Iversen, T., Kirkevåg, A., Kloster, S., Koch, D., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Lesins, G., Liu, X., Lohmann, U., Montanaro, V., Myhre, G., Penner, J. E., Pitari, G., Reddy, S., Seland, Ø., Stier, P., Takemura, T., and Tie, X.: An AeroCom initial assessment Optical properties in aerosol component modules of global models, Atmospheric Chemistry and Physics, https://doi.org/10.5194/acp-6-1815-2006, 2006.
- Kinne, S., O'Donnel, D., Stier, P., Kloster, S., Zhang, K., Schmidt, H., Rast, S., Giorgetta, M., Eck, T. F., and Stevens, B.:
 MAC-v1: A new global aerosol climatology for climate studies, Journal of Advances in Modeling Earth Systems, 5, 704–740, https://doi.org/10.1002/jame.20035, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jame.20035, 2013.
 - Kirkevåg, A., Grini, A., Olivié, D., Seland, Ø., Alterskjær, K., Hummel, M., Karset, I. H. H., Lewinschal, A., Liu, X., Makkonen, R., Bethke, I., Griesfeller, J., Schulz, M., and Iversen, T.: A production-tagged aerosol module for Earth system models, OsloAero5.3 – extensions and updates for CAM5.3-Oslo, Geoscientific Model Development, 11, 3945–3982, https://doi.org/10.5194/gmd-11-3945-2018, https://www.geosci-model-dev.net/11/3945/2018/, 2018.
 - Klimont, Z., Smith, S. J., and Cofala, J.: The last decade of global anthropogenic sulfur dioxide: 2000-2011 emissions, Environmental Research Letters, 8, 14 003, https://doi.org/10.1088/1748-9326/8/1/014003, https://doi.org/10.1088/1748-9326/8/1/014003, 2013.
 - Klingmüller, K., Steil, B., Brühl, C., Tost, H., and Lelieveld, J.: Sensitivity of aerosol radiative effects to different mixing assumptions in the AEROPT 1.0 submodel of the EMAC atmospheric-chemistry–climate model, Geoscientific Model Development, 7, 2503–2516, https://doi.org/10.5194/gmd-7-2503-2014, https://www.geosci-model-dev.net/7/2503/2014/, 2014.
 - Koch, D., Schmidt, G. A., and Field, C. V.: Sulfur, sea salt, and radionuclide aerosols in GISS ModelE, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2004JD005550, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD005550, 2006.
 - Koch, D., Bond, T. C., Streets, D., Unger, N., and van der Werf, G. R.: Global impacts of aerosols from particular source regions and sectors, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2005JD007024, https://agupubs.onlinelibrary.wiley.com/doi/ abs/10.1029/2005JD007024, 2007.
 - Koch, D., Schulz, M., Kinne, S., McNaughton, C., Spackman, J. R., Balkanski, Y., Bauer, S., Berntsen, T., Bond, T. C., Boucher, O., Chin, M., Clarke, A., De Luca, N., Dentener, F., Diehl, T., Dubovik, O., Easter, R., Fahey, D. W., Feichter, J., Fillmore, D., Freitag, S., Ghan, S., Ginoux, P., Gong, S., Horowitz, L., Iversen, T., Kirkevaring;g, A., Klimont, Z., Kondo, Y., Krol, M., Liu, X., Miller, R., Montanaro, V., Moteki, N., Myhre, G., Penner, J. E., Perlwitz, J., Pitari, G., Reddy, S., Sahu, L., Sakamoto, H., Schuster, G., Schwarz, J. P.,
- 2310 Seland, Ø., Stier, P., Takegawa, N., Takemura, T., Textor, C., van Aardenne, J. A., and Zhao, Y.: Evaluation of black carbon estimations in global aerosol models, Atmospheric Chemistry and Physics, 9, 9001–9026, https://doi.org/10.5194/acp-9-9001-2009, https://www. atmos-chem-phys.net/9/9001/2009/, 2009.
 - Kok, J. F.: A scaling theory for the size distribution of emitted dust aerosols suggests climate models underestimate the size of the global dust cycle, Proceedings of the National Academy of Sciences, 108, 1016–1021, https://doi.org/10.1073/pnas.1014798108, https://www.
- 2315 pnas.org/content/108/3/1016, 2011.

2295

2300

2305

Kok, J. F., Ridley, D. A., Zhou, Q., Miller, R. L., Zhao, C., Heald, C. L., Ward, D. S., Albani, S., and Haustein, K.: Smaller desert dust cooling effect estimated from analysis of dust size and abundance, Nature Geoscience, 10, 274–278, https://doi.org/10.1038/ngeo2912, https://doi.org/10.1038%2Fngeo2912, 2017.

Kokkola, H., Kühn, T., Laakso, A., Bergman, T., Lehtinen, K. E., Mielonen, T., Arola, A., Stadtler, S., Korhonen, H., Ferrachat, S., Lohmann, L. Naubauer, D., Tagan, L., Siaganthalar, La Drian, C., Sabultz, M. C., Pay, L., Stiar, P., Dackalakia, N., Haeld, C. L., and Bamekkaniami

2320 U., Neubauer, D., Tegen, I., Siegenthaler-Le Drian, C., Schultz, M. G., Bey, I., Stier, P., Daskalakis, N., Heald, C. L., and Romakkaniemi,

S.: SALSA2.0: The sectional aerosol module of the aerosol-chemistry-climate model ECHAM6.3.0-HAM2.3-MOZ1.0, Geoscientific Model Development, https://doi.org/10.5194/gmd-11-3833-2018, 2018.

- Lafon, S., Sokolik, I. N., Rajot, J. L., Caquineau, S., and Gaudichet, A.: Characterization of iron oxides in mineral dust aerosols: Implications for light absorption, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD007016, https: //agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD007016, 2006.
- Laj, P., et, and al.: A global analysis of climate-relevant aerosol properties retrieved from the network of GAW near-surface observatories, https://doi.org/10.5194/amt-13-4353-2020, https://doi.org/10.5194/amt-13-4353-2020, 2020.
- Latimer, R. N. C. and Martin, R. V.: Interpretation of measured aerosol mass scattering efficiency over North America using a chemical transport model, Atmospheric Chemistry and Physics, 19, 2635-2653, https://doi.org/10.5194/acp-19-2635-2019, https://www. atmos-chem-phys.net/19/2635/2019/, 2019.
- Laurent, B., Marticorena, B., Bergametti, G., Léon, J. F., and Mahowald, N. M.: Modeling mineral dust emissions from the Sahara desert using new surface properties and soil database, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009484, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD009484, 2008.

Leaitch, W. R., Kodros, J. K., Willis, M. D., Hanna, S., Schulz, H., Andrews, E., Bozem, H., Burkart, J., Hoor, P., Kolonjari, F., Ogren, J. A.,

- 2335 Sharma, S., Si, M., von Salzen, K., Bertram, A. K., Herber, A., Abbatt, J. P. D., and Pierce, J. R.: Vertical profiles of light absorption and scattering associated with black-carbon particle fractions in the springtime Arctic above 79° N. Atmospheric Chemistry and Physics Discussions, 2019, 1–37, https://doi.org/10.5194/acp-2019-785, https://acp.copernicus.org/preprints/acp-2019-785/, 2019.
 - Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmospheric Measurement Techniques, 6, 2989-3034, https://doi.org/10.5194/amt-6-2989-2013, https: //www.atmos-meas-tech.net/6/2989/2013/, 2013.
- 2340

2325

2330

- Li, F., Ginoux, P., and Ramaswamy, V.: Distribution, transport, and deposition of mineral dust in the Southern Ocean and Antarc-//agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD009190, 2008.
- Liu, D., Whitehead, J., Alfarra, M. R., Reyes-Villegas, E., Spracklen, D. V., Reddington, C. L., Kong, S., Williams, P. I., Ting, Y.-C., Haslett,
- 2345 S., et al.: Black-carbon absorption enhancement in the atmosphere determined by particle mixing state, Nature Geoscience, 10, 184–188, 2017.
 - Liu, F., Yon, J., Fuentes, A., Lobo, P., Smallwood, G. J., and Corbin, J. C.: Review of recent literature on the light absorption properties of black carbon: Refractive index, mass absorption cross section, and absorption function, Aerosol Science and Technology, 54, 33-51, https://doi.org/10.1080/02786826.2019.1676878, https://doi.org/10.1080/02786826.2019.1676878, 2020.
- 2350 Long, M. S., Keene, W. C., Kieber, D. J., Erickson, D. J., and Maring, H.: A sea-state based source function for size- and composition-resolved marine aerosol production, Atmospheric Chemistry and Physics, 11, 1203–1216, https://doi.org/10.5194/acp-11-1203-2011, https://www. atmos-chem-phys.net/11/1203/2011/, 2011.
 - Lund, M. T., Myhre, G., Haslerud, A. S., Skeie, R. B., Griesfeller, J., Platt, S. M., Kumar, R., Myhre, C. L., and Schulz, M.: Concentrations and radiative forcing of anthropogenic aerosols from 1750 to 2014 simulated with the Oslo CTM3 and CEDS emission inventory, Geoscien-
- 2355 tific Model Development, 11, 4909–4931, https://doi.org/10.5194/gmd-11-4909-2018, https://www.geosci-model-dev.net/11/4909/2018/, 2018.

- Mårtensson, E. M., Nilsson, E. D., de Leeuw, G., Cohen, L. H., and Hansson, H.-C.: Laboratory simulations and parameterization of the primary marine aerosol production, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD002263, https: //agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002263, 2003.
- 2360 Matsui, H.: Development of a global aerosol model using a two-dimensional sectional method: 1. Model design, Journal of Advances in Modeling Earth Systems, 9, 1921–1947, https://doi.org/10.1002/2017MS000936, https://agupubs.onlinelibrary.wiley.com/doi/abs/10. 1002/2017MS000936, 2017.
- Matsui, H. and Mahowald, N.: Development of a global aerosol model using a two-dimensional sectional method: 2. Evaluation and sensitivity simulations, Journal of Advances in Modeling Earth Systems, 9, 1887–1920, https://doi.org/10.1002/2017MS000937, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS000937, 2017.
 - Matsui, H., Koike, M., Kondo, Y., Fast, J. D., and Takigawa, M.: Development of an aerosol microphysical module: Aerosol Twodimensional bin module for foRmation and Aging Simulation (ATRAS), Atmospheric Chemistry and Physics, 14, 10315–10331, https://doi.org/10.5194/acp-14-10315-2014, https://www.atmos-chem-phys.net/14/10315/2014/, 2014a.
- Matsui, H., Koike, M., Kondo, Y., Takami, A., Fast, J. D., Kanaya, Y., and Takigawa, M.: Volatility basis-set approach simulation of organic
 aerosol formation in East Asia: implications for anthropogenic–biogenic interaction and controllable amounts, Atmospheric Chemistry and Physics, 14, 9513–9535, https://doi.org/10.5194/acp-14-9513-2014, https://www.atmos-chem-phys.net/14/9513/2014/, 2014b.
 - Matsui, H., Hamilton, D. S., and Mahowald, N. M.: Black carbon radiative effects highly sensitive to emitted particle size when resolving mixing-state diversity, Nature Communications, 9, 3446, https://doi.org/10.1038/s41467-018-05635-1, https://doi.org/10.1038/ s41467-018-05635-1, 2018.
- 2375 Mielonen, T., Portin, H., Komppula, M., Leskinen, A., Tamminen, J., Ialongo, I., Hakkarainen, J., Lehtinen, K., and Arola, A.: Biomass burning aerosols observed in Eastern Finland during the Russian wildfires in summer 2010 – Part 2: Remote sensing, Atmospheric Environment, 47, 279 – 287, https://doi.org/https://doi.org/10.1016/j.atmosenv.2011.07.016, http://www.sciencedirect.com/science/article/pii/ S1352231011007321, 2012.
- Monahan, E. C., Spiel, D. E., and Davidson, K. L.: A Model of Marine Aerosol Generation Via Whitecaps and Wave Disruption BT Oceanic
 Whitecaps: And Their Role in Air-Sea Exchange Processes, pp. 167–174, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-009-4668-2_16, https://doi.org/10.1007/978-94-009-4668-2_16, 1986.
 - Morcrette, J.-J., Boucher, O., Jones, L., Salmond, D., Bechtold, P., Beljaars, A., Benedetti, A., Bonet, A., Kaiser, J. W., Razinger, M., Schulz, M., Serrar, S., Simmons, A. J., Sofiev, M., Suttie, M., Tompkins, A. M., and Untch, A.: Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modeling, Journal of Geophysical Research: Atmospheres,
- 2385 114, https://doi.org/10.1029/2008JD011235, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JD011235, 2009.
 Mortier, A., Gliss, J., and Schulz, M.: AeroCom web visualisation (Satellite evaluation), https://aerocom-evaluation.met.no/overall.php?
 - project=satellite&exp=eval2019, last access: 2020-09-14, 2020a. Mortier, A., Gliss, J., Schulz, M., Aas, W., Andrews, E., Bian, H., Chin, M., Ginoux, P., Hand, J., Holben, B., Hua, Z., Kipling, Z., Kirkevåg,
- A., Laj, P., Lurton, T., Myhre, G., Neubauer, D., Olivié, D., von Salzen, K., Takemura, T., and Tilmes, S.: Evaluation of climate model aerosol trends with ground-based observations over the last two decades an AeroCom and CMIP6 analysis, Atmospheric Chemistry and Physics Discussions, 2020, 1–36, https://doi.org/10.5194/acp-2019-1203, https://www.atmos-chem-phys-discuss.net/acp-2019-1203/, 2020b.

- Mortier, A., Gliss, J., Schulz, M., and community, A.: AeroCom web visualisation (AP3-CTRL results), https://aerocom-evaluation.met. no/main.php?project=aerocom&exp=PIII-optics2019-P, online visualisation of results from Gliss et al. (acp-2019-1214), Last access: 2020-09-14, 2020c.
- 2395

2415

- Myhre, G., Bellouin, N., Berglen, T. F., Berntsen, T. K., Boucher, O., Grini, A. L. F., Isaksen, I. S. A., Johnsrud, M., Mishchenko, M. I., Stordal, F., and Tanré, D.: Comparison of the radiative properties and direct radiative effect of aerosols from a global aerosol model and remote sensing data over ocean, Tellus B, 59, 115–129, https://doi.org/10.1111/j.1600-0889.2006.00226.x, https://onlinelibrary.wiley. com/doi/abs/10.1111/j.1600-0889.2006.00226.x, 2007.
- 2400 Myhre, G., Berglen, T. F., Johnsrud, M., Hoyle, C. R., Berntsen, T. K., Christopher, S. A., Fahey, D. W., Isaksen, I. S. A., Jones, T. A., Kahn, R. A., Loeb, N., Quinn, P., Remer, L., Schwarz, J. P., and Yttri, K. E.: Modelled radiative forcing of the direct aerosol effect with multi-observation evaluation, Atmospheric Chemistry and Physics, 9, 1365–1392, https://doi.org/10.5194/acp-9-1365-2009, https: //www.atmos-chem-phys.net/9/1365/2009/, 2009.
- Myhre, G., Samset, B. H., Schulz, M., Balkanski, Y., Bauer, S., Berntsen, T. K., Bian, H., Bellouin, N., Chin, M., Diehl, T., Easter, R. C.,
- Feichter, J., Ghan, S. J., Hauglustaine, D., Iversen, T., Kinne, S., Kirkevåg, A., Lamarque, J.-F., Lin, G., Liu, X., Lund, M. T., Luo, G., Ma, X., van Noije, T., Penner, J. E., Rasch, P. J., Ruiz, A., Seland, Ø., Skeie, R. B., Stier, P., Takemura, T., Tsigaridis, K., Wang, P., Wang, Z., Xu, L., Yu, H., Yu, F., Yoon, J.-H., Zhang, K., Zhang, H., and Zhou, C.: Radiative forcing of the direct aerosol effect from AeroCom Phase II simulations, Atmospheric Chemistry and Physics, 13, 1853–1877, https://doi.org/10.5194/acp-13-1853-2013, https://www.atmos-chem-phys.net/13/1853/2013/, 2013.
- 2410 Nabat, P., Solmon, F., Mallet, M., Kok, J. F., and Somot, S.: Dust emission size distribution impact on aerosol budget and radiative forcing over the Mediterranean region: a regional climate model approach, Atmospheric Chemistry and Physics, 12, 10545–10567, https://doi.org/10.5194/acp-12-10545-2012, https://acp.copernicus.org/articles/12/10545/2012/, 2012.
 - North, P. R. J.: Estimation of aerosol opacity and land surface bidirectional reflectance from ATSR-2 dual-angle imagery: Operational method and validation, Journal of Geophysical Research: Atmospheres, 107, 4–1, https://doi.org/10.1029/2000JD000207, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD000207, 2002.
 - North, P. R. J. and Heckel, A.: SU-ATSR Algorithm Theoretical Basis Document (ATBD) v4.3, Tech. rep., http://cci.esa.int/sites/default/ files/Aerosol_cci2_ATBD_ATSR_SU_v4.3.pdf, 2017.

- O'Neill, N. T., Eck, T. F., Smirnov, A., Holben, B. N., and Thulasiraman, S.: Spectral discrimination of coarse and fine mode optical depth,
 Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD002975, https://agupubs.onlinelibrary.wiley.com/doi/ abs/10.1029/2002JD002975, 2003.
 - Pan, X., Chin, M., Gautam, R., Bian, H., Kim, D., Colarco, P. R., Diehl, T. L., Takemura, T., Pozzoli, L., Tsigaridis, K., Bauer, S., and Bellouin, N.: A multi-model evaluation of aerosols over South Asia: common problems and possible causes, Atmospheric Chemistry and Physics, 15, 5903–5928, https://doi.org/10.5194/acp-15-5903-2015, https://acp.copernicus.org/articles/15/5903/2015/, 2015.
- 2425 Popp, T., de Leeuw, G., Bingen, C., Brühl, C., Capelle, V., Chedin, A., Clarisse, L., Dubovik, O., Grainger, R., Griesfeller, J., Heckel, A., Kinne, S., Klüser, L., Kosmale, M., Kolmonen, P., Lelli, L., Litvinov, P., Mei, L., North, P., Pinnock, S., Povey, A., Robert, C., Schulz, M., Sogacheva, L., Stebel, K., Stein Zweers, D., Thomas, G., Tilstra, L., Vandenbussche, S., Veefkind, P., Vountas, M., and Xue, Y.: Development, Production and Evaluation of Aerosol Climate Data Records from European Satellite Observations (Aerosol_cci), Remote Sensing, 8, 421, https://doi.org/10.3390/rs8050421, http://www.mdpi.com/2072-4292/8/5/421, 2016.

Olivié et al.: Evaluation of the forcing and response in the Norwegian Earth System Model NorESM2-LM, in preparation, 2020.

- 2430 Rémy, S., Kipling, Z., Flemming, J., Boucher, O., Nabat, P., Michou, M., Bozzo, A., Ades, M., Huijnen, V., Benedetti, A., Engelen, R., Peuch, V.-H., and Morcrette, J.-J.: Description and evaluation of the tropospheric aerosol scheme in the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS-AER, cycle 45R1), Geoscientific Model Development, 12, 4627–4659, https://doi.org/10.5194/gmd-12-4627-2019, https://www.geosci-model-dev.net/12/4627/2019/, 2019.
- Riemer, N., West, M., Zaveri, R. A., and Easter, R. C.: Simulating the evolution of soot mixing state with a particle-resolved aerosol model, 2435 Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2008JD011073, https://agupubs.onlinelibrary.wiley.com/doi/ abs/10.1029/2008JD011073, 2009.
 - Rienecker, M. M., Suarez, M., Todling, R., Bacmeister, J., Takacs, L., Liu, H., Gu, W., Sienkiewicz, M., Koster, R., Gelaro, R., et al.: The GEOS-5 Data Assimilation System: Documentation of Versions 5.0. 1, 5.1. 0, and 5.2. 0, 2008.
- Rissler, J., Vestin, A., Swietlicki, E., Fisch, G., Zhou, J., Artaxo, P., and Andreae, M. O.: Size distribution and hygroscopic prop-2440 erties of aerosol particles from dry-season biomass burning in Amazonia, Atmospheric Chemistry and Physics, 6, 471-491, https://doi.org/10.5194/acp-6-471-2006, https://acp.copernicus.org/articles/6/471/2006/, 2006.
 - Samset, B. H., Myhre, G., Schulz, M., Balkanski, Y., Bauer, S., Berntsen, T. K., Bian, H., Bellouin, N., Diehl, T., Easter, R. C., Ghan, S. J., Iversen, T., Kinne, S., Kirkevåg, A., Lamarque, J.-F., Lin, G., Liu, X., Penner, J. E., Seland, Ø., Skeie, R. B., Stier, P., Takemura, T., Tsigaridis, K., and Zhang, K.: Black carbon vertical profiles strongly affect its radiative forcing uncertainty, Atmospheric Chemistry and
- 2445 Physics, 13, 2423–2434, https://doi.org/10.5194/acp-13-2423-2013, https://acp.copernicus.org/articles/13/2423/2013/, 2013. Samset, B. H., Myhre, G., Herber, A., Kondo, Y., Li, S.-M., Moteki, N., Koike, M., Oshima, N., Schwarz, J. P., Balkanski, Y., Bauer, S. E., Bellouin, N., Berntsen, T. K., Bian, H., Chin, M., Diehl, T., Easter, R. C., Ghan, S. J., Iversen, T., Kirkevåg, A., Lamarque, J.-F., Lin, G., Liu, X., Penner, J. E., Schulz, M., Seland, Ø., Skeie, R. B., Stier, P., Takemura, T., Tsigaridis, K., and Zhang, K.: Modelled black carbon radiative forcing and atmospheric lifetime in AeroCom Phase II constrained by aircraft observations, Atmospheric Chemistry and Physics,
- 2450 14, 12465-12477, https://doi.org/10.5194/acp-14-12465-2014, https://acp.copernicus.org/articles/14/12465/2014/, 2014. Samset, B. H., Stjern, C. W., Andrews, E., Kahn, R. A., Myhre, G., Schulz, M., and Schuster, G. L.: Aerosol Absorption: Progress Towards Global and Regional Constraints, Current Climate Change Reports, 4, 65-83, https://doi.org/10.1007/s40641-018-0091-4, https://doi.org/ 10.1007/s40641-018-0091-4, 2018.
- Sand, M., Berntsen, T. K., Seland, Ø., and Kristjánsson, J. E.: Arctic surface temperature change to emissions of black carbon within Arctic 2455 or midlatitudes, Journal of Geophysical Research: Atmospheres, 118, 7788-7798, https://doi.org/10.1002/jgrd.50613, https://agupubs.
 - onlinelibrary.wiley.com/doi/abs/10.1002/jgrd.50613, 2013. Sato, Y., Miura, H., Yashiro, H., Goto, D., Takemura, T., Tomita, H., and Nakajima, T.: Unrealistically pristine air in the Arctic produced by current global scale models, Scientific Reports, 6, 26561, https://doi.org/10.1038/srep26561, https://doi.org/10.1038/srep26561, 2016.

Sayer, A. M. and Knobelspiesse, K. D.: How should we aggregate data? Methods accounting for the numerical distributions, with an as-

sessment of aerosol optical depth, Atmospheric Chemistry and Physics, 19, 15023-15048, https://doi.org/10.5194/acp-19-15023-2019, 2460 https://www.atmos-chem-phys.net/19/15023/2019/, 2019.

Sayer, A. M., Munchak, L. A., Hsu, N. C., Levy, R. C., Bettenhausen, C., and Jeong, M.-J.: MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations, Journal of Geophysical Research: Atmospheres, 119, 913–965, https://doi.org/10.1002/2014JD022453, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/ 2014JD022453, 2014.

- 2465
 - Schmeisser, L., Andrews, E., Ogren, J. A., Sheridan, P., Jefferson, A., Sharma, S., Kim, J. E., Sherman, J. P., Sorribas, M., Kalapov, I., Arsov, T., Angelov, C., Mayol-Bracero, O. L., Labuschagne, C., Kim, S.-W., Hoffer, A., Lin, N.-H., Chia, H.-P., Bergin, M., Sun, J., Liu, P.,

and Wu, H.: Classifying aerosol type using in situ surface spectral aerosol optical properties, Atmospheric Chemistry and Physics, 17, 12097–12120, https://doi.org/10.5194/acp-17-12097-2017, https://www.atmos-chem-phys.net/17/12097/2017/, 2017.

- 2470 Schultz, M. G., Stadtler, S., Schröder, S., Taraborrelli, D., Franco, B., Krefting, J., Henrot, A., Ferrachat, S., Lohmann, U., Neubauer, D., Siegenthaler-Le Drian, C., Wahl, S., Kokkola, H., Kühn, T., Rast, S., Schmidt, H., Stier, P., Kinnison, D., Tyndall, G. S., Orlando, J. J., and Wespes, C.: The chemistry–climate model ECHAM6.3-HAM2.3-MOZ1.0, Geoscientific Model Development, 11, 1695–1723, https://doi.org/10.5194/gmd-11-1695-2018, https://www.geosci-model-dev.net/11/1695/2018/, 2018.
 - Schulz, M., Textor, C., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Dentener, F., Guibert, S., Isaksen, I. S.,
- 2475 Iversen, T., Koch, D., Kirkevåg, A., Liu, X., Montanaro, V., Myhre, G., Penner, J. E., Pitari, G., Reddy, S., Seland, Ø., Stier, P., and Takemura, T.: Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations, Atmospheric Chemistry and Physics, https://doi.org/10.5194/acp-6-5225-2006, 2006.
- Schulz, M., Cozic, A., and Szopa, S.: LMDzT-INCA dust forecast model developments and associated validation efforts, IOP Conference Series: Earth and Environmental Science, 7, 12014, https://doi.org/10.1088/1755-1307/7/1/012014, https://doi.org/10.1088/
 2F1755-1307%2F7%2F1%2F012014, 2009.
 - Schulz, M., Steensen, B. M., Tsyro, S., and Koffi, B.: First results of comparison modelled aerosol extinction profiles with CALIOP data, in: Transboundary acidification, eutrophication and ground level ozone in Europe in 2010. EMEP Status Report 1/2012, pp. 113–121, Norwegian Meteorological Institute, https://emep.int/publ/reports/2012/status_report_1_2012.pdf, 2012.
- Schulz, M., Griesfeller, J., and community, A.: AeroCom web interface (all AeroCom phase results), https://aerocom.met.no/cgi-bin/surfobs_
 annualrs.pl?PROJECT=AEROCOM&MODELLIST=Glissetal.2020, online visualisation of results from Gliss et al. (acp-2019-1214),
 Last access: 2020-09-10, 2020.
 - Schuster, G. L., Dubovik, O., and Holben, B. N.: Angstrom exponent and bimodal aerosol size distributions, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD006328, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/ 2005JD006328, 2006.
- 2490 Schutgens, N., Tsyro, S., Gryspeerdt, E., Goto, D., Weigum, N., Schulz, M., and Stier, P.: On the spatio-temporal representativeness of observations, Atmospheric Chemistry and Physics, 17, 9761–9780, https://doi.org/10.5194/acp-17-9761-2017, https://www.atmos-chem-phys.net/17/9761/2017/, 2017.
 - Schutgens, N. A. J., Partridge, D. G., and Stier, P.: The importance of temporal collocation for the evaluation of aerosol models with observations, Atmospheric Chemistry and Physics, 16, 1065–1079, https://doi.org/10.5194/acp-16-1065-2016, https://www.atmos-chem-phys.net/16/1065/2016/, 2016.
 - Schwarz, J. P., Spackman, J. R., Fahey, D. W., Gao, R. S., Lohmann, U., Stier, P., Watts, L. A., Thomson, D. S., Lack, D. A., Pfister, L., Mahoney, M. J., Baumgardner, D., Wilson, J. C., and Reeves, J. M.: Coatings and their enhancement of black carbon light absorption in the tropical atmosphere, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009042, https://agupubs. onlinelibrary.wiley.com/doi/abs/10.1029/2007JD009042, 2008.

2495

- 2500 Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, John Wiley & Sons, 2016. Seland, Ø., Bentsen, M., Seland Graff, L., Olivié, D., Toniazzo, T., Gjermundsen, A., Debernard, J. B., Gupta, A. K., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Schancke Aas, K., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Hafsahl Karset, I. H., Landgren, O., Liakka, J., Onsum Moseid, K., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iverson, T., and Schulz, M.: The Norwegian Earth System Model, NorESM2 – Evaluation of theCMIP6 DECK and historical simulations, Geoscientific Model De-
- velopment Discussions, 2020, 1–68, https://doi.org/10.5194/gmd-2019-378, https://gmd.copernicus.org/preprints/gmd-2019-378/, 2020.

- Simpson, D., Benedictow, A., Berge, H., Bergström, R., Emberson, L. D., Fagerli, H., Flechard, C. R., Hayman, G. D., Gauss, M., Jonson, J. E., Jenkin, M. E., Nyiri, A., Richter, C., Semeena, V. S., Tsyro, S., Tuovinen, J.-P., Valdebenito, A., and Wind, P.: The EMEP MSC-W chemical transport model – technical description, Atmospheric Chemistry and Physics, 12, 7825–7865, https://doi.org/10.5194/acp-12-7825-2012, https://www.atmos-chem-phys.net/12/7825/2012/, 2012.
- 2510 Sofiev, M., Soares, J., Prank, M., de Leeuw, G., and Kukkonen, J.: A regional-to-global model of emission and transport of sea salt particles in the atmosphere, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2010JD014713, https://agupubs. onlinelibrary.wiley.com/doi/abs/10.1029/2010JD014713, 2011.
 - Sogacheva, L., Popp, T., Sayer, A. M., Dubovik, O., Garay, M. J., Heckel, A., Hsu, N. C., Jethva, H., Kahn, R. A., Kolmonen, P., Kosmale, M., de Leeuw, G., Levy, R. C., Litvinov, P., Lyapustin, A., North, P., and Torres, O.: Merging regional and global AOD records from 15
- 2515 available satellite products, Atmospheric Chemistry and Physics Discussions, 2019, 1–62, https://doi.org/10.5194/acp-2019-446, https: //www.atmos-chem-phys-discuss.net/acp-2019-446/, 2019.
- Sogacheva, L., Popp, T., Sayer, A. M., Dubovik, O., Garay, M. J., Heckel, A., Hsu, N. C., Jethva, H., Kahn, R. A., Kolmonen, P., Kosmale, M., de Leeuw, G., Levy, R. C., Litvinov, P., Lyapustin, A., North, P., Torres, O., and Arola, A.: Merging regional and global aerosol optical depth records from major available satellite products, Atmospheric Chemistry and Physics, 20, 2031–2056, https://doi.org/10.5194/acp-20-2031-2020, https://www.atmos-chem-phys.net/20/2031/2020/, 2020.
 - Sokolik, I. N. and Toon, O. B.: Direct radiative forcing by anthropogenic airborne mineral aerosols, Nature, 381, 681–683, https://doi.org/10.1038/381681a0, https://doi.org/10.1038/381681a0, 1996.
- Søvde, O. A., Prather, M. J., Isaksen, I. S. A., Berntsen, T. K., Stordal, F., Zhu, X., Holmes, C. D., and Hsu, J.: The chemical transport model Oslo CTM3, Geoscientific Model Development, 5, 1441–1469, https://doi.org/10.5194/gmd-5-1441-2012, https://www.geosci-model-dev.
 net/5/1441/2012/, 2012.
 - Stohl, A.: Characteristics of atmospheric transport into the Arctic troposphere, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD006888, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006888, 2006.
 - Sun, J., Zhang, M., and Liu, T.: Spatial and temporal characteristics of dust storms in China and its surrounding regions, 1960–1999: Relations to source area and climate, Journal of Geophysical Research: Atmospheres, 106, 10325–10333, https://doi.org/10.1029/2000JD900665, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD900665, 2001.
 - Takemura, T., Nozawa, T., Emori, S., Nakajima, T. Y., and Nakajima, T.: Simulation of climate response to aerosol direct and indirect effects with aerosol transport-radiation model, Journal of Geophysical Research: Atmospheres, 110, https://doi.org/10.1029/2004JD005029, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD005029, 2005.

2530

- Takemura, T., Egashira, M., Matsuzawa, K., Ichijo, H., O'ishi, R., and Abe-Ouchi, A.: A simulation of the global distribution
 and radiative forcing of soil dust aerosols at the Last Glacial Maximum, Atmospheric Chemistry and Physics, 9, 3061–3073, https://doi.org/10.5194/acp-9-3061-2009, https://www.atmos-chem-phys.net/9/3061/2009/, 2009.
 - Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in
- 2540 MIROC6, Geoscientific Model Development, 12, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019, https://www.geosci-model-dev. net/12/2727/2019/, 2019.
 - Taylor, K. E., Williamson, D., and Zwiers, F.: The Sea Surface Temperature and Sea-Ice Concentration Boundary Conditions for AMIP II Simulations, Report No. 60, Tech. rep., LLNL, 2000.

Tegen, I., Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Bey, I., Schutgens, N., Stier, P., Watson-Parris, D., Stanelle, T.,

- 2545 Schmidt, H., Rast, S., Kokkola, H., Schultz, M., Schroeder, S., Daskalakis, N., Barthel, S., Heinold, B., and Lohmann, U.: The global aerosol-climate model ECHAM6.3–HAM2.3 – Part 1: Aerosol evaluation, Geoscientific Model Development, 12, 1643–1677, https://doi.org/10.5194/gmd-12-1643-2019, https://www.geosci-model-dev.net/12/1643/2019/, 2019.
 - Textor, C., Graf, H.-F., Timmreck, C., and Robock, A.: Emissions from volcanoes, in: Emissions of Atmospheric Trace Compounds, edited by Granier, C., Artaxo, P., and Reeves, C. E., pp. 269–303, Springer Netherlands, Dordrecht, 2004.
- Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Feichter, H., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I., Iversen, I., Kloster, S., Koch, D., Kirkevåg, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J., Pitari, G., Reddy, S., Seland, Ø., Stier, P., Takemura, T., and Tie, X.: Analysis and quantification of the diversities of aerosol life cycles within AeroCom, Atmospheric Chemistry and Physics, 6, 1777–1813, https://doi.org/10.5194/acp-6-1777-2006, https://www.atmos-chem-phys.net/6/1777/2006/, 2006.
- TF HTAP, title = Task Force on Hemispheric Transport of Air Pollution (TF HTAP), y. . . h. . http://www.htap.org/. n. . L.: .
 - Timmreck, C. and Schulz, M.: Significant dust simulation differences in nudged and climatological operation mode of the AGCM ECHAM, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2003JD004381, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003JD004381, 2004.
- 2560 Tsigaridis, K., Koch, D., and Menon, S.: Uncertainties and importance of sea spray composition on aerosol direct and indirect effects, Journal of Geophysical Research: Atmospheres, 118, 220–235, https://doi.org/10.1029/2012JD018165, https://agupubs.onlinelibrary.wiley.com/ doi/abs/10.1029/2012JD018165, 2013.
 - Tsigaridis, K., Daskalakis, N., Kanakidou, M., Adams, P. J., Artaxo, P., Bahadur, R., Balkanski, Y., Bauer, S. E., Bellouin, N., Benedetti, A., Bergman, T., Berntsen, T. K., Beukes, J. P., Bian, H., Carslaw, K. S., Chin, M., Curci, G., Diehl, T., Easter, R. C., Ghan, S. J., Gong, S. L.,
- Hodzic, A., Hoyle, C. R., Iversen, T., Jathar, S., Jimenez, J. L., Kaiser, J. W., Kirkevåg, A., Koch, D., Kokkola, H., Lee, Y. H., Lin, G., Liu, X., Luo, G., Ma, X., Mann, G. W., Mihalopoulos, N., Morcrette, J.-J., Müller, J.-F., Myhre, G., Myriokefalitakis, S., Ng, N. L., O'Donnell, D., Penner, J. E., Pozzoli, L., Pringle, K. J., Russell, L. M., Schulz, M., Sciare, J., Seland, Ø., Shindell, D. T., Sillman, S., Skeie, R. B., Spracklen, D., Stavrakou, T., Steenrod, S. D., Takemura, T., Tiitta, P., Tilmes, S., Tost, H., van Noije, T., van Zyl, P. G., von Salzen, K., Yu, F., Wang, Z., Wang, Z., Zaveri, R. A., Zhang, H., Zhang, K., Zhang, Q., and Zhang, X.: The AeroCom evaluation and intercomparison
- of organic aerosol in global models, Atmospheric Chemistry and Physics, 14, 10845–10895, https://doi.org/10.5194/acp-14-10845-2014, https://www.atmos-chem-phys.net/14/10845/2014/, 2014.
 - Tunved, P., Ström, J., and Krejci, R.: Arctic aerosol life cycle: linking aerosol size distributions observed between 2000 and 2010 with air mass transport and precipitation at Zeppelin station, Ny-Ålesund, Svalbard, Atmospheric Chemistry and Physics, 13, 3643–3660, https://doi.org/10.5194/acp-13-3643-2013, https://acp.copernicus.org/articles/13/3643/2013/, 2013.
- 2575 Twomey, S.: The Influence of Pollution on the Shortwave Albedo of Clouds, Journal of the Atmospheric Sciences, 34, 1149–1152, https://doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2, https://doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2, 1977.
 - van Marle, M. J. E., Kloster, S., Magi, B. I., Marlon, J. R., Daniau, A.-L., Field, R. D., Arneth, A., Forrest, M., Hantson, S., Kehrwald, N. M., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Yue, C., Kaiser, J. W., and van der Werf, G. R.: Historic global biomass burning emissions for
- 2580 CMIP6 (BB4CMIP) based on merging satellite observations with proxies and fire models (1750–2015), Geoscientific Model Development, 10, 3329–3357, https://doi.org/10.5194/gmd-10-3329-2017, https://www.geosci-model-dev.net/10/3329/2017/, 2017.

- van Noije, T., Bergman, T., and Le Sager, P.: EC-Earth3-AerChem, a global climate model with interactive aerosols and atmospheric chemistry participating in CMIP6, Geoscientific Model Development, in preparation, 2020.
- van Noije, T. P. C., Le Sager, P., Segers, A. J., van Velthoven, P. F. J., Krol, M. C., Hazeleger, W., Williams, A. G., and Chambers, S. D.:
 Simulation of tropospheric chemistry and aerosols with the climate model EC-Earth, Geoscientific Model Development, 7, 2435–2475, https://doi.org/10.5194/gmd-7-2435-2014, https://www.geosci-model-dev.net/7/2435/2014/, 2014.
 - Vignati, E., Wilson, J., and Stier, P.: M7: An efficient size-resolved aerosol microphysics module for large-scale aerosol transport models, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2003JD004485, https://agupubs.onlinelibrary.wiley.com/doi/ abs/10.1029/2003JD004485, 2004.
- 2590 Wang, R., Balkanski, Y., Boucher, O., Ciais, P., Schuster, G. L., Chevallier, F., Samset, B. H., Liu, J., Piao, S., Valari, M., and Tao, S.: Estimation of global black carbon direct radiative forcing and its uncertainty constrained by observations, Journal of Geophysical Research: Atmospheres, 121, 5948–5971, https://doi.org/10.1002/2015jd024326, https://doi.org/10.1002%2F2015jd024326, 2016.
 - Wang, R., Andrews, E., Balkanski, Y., Boucher, O., Myhre, G., Samset, B. H., Schulz, M., Schuster, G. L., Valari, M., and Tao, S.: Spatial Representativeness Error in the Ground-Level Observation Networks for Black Carbon Radiation Absorption, Geophysical Research Letters,
- 45, 2106–2114, https://doi.org/10.1002/2017GL076817, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL076817, 2018.
 Williams, J. E., Boersma, K. F., Le Sager, P., and Verstraeten, W. W.: The high-resolution version of TM5-MP for optimized satellite retrievals: description and validation, Geoscientific Model Development, 10, 721–750, https://doi.org/10.5194/gmd-10-721-2017, https://www.geosci-model-dev.net/10/721/2017/, 2017.
- Zanatta, M., Gysel, M., Bukowiecki, N., Müller, T., Weingartner, E., Areskoug, H., Fiebig, M., Yttri, K. E., Mihalopoulos, N.,
 Kouvarakis, G., Beddows, D., Harrison, R. M., Cavalli, F., Putaud, J. P., Spindler, G., Wiedensohler, A., Alastuey, A., Pandolfi, M., Sellegri, K., Swietlicki, E., Jaffrezo, J. L., Baltensperger, U., and Laj, P.: A European aerosol phenomenology-5: Climatology of black carbon optical properties at 9 regional background sites across Europe, Atmospheric Environment, 145, 346–364, https://doi.org/https://doi.org/10.1016/j.atmosenv.2016.09.035, http://www.sciencedirect.com/science/article/pii/ S135223101630735X, 2016.
- 2605 Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., Chen, J.-H., Chen, X., Donner, L. J., Dunne, J. P., Dunne, K., Durachta, J., Fan, S.-M., Freidenreich, S. M., Garner, S. T., Ginoux, P., Harris, L. M., Horowitz, L. W., Krasting, J. P., Langenhorst, A. R., Liang, Z., Lin, P., Lin, S.-J., Malyshev, S. L., Mason, E., Milly, P. C. D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps, P., Radhakrishnan, A., Ramaswamy, V., Robinson, T., Schwarzkopf, D., Seman, C. J., Shevliakova, E., Shen, Z., Shin, H., Silvers, L. G., Wilson, J. R., Winton, M., Wittenberg, A. T., Wyman, B., and Xiang, B.: The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 1. Simulation Char-
- 2610 acteristics With Prescribed SSTs, Journal of Advances in Modeling Earth Systems, 10, 691–734, https://doi.org/10.1002/2017MS001208, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001208, 2018a.
 - Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., Chen, J.-H., Chen, X., Donner, L. J., Dunne, J. P., Dunne, K., Durachta, J., Fan, S.-M., Freidenreich, S. M., Garner, S. T., Ginoux, P., Harris, L. M., Horowitz, L. W., Krasting, J. P., Langenhorst, A. R., Liang, Z., Lin, P., Lin, S.-J., Malyshev, S. L., Mason, E., Milly, P. C. D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps,
- 2615 P., Radhakrishnan, A., Ramaswamy, V., Robinson, T., Schwarzkopf, D., Seman, C. J., Shevliakova, E., Shen, Z., Shin, H., Silvers, L. G., Wilson, J. R., Winton, M., Wittenberg, A. T., Wyman, B., and Xiang, B.: The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 2. Model Description, Sensitivity Studies, and Tuning Strategies, Journal of Advances in Modeling Earth Systems, 10, 735–769, https://doi.org/10.1002/2017MS001209, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001209, 2018b.

Zieger, P., Fierz-Schmidhauser, R., Weingartner, E., and Baltensperger, U.: Effects of relative humidity on aerosol light scattering: results

2620

from different European sites, Atmospheric Chemistry and Physics, 13, 10609–10631, https://doi.org/10.5194/acp-13-10609-2013, https://www.atmos-chem-phys.net/13/10609/2013/, 2013.

Table 1. Observations used in this study, including relevant meta data information. ID: name of observation network; Source: Data source or subset; Var: Variable name; NMB: Normalised mean bias of satellite product at AERONET sites (monthly statistics); λ : Wavelength used for analysis (may be different from measurement wavelength, for details see text); Ver: Data Versionversion; Lev: Data level; Freq: Original frequency of data used to derive monthly means; Res: Resolution of gridded data product; Clim: Use of a multi-annual climatology or not; #st: Number of stations / coordinates, with observations used; Date: Retrieval date from respective database. See text in Section 2.1 for additional quality control measures that have been applied to some of these data base)sets.

Data ID and ID	Source	Variable	NMB [%]	λ [nm]	Ver.	Lev.	Freq.	Res.	Clim.	#st.	Date
		Var.									
GAW	EBAS	Abs. coeff.		3_550		2~	hourly		Y	39	20192020/12 04/ 18
GAW	EBAS	ACdrx. Sc. coeff.		3_550_		2	hourly		Y	37-39_	<u>01</u> 20192020/1204/18
AERONET <mark>Sun-</mark>		<u>SCdry</u> AE		4.2. c.1	3	2	daily		N	250	01
	AOD			4- <i>λ</i> ħť~	2	2	1.1		N	2 30 245	13
AERONEI Sun	AOD	AOD		<u>550</u>	3	2	daily		N	240 235	20192020/0908/20 13
AERONET	SDA	AOD <mark><1um</mark> £		550	3	2	daily		Ν	~~~ 226	~~ 2019 2020/ 09 08/ 20
AERONET	SDA	AOD <mark>>1um</mark> c		<u>550</u>	3	2	daily		Ν	222 226	13 20192020/0908/20
MODISt — Terra	Terra DT/DB	AOD	+16.5	<u>550</u>	6.1	3	daily	1x1	N	222 2235	1 <u>3</u> 2019/11/22
MODIS-T MODISa Aqua	Aqua DT/DB	AOD	+8.9	<u>550</u>	6.1	3	daily	1x1	N	246 2241	2019/11/25
MODIS-A AATSR-SU	Swansea	AOD	-4.2	550	4.3	3	daily	1x1	Ν	246 2055	2016/09/30
AATSR-SU	Swansea	AE	+14.3	550-865	4.3	3	daily	1x1	N	246 2055	2016/09/30
AATSR-SU	Swansea	AOD <mark><1um</mark>	+1.6	550	4.3	3	daily	1x1	N	2 <u>57</u> 2055	2016/09/30
AATSR-SU	Swansea	£ AOD >1um	-14.7	80 550	4.3	3	daily	1x1	Ν	233 2055	2016/09/30
		R.	~~~~~	~~~						233	

Table 2. Models used in this study including relevant additional information. Kinne et al., along with 2006: name of model in Kinne et al. (2006) (see table 2 therein; where applicable); Lat./Lon.: horizontal grid resolution,-; Levs.: number of vertical levels; Type: type of atmospheric model; Aerosol module: name of aerosol module: Scheme: type of aerosol scheme; Meteorology: meteorological data set used for the simulated year 2010; CS: clear-sky optics available (LevsY/N) and-; AC:: availability of dry surface absorption coefficient fields for comparison with GAW observations; References: key references. More details about the models can be found in the supplementary material 1&2.

Name-This study	Kinne et al., 2006
CAM5-ATRAS	N/A
EC-Earth3-AerChem	N/A
TM5	TM5
ECHAM-HAM	MPI-HAM
ECHAM-SALSA	N/A ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
ECMWF-IFS	N/A
EMEP	N/A
GEOS	GOCART
INCA L3 x 25.79 GFDL-AM4	Balkanski et al. (2004), Schulz et al. (2009)NorESM2 N/A
Temporal AOD ECMWF-IFS 3 hourly AERONET GISS-OMA	+6.9 GISS
Model names and corresponding AeroCom database IDs. Also indicated is whether models diagnosed a clear-sky AOD (CS) or not and which models were included in the AeroCom mean and median ensemble. Name AeroCom ID-INCA	CS AODLISCE
ECHAM-HAM-NorESM2 (CAM6-Nor)	ECHAM6.3-HAM2.3-met2010_AP3-CTRL_UIO_GCM
EMEP OsloCTM3	EMEP_rvd_33_Glob CTRL-UIO_CTM
INCA INCA_AP3-CTRL_MIROC_SPRINTARS	¥KYU

Table 3. Global annual averages for each aerosol species, grouped by aerosol emissions, lifetimes , burdens, optical depths (ODs), mass extinction coefficients (MEC) and mass absorption coefficient (MAC), for models participating in the AP3-CTRL experiment, year 2010. Also shown in the OD section are total AOD for all-sky (AOD (AS)) and clear-sky (AOD (CS)) conditions as well as AOD due to water (H₂O). The following columns show the median from all model values (MED) and associated diversities as interquartile range and standard deviation (δ_{IQB} , δ_{std}). AP1 median and standard deviation are based on values given in Table 10 in Textor et al. (2006) and Table 4 in Kinne et al. (2006). Colors illustrate the bias of individual model and AP1 median values with respect to the AP3 median. Units of emissions and burdens are full molecular weight (for OA and POA, the total organic weight is used). Note that the "emissions" of SO₄, NO₃ and OA are really secondary chemical formation in the atmosphere plus primary particle emissions. They are computed using total deposition as a proxy (indicated with \downarrow). For BC↑, DU↑, POA↑ and SS↑ the provided emission data were used. For OsloCTM3 an additional OD of 0.0086 due to biomass burning was reported and is not included here. See further details on parameter computation in section 2.2. Values in brackets indicate erroneous or inconsistent values (i.e., BC OD, MEC and MAC from some models) and are not included in the corresponding AP3 median value (MED) and diversities (details are discussed in Sect. 3.1).

	CAMS'R,	ins the	any	ECHANA INS	ECHANN'S	ECMM SALSA		Mrs (KOS GROU	CISS MA	OMA	INCS NOT	Shi Osloc	SPRIN.	ARS .	AP3 MED	(this st δ _{IQR}	udy) δ _{std}	AF MED	ν 1 δ _{std}			
[Tg/yr]	BC (†) - DU (†) - NO2 (1) -	9.5 1930 20.2	9.7 848	9.7 1120	9.8 1170	9.8 1380	9.6 5650	8.4 1640	9.7 1480 128.0	9.7 1360	9.7 1400	9.7 1560 44.8	9.7 1090	9.1 2150	9.7 1880	9.7 1440 32.5	0.4% 42% 286%	3.8% 64% 90%	11.3 1640	23% 49%			
nissions"	OA (↓) - POA (↑) - SO₄ (↓) -	117.0	108.0 142.0	108.0 146.0	69.2 218.0	76.6	160.0 122.0	246.0 94.3	158.0 62.0 95.3	170.0 71.2 110.0	75.6	48.0 144.0	177.0 81.5 124.0	114.0 77.9 188.0	134.0 240.0	116.0 74.6 143.0	64% 41%	40%	96.0 186.0	26% 22%			
Ē	SS (↑) - BC -	5090	6640 8 7	6620 8.4	5920 6.4	4160 9.6	5e+04 उष	2030	5300 4 1	7100 5.9	2760 4 2	4030	3450 6.4	4880 4 4	3650	4980	54% 41%	150% 35%	6280 6.5	199% 33%			
ne [d]	DU - NO3 -	3.0 3.0	3.9 9.9	4.0 10.4	6.0	7.0	1.4 2.5	3.2 3.1	5.4 2.7	3.5	5.3 4.7	4.5 5.4	1.9	3.4	2.2	3.7 3.9	56% 94%	39% 58%	4.0	43%	Î	- 100%	
Lifetir	OA - SO ₄ - SS -	6.1 3.3 0.61	9.3 6.7 0.36	8.8 7.0 0.41	6.0 4.3 0.63	8.2 5.0	4.3 3.1 0.19	4.3 2.6 0.50	4.6 4.9 0.67	4.5 6.3 0.45	6.3 4.2	6.0 4.9	6.2 4.9	5.3 5.3 0.46	3.4 1.8 0.24	6.0 4.9 0.56	29% 36% 92%	29% 32% 91%	6.2 4.1 0.41	27% 18% 58%		- 75%	
Tg]	BC - DU -	0.119 16.1	0.231 9.1	0.224 12.3	0.174 19.5	0.260 17.1	0.104 22.3	0.068 14.3	0.105 21.9	0.157 12.9	0.111 20.5	0.143 19.4	0.171 5.7	0.117 20.1	0.119 11.7	0.131 16.6	46% 45%	36% 31%	0.210 20.5	42% 40%		- 50%	
Burden [NO3 - OA - SO4 -	0.164 1.95 1.41	0.167 2.73 2.62	0.154 2.61 2.78	1.14 2.58	1.71 2.95	0.690 1.87 1.02	0.926 2.87 0.66	0.929 1.98 1.29	2.07 1.92	0.142 1.30 1.39	0.662 0.79 1.92	<mark>2.99</mark> 1.67	0.084 1.65 2.76	1.24 1.18	0.167 1.91 1.80	321% 57% 72%	78% 34% 39%	1.76 1.98	27% 25%		- 25%	
	SS - BC -	8.6 0.002	6.5 (0.003)	7.5 (0.002)	10.3 (0.002)	10.2 (0.003)	26.4 0.002	2.8 (0.001)	9.8 0.002	8.8 0.004	8.3 0.003	10.9 (0.001) 0.031	9.7 0.002	6.2 0.002	2.5 0.001	8.7 0.002	38% 20% 32%	59% 33% 27%	6.4 0.004	54% 45% 42%		- 10%	EDIAN
0	NO3 - OA -	0.002	0.003	0.002	0.013	0.018	0.005	0.010	0.007	0.031	0.005	0.005	0.029	0.002	0.034	0.005	60% 82%	57% 43%	0.018	34%		10%	Δ_{ME}
ы	- SO ₄ SS - H ₂ O -	0.010	0.035 0.029 0.055	0.036 0.033 0.058	0.045 0.058 0.096	0.045 0.065 0.087	0.017	0.018	0.027	0.046	0.024	0.047	0.015 0.078 0.089	0.044 0.038 0.048	0.040 0.013 0.070	0.035	70% 60% 45%	39% 41% 24%	0.034	32% 41%		25%	
	AOD (AS) - AOD (CS) -	0.106	0.132	0.138	0.148	0.155	0.110	0.097	0.125	0.140	0.127	0.156	0.144 0.130	0.130	0.104 0.060	0.128	26% 9%	14% 22%	0.127	18%		50%	
[m ² /g]	BC - DU - NO3 -	10.2 0.39 5.2	(5.8) 1.24 8.1	(5.3) 1.21 8.1	(6.0) 0.80	(5.7) 0.73	9.4 0.46 3.8	(8.2) 0.40 5.5	10.5 0.60 3.9	11.9 0.74	13.7 0.52 18.2	(4.8) 0.82 4.2	7.1 1.68	13.4 0.62 9.7	3.7 0.74	10.3 0.74 5.5	34% 37% 71%	31% 45% 58%	9.1 0.92	35% 44%		75%	
] MEC	OA - SO ₄ - SS -	4.3 3.8 3.58	7.6 6.9 2.28	7.0 6.6 2.23	5.6 8.8 2.86	5.2 7.7 3.25	4.6 8.2 0.93	6.2 13.8 2.82	7.0 9.2 1.85	7.8 12.3 2.33	5.5 8.9 3.68	6.0 12.5 2.86	5.0 4.7 4.11	6.3 8.2 3.11	13.3 17.2 2.62	6.1 8.5 2.84	28% 52% 33%	33% 38% 28%	5.4 9.0 2.94	28% 54% 53%		100%	
AC [m ² /g	BC - DU - OA -	9.1 0.027 0.163			(10.2) 0.015 0.080	(15.0) 0.033 0.112		(10.4)	7.8 0.046		10.0 0.053 0.274	(7.5) 0.048 0.142	3.2 0.050 0.265	13.0 0.042 0.110	3.1 0.031 0.122	8.5 0.042 0.132	40% 58%	31% 43%		·			
Σ	CAMPAN	RAS EC.S	attr	TMS FCHAM	HAM - CHAMS	ALSA ECANY	F.IFS F	IMER	SFOS GFOL	AMA GISS	DMA V	NOR NOT	5M2 OSIOC	TM3 SPRINT	ARS	MED AP3	δ _{IQR} (this st	δ _{std} udy)	MÉD AF	δ _{std}			

Table 4. Normalised mean bias (NMB, top) and Pearson correlation coefficients (bottom), computed from the monthly co-located data for each model (columns) and observation/variable combination (rows). For the comparison with the gridded $5^{\circ} \times 5^{\circ}$ satellite products, area weights were applied to compute the average metrics shown here. Please note that the biases do not represent global averages but the site/sampling locations of each data set with more weight given to regions with higher spatial density (e.g., Fig. 1). Please also note potential offsets in the absolute biases arising from uncertainties in the observation retrievals, particularly for the satellite products (see Tab. 1). The rightmost columns show evaluation results from the ensemble median (ENS-MED) as well as first and third quantile fields (Q1, Q3). The latter indicate the spread of the results. Model values for AODs are for clear-sky conditions unless only all-sky were available.

CANAS	¢.	<	Cria C	HAN	FCA.			ଦ୍ନ	୍ତ୍ୟ		Va	Q.	Sop	€N.					
(FF		74	INS INS	AND SA	SA	1.5 °	MES CI		My SC	MA II	NCA CK	SM2 OC	IN IN	ARS S.N.	KD (Q,	<i>6</i> 3		
ΔΕ (ΔERONET) -	2	2	4	-23	-25	-9	35	-2	4	-20	-61	-11	-21	-11	-10	-27	6	- 100	
AE (AATSB-SU) -	-37	-2	1	-25	-30	-35	40	-22	-2	-43	-68	-34	-23	-9	-22	-44	-1	- 75	
AOD (AFRONET) -	-37	-6	1	-27	-15	-17	-10	-1	-5	-26	3	-46	-6	-50	-21	-41	-1	- 75	
AOD (MERGED-EMI) -	-22	-7	-3	-4	11	-20	-34	-14	-4	-12	11	-6	-11	-57	-16	-36	8	- 50	
AOD (MERGED-FMI-LAND) -	-26	1	3	-31	-13	-15	-16	-1	2	-25	9	-26	0	-38	-17	-37	7		
AOD (MERGED-FMI-OCN) -	-19	-10	-5	11	25	-22	-43	-21	-7	-5	13	6	-17	-67	-15	-35	10	- 25	_
AOD (AATSR-SU) -	-18	-5	-2	-6	12	-18	-34	-15	-8	-18	10	-9	-12	-55	-17	-37	7	- 10	%
AOD (MODIS-A) -	-37	-22	-19	-21	-9	-35	-45	-28	-19	-26	-8	-22	-25	-64	-31	-48	-10	10	m
AOD (MODIS-T) -	-42	-29	-26	-28	-17	-40	-49	-34	-26	-33	-16	-29	-32	-68	-37	-52	-18	10	Σ
AOD _c (AERONET) -	-48	-23	-14	-57	-2	-33	-67	25	-65	-43		-18	-46	-84	-46	-67	-17	25	z
AOD _c (AATSR-SU) -	-2	-20	-10	-28	24	15	-70	5	-54	-8		44	-39	-82	-27	-55	7	25	
AOD _f (AERONET) -	-31	5	8	11	-21	-11	20	-11	27	-19		-36	10	-36	-13	-34	9	50	
AOD _f (AATSR-SU) -	-28	4	3	20	3	-40	-11	-28	22	-24		-29	5	-38	-17	-39	8		
SC _{dry} (GAW) -	-32	-13	-14	-52	-64	12	-46	-4	-18	18		-62	-12	-63	-35	-59	-11	75	
AC _{dry} (GAW) -	-40	6	1	-15	-28	16	-28	-28	21	-5		-48	58	-55	-20	-38	2	100	
AE (AEDONET)	0.66	0.69	0.69	0.49	0.50	0.77	0.71	0.76	0.52	0.68	0.65	0.58	0.68	0.53	0.73	0 73	0.72	1.0	
	0.00	0.03	0.03	0.49	0.50	0.77	0.65	0.70	0.52	0.00	0.05	0.50	0.00	0.55	0.75	0.75	0.72	- 0 0	
ACD (AEPONET) -	0.72	0.72	0.79	0.65	0.63	0.00	0.05	0.84	0.72	0.81	0.74	0.75	0.70	0.00	0.70	0.75	0.87	0.9	
AOD (MERGED-EMI) -	0.65	0.83	0.82	0.58	0.60	0.80	0.72	0.86	0.80	0.69	0.72	0.55	0.82	0.77	0.87	0.86	0.86	- 0.8	
AOD (MERGED-EMI-LAND) -	0.70	0.80	0.79	0.67	0.67	0.83	0.70	0.84	0.77	0.85	0.72	0.67	0.84	0.75	0.88	0.87	0.88	- 0 7	
AOD (MERGED-FMI-OCN) -	0.53	0.82	0.82	0.52	0.52	0.69	0.70	0.86	0.78	0.44	0.68	0.42	0.75	0.77	0.82	0.80	0.79	0.7	ि
AOD (AATSB-SU) -	0.66	0.75	0.75	0.49	0.57	0.82	0.73	0.85	0.79	0.74	0.70	0.57	0.83	0.77	0.85	0.84	0.83	- 0.6	105
AOD (MODIS-A) -	0.57	0.79	0.79	0.53	0.53	0.71	0.68	0.81	0.77	0.64	0.66	0.49	0.76	0.71	0.80	0.78	0.80	- 0.5	are
AOD (MODIS-T) -	0.55	0.78	0.76	0.52	0.51	0.69	0.66	0.78	0.76	0.63	0.63	0.49	0.74	0.70	0.78	0.76	0.78		Pe
AOD _c (AERONET) -	0.73	0.61	0.62	0.39	0.48	0.79	0.69	0.73	0.71	0.78		0.48	0.78	0.67	0.80	0.78	0.80	- 0.4	-
AOD _c (AATSR-SU) -	0.63	0.56	0.64	0.26	0.59	0.78	0.64	0.81	0.62	0.70		0.12	0.82	0.75	0.74	0.68	0.76	- 0.3	ш.
AOD _f (AERONET) -	0.74	0.82	0.79	0.56	0.78	0.79	0.76	0.79	0.75	0.76		0.78	0.75	0.79	0.83	0.84	0.82		
AOD _f (AATSR-SU) -	0.68	0.85	0.81	0.53	0.62	0.81	0.74	0.83	0.83	0.76		0.79	0.77	0.79	0.85	0.86	0.85	- 0.2	
SC _{drv} (GAW) -	0.44	0.71	0.70	0.27	0.51	0.59	0.74	0.73	0.56	0.66		0.51	0.69	0.61	0.72	0.66	0.73	- 0.1	
AC _{dry} (GAW) -	0.60	0.68	0.72	0.80	0.79	0.81	0.79	0.75	0.63	0.69		0.80	0.74	0.62	0.75	0.72	0.71	0.0	
-	5	~	5	4		<i>с</i>	,o'	5	. >			. <u>.</u> .	<u>`</u>	5	ດ່	~	3		
S	\$ 42	\$~` ~	N. Y	AL	S' NK	×´ d	Nr. G	N. X	ZH. C	hu. H	ی بر ک	Mr. C	M. MA	N. N	e.	α,	0,		
NST	40		HAM	, chi?	CNNA	•	•	GEDY	GKS7		401	05102	PHIL	ENS					
CAN		~	ن ب	· ۲	×-			-	-			-	Э.	*					

Table 5. Comparison of statistics (NMB and R) retrieved when co-locating models with satellite data a) in monthly resolution and $5^{\circ} \times 5^{\circ}$ horizontally with requirement of at least 7 daily values Results from sensitivity studies related to compute a monthly mean, as done in this study spatio-temporal representation errors. AERONET* indicates that two different site selection schemes were used (*Low*) and b) in daily resolution and see text in highest available horizontal resolution from both data-sets (*High*Sect. 4.5 for details).

CAM5-ATRAS AATSR4.3-SU AOD -1.8 -2.1 0.67 0.51 MODIS6.1-aqua AOD -25.9 -20.4 0.58 0.36 MODIS6.1-terra AOD -33.3 -28.2 0.58 0.36 ECM

OsloCTM3-2

TM5-

3____



Figure 1. <u>Overview of data used for model evaluation.</u> Yearly averages of AODs from <u>a</u>) AERONETand; <u>b</u>) merged satellite data-set (top panel)<u>data set</u>, <u>c</u>) fine and <u>d</u>) coarse AOD from AERONET(2nd panel), <u>e</u>) AE from AERONETand, <u>f</u>) AATSR(3rd panel, g) as well as surface in-situ observations of dry scattering and <u>h</u>) dry absorption coefficients from surface in situ observations.


Figure 2. NMBs from satellite evaluation against AERONET Relation between aerosol lifecycle and optical parameters for different individual models along with model diversity. The individual panels show model spread of global annual averages for each of the considered lifecycle and optics variables (x-axis) and for each model (different colors). The y-axis corresponds to the percentage bias from the ensemble median. Also plotted are the corresponding correlation coefficients model spread (gray shaded area, IQR) as well as the numerical values of median and IQR (in green gray colors at the bottom of each subplot; values correspond to Table 3 but may differ due to rounding errors). Note that fine some models reported erroneous BC MECs and coarse AOD from MODIS terra is ODs which are not further used in this studyincluded here (for details see Tab. 3).



Figure 3. Global emissions of major aerosol species and precursors. Units are full molecular weight and Species contributions to total AOD for OA, the each model (annual global average). The type of total organic weight AOD (AS or CS) is used indicated at the top of each bar. Note that only major species are included and that other potentially provided species (eBB* denotes biomass burning OD in OsloCTM3. gModels with ** in their name submitted speciated ODs for AS conditions and total AOD for both CS and AS. or , VOCs) The corresponding CS AODs are not shownindicated in red with a + symbol. The rightmost columns show mean, median and spread Also shown are estimates of the results total global CS AOD from the individual models, the latter being computed as the half difference between 1st AERONET and 3rd quantiles. Colors are scaled MERGED-FMI (see main text for details), similar to min and max row-wise Fig. 3 in order to highlight differences between the modelsKinne et al. (2006).

Global lifetimes in days for all major aerosol species, computed from burdens (Fig. ??) and total deposition (wet + dry). A more detailed description of this plot type is provided in Fig. ??. Global annual burdens of major aerosol species in units of Tg. A more detailed description of this plot type is provided in Fig. ??. Globally averaged columnar MECs of models for all major aerosol species. The MEC for each species *i* is computed from OD_{*i*}/LOAD_{*i*} (Figs. ??, ??). Note that the two ECHAM models reported the OD_{*i*} fields at dry conditions (indicated with a star, Fig. ??) and thus, show comparatively small MECs for the hydrophilic species. Therefore, they were excluded for the computation of mean, median and diversity shown in the rightmost columns. Hence, A more detailed description of this plot type is

provided in Fig. ??.

ODs from individual species as well as the sum and, dependent on availability clear-sky and all-sky AOD. Please note that for OsloCTM3 an additional OD of 0.0086 due to biomass burning was reported (combination of OA and BC) which is not included here. Like in Fig. ??, the two ECHAM models were excluded for the computation of mean, median and diversity, since dry speciated ODs were reported. A more detailed description of this plot type is provided in Fig. ??.



Figure 4. LeftLeft: maps Maps showing yearly averages of relevant optical property variables from the <u>AeroCom</u> ensemble model as well as mean values from corresponding ground-based network used median (circlesENS-MED). Also shown are The number in the lower left corner of each map represents the yearly mean global average values from model (both global and at obs. stations) as well as the observation mean from all stationsensemble. RightRight: Corresponding diversity fields of ensemble mean calculated using standard deviation of the individual results normalised by the mean (Textor et al. $(2006)\delta_{IQR}$) for each variable, including global average diversity in lower right corner.



Figure 5. Figure showing NMB in percent-Yearly average NMBs of the ensemble median AOD against the several observation records. 1st row: AODs from AERONET and merged merged satellite AOD data-set (circles, only ocean locations are displayed) data set. 2nd row: fine and coarse AODs from AERONET(triangles) as well as surface scattering coefficient against the in-situ sites (diamond). The edge colors of the markers correspond to the respective global average NMB, which is also indicated in the legend as well as Pearson correlation coefficient 3rd row: AE from AERONET and total number of monthly-AATSR-SU satellite data points set. 4th row: surface dry scattering and number of stations respective grid points for the Merged FMI productabsorption coefficients from GAW in situ sites.

Pearson correlation coefficients (R) computed from the monthly colocated data for each model (columns) and observation / variable

combination (rows). For the $5^{\circ} \times 5^{\circ}$ satellite products, area weights were applied to the monthly values. Please note further remarks on



Figure 6. Overall Summary of results from comparison of optical properties evaluation for all models with ground-based observation networks and the AeroCom ensemble model/MERGED-FMI satellite AOD data set. NMB Biases of all variables for AeroCom median (blue triangles) and mean The y-axis indicates the retrieved biases (red triangles/NMBs) as well as those from for individual models (indicated as circles). Pearson correlation coefficients are plotted in red-yellow-green colors The black boxes indicate results from the ensemble median (same as in Fig. ??ENS-MED). Also included is the standard deviation of NMBs from, together with the models for each variable associated spread (red and blue error bars δ_{IQR}).

Normalised mean bias (NMB) computed from the monthly collocated data for each model (columns) and observation / variable combination (rows). For the $5^{\circ} \times 5^{\circ}$ satellite products, area weights were applied to compute the average bias. Please note that the biases do not represent global averages but the site / sampling locations of each data-set with more weight given to regions with higher spatial density (see e.g.Fig. 1). Please also note potential offsets in the absolute biases arising from uncertainties in the observation retrievals, particularly for the satellite products (Seet. ?? and Fig. ??).

Model biases of surface dry scattering at all GAW in-situ sites that had sufficient temporal coverage to compute monthly elimatology. Model biases of surface absorption coefficient at all GAW in-situ sites that had sufficient temporal coverage to compute monthly elimatology. Scatter plot showing results of 3-hourly (left) vs. monthly (right) colocation of AOD from

2630 ECMWF-IFS model against AERONET all points data. Also included are statistical results. Scatter plot showing results of hourly (left) vs. monthly (right) colocation of in-situ surface absorption from TM5 model (from AeroCom INSITU experiment, i.e. different version than the one used in this study) evaluated at GAW stations. Also included are statistical results.



Figure 7. Scatter plot showing colocation results of the ENSEMBLE AE model biases in different AE regimes. The biases for each model AOD evaluated at all available AERONET stations (lefty-axis) and evaluated only at stations are retrieved by co-location with small spatial representativity errors AERONET observations, selected based on using only measurements that fall into the results from Wang et al. (2018) respective AE bin.