AEROCOM/AEROSAT ~~AAOT~~ AAOD & SSA study, part I: evaluation and intercomparison of satellite measurements

Nick Schutgens[1], Oleg Dubovik[2], Otto Hasekamp[3], Omar Torres[4], Hiren Jethva[5], Peter J.T. Leonard[6], Pavel Litvinov[2], Jens Redemann[7], Yohei Shinozuka[8,9], Gerrit de Leeuw[10,11], Stefan Kinne[12], Thomas Popp[13], Michael Schulz[14], and Philip Stier[15]

[1]Department of Earth Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, the Netherlands
[2]Laboratoire d'Optique Atmosphérique, CNRS/Université Lille, Villeneuve d'Ascq, France
[3]SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
[4]Atmospheric Chemistry and Dynamics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[5]Universities Space Research Association-GESTAR, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[6]ADNET Systems, Inc., Suite A100, 7515 Mission Drive, Lanham, MD 20706, USA
[7]School of Meteorology , University of Oklahoma, Norman, USA
[8]Universities Space Research Association, Columbia, Maryland, USA
[9]NASA Ames Research Center, Moffett Field, California, USA
[10]Finnish Meteorological Institute (FMI), Climate Research Programme, Helsinki, Finland
[11]currently at: Royal Netherlands Meteorological Institute (KNMI), R&D Satellite Observations, De Bilt, the Netherlands
[12]Max-Planck-Institut für Meteorologie, D-20146 Hamburg, Germany
[13]German Aerospace Center (DLR), German Remote Sensing Data Center Atmosphere, Oberpfaffenhofen, Germany.
[14]Norwegian Meteorological Institute, P.O.Box 43, Blindern, 0313 Oslo, Norway.
[15]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, UK.

Correspondence: Nick Schutgens (n.a.j.schutgens@vu.nl)

Abstract. Global measurements of ~~absorptive~~ absorbing aerosol optical depth (AAOD) are scarce and mostly provided by the ground network AERONET (AErosol RObotic NETwork). In recent years, several satellite products of AAOD have ~~appeared~~been developed. This study's primary aim is to establish the usefulness of these datasets for AEROCOM (AEROsol Comparisons between Observations and Models) model evaluation with a focus on the years 2006, 2008 and 2010. The satellite products are super-observations consisting of $1^o \times 1^o \times 30^{\min}$ aggregated retrievals.

This study ~~consist of two parts: 1) an~~ consists of two papers, the current one that deals with the assessment of satellite ~~datasets; 2) their application to~~ observations and a second paper that deals with the evaluation of ~~AEROCOM models . The current paper describes the first part and~~ models using those satellite data. In particular, the current paper details an evaluation with AERONET observations from the sparse AERONET network as well as a global intercomparison of satellite datasets, with a focus on how minimum AOD (Aerosol Optical Depth) thresholds and temporal averaging may improve agreement between satellite observations.

All satellite datasets are shown to have reasonable skill for AAOD (3 out of 4 datasets show correlations with AERONET ~~to be $r > 0.6$~~in excess of 0.6) but less skill for SSA (Single Scattering Albedo; only 1 out of 4 datasets shows correlations with AERONET ~~to be $r > 0.6$). Nevertheless,~~ in excess of 0.6). In comparison, satellite AOD shows correlations from 0.72 to 0.88 against the same AERONET dataset. However, we show that performance vs. AERONET and ~~dataset~~ inter-satellite agreements

for SSA ~~significantly improve~~ improve significantly at higher AOD. Temporal averaging also improves agreements between satellite datasets. Nevertheless multi-annual averages still show systematic differences, even at high AOD. In particular, we show that two POLDER products appear to have a systematic SSA difference over land of ~ 0.04, independent of AOD. Identifying the cause of this bias offers the possibility of substantially improving current datasets.

We also provide evidence that suggests that evaluation with AERONET observations leads to an underestimate of true biases in satellite SSA.

In the second part of this study we show that~~notwithstanding the~~ , notwithstanding these biases in satellite AAOD and SSA, ~~these~~ the datasets allow meaningful evaluation ~~with~~ of AEROCOM models.

1 Introduction

Aerosol is an important component of the Earth's atmosphere that affects the planet's climate, the biosphere, and human health. Aerosol particles scatter and absorb sunlight as well as modify clouds. Anthropogenic aerosol changes the radiative balance and influences global warming (**??????**). It may negatively affect solar power generation (**??**). Aerosol can transport soluble iron, phosphate and nitrate over long distances and provide nutrients for the biosphere (**?????**) . Aerosol can penetrate deep into lungs and may carry toxins or serve as disease vectors (**??????**).

Aerosol reflects visible radiation from the Sun, and some aerosol also absorbs it (**??**). The species that absorb the most visible sunlight are, in order of importance: black carbon, dust and brown carbon. Of these, black carbon is expected to exert a significant positive radiative forcing on the climate (**??**). Absorbing aerosol's impact is mostly through heating of the atmospheric profile (direct effect) and subsequent stabilisation or instabilisation (**?**) of the boundary layer (semi-direct effect). This affects cloud formation (**??**) and precipitation (**???**). In particular over bright surfaces (ice, deserts, clouds) ~~can~~ the forcing due to absorbing aerosol can be significant (**???**).

On regional scales, biomass burning smoke has been implicated in increased tornado severity (**?**) while dust was observed to reduce cyclones (**?**), black carbon may affect the Hadley cell circulation (**??**), and black carbon deposition can reduce glacier albedo (**???**) which may speed up glacier melt.

Currently, ~~absorptive~~ absorbing aerosol can be measured in a number of ways. AERONET (**?**) is a global but spatially sparse network of sun photometers that includes two scanning protocols (almucantar and hybrid) that allow inversion of measured radiances into particle size distributions and refractive indices (**?**). From this inversion, columnar ~~properties AOD and~~ AAOD can be derived. There are also networks ~~of~~ (**?**) of (filter-based) absorption photometers, as used in EMEP (European Monitoring and Evaluation Programme), ACTRIS (Aerosol, Clouds and Trace Gases Research Infrastructure) and IMPROVE (Interagency Monitoring of Protected Visual Environments). These networks are concentrated in Europe and North America, and there is no global coverage. Moreover, these are surface measurements that do not measure the full atmospheric column. Finally,

absorption photometers like the SP2 were used on flight campaigns like HIPPO (**???**). Again, this yields spatially sparse in-situ observations of absorbing aerosol. While these measurement networks have proven to be very important to our understanding of absorbing aerosol, a satellite derived AAOD would contribute greatly by adding spatial context in regions with ground-based instruments, and measurements in regions without such instruments. As it now stands, we have almost no observations of ~~absorptive~~ absorbing aerosol over the oceans, in particular in continental outflow regions.

However, in recent years a number of satellite AAOD products have ~~appeared~~been developed, often based on POLDER (Polarization and Directionality of the Earth's Reflectances) measurements. For example, **?** used POLDER data to evaluate SSA from AEROCOM models over oceans; **?** evaluated over ocean above-cloud SSA in AEROCOM models for the African fire season; **?** estimated the global direct radiative effect of aerosol and ~~**?? assimilated ?**~~ estimated aerosol-cloud interactions. **??** assimilated POLDER AOD and AAOD observations to estimate aerosol emissions while **?** showed the benefit of jointly assimilating POLDER AOD, AAOD and SSA observations. **?** used combinations of A-TRAIN sensors to infer AAOD over clouds and estimate short-wave direct aerosol effects.

The challenge in retrieving AAOD from satellite is made clear by the challenge ~~of~~ in retrieving AAOD from AERONET measurements. AERONET AAOD observations are known to be more ~~error prone~~ uncertain than AOD observations. **?** estimated that AERONET SSA ~~errors~~ uncertainties for AOD ≤ 0.2 at 440 nm would be at least 0.05, using numerical sensitivity tests. A recent in-depth estimate of the uncertainty in Inversion V3 data (**?**) ~~suggested those thresholds to be 440nm AOD > 0.3 and ≥ 0.45, respectively.~~ for four different sites suggested SSA uncertainties at AOD (at 440 nm) $= 0.2$ from 0.037 to 0.048 at 440 nm and from 0.035 to 0.045 at 675 nm. It is not clear whether these uncertainties should be interpreted as site-specific biases or random errors. This distinction matters as random errors can be reduced through appropriate averaging of data. Large differences between AERONET SSA at low AOD and in-situ measurements were indeed confirmed by **?**. Even at higher AOD (≥ 0.5), **?** suggested SSA errors of at least 0.03. **?** suggest smaller SSA uncertainties of 0.017 to 0.023 at 440nm and 0.015 to 0.026 at 675 nm for AOD (at 440 nm) = 0.6 . Given the challenges in satellite remote sensing compared to ground-based remote sensing, satellite AAOD and SSA products ~~may~~ can be expected to have ~~larger errors~~ large errors as well.

GCOS requirements (**?**) for SSA specify an accuracy within 0.03 and a stability per decade within 0.01, for a horizontal resolution of 5–10 km and a temporal resolution of 4^{hr}. These requirements appear based on typical regional and yearly variations in SSA. However, SSA requirements are different for different applications (monitoring, trends, model evaluation, process studies) while the GCOS requirements are meant to provide a general broad estimate (**?**). In part 2 of our study we will show that current satellite AAOD and SSA capabilities allow useful evaluation of models.

For measurements to be useful in model evaluation, their errors after averaging (spatially, temporally) need to be smaller than the model errors the observations should be able to identify. A traditional evaluation of satellite datasets with AERONET data is unlikely to establish this, partly because the model aspect is ignored, partly because AERONET ~~hardly~~ covers some very interesting aerosol source regions (e.g. oceans, most deserts and boreal fire scapes) only sparsely. In the first part of this study (the current paper) we complement the traditional evaluation with a satellite intercomparison (in itself not unusual) to broaden our understanding of satellite performance over diverse regions. In the second part (a follow-up paper), we present a

novel analysis that combines satellite evaluation & intercomparison with model evaluation, and allows assessment of model biases in the ~~satellite data in the context of model~~ context of satellite biases.

We will use satellite data aggregated over $1^o \times 1^o \times 30^{\min}$ as it allows spatio-temporal collocation amongst datasets (satellite, AERONET, AEROCOM) which should strongly reduce representation errors in our analyses (**??**). All analyses, even of multi-year averages, will start from spatio-temporally collocated datasets.

This paper is the result of discussions in the AeroCom (AEROsol Comparisons between Observations and Models, https://aerocom.met.no) and AeroSat (International Satellite Aerosol Science Network, https://aero-sat.org) communities. Both are grass-roots communities, the first organised around aerosol modellers, and the second around retrieval groups. They meet every year to discuss common issues in the field of aerosol studies.

The observational ~~model~~ datasets used in this study are described in Sect. **??**. The collocation and analysis methodology are described in Sect. **??**. A first look at the satellite datasets is presented in Sect. **??**. Evaluation of satellite AOD, AAOD and SSA with AERONET is performed in Sect **??** and a more detailed intercomparison of satellite data is shown in Sect. **??**. A summary and conclusions can be found in Sect. **??**.

2 Datasets

2.1 Remote sensing data

Original satellite L2 data (estimates of geophysical variables on the spatio-temporal sampling pattern of the radiances, see also **?**) were aggregated unto a regular spatio-temporal grid with spatio-temporal grid-boxes of $1^o \times 1^o \times 30^{\min}$. The resulting super-observations ($1^o \times 1^o \times 30^{\min}$ aggregates) are more representative of global model grid-boxes ($\sim 1^o - 3^o$ in size) while allowing accurate temporal collocation with other datasets. At the same time, the use of super-observations significantly reduces data amount without much loss of information (at the scale of global model grid-boxes). A list of products used in this paper is given in Table **??**. A colour legend to the different products can be found in Fig. **??**. More explanation of the aggregation procedure can be found in Appendix **??**.

Super-observations of AOD and AAOD at the same location and time were derived from the same set of L2 data and ~~therefor~~ therefore measure the exact same scene ~~. The exception is the POLDER-GRASP-M dataset which provides aggregate AOD and SSA for slightly different samplings (there is an additional minimum AOD threshold for the calculation of the AAOD that will be aggregated and the resulting aggregated SSA). We assumed that this SSA nevertheless represents the same scene as the AOD aggregate and recalculated an AAOD from that AOD and SSA. Consequently, The POLDER-GRASP-M AAOD presented in this paper is different from the AAOD found in the official L3 product. The latter shows a high bias vs. AERONET due to same aforementioned minimum AOD threshold. Note that in-situ measurements (????) have suggested a change in SSA at lower AOD so our SSA assumption may introduce additional biases.~~ (note an exception for GRASP dataset described below).

The main data are AOD and AAOD at 550 nm, the wavelength at which models typically provide (A)AOD. If (A)AOD was not retrieved at this wavelength, it was logarithmically interpolated or extrapolated from surrounding wavelengths.

2.1.1 FL-MOC

~~MOC~~ FL-MOC (Fu Liou - MODIS, OMI, CALIOP) is a technique for combining CALIOP aerosol backscatter, MODIS spectral AOD, and OMI AAOD retrievals for estimating full spectral sets of aerosol radiative properties (SSA, asymmetry parameter and AOD). It is not a retrieval per se but a consistent reinterpretation of the combined data within their stated uncertainties. Details are given in **?**, Appendix A. In brief, ~~MOC uses the level 2~~ FL-MOC uses the L2 retrieved aerosol properties as input to a simple look-up table retrieval of aerosol types and concentrations, under the assumption that aerosol properties are consistent with the L2 aerosol observations within the stated uncertainties of each sensor's retrieval. This technique also assumes that the surface reflectance and clouds are properly treated in the underlying retrievals.

Over land, ~~MOC~~ FL-MOC uses OMAERUV AAOD, over ocean OMAERO AAOD. OMAERO is an advanced multi wavelength UV-VIS algorithm that uses 17 wavelengths in the 331-500 nm range in order to calculate the aerosol optical depth and to discriminate between various types of aerosols. It is an extension of the near UV TOMS method (see the OMAERUV product) to a wider wavelength range. The OMAERO algorithm is applied over all surface types, however its primary objective is to derive aerosol properties over the oceans due to the limited availablity of spectral surface reflectivity databases over land.

2.1.2 OMAERUV

The Ozone Monitoring Instrument (OMI) on the EOS-Aura satellite was deployed in July 2004. It is a high resolution spectrograph that measures the upwelling radiance at the top of the atmosphere in the ultraviolet and visible (270–500 nm) regions of the solar spectrum (**?**). It ~~has~~ had a 2600 km wide swath and provides daily global coverage at a spatial resolution varying from 13×24 km at nadir to 28×150 km at the extremes of the swath. OMI hyperspectral measurements are used as input to inversion algorithms to retrieve ozone vertical distribution and column amounts of O_3, NO_2, SO_2, HCHO, BrO, and OClO. OMI observations are also used to retrieve information on aerosols and clouds.

Aerosol properties in the near UV are derived from OMI observations at 354 and 388 nm (**?**). The OMI UV aerosol algorithm (OMAERUV) takes advantage of the large sensitivity to aerosol absorption in the near UV discovered in the mid-90's using heritage TOMS instruments (**?**), and the low reflectance of all ice/snow free terrestrial surfaces, which facilitates the aerosol characterization over all arid and semi-arid regions of the world. The OMAERUV two-channel algorithm simultaneously retrieves AOD and SSA at 388 nm. The main sources of uncertainty are assumed aerosol layer height, and cloud contamination, the latter associated with the sensor's coarse spatial resolution. The OMAERUV fifteen-year record of AOD has been validated with AERONET observations (**??**). The SSA record has also been evaluated by comparisons to AERONET and SKYNET (`https://www.skynet-isdc.org/index.php`) ground-based retrievals (**??**).

2.1.3 POLDER-SRON

The POLDER-3 instrument was a multi-angle, multi-wavelength polarimeter flying aboard the Polarization & Anisotropy of Reflectances for Atmospheric Sciences coupled with Observations from a Lidar (PARASOL) satellite. It was launched in 2004 and was a part of the satellite constellation A-Train until 2009. Initially designed to be operated for 2 years, POLDER-3

performed its measurements until late 2013, when it was decommissioned. PARASOL provides measurements of a ground scene under (up to) 16 different viewing geometries in 9 spectral bands (443, 490, 565, 670, 763, 765, 865, 910, 1020 nm). Linear polarization measurements (Stokes parameters Q and U) are performed in 3 spectral bands (490, 670, 865 nm). Its spatial resolution at the nadir was about 6 km, and its swath width was 2400 km.

An advanced retrieval algorithm making full use of the information content of the multi-angle photopolarimetric observations from POLDER-3/PARASOL has been developed at SRON-Netherlands Institute for Space Research. ~~This algorithm yields the different microphysical characteristics of a bi-modal aerosol size distribution~~The algorithm has large flexibility in defining the aerosol properties included in the retrieval state vector (**?**). The aerosol ~~parameters of each mode included in the state vector are the effective radius~~ size distribution is described by the sum of an arbitrary number log-normal functions, called modes, where for each mode the effective radius (reff), effective variance ~~,~~ (veff), aerosol column number, ~~and~~ real and imaginary parts of the refractive index ~~. For the coarse mode, also the fraction of spheres is included in the state vector,~~ (in the form of coefficients of spectrally dependent functions), fraction of spherical particles assuming the mixture of spheres and spheroids proposed by ~~?.~~ **?**, and the Aerosol Layer Height can (in principle) be retrieved. In the setup used in the present study, the POLDER-SRON algorithm yields the different microphysical characteristics of a bi-modal aerosol size distribution (fine and coarse mode), with the fraction of spheres only be retrieved for the coarse mode (fine mode assumed to consist only of spheres) and the Aerosol Layer Height is fixed to 1km. For retrievals over ocean, the state vector also includes the wind speed, chlorophyll-a concentration, and white-cap fraction, while for retrievals over land, the state vector includes the parameters describing the surface BRDF ~~(?)~~(Bidirectional Reflectance Distribution Function) (**?**). The retrieval is based on an iterative fitting of a linearized radiative transfer model (**?**) to the PARASOL data, using a cost function containing a misfit term between the forward model and measurement and a regularization term using a priori estimates of values of some of the retrieved parameters. The algorithm, including an application to PARASOL measurements over ocean, is described in **?**. More recent refinements are described by **???????**. Retrieval results from the SRON algorithm have been used for aerosol type determination by **?**, in studies related to aerosol absorption and direct radiative effect by **??**, and aerosol-cloud interactions by **??**, and data assimilation by **?**. Currently, the algorithm has been applied to one year (2006) of global aerosol data.

2.1.4 POLDER-GRASP

For a description of the POLDER instrument, see the previous subsection.

GRASP (Generalized Retrieval of Aerosol and Surface Properties) is a unified retrieval algorithm for atmosphere properties from diverse remote sensing observations (**??**), based on earlier work by **???** for AERONET Inversions.

In the current paper, retrievals from the so-called "models" dataset ~~(here: GRASP-M) are presented~~are used. Aerosol is assumed to be an external mixture of five different aerosol components ~~and~~ which are retrieved together with spectral parameters of surface BRDF and BPDF (Bidirectional Polarisation Distribution Function). The aerosol is assumed to be a mixture of spherical and non-spherical particles. Each fraction is characterized by particle size distributions similarly to AERONET retrievals. The non-spherical component is modeled as a mixture of randomly oriented spheroids with fixed shape distribution

(**?**). The details of the "models" approach are discussed by **?** and **?**.The actual inversion uses multi-pixel retrieval (**?**) where horizontal pixel-to-pixel variations of aerosol and day-to-day variations of surface reflectance are enforced to be smooth.

The full archive of POLDER/PARASOL observations was retrieved using GRASP and can be found at https://www.grasp-open.com. In addition to the "models" dataset, two other datasets are available ("~~improved~~optimized" and "high-precision") that use slightly different assumptions in the retrieval. The detailed discussion and validation of all three 0.1 degree PARASOL/GRASP retrievals are provided by **?**. The "models" dataset used in this paper is considered the most applicable for a wide range of circumstances.

The dataset used in the current paper is aggregated to 1 degree spatial resolution (details are listed at https://www.grasp-open.com). The "models" dataset provides AOD and AAOD aggregated from slightly different L2 samplings: an additional minimum AOD threshold is used when aggregating AAOD. To select data of higher quality, AAOD retrievals were used only for cases with sufficient aerosol loading. The same AOD threshold is used for SSA as well. Specifically, a minimum AOD (at 440 nm) threshold of 0.3 over land and 0.02 over ocean were applied (the threshold over ocean is probably too low to assure high quality AAOD but higher thresholds result in significant data loss).

In the current study we prefer to use aggregated AOD and AAOD data that describe the exact same scene, and this is the case for the FL-MOC, OMEARUV and POLDER-SRON datasets mentioned earlier. For the GRASP product, we decided to assume that the aggregated SSA represents the same scene as the AOD aggregate and recalculated an AAOD from that AOD and SSA. Consequently, the AAOD product (indicated as GRASP-M) presented in this paper is different from the AAOD found in the official L3 "models" product. In-situ measurements (**????**) have suggested a change in SSA at lower AOD so our SSA assumption may introduce additional biases. However, GRASP-M AAOD evaluated better against AERONET than "models" AAOD which showed a high bias vs. AERONET due to the aforementioned minimum AOD threshold.

For this study the L3 GRASP data were additionally filtered based on the FittingResidual field which was required to be smaller than 0.05 (over Land) or 0.1 (over Ocean). This subset evaluates substantially better for AOD retrievals and somewhat better for AAOD retrievals than the full dataset.

2.1.5 AERONET

AERONET (**?**) DirectSun V3 L2.0 (**??**) and Inversion V3 L1.5 & 2.0 data were downloaded from `https://aeronet.gsfc.nasa.go` logarithmically interpolated to values at 550 nm and aggregated by averaging over 30 minutes. The DirectSun dataset contains only AOD (at multiple wavelengths). These observations are based on direct transmission measurements of solar light and have a low uncertainty of ± 0.01 (**??**), at 400nm and larger.

The Inversion dataset contains ~~both AOD and AAOD~~ AAOD and SSA (at multiple wavelengths) ~~and these observations are~~ based on measurements of scattered solar light from multiple directions. This inversion uses radiative transfer calculations (**?**) and yields larger errors than the DirectSun measurements. In particular, **?** showed that SSA errors decrease with increasing AOD and estimated 440nm SSA errors of ± 0.03 for water-soluble aerosol at 440nm AOD ≥ 0.2 although for dust and biomass burning aerosol higher AOD ≥ 0.5 were needed. These error estimates were based on numerical calculations. A recent in-depth estimate of the uncertainty in Inversion V3 data (**?**) suggested those thresholds to be 440nm AOD > 0.3 and

≥ 0.45, respectively. For an examination of the impact of geometrical configuration on SSA observations, see **?**. **?** showed that AERONET SSA retrievals were lower by 0.011 than flight campaign data (on average). **?** also ~~used~~ compared flight campaign measurements to ~~evaluate~~ AERONET SSA and found that ~~it was~~ the data were usually within the expected errors, although at low AOD ≤ 0.2 ~~significant underestimation by AERONETwas observed. only had observations over two sites~~significantly lower SSA values were observed by AERONET. A confounding issue for the evaluation of SSA (or, for that matter, AAOD) datasets is that there is no established gold standard.

The Inversion dataset also contains AOD (from Direct Sun retrievals) which is actually used in the inversion. Here we only use those AOD values in the Inversion dataset that have corresponding AAOD and SSA values, so that aggregate values always describe the same scene.

Inversion L2.0 is a subset of L1.5 (which contains almost $30\times$ more observations), based on further cloud screening and the requirement that AOD at 440nm ≥ 0.4. This last criterion results in a minimum AOD at 550nm of 0.25 in the Inversion L2.0 product.

Since an individual AERONET site ~~cannot be expected to be representative for~~ is not necessarily representative of a $1^o \times 1^o$ grid-box, satellite evaluation may be negatively affected. To select only sites with high representativity we use a list published in **?** as described in ~~**??**~~, where we also ~~describe some tests for its suitability (based on~~ tested this representativity (using 14 satellite AOD products). The Kinne list was developed with the AERONET ~~Direct Sun~~ DirectSun product (i.e. AOD) in mind but a high-resolution modelling study by ~~**?** suggests that~~ **?** suggests that spatial representativity for AOD and AAOD observations can differ substantially for individual sites. We chose to use the Kinne list because it also includes information on maintenance quality, likely more important for Inversion than ~~Direct Sun~~ DirectSun retrievals.

2.1.6 How independent are these satellite products?

An interesting question is how independent these satellite products are.

The GRASP and SRON algorithms are independent retrieval codes with many specific differences in the implementation. First, in the present study POLDER-SRON retrieves parameters of bi-modal lognormal size distribution and complex refractive index for each size mode, while POLDER-GRASP-M retrieves the concentrations of five aerosol components with assumed properties of each component (**??**). Second, GRASP and SRON use the same mathematical function for the BRDF over land (**?**) but estimate the parameters to this function independently. In both algorithms, aerosol and surface properties are estimated simultaneously. Third, there are significant differences in use of a priori constraints. POLDER-SRON follows Phillips-Tikhonov regularization (**??**) including a priori estimates for most of the retrieved state vector parameters (a globally constant value is used) and a flexible strength of the regularization term. The GRASP algorithm is based on the least-square multi-term approach (see **?**) and uses several a priori constraints simultaneously. Specifically, GRASP "models" uses smoothness constraints on the spectral dependence of surface BRDF parameters. Fourth, the SRON algorithm retrieves from measurements of individual pixels while the GRASP algorithm retrieves from measurements of multiple pixels simultaneously, applying spatio-temporal constraints in the process. For example, over land constraints were used to limit temporal variability of retrieved BRDF parameters as well as spatial variability of aerosol retrieved parameters (see **??**).

The FL-MOC product uses OMAERUV AAOD as input over land but FL-MOC only uses OMAERUV AAOD as an a-priori and assigns this a sizeable uncertainty. CALIOP backscatter is expected to provide a constraint on SSA, and consequently AAOD. As a matter of fact, our analysis shows that FL-MOC and OMAERUV exhibit rather low correlations for AAOD (and SSA). This suggests that the OMAERUV a-priori does not lead to a strong dependency of FL-MOC on OMAERUV. On the other hand, it also suggests that at least one of these products contains sizeable errors.

3 Collocation & analysis methodology

To evaluate and intercompare the remote sensing datasets, they will need to be collocated in time and space to reduce representation errors (**???**). In practice this collocation is another aggregation (performed for each dataset individually) to a spatio-temporal grid with slightly coarser temporal resolution (1 or 3 hours, the spatial grid-box size remains $1^o \times 1^o$). This is followed by a masking operation that retains only aggregated data if it exists in the same grid-boxes for all involved datasets. More details can be found in Appendix **??**.

We need to allow some flexibility in the time separation between data (here 3 hours) to ensure sufficient numbers of collocated data pairs for further analysis. **?** showed that shorter time separations greatly limited the number of pairs but did not substantially alter the correlation of satellite AOD with AERONET. On the other hand, longer time separations appear to negatively affect the correlation of satellite AAOD with AERONET, see Fig. **??**. The analysis shows that satellite AOD correlation with AERONET Inversion data slowly decreases as the collocation ~~criterium~~ criterion is relaxed from 3 to 24 hours. However, satellite AAOD shows a sharp drop in correlation with AERONET at 6 hours (OMAERUV is the exception, the correlation is already low and barely changes). We surmise this is due to plumes of absorbing aerosol drifting over the sites, requiring tight temporal constraints on collocation. Consequences of this finding will be further discussed in Sect. **??**.

As the FL-MOC dataset, based on CALIOP measurements, is smaller than the other satellite datasets, we were compelled to collocate FL-MOC with AERONET within 2^o instead of 1^o. Even so, the data count for the FL-MOC evaluation is low~~and this results in significant statistical noise~~.

After spatio-temporally collocating two or more datasets, the data may be further averaged in space and/or time for analysis purposes. Spatio-temporally averaged SSA is *always* derived from averaged AOD ~~&~~ and AAOD:

$$\overline{\mathrm{SSA}} = 1 - \overline{\mathrm{AAOD}}/\overline{\mathrm{AOD}}. \quad (1)$$

During the evaluation of products with AERONET, a distinction will be made between either land or ocean grid-boxes in the common grid. A high resolution land mask was used to determine which $1^o \times 1^o$ grid-box contained at most 30% land (designated an ocean box) or water (designated a land box). Most ocean boxes with AERONET observations will be in coastal regions, with some over isolated islands.

3.1 Taylor diagrams

A suitable graphic for displaying multiple datasets' correspondence with a reference dataset ('truth'), is provided by the Taylor diagram (**?**). In this polar plot, each data point (r, ϕ) shows basic statistical metrics for an entire dataset. The distance from the origin (r) represents the internal variability (standard deviation) in the dataset. The angle ϕ through which the data point is rotated away from the horizontal axis represents the correlation with the reference dataset, which is conceptually located on the horizontal axis at radius 1 (i.e. every distance is normalised to the internal variability of the reference dataset). It can be shown (**?**) that the distance between the point (r, ϕ) and this reference data point at $(1, 0)$ is a measure of the Root Mean Square Error (RMSE, unbiased). A line extending from the point (r, ϕ) is used to show the bias versus the reference dataset (positive for pointing clock-wise).The distance from the end of this line to the reference data point is a measure of the Root Mean Square Difference (RMSD, no correction for bias).

3.2 Uncertainty analysis using bootstrapping

Our estimates of error metrics are inherently uncertain due to finite sampling. If the sampled error distribution is sufficiently similar to the underlying true error distribution, bootstrapping (**?**) can be used to assess uncertainties in e.g. biases or correlations due to finite sample size. Bootstrapping uses the sampled distribution to generate a large number of synthetic samples by random draws *with replacement*. For each of these synthetic samples, a bias etc. can be calculated and the distribution of these biases provides measures of the uncertainty, e.g. a standard deviation, in the bias due to statistical noise. Bootstrapping has been shown to be reliable even for relatively small sample sizes (that is the size of the original sample, not the number of bootstraps), see **?**. In this study, the uncertainty bars in some figures were generated by bootstrap analysis.

If the sampled error distribution is different from the true error distribution, bootstrapping will likely underestimate uncertainties. Sampled error distributions may be different from the true error distribution because the act of collocating satellite and AERONET data favours certain conditions. E.g. the effective combination of two cloud screening algorithms (one for the satellite product, the other for AERONET) may favour clear sky conditions and ~~limit~~ reduce our sampling of errors ~~in case of~~ due to cloud contamination. This uncertainty due to sampling is unfortunately hard to assess~~(~~, see e.g. ~~?)~~**?**.

As an example of uncertainty due to sampling, we present Fig. **??** in which an evaluation of the current satellite AOD data with Inversion L2.0 data (only those AOD that have corresponding AAOD inversions, which constrains AOD at 440nm > 0.4) shows substantial shifts compared to ~~Direct Sun~~ DirectSun L2.0. As the uncertainty ranges indicate, the changes in biases are *not* due to statistical noise. Neither is this due to differences in collocated DirectSun and Inversion L2.0 AOD values, that agree very well. Rather, the issue is that AERONET Inversion data are an unrepresentative subsample of the DirectSun data (Inversion data are skewed to high AOD). It is unclear what this means for the AAOD and SSA evaluation but readers should be aware of this unaccounted-for sampling issue that may introduce biases.

3.3 Error metrics for evaluation

We will use the usual global error statistics (bias, standard deviation, Pearson correlation, regression slopes), treating all data as independent. Regression slopes were calculated with a robust Ordinary Least Squares regressor (OLS bisector from the IDL `sixlin` function, **?**). This regressor is recommended when there is no proper understanding of the errors in the independent variable, see also **?**.

4 A first look at the satellite products

Multi-year averages of satellite AAOD and their differences are shown in Fig. **??**. The AAOD maps can only be compared with ~~some~~ caution, as they are derived from products with different temporal sampling. The differences, on the other hand, are based on collocated data and confirm major features. The products all agree on a major AAOD hotspot from (likely) African Savannah biomass burning. Three products agree on AAOD hotspots in China and India, that are known polluted regions ~~like India and China also being AAOD hotspots~~ (OMAERUV, which is relatively featureless, is the exception. We surmise this is due to the large pixelsize of the OMI instrument, see Table **??**, which will not resolve small scale structure in AAOD. The existence of such small scale structure was inferred from Fig. **??**). POLDER-GRASP-M and OMAERUV show a clear AAOD hotspot due to Amazonian biomass burning. POLDER-GRASP-M estimates relatively high values over land, and the ocean at high northern latitudes. OMAERUV shows relatively low AAOD over land but high over the entire ocean. FL-MOC clearly estimates higher AAOD over the Sahara than either POLDER-GRASP-M or OMAERUV. POLDER-SRON estimates relatively high AAOD over the Rocky ~~mountains~~Mountains, the Andes and Australia. Unfortunately, even in multi-year averages significant differences in regional AAOD between the products are observed, in excess of 50%. Figure S1 shows the corresponding SSA maps. As expected, POLDER-GRASP-M has relatively low SSA and OMAERUV relatively high SSA over land. FL-MOC has the highest SSA over ocean of all products. As the satellite AOD are fairly similar, lower values of AAOD translate into higher values of SSA.

One caveat is that AAOD and SSA retrievals are likely to be better (more accurate and precise) at high AOD. In the above analysis, no account was taken of AOD levels and the products were discussed as they are. The impact of AOD will ~~later be discussed .~~ be discussed later, when discussing the evaluation with AERONET in Sect. **??** and the satellite intercomparison in Sect. **??**.

5 Evaluation of satellite products with AERONET

Taylor plots of the performance of the satellite products are shown in Fig. **??**. Satellite AOD is evaluated against AERONET ~~Direct Sun~~ DirectSun L2.0. Satellite AAOD & SSA, are evaluated against AERONET Inversion L2.0 (which constrains AOD at 440nm > 0.4 and provides much less data than ~~Direct Sun~~DirectSun). All products show high correlation with AERONET AOD ($r \geq 0.76$), although the correlations found are lower than ~~found in ?~~ those found in **?** for several MODIS Aqua products (0.87-0.88). Correlations for AAOD and SSA are lower than for AOD suggesting that it is more challenging to retrieve ~~absorptive qualities.~~ absorbing qualities.

Interestingly, POLDER-SRON's SSA correlates significantly better with AERONET than POLDER-GRASP-M's ~~(their AOD and AAOD perform similarly)), suggesting balancing errors in AODand AAOD in the first product~~but this is a sampling effect: once both products are collocated together, POLDER-GRASP-M's SSA correlation with AERONET increases from 0.41 to 0.69. The explanation for this is not entirely clear, although it turns out that POLDER-GRASP-M evaluates poorer with AERONET for 2010 than for 2006 and 2008 (POLDER-SRON is currently limited to 2006, see Table **??**). Although the poorer evaluation for 2010 can be seen in AOD, AAOD and SSA, it is only statistically significant for SSA.

The impact of statistical noise on the AAOD evaluation is explored in Fig. **??**. Using a bootstrapping technique, the spread in correlation and standard deviation were explored. For most datasets, the results seem fairly robust, except for FL-MOC which ~~uses~~ yielded only 24 data points. A proper intercomparison of products ~~, however,~~ requires collocation (of *all* the satellite data), which reduces available cases even further. Figure S2 shows that results are not very different from Fig. **??**, but the statistical noise increases substantially. The sampling noise on such a small subset should be even larger, see also Fig. **??** and **??**. For a sense of perspective, ~~53~~ 48 data points represents less than ~~0.0006%~~ 0.0008% of the total POLDER-GRASP-M data amount used in this paper.

5.1 Evaluation and intercomparison of AOD

In Fig. **??**, we provide more detail on the satellite AOD products and their evaluation against AERONET ~~Direct Sun~~ DirectSun L2.0 AOD. In the central column, we show the products themselves, averaged over ~~several years~~1, 2 or 3 year(s), depending on availability (see Table **??**). Note that the products exist for different years and even ~~within~~ for the same years ~~have different~~ products will have different temporal samplings so comparisons should be made with caution (**??**). In the left and right column, we show satellite data collocated with AERONET. On the left-hand side is a scatterplot of the ~~raw~~ data (with associated statistics provided) and on the right-hand side is a map of multi-year difference with AERONET (provided at least 32 data points were available per site).

The scatter plots show good correlation with AERONET. The POLDER products show higher correlations and slopes closer to one (1~~(one~~) than FL-MOC and OMAERUV. Nevertheless, differences in evaluation seem rather small, which unfortunately cannot be said for the global distributions of AOD. POLDER-GRASP-M has rather high AOD over land and OMAERUV has rather high AOD over ocean (note that the satellite data themselves are not collocated). The multi-year differences with AERONET suggest that ~~POLDER-GRASP-M mostly overestimates AOD (several sites show small underestimates) while~~ OMAERUV overestimates everywhere except in some regions with strongly absorbing aerosol. ~~In ? we evaluated 14 satellite AOD products (see also list in Fig. **??**), and most showed both positive and negative biases varying with region. Compared to those products, POLDER-GRASP-M and OMAERUV show a more globally consistent positive bias. Note that POLDER-SRON provides fewer observations and hence collocated data than the other two products. There is however a suggestion it is less biased.~~ An intercomparison of satellite AOD with Aqua-DT is presented in Fig. S3 and ~~also suggests typically high estimates over land~~ suggests typically higher estimates over (Southern Hemisphere) Land for the POLDER products and ~~also over ocean~~ over Ocean for OMAERUV. Note that Aqua-DT is not without ~~biases either, see ?, but this analysis confirms the evaluation with AERONET and adds spatial context to it~~significant regional biases, see **?**.

Figure **??** shows results when bias (sign-less) and correlation per site (that yielded at least 32 collocations) are averaged over all sites, for ~~all satellite products~~each satellite product. The same ~~92~~ 52 sites are used for all datasets although each product is individually collocated with AERONET. For FL-MOC, no site provided at least 32 observations and it is not included in the analysis. For POLDER-SRON, only 18 sites provided at least 32 collocated observations ~~. The POLDER-SRON result should therefore not be~~ and it was similarly excluded. As was also shown in **?**, OMAERUV shows rather large biases compared to the other ~~datasets. In any case,~~ AOD products. POLDER-GRASP-M~~and OMAERUV appear to have larger biases and lower correlations per site than most of the datasets studied in **?**.~~, on the other hand, shows the smallest bias. The filtering of GRASP retrievals described in Sect. **??** plays a significant role in this result (without filtering, POLDER-GRASP-M shows a bias twice as large).

5.2 Evaluation of AAOD and SSA

Figure **??** provides more detail on the evaluation of satellite (A)AOD & SSA products against AERONET Inversion L2.0 (which constrain AOD at 440nm > 0.4). In the first three columns, we show scatter plots for respectively AOD, AAOD and SSA. In the last column we show SSA differences with AERONET as a function of AERONET AOD (Inversion L1.5). All products underestimate AERONET AOD and AAOD, although only by a small amount in the case of POLDER-GRASP-M. More importantly, AAOD correlations can be low as 0.34 (OMAERUV) and regression slope can deviate substantially from 1 (0.6 for OMAERUV). In contrast, some ~~product~~ products underestimate SSA while others over-estimate it. Due to data sparsity (e.g. for POLDER-GRASP-M, the count dropped from ~~17692 to 529~~10454 to 423), it is not possible to do an analysis per AERONET site (as was done for AOD) and see how the global bias relates to regional biases. Bootstrap analysis suggest that ~~the global statistics~~ results are fairly robust against statistical noise (except FL-MOC, see also Fig. **??**).

The right-most column in Fig. **??** shows SSA difference as a function of (AERONET) AOD. To ensure the largest possible range in AOD values Inversion L1.5 instead of L2.0 is used. Especially at lower AOD, this dataset will have larger errors in AAOD and SSA than L2.0. Interestingly, as AOD increases, all satellite products seem to agree better with AERONET (for FL-MOC, the bin with largest AOD values is affected by a very low data count). This is of course as one would expect. For smaller AOD, there is increasingly more spread although the difference distribution remains fairly unbiased. The exception is POLDER-GRASP-M which shows increasingly lower SSA than AERONET at low AOD. We suggest that it is rather unlikely that three different satellite products have a similar SSA bias at low AOD as AERONET (and hence show no bias in the difference with AERONET) and that this low bias in POLDER-GRASP-M analysis is real. ~~A~~ However, a better understanding of the nature of errors (bias vs. random) in AERONET SSA at low AOD is desirable.

Summarizing, there is skill in satellite AAOD and SSA but compared to AOD the correlations with AERONET are substantially lower. POLDER-SRON is the exception, with similar and fairly high correlations (~ 0.75) for all three parameters. However, it seems to underestimate AAOD by $\sim 25\%$ at high ~~AOD~~ AAOD (slope of 0.76 in the AAOD scatter plot). OMAERUV appears to show the largest deviations from AERONET (low correlations and slopes) but its overall error statistics (mean and standard deviation) is not too different from the other products. Results for FL-MOC may be a statistical fluke due to the low data count. POLDER-GRASP-M ~~overal performs rather nicely~~ shows quite high correlations for AOD (0.86) and AAOD (0.6)

with reasonable slopes but has a very low correlation with AERONET for SSA (~~0.37)~~0.41), but this seems to depend strongly on sampling as discussed at the start of this section. In addition, it appears to systematically underestimate SSA at low AOD. Yet another aspect to this dataset (not visible in any of the analysis shown) is that it appears to have ~~hard~~ a hard SSA cut-off as SSA values larger than 0.99 do not occur.

A profound problem is the paucity of data. Even for POLDER-GRASP-M, we can only evaluate its performance (against AERONET) for less than 0.006% the total number of available observations. Is this sufficient to make meaningful statements about the performance of a product *at large*? In ~~Schutgens et al. 2019a~~**?**, we showed that the process of collocation can skew error statistics (by changing the sampling) to the point that it becomes hard to meaningfully distinguish performance of several products. That study was done for AOD which allows much higher numbers of collocated data with AERONET than AAOD.

To elucidate this, we compare the difference in SSA between the two POLDER products (collocated within 3 hours, considering AOD ≥ 0.25 only) for three different samplings. First, we look at global POLDER SSA statistics~~, second,~~. Secondly, we look at POLDER SSA statistics over AERONET sites only. Thirdly, we look at POLDER SSA statistics that are collocated with AERONET observations. Figure **??** shows the associated difference distributions. Using various non-parametric statistical tests (Mann-Whitney U, Student's t, Kolmogorov-Smirnov) we can show that the distribution means for the first and third sampling are ~~fundamentally~~ significantly different. Not only that, but the mean difference in SSA for the first sampling is 2.6 as large (~~-0.044~~ -0.043 vs. -0.017) as for the third sampling~~(and is statistically significant)~~. As POLDER-SRON is biased high and POLDER-GRASP-M is biased low vs AERONET, the ~~corrollary~~ corollary to this is of course that at least one of the products has a larger bias vs the truth globally than can be seen in the AERONET observations. Conversely this suggests that the AERONET Inversion dataset does not allow a truly global evaluation of satellite datasets: it provides a sub-sample with skewed statistics of SSA errors. Incidentally, it is the temporal sub-sampling enforced by collocation with AERONET observations that causes the largest shift in the difference distribution (POLDER measurements over AERONET sites show a similar SSA distribution as the global dataset). It is possible that the SSA difference is partly driven by cloud contamination which we know is present in these satellite datasets ~~(?)~~ (**?**) and may be ameliorated when a third cloud masking (from AERONET) is applied (through the collocation of data).

6 Intercomparison of satellite AAOD and SSA

To get a better appreciation of the satellite products, we now present a global intercomparison. To start with, Fig. **??** shows SSA differences between two products as a function of their mean AOD. As in Fig. **??**, these differences become smaller (i.e. show a smaller spread) at higher AOD, as expected (intercomparisons with FL-MOC are the exception). However, ~~while the spread in the difference distribution may become narrower, substantial biases remain: -0.012 or 0.022 for either OMAERUV intercomparisons and 0.037 for an intercomparison of the POLDER products. Even ignoring these biases and concentrating on the spreads: satellite SSA~~ satellite SSA values still exhibit random differences of 0.03 or larger for AOD $\gtrsim 1$, as also confirmed by the AERONET evaluation. In addition, substantial biases remain.

The previous analysis was global but substantial differences can be seen between land and ocean scenes. For instance, the SSA bias between the POLDER products over land, does not decrease at lower AOD but remains fairly constant. A more detailed analysis can be found in Fig. **??** which shows biases, correlations and regression slopes for different products. Unsurprisingly, correlations and slopes tend to improve with minimum AOD, while biases may remain fairly constant (POLDER products), decrease (OMAERUV vs POLDER-GRASP-M) or even increase (FL-MOC). As a consequence it should be challenging to determine an AOD threshold above which products can be expected to perform within certain parameters. A similar analysis for AAOD can be found in Fig. S4. ~~Note that the improvement in correlation with increasing AOD threshold is now only seen for the POLDER intercomparison, again suggesting SSA may be positively affected by balancing errors.~~

A final analysis concerns multi-year averages of these products. Model evaluation will be done on such averages and it may be useful to better understand the agreement (or lack thereof) between products in that case, even though the aforementioned biases are unlikely to be much reduced. Figure **??** shows an intercomparison of three products (FL-MOC is excluded due to its low data count). The analysis shows statistics of the intercomparison of multi-year averages of SSA, as a function of two thresholds: a minimum AOD and a minimum number of ~~observations~~ super-observations during three years (per $1^o \times 1^o$ grid-box). The underlying super-observations were always collocated (to within 3 hours) before temporal averaging took place. We see that~~in general~~ , in general, correlations increase and standard deviation in the difference decrease when either threshold increases. The improvement with increasing AOD has already been discussed and is due to better signal-to-noise conditions for the retrieval schemes. The improvement with increasing number of observations (used in the temporal averaging) can be interpreted as a significant random error in either product being lessened through averaging. In general, the AOD threshold has a more profound impact but the number of observations threshold allows more flexibility (by choosing a longer time-series to work with, smaller SSA differences (up to a point!) may be achieved).

However, biases between products can be quite robust as is particularly clear for the POLDER products. The decreasing bias for OMAERUV vs. POLDER-SRON (and, incidentally, the sudden jump in correlation for AOD > 0.4) is not really a sign of a better agreement between products at high AOD. Under these conditions, most observations come from the African dust and biomass burning regions. POLDER-SRON retrieves very reflective dust and very ~~absorptive~~ absorbing biomass burning aerosol while OMAERUV retrieves fairly reflective dust and fairly ~~absorptive~~ absorbing biomass burning aerosol. Consequently, global SSA bias decreases due to a balancing of very different biases over these regions while similar spatial patterns yield high correlations. Maps of the SSA difference between the POLDER products as a function of minimum AOD can be seen in Fig. S5. A higher minimum AOD mostly constrains data to a smaller portion of the globe but does not affect local biases greatly.

7 Conclusions

In this study, we evaluate several remote sensing datasets of AAOD and SSA, from a variety of sensors (CALIOP on CALYPSO, OMI on Aura, POLDER on PARASOL), ~~and use them to evaluate AEROCOM models~~in preparation of an AEROCOM model evaluation. This is the first global study to intercompare satellite remotely sensed products of AAOD (and SSA).

The evaluation of the products (daily aggregates over $1^o \times 1^o$) is done through comparison with AERONET ~~Direct Sun~~ DirectSun (AOD) and Inversion (AAOD and SSA) observations. To minimize sampling issues, satellite products and AERONET data are collocated in time and space, within 3 hours and 1 degree. One interesting finding is that AAOD evaluation requires a tighter temporal collocation ~~criterium~~ criterion than AOD, with steep declines in correlation found for temporal collocation after 3 hours or more. We interpret this to be due to ~~absorptive~~ absorbing aerosol primarily being found in plumes. While we do not explore this further, this high temporal variability in observed AAOD may affect model evaluation as well. It could ~~suggests~~ suggest that models need emissions with diurnal profiles, and output at higher frequencies than daily to obtain the best possible agreement with observations.

All satellite AOD products show significant correlation with AERONET ($0.76 \leq r \leq 0.86$)~~but their biases tend to be fairly large and more systematically positive compared to traditional products~~. Global biases are not very different from those found in an earlier study of traditional products (**?**). However, when considering typical multi-year biases per AERONET site, there is a suggestion that POLDER-GRASP-M has smaller biases than these traditional products (there is a hint this may also be true for POLDER-SRON but paucity of data makes this analysis less certain). In contrast, OMAERUV shows the largest (and mostly positive) biases in AOD. Compared to Aqua-DT (~~e. g. MODIS~~ Dark Target)~~that are only used for AOD retrievals (**?**). The exception is FL-MOC over ocean, which actually relies on Dark Target retrievals~~, the four products studied in this paper tend to estimate higher AOD over most of the land.

Results for AAOD are more diverse, with generally lower correlations ($0.34 \leq r \leq 0.78$) than for AOD. For most products~~(POLDER-SRON is the exception)~~, SSA correlates significantly worse with AERONET than AAOD~~(exception is POLDER-SRON)~~. All products show an improvement in SSA with regards to AERONET at higher AOD. POLDER-GRASP-M is noted for a low bias in SSA at low AOD.

The two POLDER products perform better against AERONET than the other two products, with typically (but not always) higher correlations, smaller biases and regression slopes closer to one (1) for all three parameters AOD, AAOD and SSA. However, dearth of measurements makes it very difficult to 1) meaningfully compare evaluation metrics amongst the products and 2) draw global conclusions. Theoretical evidence (**???**) suggests that retrieval schemes for absorptive properties will benefit from using polarisation measurements at multiple view angles which would support the idea that the POLDER products perform better. In addition, the OMAERUV product is based on measurements from a sensor with substantially larger pixels than POLDER and will struggle to resolve the fine-scale structure of aerosol plumes.

An intercomparison of multi-year satellite AAOD and SSA suggests significant biases across the globe. Differences of ~~75~~50% in multi-year averages of AAOD are not unusual. OMAERUV shows lower AAOD over land than the other products, but slightly higher AAOD over ocean. FL-MOC shows significantly higher AAOD over the Sahara and POLDER-GRASP-M is noted for a high AAOD at high Northern latitudes, both over land and ocean. POLDER-SRON has much higher AAOD than the other products over high-altitude regions. Many of these regions are unfortunately poorly instrumented with AERONET sites. Satellite SSA does agree better at high AOD, as was also observed for AERONET, although dearth of data means this can not be firmly concluded for FL-MOC. However, correlations for super-observations are often lower than 0.6, even at high AOD (0.75). Over ocean, SSA products tend to correlate better than over land. The two POLDER products correlate better than

any other satellite pair ($r =\sim 0.8$ over ocean for AOD > 0.75). In addition to high AOD, we show that temporal averaging also improves agreement between satellite products, although it is ~~difficult~~ not possible to give recommendations that work well with all products and for all regions. Even so, biases between products exist at high AOD after substantial temporal averaging.

Most surprisingly, POLDER-GRASP-M and POLDER-SRON show a fairly systematic difference in SSA (-0.04), independent of AOD (there are regional variations). ~~A major exception would be cases over the deep ocean at~~ For low AOD (< 0.1) ~~. Especially at high AOD (> 0.4), over land, this bias is pronounced.~~cases over ocean, this systematic difference becomes small in the global average because of two opposite biases organised roughly (!) by hemisphere (see also Fig. S1). Identifying the cause of this bias may lead to substantial improvements of both products (or at least one of them). Based on a comparison with AERONET data, we suggest that cloud contamination is a possible candidate.

Throughout the paper, we have given examples of how limited sampling of observations (especially AERONET) constrains our ability to understand the true error statistics of satellite AAOD and SSA. The most prominent example is a much reduced systematic difference (-0.017) between POLDER-GRASP-M and POLDER-SRON SSA as seen in an evaluation with AERONET Inversion L2.0 observations, ~~than is present in~~ as compared to the global satellite dataset (-0.04). This suggest that biases inferred from an AERONET evaluation will be smaller than those actually present in the satellite products. ~~It will not be easy to increase Inversion L2.0 observationsdue to the technical limitations in the sensors (?). However, an alternative~~ To increase available SSA observations, one could use Inversion L1.5 ~~product with similar cloud screening as~~ data (which includes SSA at low AOD) and sample it to L2.0 ~~might go a long way, especially~~ AOD measurements (which, unlike SSA, exist at low AOD), thereby benefitting from the better L2.0 cloud screening. Especially if follow-up studies can show that inversion errors at individual sites behave as random errors (amenable to temporal averaging) and not systematic biases such an intermediate product might be very useful.

This paper is one part of a two paper study into the use of satellite AAOD and SSA for aerosol model evaluation. In its companion paper, we use the datasets introduced in the current paper to evaluate AEROCOM (AEROsol Comparisons between Observations and Models) models. It turns out that ~~, notwithstanding serious biases in the satellite data,~~ robust and consistent evaluation of the models is possible~~. In~~ , notwithstanding the biases in the satellite data we have detailed in the current paper. The main reason seems to be that model biases (and the diversity in those biases) are even larger than satellite biases. Hence these satellite AAOD and SSA products are very useful: in regions with AERONET sites, they provide spatial detail lacking in a surface network~~. In~~ ; in regions without AERONET sites, they are the only datasets of observed AAOD and SSA available.

Code and data availability. All remote sensing data is freely available. Analysis code was written in IDL and is available from the author upon request.

Author contributions. NS designed the study, with the help of GL, TP, SK, MS and PS, and carried it out. OD, OH, OT, HJ, PL, JR and YS provided the remote sensing data. NS prepared the manuscript, with the help of all co-authors.

Competing interests. No competing interests are present

Acknowledgements. We thank the PI(s) and Co-I(s) and their staff for establishing and maintaining the many AERONET sites used in this investigation. The figures in this paper were prepared using David W. Fanning's Coyote Library for IDL. The work by N. Schutgens is part of the Vici research programme with project number 016.160.324, which is (partly) financed by the Dutch Research Council (NWO). NS thanks Tom Eck, Greg Schuster and Kostas Tsigaridis for insightful discussions on the use of AERONET observations. We would also like to thank four anonymous reviewers for attentive reading of our manuscript and many useful comments.

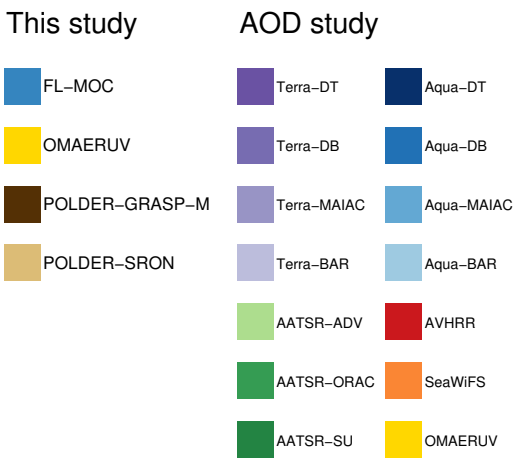

Figure 1. Colour legend used throughout this paper to designate the different satellite products, for both this study and the AOD study in ??.

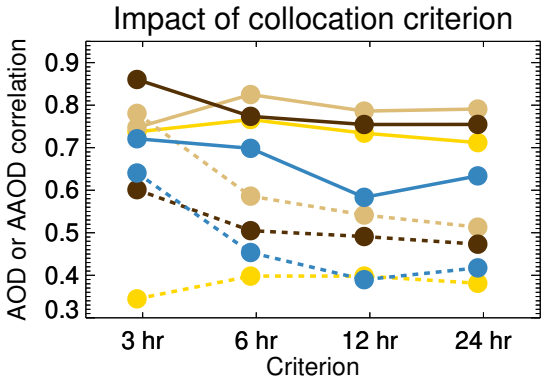

Figure 2. Correlation of satellite AOD (solid) and AAOD (dashed) with AERONET Inversion L2.0 data, as a function of temporal collocation ~~criterium~~criterion. Colours indicate satellite product, see also Fig **??**. Satellite products were individually collocated with AERONET.

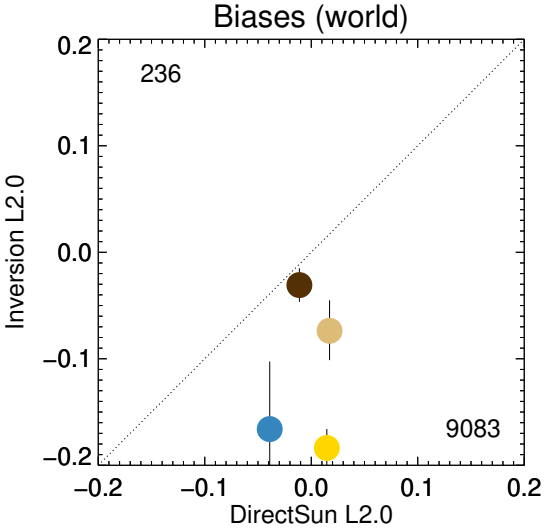

Figure 3. Global biases in four satellite AOD datasets depending on the chosen reference dataset (~~Direct Sun~~ DirectSun or Inversion). Colours indicate satellite product, see also Fig **??**. Numbers in upper left and lower right corner indicate amount of collocated data, averaged over all products. Error ranges indicate 5-95% uncertainty ranges based on a bootstrap analysis, see Sect. **??**. Satellite products were individually collocated with AERONET, within 3 hours.

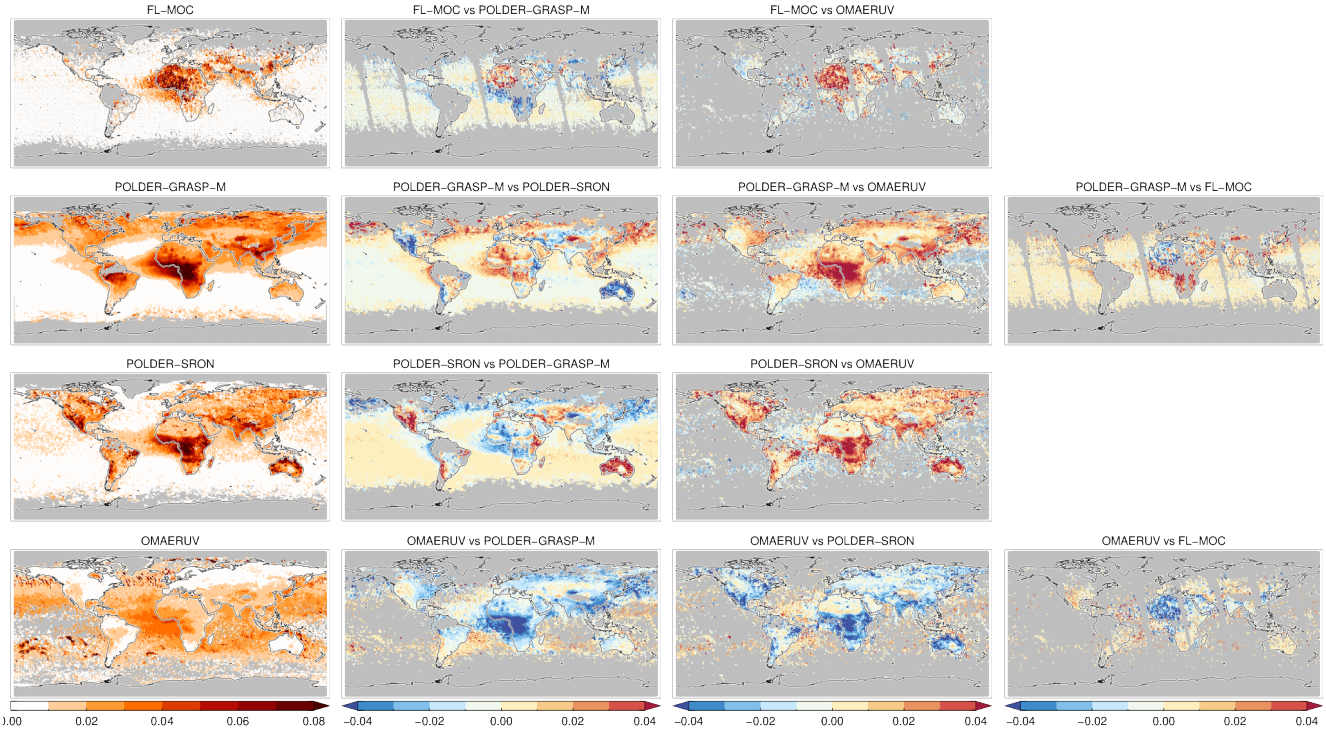

Figure 4. Global maps of AAOD for four products, and their differences. AAOD differences are based on collocated data (within 3 hours). Note that the products are available for different years, e.g. POLDER-SRON and FL-MOC do not overlap. No minimum AOD was required.

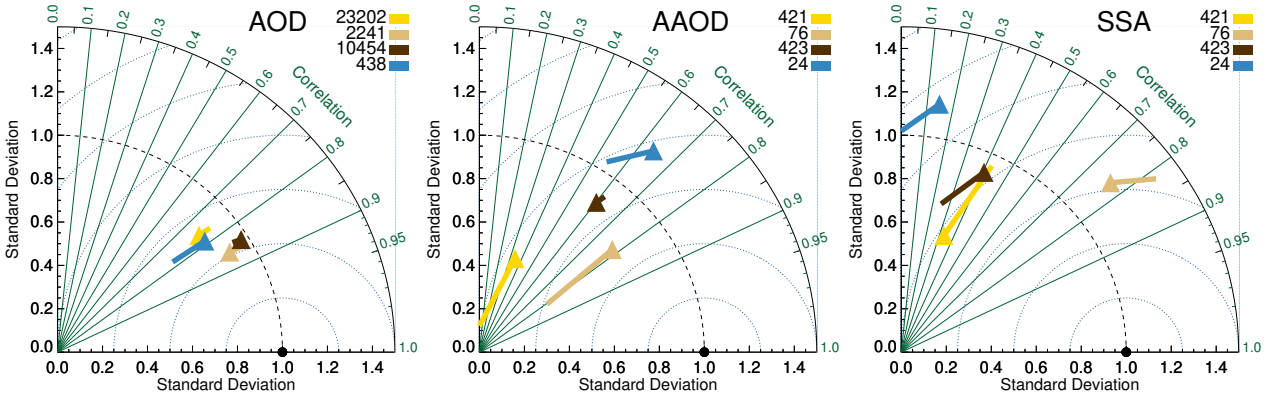

Figure 5. Taylor diagrams ?? for the satellite products. AOD is evaluated against AERONET ~~Direct Sun~~ DirectSun L2.0, AAOD and SSA are evaluated against AERONET Inversion L2.0. Colours indicate satellite product (see also Fig. 1), numbers next to coloured blocks indicate amount of collocated data. The lines extending from the data points indicate the bias. Products were individually collocated with AERONET, within 3 hours.

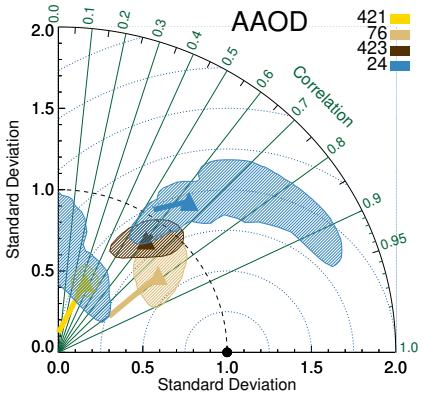

Figure 6. Impact of statistical noise on the correlation and internal variability of satellite AAOD products, using bootstrapping. Shaded regions indicate $5\% - 95\%$ uncertainty range of correlation and standard deviation (uncertainty in bias is not shown). Colours indicate satellite product, see also Fig **??**, numbers next to coloured blocks indicate amount of collocated data. Satellite products were individually collocated with AERONET Inversion L2.0 within 3 hours.

Appendix A: Generic aggregation and collocation

The aggregation of satellite L2 products into super-observations in this paper, and the subsequent collocation of different datasets for intercomparison and evaluation used the following scheme.

Assume a homogenous L2 dataset with times and geo-locations and observations of AOD and AAOD. Homogenous means that AOD and AAOD are available for the same times, geo-locations and wavelengths. Each observation has a known spatio-temporal foot-print, e.g. in the case of satellite L2 retrievals that would be the L2 retrieved pixel size and the short amount of time (less than a second) needed for the original measurement.

Satellite L2 data are aggregated into super-observations as follows. A regular spatio-temporal grid is defined as in Fig. **??**. The spatio-temporal size of the grid-boxes (here ~~$30^{\min} \times 1^o \times 1^o$~~$1^o \times 1^o \times 30^{\min}$) exceeds that of the footprint of the L2 data that will be aggregated. All observations are assigned to a spatio-temporal grid-box according to their times and geo-locations. Once all observations have been assigned, observations are averaged by grid-box. It is possible to require a minimum number of observations to calculate an average. Finally, all grid-boxes that contain observations are used to construct a list of super-observations as in Fig. **??**. Only times and geo-locations with aggregated observations are retained. As the original L2 dataset was homogeneous, so is the resulting L3 dataset.

Station data is similarly aggregated over ~~$30^{\min} \times 1^o \times 1^o$~~$1^o \times 1^o \times 30^{\min}$. Point observations will suffer from spatial representativeness issues (**???**), but the representativity of AERONET sites for $1^o \times 1^o$ grid-boxes is fairly well understood ~~(?)~~(?), see also Section **??**. These aggregated L3 AERONET ~~and MAN~~ data will also be called super-observations.

Different datasets of super-observations can be collocated in a very similar way. Again a regular spatio-temporal grid is defined as in Fig. **??** but now with grid-boxes of larger temporal extent (typically ~~$3^{\text{hr}} \times 1^o \times 1^o$~~$1^o \times 1^o \times 3^{\text{hr}}$). Because this temporal extent is short compared to satellite revisit times, either a single satellite super-observation or none is assigned to each

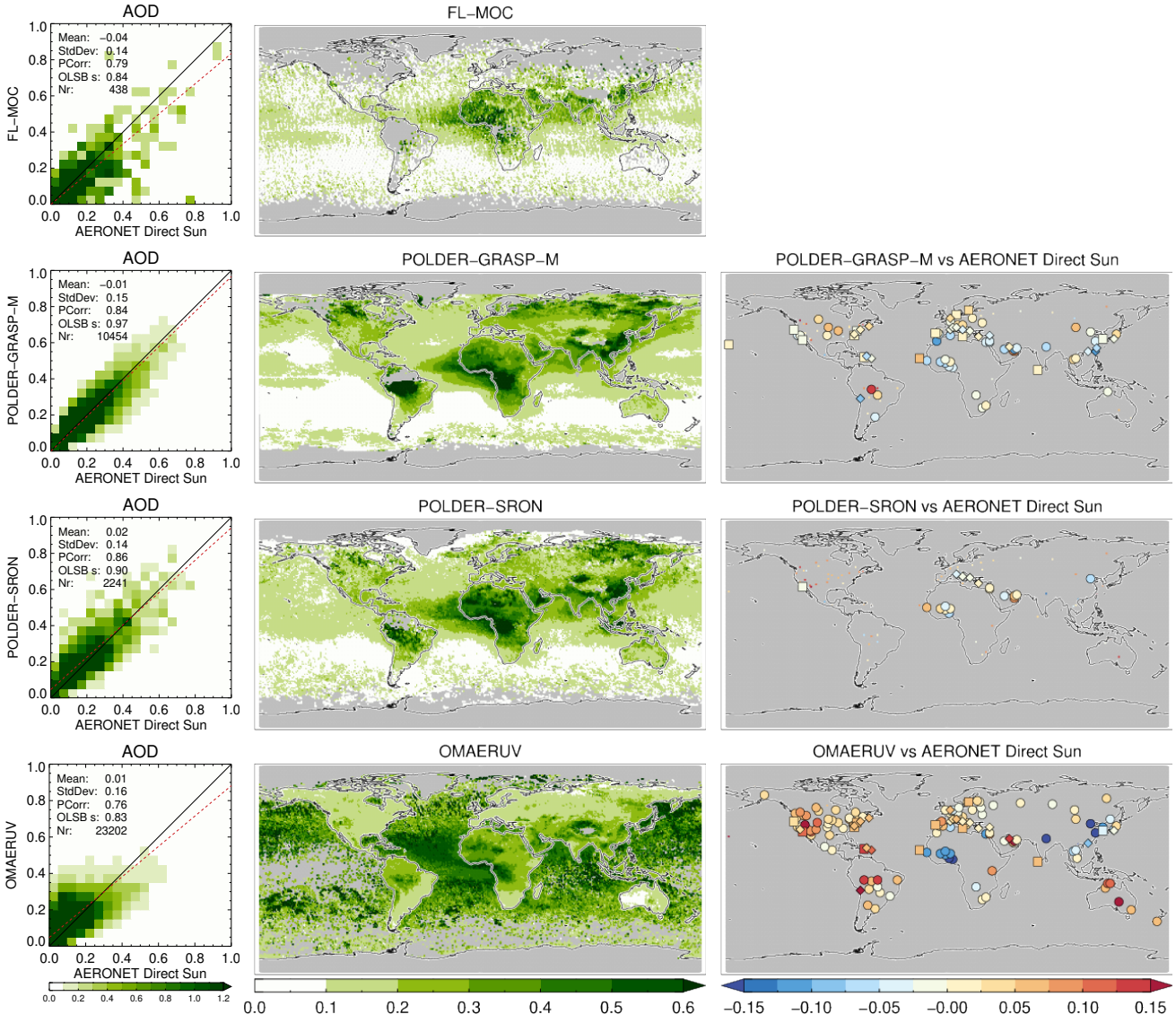

Figure 7. For the four satellite products are shown: a scatter plot of individual super-observations versus AERONET (the colour indicates amount of data in percentages, see Sect. **??** for an explanation of the metrics); a global map of the three-year AOD average; a global map of the three-year AOD difference average with AERONET (if site provided at least 32 observations; land sites are circles, ocean sites are squares, diamonds are the remainder). For FL-MOC, insufficient data prevent the plotting of a difference map. Products were individually collocated with AERONET DirectSun L2.0 within 3 hours.

grid-box. A single AERONET site however may contribute up to 6 super-observations per grid-box (in which case they are averaged). After two or more datasets are thus aggregated *individually*, only grid-boxes that contain data for both datasets will be used to construct two lists of aggregated data as in Fig. **??**. Those two lists will have identical size and ordering of times and

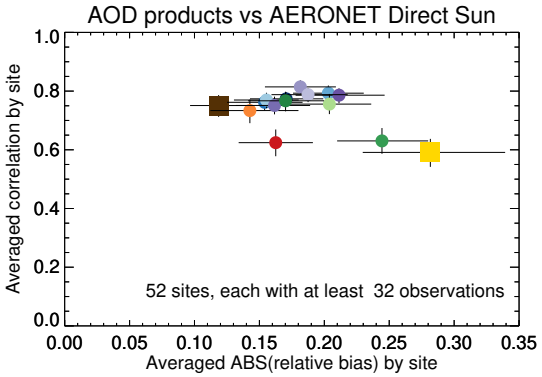

Figure 8. Evaluation of satellite products with AERONET per site, averaged over all sites. Squares indicate products used in the present study, circles indicate products used ~~?~~in **?**. Error bars indicate 5-95% uncertainty range based on a bootstrap analysis (see Sect. **??**) of sample size ~~1000.~~ 1000 (the bootstrap was performed on the contributing AERONET sites). Colours indicate satellite product, see also Fig. **??**. Products were individually collocated with AERONET DirectSun L2.0 within 3 hour. All products use the same sites, each of which produced at least 32 collocations. POLDER-SRON and FL-MOC were excluded from this analysis due to lack of data.

565 geo-locations and are called collocated datasets. By choosing a larger temporal extent of the grid-box, the collocation criterion can be relaxed.

As the super-observations are on a regular spatio-temporal grid and collocation requires further aggregation to another regular but coarser, grid, the whole procedure is very fast. It is possible to collocate ~~all~~ 7 products from afternoon platforms over three years using an IDL (Interactive Data Language) code (that served as a prototype for CIS) and a single processing 570 core in just 30 minutes (**?**). This greatly facilitates sensitivity studies.

Starting from super-observations, a 3-year average can easily be constructed by once more performing an aggregation operation but now with a grid-box of ~~$3^{\mathrm{yr}} \times 1^{o} \times 1^{o}$~~ $1^{o} \times 1^{o} \times ^{\mathrm{yr}}$. If two *collocated* datasets are aggregated in this fashion, their 3-year average can be compared with minimal representation errors. This allows us to construct global maps of e.g. multi-year AOD difference between two sets of super-observations.

575 A software tool (the Community Intercomparison Suite) is available for these operations at `www.cistools.net` (last accessed on December 20, 2019) and is described in great detail in **?**.

Table 1. Remote sensing products used in this study

Platform	Overpass [hr]	Sensor	Swath [km]	Pixel [km]	Product	(A)AOD[1] 550nm	Years	References
Aqua/AURA/ CALIPSO	1:30PM	MODIS/OMI/ CALIOP	1	1	FL-MOC[1]	R	2007, '08	?
AURA	1:30PM	OMI	2600	18	OMAERUV v1.8.9.1	E	2006, '08, '10	??
PARASOL	1:30PM[2]	POLDER	1600	6.18	POLDER-GRASP-M v1.2	I	2006, '08, '10	??
PARASOL	1:30PM[2]	POLDER	1600	6.18	POLDER-SRON	I	2006	? ?

[1] This product uses a combination of Aqua-MODIS, OMI and CALIOP observations

[2] PARASOL started drifting away from Aqua at the end of 2009.

[3] Interpolated or Extrapolated to 550 nm, depending on surface type; or Retrieved at 550 nm

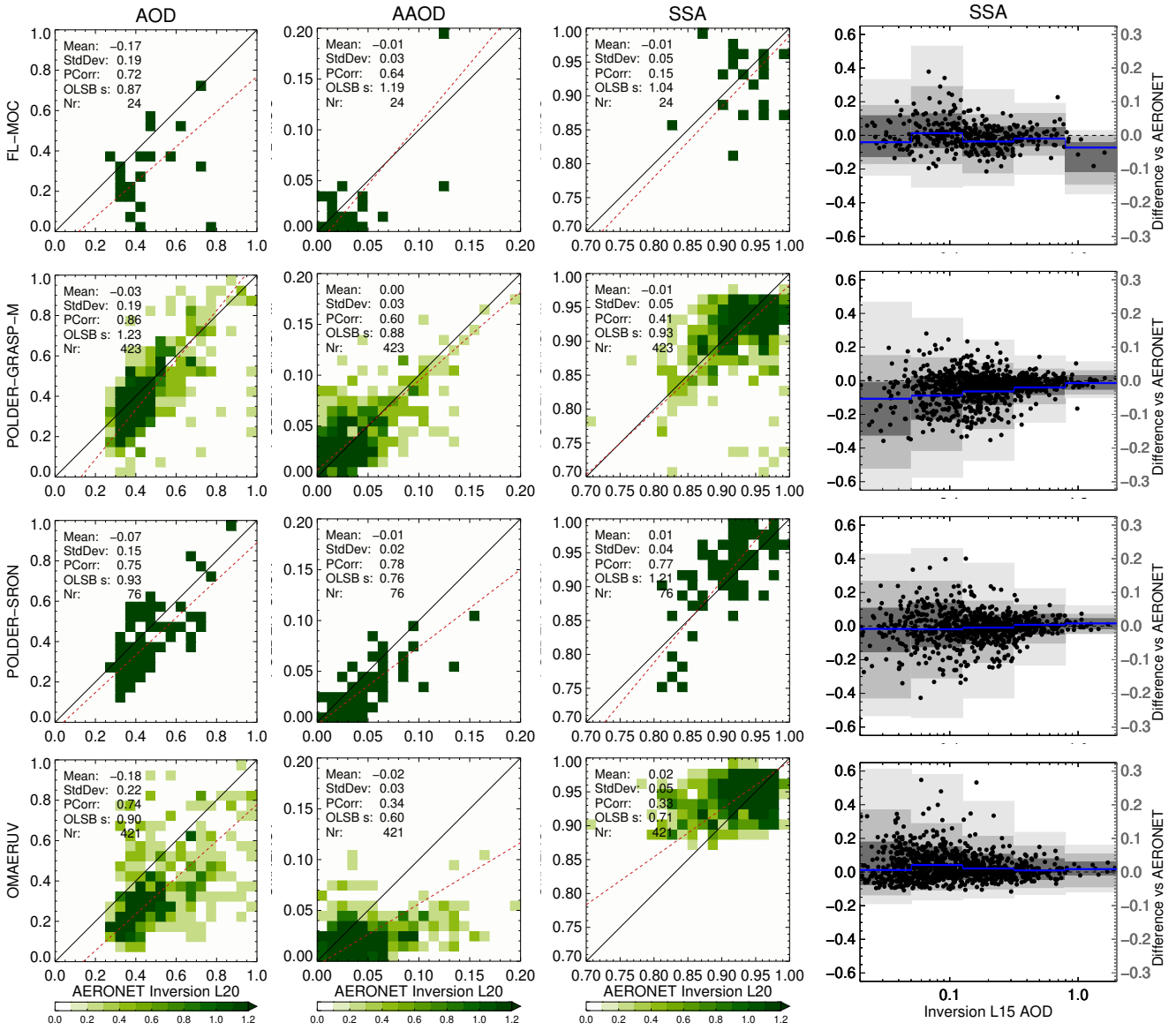

Figure 9. Evaluation of super-observations of AOD, AAOD and SSA for the satellite products. SSA is also evaluated as a function of AOD (binned). In the three left-most figures, the colour indicates amount of data in percentages; for an explanation of the metrics, see Sect. **??**. The right-most column uses two vertical axes: the left-hand side is used for individual data points (sub-sampled), the right-hand axis is used for the grey-scale distribution ($9, 25, 50, 75, 91\%$ quantiles) and the median difference (blue line). Products were individually collocated with AERONET Inversion L2.0 within 3 hour, except the right-most column which used Inversion L1.5.

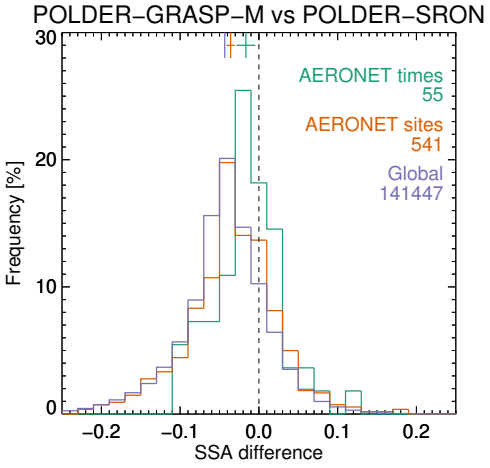

Figure 10. SSA differences POLDER-GRASP-M vs. POLDER-SRON for three different samplings: all available data, data available over AERONET sites that provide Inversion L2.0 data, data available at the times and locations of Inversion L2.0 data. The vertical coloured lines at the top show distribution means and the short horizontal lines extending from the middle show 2σ ranges. The dashed vertical line shows zero difference. Number of collocated data are indicated in the figure as well. This analysis suggests that an evaluation with AERONET would underestimate the actual difference between the two products. In all cases, data was collocated within 3 hours and a minimum AOD > 0.25 was required.

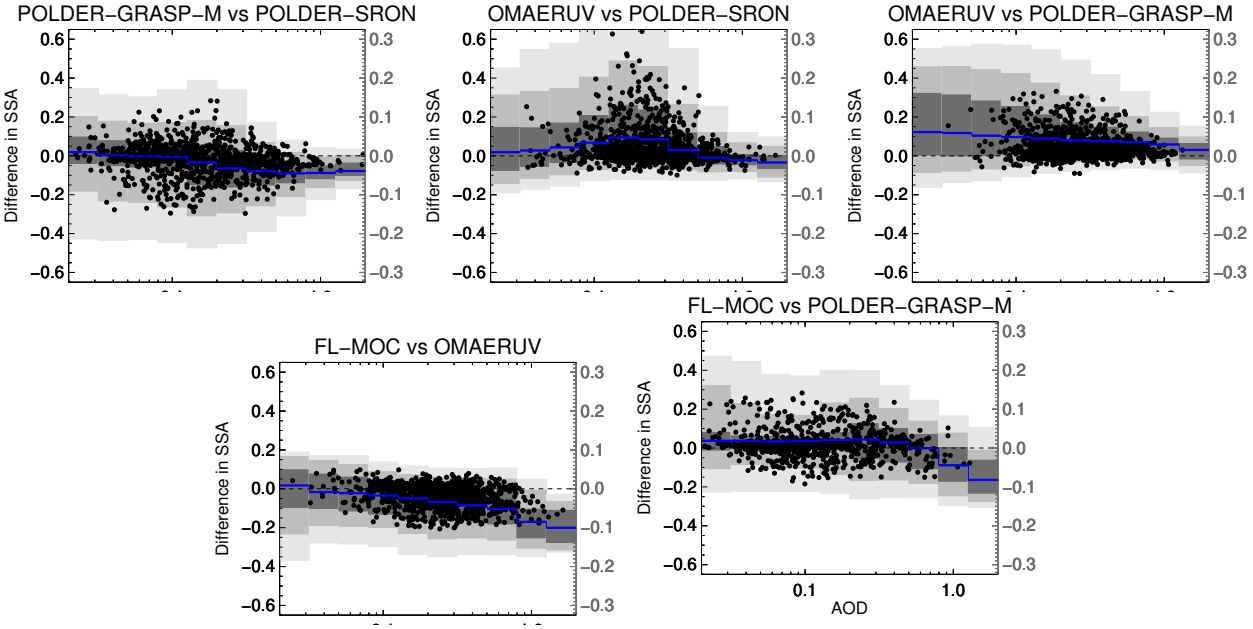

Figure 11. Difference in satellite product SSA as a function of AOD (averaged over both products). Two vertical axes are used: the left-hand side is used for individual data points (sub-sampled), the right-hand axis is used for the grey-scale distribution ($9, 25, 50, 75, 91\%$ quantiles) and the median difference (blue line). Data were collocated within 3 hours.

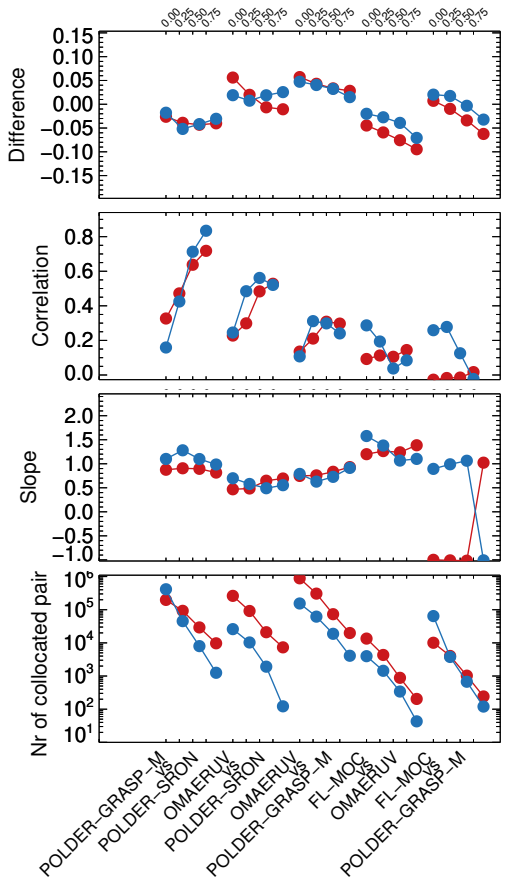

Figure 12. Comparison of different pairs of satellite SSA, over land (red) and ocean (blue), for different thresholds of minimum AOD (0.0, 0.25, 0.5, and 0.75). The data were collocated within 3 hours.

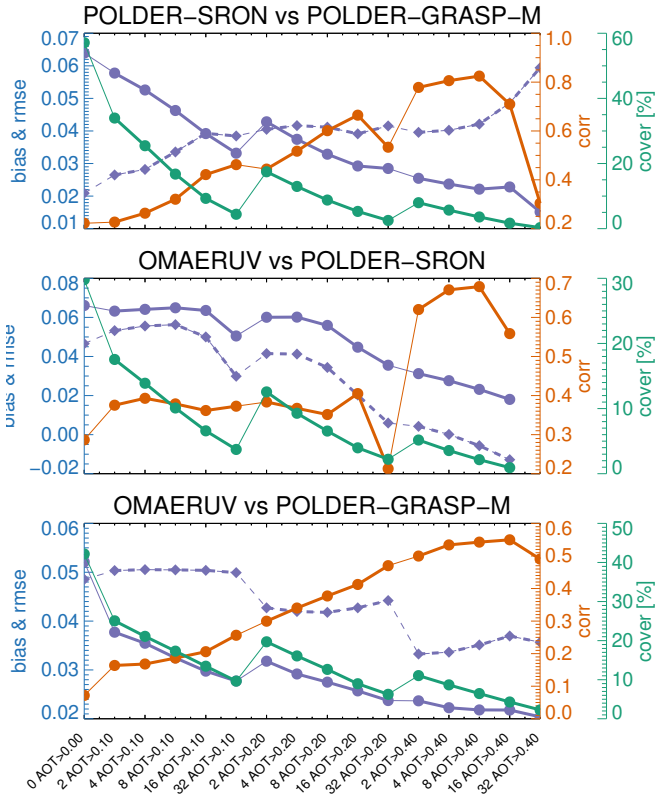

Figure 13. Intercomparison of SSA satellite products after multi-year averaging, as a function of minimum AOD and number of collocated observations (thicker lines group cases with the same minimum AOD but increasing number of observations). Bias uses a dashed line, and RMSE a solid line. Cover is defined as fraction of surface area covered by data. FL-MOC is not present due to scarcity of observations. The data were collocated within 3 hours.

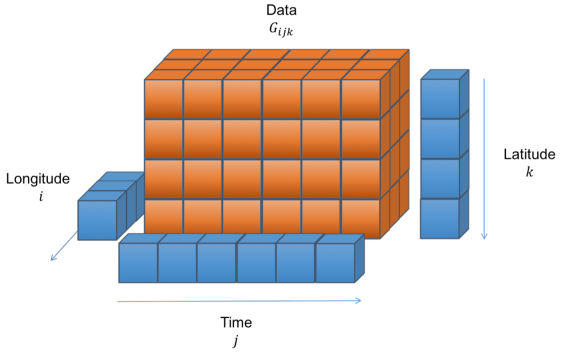

Figure A1. A regular spatio-temporal grid in time, longitude and latitude. Such a grid is used for the aggregation operation that is at the heart of the collocation procedure used in this paper. Grid-boxes may either contain data or be empty. Note that data may refer to any combination of observations, e.g. AOD at multiple wavelengths or AOD and AAOD at 550 nm. However, the dataset is homogenous. Reproduced from **?**.

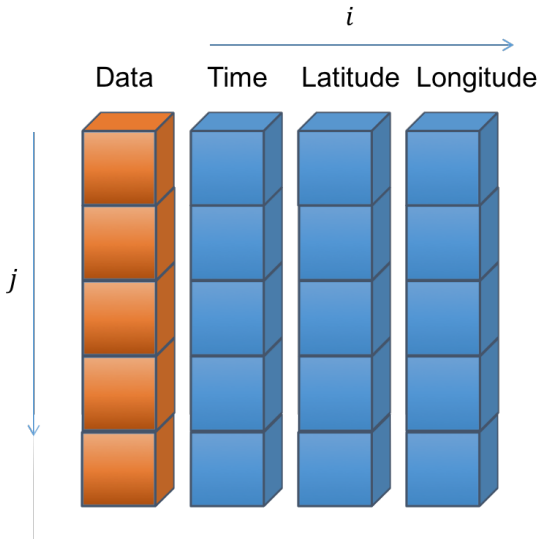

Figure A2. A list of data. Such a list is the primary data format used for ~~both~~ the observations ~~and model data~~ in this paper. Note that data may refer to any combination of observations, e.g. AOD at multiple wavelengths or AOD and AAOD at 550 nm. However, the dataset is homogenous. Reproduced from **?**.