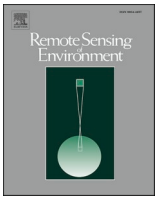
Remote Sensing of Environment 285 (2023) 113412

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse



Atmospheric Correction Inter-comparison eXercise, ACIX-II Land: An assessment of atmospheric correction processors for Landsat 8 and Sentinel-2 over land

Georgia Doxani [a,*], Eric F. Vermote [b], Jean-Claude Roger [b,c], Sergii Skakun [b,c], Ferran Gascon [d], Alan Collison [e], Liesbeth De Keukelaere [f], Camille Desjardins [g], David Frantz [h,i], Olivier Hagolle [g], Minsu Kim [j], Jérôme Louis [k], Fabio Pacifici [l], Bringfried Pflug [m], Hervé Poilvé [n], Didier Ramon [o], Rudolf Richter [p], Feng Yin [q,r]

[a] SERCO SpA c/o European Space Agency ESA, European Space Research Institute (ESRIN), 00044 Frascati, Italy
[b] NASA Goddard Space Flight Center Code 619, Greenbelt, MD 20771, USA
[c] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA
[d] European Space Agency ESA, European Space Research Institute (ESRIN), 00044 Frascati, Italy
[e] Planet Labs PBC, San Francisco, USA
[f] VITO, Boeretang 200, 2400 Mol, Belgium
[g] Centre National d'Etudes Spatiales CNES, 31401 Toulouse Cedex 9, France
[h] Geography Department, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
[i] Earth Observation and Climate Processes, Trier University, 54286 Trier, Germany
[j] KBR, Contractor to U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS), Sioux Falls, SD 57198, USA
[k] Telespazio France, 31023 Toulouse, France
[l] Maxar, Denver, CO, United States
[m] DLR, German Aerospace Center, Berlin, Germany
[n] Airbus Defence and Space, 5, rue des Satellites, 31400 Toulouse, France
[o] HYGEOS, Euratechnologies, 165 avenue de Bretagne, 59000 Lille, France
[p] DLR, German Aerospace Center, Wessling, Germany
[q] Department of Geography, University College London, United Kingdom
[r] National Centre for Earth Observation (NCEO), NERC, United Kingdom

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Atmospheric correction
Intercomparison
Landsat 8
Copernicus Sentinel-2
Surface reflectance
Aerosol optical depth
Water vapour

ABSTRACT

The correction of the atmospheric effects on optical satellite images is essential for quantitative and multi-temporal remote sensing applications. In order to study the performance of the state-of-the-art methods in an integrated way, a voluntary and open-access benchmark Atmospheric Correction Inter-comparison eXercise (ACIX) was initiated in 2016 in the frame of Committee on Earth Observation Satellites (CEOS) Working Group on Calibration & Validation (WGCV). The first exercise was extended in a second edition wherein twelve atmospheric correction (AC) processors, a substantially larger testing dataset and additional validation metrics were involved. The sites for the inter-comparison analysis were defined by investigating the full catalogue of the Aerosol Robotic Network (AERONET) sites for coincident measurements with satellites' overpass. Although there were more than one hundred sites for Copernicus Sentinel-2 and Landsat 8 acquisitions, the analysis presented in this paper concerns only the common matchups amongst all processors, reducing the number to 79 and 62 sites respectively. Aerosol Optical Depth (AOD) and Water Vapour (WV) retrievals were consequently validated based on the available AERONET observations. The processors mostly succeeded in retrieving AOD for relatively light to medium aerosol loading (AOD < 0.2) with uncertainties <0.08, while the overall uncertainty values were typically 0.23 ± 0.15. Better performances were observed for WV retrievals with $>90\%$ of the results falling within the suggested empirical specifications and with the Root Mean Square Error (RMSE) being mostly <0.25 g/cm^2. Regarding Surface Reflectance (SR) validation two main approaches were followed. For the first one, a simulated SR reference dataset was computed over all of the test sites by using the 6SV (Second Simulation of the

* Corresponding author at: SERCO SpA c/o European Space Agency, ESA-ESRIN, Largo Galileo Galilei, 00044 Frascati, Italy.
E-mail address: georgia.doxani@esa.int (G. Doxani).

https://doi.org/10.1016/j.rse.2022.113412
Received 14 July 2022; Received in revised form 25 October 2022; Accepted 6 December 2022
Available online 21 December 2022

Satellite Signal in the Solar Spectrum vector code) full radiative transfer modelling (RTM) and AERONET measurements for the required aerosol variables and water vapour content. The performance assessment demonstrated that the retrievals were not biased for most of the bands. The uncertainties ranged from approximately 0.003 to 0.01 (excluding B01) for the best performing processors in both sensors' analyses. For the second one, measurements from the radiometric calibration network RadCalNet over La Crau (France) and Gobabeb (Namibia) were involved in the validation. The performance of the processors was in general consistent across all bands for both sensors and with low standard deviations (<0.04) between on-site and estimated surface reflectance. Overall, our study provides a good insight of AC algorithms' performance to developers and users, pointing out similarities and differences for AOD, WV and SR retrievals. Such validation though still lacks of ground-based measurements of known uncertainty to better assess and characterize the uncertainties in SR retrievals.

1. Introduction

The quantitative use of Earth Observation (EO) optical satellite data entails the correction of the atmospheric effects on the top of atmosphere radiance values. The atmospheric correction mainly involves the computation of atmospheric variables using a radiative transfer model, the estimation of the two main scattering and absorption contributors, i.e., aerosols and water vapour, and the conversion of Top-Of-Atmosphere (TOA) to Bottom-Of-Atmosphere (BOA) reflectance — also referred as Surface Reflectance (SR) (Liang and Wang, 2020).

Since free and open access to a large volume of EO data increased considerably the number of applications, particularly those based on time series, an accurate atmospheric correction (AC) became an essential data pre-processing step for precise land monitoring and land products retrieval, e.g., leaf chlorophyll, mineral concentration in soil, water quality parameters, etc. Consequently, researchers have focused on developing innovative AC approaches utilizing various radiative transfer models (RTMs), single or multitemporal images, constant or various aerosol models, new sources of ancillary data, etc. In addition, data users request the provision of corrected data, i.e., geometrically, radiometrically and atmospherically, which are consistent and can be used for multitemporal analysis. Towards this end, several entities have started to investigate ways to generate Analysis Ready Data (ARD) (Dwyer et al., 2018; Frantz, 2019; Potapov et al., 2020), described by CEOS as "satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets" (http://ceos.org/ard).

The validation of the available AC methodologies is usually performed independently by developers and/or users either based on reference data over a certain, typically small, number of sites or by comparing with other processors' outcomes (Claverie et al., 2015; Vermote et al., 2016; Rouquié et al., 2017; Li et al., 2018). With the intention to investigate all the AC aspects and issues in a consolidated way, a benchmark Atmospheric Correction Inter-comparison eXercise (ACIX) was initiated in order to compare the state-of-the-art AC processors. ACIX is a voluntary and open-access initiative to which every AC processor's developer is invited to participate. ACIX-I was an initial attempt to study the variability of AC performances over diverse atmospheric and land cover conditions using Landsat 8 and Sentinel-2A input data. The description and the results of the exercise are summarized in Doxani et al. (2018). ACIX-I was endorsed by the participants and considered as a useful tool to discover not only the assets and flaws of the approaches, but also ways to improve them. Thus, a second implementation of the experiment was requested to inter-compare the enhanced versions of the participating processors, but also to be expanded by including additional AC processors and validation approaches. Moreover, ACIX was split in two categories: -Land and -Aqua, with focus on the processors' performance over land and lakes, rivers, and coastal waters accordingly. In this paper, attention is given only to the -Land part of the exercise, while the description and results of ACIX-Aqua can be found in Pahlevan et al. (2021).

The main objective of this paper is to give an overview of the performance of the most commonly used AC processors over a globally distributed dataset of Sentinel-2 Multispectral Instrument (MSI) and Landsat 8 Operational Land Imager (OLI) acquisitions. A detailed description of all the guidelines, i.e., datasets, sites, metrics and reference sources, is given in Section 2 of this paper. In Section 3, the main results are highlighted in two sub-sections dedicated to the processors applied on Sentinel-2 and Landsat 8 data. A discussion on the overall analysis is presented in Section 4, while conclusions are given in the final section.

2. Methods

In this section the validation strategy is presented thoroughly per AC product, with the aim to define a set of guidelines for future similar activities. The processors involved in the analysis are described in Section 2.1, while datasets and sites of interest in Section 2.2. The validation of AOD and WV with the help of AERONET measurements (Holben et al., 1998; Giles et al., 2019) is given in Section 2.3.1. As an equivalent network for surface reflectance validation is not available yet, several alternative approaches had to be engaged as outlined in the rest of Section 2.3. In particular, simulated reference SRs were produced similarly to the methodology introduced by Vermote and Kotchenova (2008) and Roger et al. (2022) and ground measurements were employed by Centre National d'Études Spatiales (CNES) RObotic Station for Atmosphere and Surface characterization (ROSAS) stations (Meygret et al., 2011), part of the Radiometric Calibration Network (RadCalNet) (Bouvet et al., 2019).

2.1. Atmospheric correction processors

The features of the processors which participated in ACIX-II Land are described in Table 1. In total, eleven AC processors were applied to Sentinel-2A, -2B and nine to Landsat 8 data. The developers were responsible for their processor's implementation and results' submission. Due to the fact that terrain and BRDF (Bidirectional Reflectance Distribution Function) corrections were not necessarily part of the operational processing chain of the participating AC processors, they were omitted in order to acquire consistent, homogeneous and comparable results overall. The correction of adjacency effects, conversely, could be included in the process, since in the previous exercise minor differences related to these corrections were observed at the prescribed ACIX image subset, i.e., 9 km × 9 km (Doxani et al., 2018). Therefore, surface reflectance in our study refers mainly to the top of atmosphere (TOA) radiance corrected for the scattering and absorbing effects of atmospheric gases and aerosols (Vermote et al., 2016).

Quality description layers indicating the quality attributes of the pixels, i.e., cloud contamination, instrument's artifacts and surface conditions, accompanied the results submitted by most of the participants (Table 1). These quality layers were used either altogether or in combinations as provided by the developers and without being validated by the ACIX-II team. For the assessment of quality layers and particularly the ones associated to clouds, a Cloud Masking Intercomparison eXercise (CMIX) was performed along with ACIX-II (Skakun et al.,

Table 1
The features of the atmospheric correction processors as implemented in ACIX-II Land.

	AComp	ATCOR	FORCE	EMBAC	LaSRC	MAJA	Overland	Planet_SR	SIAC	SMACG	Sen2Cor	iCOR
Sensor Data	S2/L8	S2/L8	S2/L8	L8	S2/L8	S2	S2/L8	S2/L8	S2/L8	S2	S2	S2/L8
RTM	MODTRAN5	MODTRAN5.4.0	modified 5S, multiple scattering, water vapour absorption coefficients from HITRAN2016	6SV2.1	6SV2.1	Successive Orders of Scattering (SOS)	LOWTRAN7 plus cloud model	6SV2.1	6SV2.1	SMAC	libRadtran 2.0.1	MODTRAN5
AOD	**Proprietary**	Dense Dark Vegetation and Dark Soils	triangle-shape-based Dense Dark Vegetation approach, spectral-library-based Dark Water approach, interpolation, estimation of elevation-dependency	End-Member optimization & local max AOD estimation	Blue/Red/SWIR ratios spatially and temporally variable derived from MODIS/MISR	Cost function inversion combining several multi-spectral and multi-temporal criteria	Inversion with coupled RTM/scene model plus weighted spatial filtering	MOD09CMA (NRT and standard)	Inversion with MCD43 simulated surface reflectance	From ancillary data: MERRA-2 reanalysis	Dense Dark Vegetation, CAMS as a fall-back solution	Dense Dark Vegetation and a multi-parameter end member inversion technique, CAMS as a fall-back solution
Aerosol Model	Rural, urban, maritime, desert, or mix	Rural	continental model in two-term Henyey–Greenstein (TTHG) function, estimation of Angstrom coefficients (linear model for vegetation, second-order polynomial for water)	Urban clean	Urban clean with variable Angstrom coefficient (retrieved)	CAMS used to set aerosol type	Rural with adaptative parameters (angstrom coef., relative humidity)	Continental	Continental	Mix of SU, DU, BC, OC and SS components	Rural	Rural or desert
Water Vapour	Proprietary	Atmospheric Pre-corrected Differential Absorption Algorithm (APDA)	Sentinel-2: optimization of water vapour until SR of bands 8A and 9 match; scene average for water and shadows. Landsat: day-specific MODIS scene average if available; climatology otherwise	MOD09CMA (NRT and standard)	MODIS retrieval in priority from Terra then Aqua then GDAS (assimilation) when none of the previous are available	Inversion assuming B8a and B9 have the same surface reflectance, using an exponential fit based on 6S	Inversion with coupled RTM/scene model	MOD09CMA (NRT and standard)	Prior information from CAMS and observations from S2/L8 bands	From ancillary data: MERRA-2 reanalysis	Atmospheric Pre-corrected Differential Absorption Algorithm (APDA)	Atmospheric Pre-corrected Differential Absorption Algorithm (APDA)
Corrections for ACIX-II Land												
Cirrus	Yes	Yes	No	Yes		No	Yes	No	No	No	No	No
Adjacency Effects	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Yes	No
Terrain Correction	No	No	No	No	No	No	No	No	No	No	No	No
BRDF	No	No	No	No	No	No	No	No	No	No	No	No
Auxiliary Data												
DEM		7.5 Arcsec GMTED 2010 DEM	30 m SRTM DEM, filled with the 30 m ASTER DEM	–	ETOPO5	30 m SRTM DEM	90 m SRTM DEM - GLSDEM for lat > 65°	–	ASTER GDEM	30 STRM DEM	Planet DEM	GLOBE DEM
other data			Landsat 8: water vapour from MODIS	–	MOD09CMA (WV)	CAMS data to get information		MOD09CMA (AOD, WV) and MOD09CMG	MODIS MCD43 BRDF descriptor product,		ESA CCI Maps a priori information, a snow	Monthly ozone climatology data derived

(*continued on next page*)

Table 1 (*continued*)

	AComp	ATCOR	FORCE	EMBAC	LaSRC	MAJA	Overland	Planet_SR	SIAC	SMACG	Sen2Cor	iCOR
						on the aerosol type		(OZ) (NRT and standard)	CAMS priors on AOT, TCWV and TCO3		climatology included in the installation package, CAMS for the AOT fall back	from TOMS, ECMWF Water vapour data (only for Landsat 8), CAMS for the AOT fall back
Quality Flags	–	Yes	Yes	Per-pixel uncertainty information	Yes	Yes	Yes	–	Per-pixel uncertainty information	Per-pixel uncertainty information	Yes	Yes
Version	2.0.0	9.3.0	3.0-dev	1	3.5	3.4	2019.1.0	1	2.3.6		2.8	
Open source access	No	No	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Organisation	Maxar	DLR	Universität Trier, Humboldt-Universität zu Berlin	KBR/USGS	NASA	CNES, CESBIO, DLR	Airbus	Planet	University College London	Hygeos	ESA, Telespazio France, DLR	Vito
References	–	Richter, 1998	Frantz, 2019, Frantz et al., 2016	–	Vermote et al., 2016	Hagolle et al., 2017	–	white paper	Yin et al., 2019	–	Level-2A Algorithm Theoretical Basis Document (ATBD) version 2.10	De Keukelaere et al., 2018

2022). Although different datasets, in terms of time and location, were involved in these two exercises, CMIX outcomes are good indicators of how the clouds are defined in each case. In particular, in ACIX-II the union of all the 'good quality pixels' as indicated by each processor defined the pixels to participate in the analysis, i.e., 'clear' (meaning not a 'cloud' or 'cloud shadow') and 'snow'. Processors without any quality descriptor in their processing chain, such as EMBAC, SIAC and SMAC-G, follow an error propagation scheme to assess the uncertainty on surface reflectance per pixel. However, the uncertainty layers were not taken into consideration in this current experiment. Despite the provision of quality information, all of the processors' results were evaluated based on the common quality approved pixels.

2.2. Input and validation datasets

For the first ACIX experiment the AC performances were analysed over a small set of 19 sites, which nonetheless included a variety of land cover and aerosol types (Doxani et al., 2018). The number of sites was considered adequate for the first implementation of such a type of exercise, considering the processing effort for the participants, who were responsible for running the processors in their own premises, and suitable to achieve reliable performance conclusions. The sites were selected as part of AERONET, which is regarded as a reliable, ground-based measurements network for atmospheric variables, such as the Aerosol Optical Depth (AOD) (Holben et al., 1998; Giles et al., 2019). In this second exercise, organisers and participants agreed that a wider selection of sites was necessary to acquire more global and robust outcomes. To that end, the full catalogue of AERONET sites was investigated in order to identify the ones with available measurements within 30 min (±15 min) from the satellites' overpass. Eventually a total of 123 and 110 AERONET sites distributed globally and representing various land cover types were selected for Sentinel-2A, -2B and Landsat 8 acquisitions respectively (Fig. 1). 'Urban' is the main class (41%) and almost equally representative to green areas, taken together 'cropland' (28%) and 'forest' (14%). 'Arid/desert' (10%) sites and 'water' (7%) are also part of the study areas but to a lesser extent.

The experiment was implemented on a yearly time series acquired from October 2017 to September 2018. For this time period, there were around 2500 Sentinel-2 and 1250 Landsat 8 scenes with available coincident AERONET observations. The set of Sentinel-2 and Landsat 8 L-1C image scenes were in their original format and structure as provided by ESA (Sentinel-2 L-1C User guide) and USGS (Landsat Level-1 Processing Details) respectively.

The original scene size remained the same, i.e., 10,980 × 10,980 pixels for Sentinel-2 10 m bands, 7791 pixels x 7671pixels for Landsat 8. However, due to the big data volume and the computational demands implied, all the pixels outside a 30 km × 30 km area centered on AERONET station location are assigned to 'No data' values. The size of the 'valid' pixel area (30 × 30 km) was selected as the minimum adequate for the processors that require the detection of Dark Dense Vegetation (DDV) pixels in the scene in order to estimate AOD. Eventually, the intercomparison analysis was performed on 9 km × 9 km image subsets, always centered on the AERONET station. This image size was chosen to assure a spatial atmospheric homogeneity and to contain a whole number of pixels for Sentinel-2 and Landsat 8 spatial resolutions, i.e., 10 m, 20 m, 60 m and 30 m accordingly.

2.3. Validation methodology

2.3.1. AOD and WV validation using AERONET measurements

AOD retrieval is a key component of the atmospheric correction processing chain, as aerosols can reduce the solar radiation reaching an optical sensor through absorption and/or scattering in the atmosphere (Liang and Wang, 2020). The validation was performed with the help of AERONET measurements collected at a temporal window of ±15 min from the satellites' overpass, interpolated at 550 nm using the Ångström

Fig. 1. Location of the AERONET sites over which Sentinel-2A, -2B and Landsat 8 scenes were acquired. The common sites are indicated with green circles, while red diamonds correspond to sites with only Sentinel-2 acquisitions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exponent and averaged. The specific wavelength was chosen as a reference, because it is the common reporting wavelength for AOD. At least two AERONET measurements were required in the 30 min temporal window in order to consider it as a valid reference measurement and participate in the analysis. Although AERONET delivers the measurements at three processing levels, i.e., unscreened (Level 1.0), cloud screened (Level 1.5), and cloud screened and quality assured (Level 2.0), Level 1.5 data were used in this exercise. Unfortunately, the better-quality Level 2.0 data for the years 2017 and 2018 were not available for all the sites, when the exercise was running. The AOD estimation assessment was achieved by validating the averaged retrieved AOD over the analysis image subset of 9 km × 9 km with the AERONET reference value. The two variables were plotted including all the sites and the statistic sample, Root Mean Square Error (RMSE), bias and Coefficient of determination (R^2) were calculated.

Additional quantitative analysis was performed by estimating the Accuracy (A), Precision (P) and Uncertainty (U) metrics [Eq.1, 2, 3] per 0.1 AOD value bin, where n is the number of valid samples used for the validation and Δ_{AOD} is the difference between the estimate and the reference observation.

$$Accuracy\ (A) : A = \frac{1}{n}\left(\sum_{i=1}^{n} \Delta_{AOD}\right) \tag{1}$$

$$Precision\ (P) : P = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(\Delta_{AOD} - A)^2} \tag{2}$$

$$Uncertainty\ (U) : U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta_{AOD}\Big)^2} \tag{3}$$

WV was computed and validated only for Sentinel-2 MSI data, thanks to the available B09 spectral band in the WV absorption region (Central Wavelength: 945 nm) that is suitable for WV retrieval. The Landsat 8 OLI instrument does not collect data in this spectral region, so WV could not be retrieved directly. The validation analysis was similar to the one for the AOD, comparing the averaged estimated WV over the 9 km × 9 km image subset with the corresponding AERONET reference measurements, which were collected at a temporal window of ±15 min from the satellites' overpass and averaged.

2.3.2. Surface reflectance validation using AERONET-derived reference

Considering the limitations to a direct SR validation, an indirect approach was followed by retrieving simulated surface reflectance based on 6SV radiative transfer code (Kotchenova et al., 2006) with the required atmospheric variables, i.e. water vapour and aerosol information, to have been obtained from AERONET measurements (Roger et al., 2022; Vermote and Kotchenova, 2008; Claverie et al., 2015; Vermote et al., 2016; Doxani et al., 2018). The simulated SRs were computed for 9 km × 9 km image subsets over the AERONET sites of interest based on Level-1C products of both sensors. These simulated, AERONET-derived values were considered as the reference in our analysis, which was performed on a per pixel basis and only for the pixels that were not labeled as clouds, cloud shadows, snow, water, high aerosol loads. In this analysis, the quality pixels were the common quality-approved pixels by the Quality Assessment (QA) band estimated by LaSRC and the analysed processor's quality mask. If the processor did not provide any quality information, only the LaSRC QA was taken into consideration. LaSRC QA layer provided a high cloud producer's accuracy when compared to other cloud masks in various datasets in CMIX (Skakun et al., 2022), so it was considered adequate for accurately detecting most of the clouds and defining the quality pixels. Additional thresholds to OLI B6 and B7 pixel values, as well as to MSI B11 and B12 ensured that water pixels were excluded from the analysis. The quantitative analysis was performed estimating Accuracy (A), Precision (P) and Uncertainty (U) metrics per wavelength based on the Eq.1, 2 and 3 per 0.01 SR value bin. The statistical results were compared to suggested specifications for Sentinel-2 and Landsat 8 SR, i.e., specs = 0.05ρ + 0.005, for the corresponding reflectance value ρ (Vermote and Kotchenova, 2008; Doxani et al., 2018).

2.3.3. Surface reflection validation with RadCalNet measurements

Due to the absence of an operational network to deliver in-situ observations suitable for the validation of satellite-derived surface reflectance, measurements from the Radiometric Calibration Network (RadCalNet) (Bouvet et al., 2019) were involved in the analysis. In particular, observations from La Crau (France) and Gobabeb (Namibia)

sites were used as references to validate the retrieved surface reflectance. Both stations are equipped with a ROSAS system (RObotic Station for Atmosphere and Surface), which is comprised of a CIMEL photometer, similar to the AERONET one. However, besides the atmospheric properties, the ROSAS instruments also measure the upwelling radiance over the ground to fully characterize the surface reflectance properties (BRDF). The photometers measure from 414 nm to 1600 nm in 12 narrow spectral bands (Meygret et al., 2011; Marcq et al., 2018). The BRDF values for the Sentinel-2 and Landsat 8 sun and sensor geometries were extracted, and spectrally integrated to the corresponding sensor spectral bands of Sentinel-2 and Landsat 8 (Rouquié et al., 2017). Both La Crau and Gobabeb stations are operated by the French National Centre for Space Studies (CNES), either entirely (La Crau) or with ESA and NPL (National Physical Laboratory) collaboration (Gobabeb). The data pre-processing was also performed by CNES to fit the exercise purposes. It should be noted that the Gobabeb ROSAS system was unfortunately out of service from June to the beginning of December 2018, meaning that there were no available measurements for four out of the 12-months ACIX study period. In addition, the observations for the 60 m Sentinel- 2 bands, B01 and B09, were not provided by CNES due to the corresponding strong aerosol impact and water vapour absorption, so they were excluded from this analysis. The SWIR bands, i.e., Sentinel-2 B12 and Landsat 8 Band 7, are also missing as they are not covered by the spectral range observed by ROSAS photometers.

La Crau station is located in a flat area of south-eastern France with dry and sunny Mediterranean climate conditions. The area is mainly covered by pebbles and sparsely by low vegetation (Bouvet et al., 2019). The average measured AOD value at 550 nm for the study period was around 0.1 and water vapour around 1.06 g/cm^2. As a study area CNES team recommended a square of 60 m × 60 m for the 20 m bands, i.e., a 3 × 3 pixels area centred at the station's location, and 70 × 70 m for the 10 m bands (7 × 7 pixels). Regarding LANDSAT 8, a square of 3 × 3 pixels was selected, meaning an area of 90 m × 90 m. Gobabeb is a flat, arid site with high spatial homogeneity, so a larger study area of 5 × 5 pixels was selected for 20 m bands, and 9 × 9 pixels for 10 m bands. Regarding the atmospheric conditions for the study period, the average measured AOD at 550 nm was around 0.06 and water vapour around 1 g/cm^2.

The validation of the retrieved Sentinel-2 and Landsat-8 surface reflectance was performed by calculating the mean and standard deviation of relative differences according to Eq. 4. The estimated reflectances $X_{processor}$ were averaged over the aforementioned areas by site and spatial resolution.

$$Relative\ Difference = \frac{X_{processor} - X_{RadCalNet}}{X_{RadCalNet}} \quad (4)$$

2.3.4. Noise of surface reflectance time series

The assessment of the surface reflectance variation between three consecutive scenes a few days apart can be considered as a performance indicator for an atmospheric correction processor. The metric was chosen on account of the fact that surface reflectance should vary smoothly over a short time period, while the atmospheric effects can change quickly. Therefore, the smaller the difference in successive surface reflectance is, the more accurate the atmospheric correction approach should be. To estimate the time series variation, the root mean square of the differences between the linearly interpolated and actually observed values of three successive days few days apart was estimated as introduced by Vermote et al. (2009) (Eq.4) and implemented in Rouquié et al. (2017):

$$Noise = \sqrt{\frac{\sum_{i=1}^{n-2}\left(\rho_{i+1} - \frac{\rho_{i+2}-\rho_i}{d_{i+2}-d_i}(d_{i+1}-d_i) - \rho_i\right)^2}{n-2}} \quad (5)$$

where ρ_i, ρ_{i+1}, and ρ_{i+2} are the surface reflectance of days d_i, d_{i+1}, and d_{i+2}, respectively. The calculations were performed for a set of 81 pixels per scene, selected with a step of 100 pixels for every row and column of the 9 km × 9 km image subset. The mean surface reflectance was computed on a 7 × 7 pixel window around these 81 pixels, in order to reduce the impact of image registration errors (Rouquié et al., 2017). Similar to the rest of ACIX analyses, only the pixels that were quality approved by all processors participated in the computations. A weight was assigned to every pixel based on the number of dates at which the pixel was quality accepted. The weighted average was calculated for all the 81 pixels. As noise estimation assumes a linear variation of observations a few days apart, a threshold of 20 days was set as the maximum difference between the two extreme dates d_i and d_{i+2}.

3. Results

The presentation of the performance analysis is organised by sensor, atmospheric correction product and metric. For the sake of brevity, just the most representative results will be shown in this paper. An extensive presentation of the exercise and the results can be found on the ACIX-II Land web page that is hosted in CEOS Cal/Val portal (http://calvalportal.ceos.org/acix-ii-land).

3.1. Sentinel-2 MSI

The atmospheric correction products of 11 processors are analysed in this section. Although there were initially >2500 Sentinel-2 scenes acquired over 120 AERONET sites spread globally (Fig. 1), some processors either did not provide any outputs over certain sites or they provided outputs of incorrect type in terms of file format, image size, projection or tile, resulting this way in a variation of the sample size per processor. Therefore, the AOD, WV and SR analysis presented in this paper concerns only the common matchups (sites and dates) amongst all the processors involved in the corresponding analysis, in order to ensure a fair comparison. This option limited the number of sites for Sentinel-2 data to 79. The full analysis regarding all the samples as provided by processor can be found in CEOS Cal/Val portal.

3.1.1. AOD validation over AERONET sites

Four of the processors initiate the AOD retrieval with the Dense Dark Vegetation (DDV) method and then follow various refining approaches based on spectral criteria (Table 1). Planet_SR does not retrieve AOD and uses the equivalent MODIS product (MOD09CMA), so it is not included in this analysis. The scatterplots of AOD as estimated by the participants at 550 nm wavelength versus the corresponding AERONET measurements are presented in Fig. 2. For validating the AOD estimates, the statistic sample, Root Mean Square Error (RMSE), bias and Coefficient of determination (R^2) were estimated and shown in the plots and in Table 2. In the ideal case where the estimates are identical to the reference, the results would align to the 1:1 line indicated with a black dashed line. The black solid lines represent the uncertainty limits specifications, specs = 0.15*AOD_{550ref} + 0.05, according to the empirical uncertainty of MODIS land AOD retrievals (Remer et al., 2009). The blue solid line is the least-squares regression line for the two compared datasets.

The results are mostly in moderate agreement with the reference values, with the correlation R^2 to range from ~ 0.65 to 0.77 and RMSE from ~ 0.115 to 0.2. Overall, SMACG and iCOR demonstrate superior performances with relatively high correlation and low RMSE. AComp, SIAC and MAJA are also found to perform well with retrievals in good agreement with the reference AOD in terms of correlation, RMSE and bias. More than ~60% of the retrievals fall within the suggested uncertainty specifications for most of the processors, with SMACG to achieve the greatest number of points (78%), while Overland the lowest (40%). ATCOR and FORCE use a default AOD value as a fallback solution, when they are not able to detect pixels suitable for AOD retrievals, a fact that leads to large errors and impacts their general statistical performance. However, quality flags exist in FORCE, with which these conditions can be easily filtered out. In addition, this usually happens in

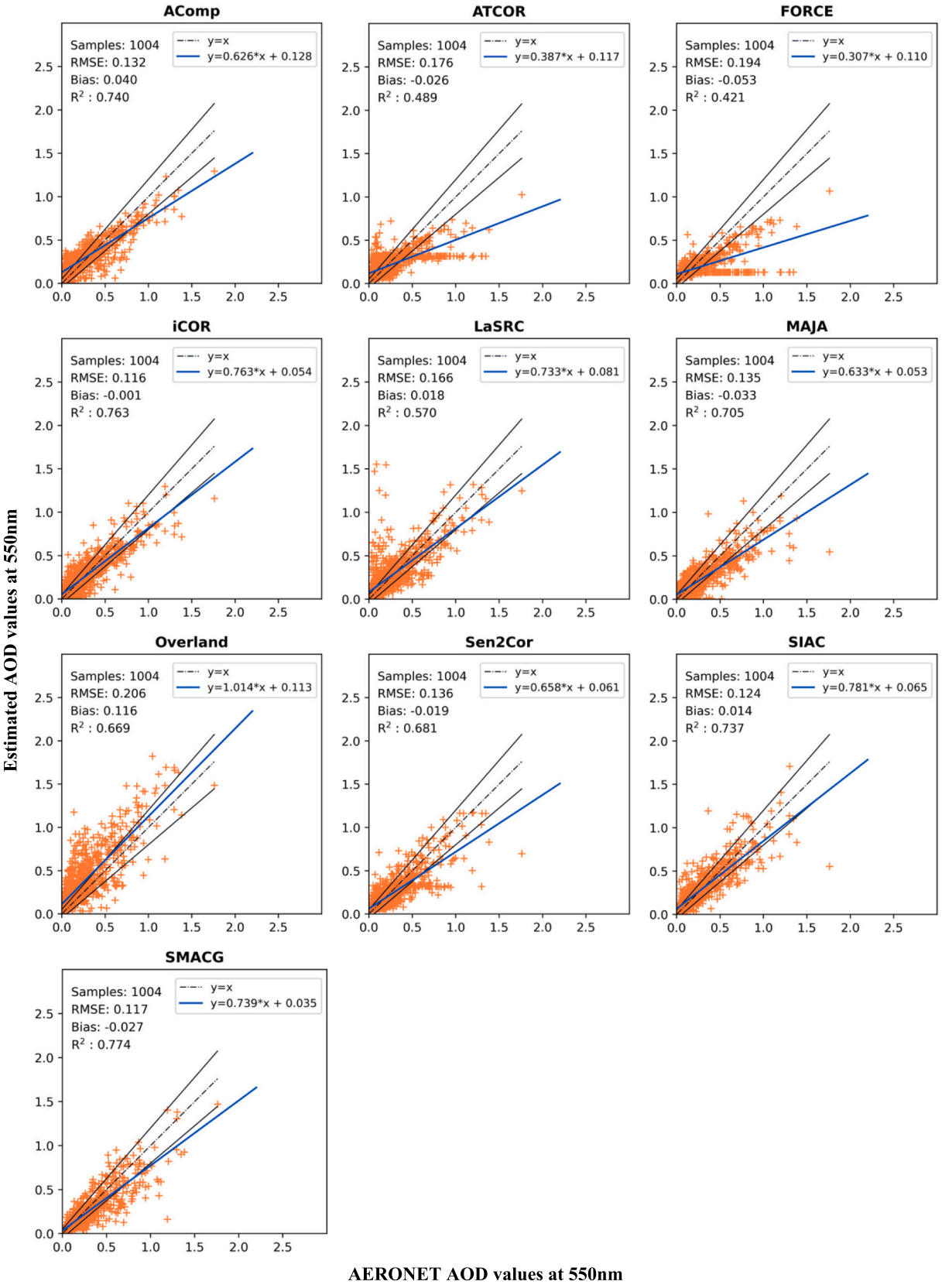


Fig. 2. Scatterplots of AOD estimates versus AERONET measurements. The 1:1 agreement line is indicated with a dashed line, while the black solid lines represent the uncertainty limits specifications, specs = $0.15*AOD_{550ref} + 0.05$. The blue solid line represents the least-squares regression line for the estimated datasets and AERONET reference measurements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
RMSE, Bias and R^2 for the AOD estimates of all the processors. The best performance per metric is highlighted in bold.

	AComp	ATCOR	FORCE	iCOR	LaSRC	MAJA	Overland	Sen2Cor	SIAC	SMACG
RMSE	0.13	0.176	0.194	**0.116**	0.166	0.135	0.206	0.136	0.124	0.117
Bias	0.04	−0.026	−0.053	**−0.001**	0.018	−0.033	0.116	−0.019	0.014	−0.027
R^2	0.74	0.489	0.421	0.763	0.57	0.705	0.669	0.681	0.737	**0.774**

bright landscapes, where scattering effects have a smaller relative effect on the final SR. When AOD fallback cases were excluded from FORCE's performance analysis, the correlation R^2 and RMSE were improved with values of 0.793 and 0.116 respectively. In general, all the processors, but Overland, underestimated AOD. Overland most likely overestimated AOD due to its approach to make the difference between aerosols and cloud veils (these with much higher spatial variability) and estimate both of them with distinct models. However, for the exercise purposes, aggregating the two estimates to a single AOD value resulted in AOD overestimation and spurious inclusion of cloud veils over sandy and/or bright surfaces, so over AERONET sites located in deserts and/or dense urban areas.

The outcomes of the APU analysis for the AOD validation, as described in Section 2.3.1, were compared to MODIS specifications (Remer et al., 2009). Results in Fig. 3 confirm the consistent performance of most of the processors in retrieving light to medium aerosol loading (AOD < 0.2), with uncertainties within or very close to the suggested specifications, i.e., shaded area below the black solid line. This range also contains >60% of the collocated AERONET matches for all the processors. Concerning high AOD values (AOD > 0.6), the retrievals were mostly beyond the uncertainty specifications. Nevertheless, the atmospheric correction is unlikely to be correct, even with

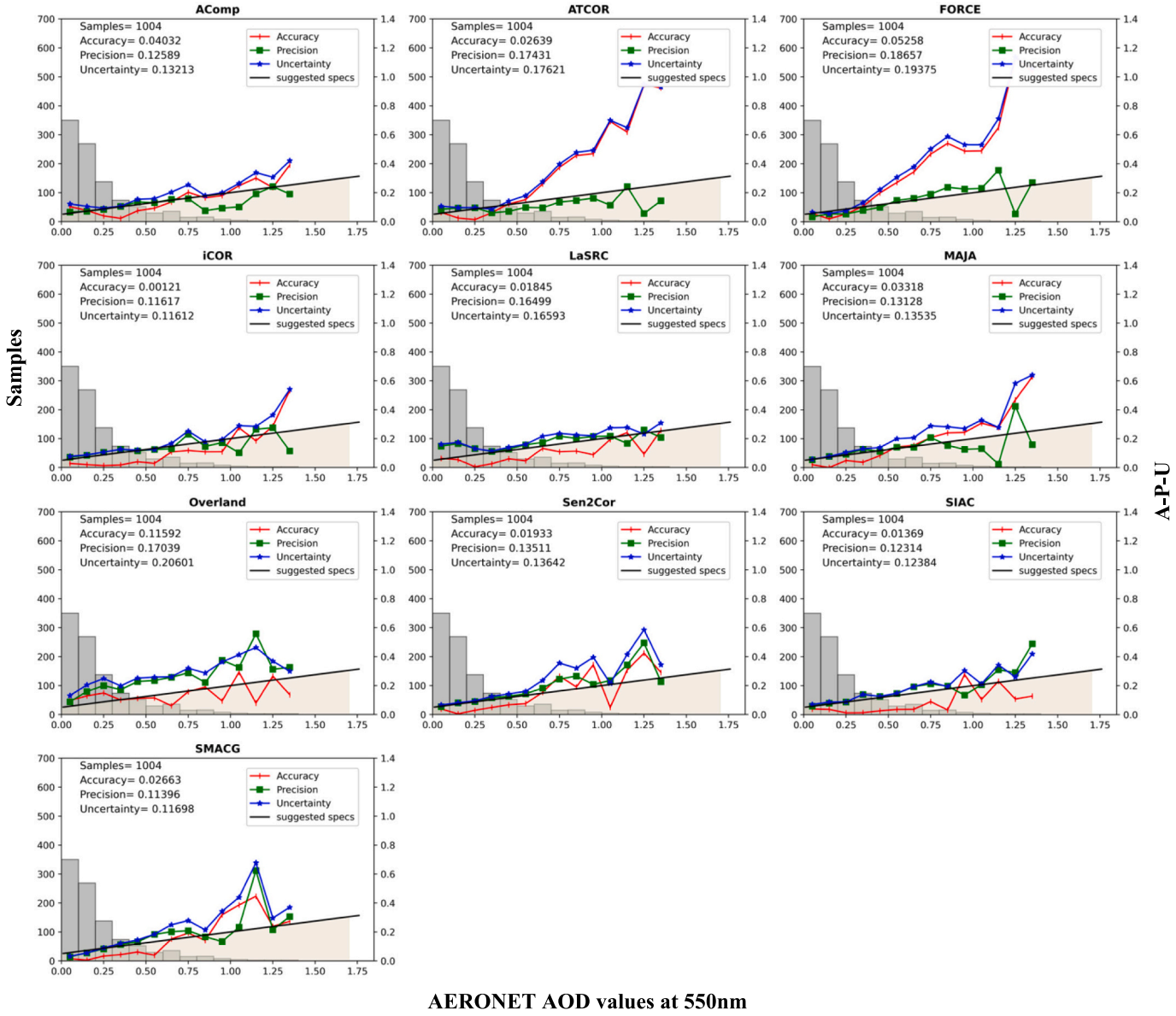


Fig. 3. Accuracy (Red), Precision (Green) and Uncertainty (Blue) plots of AOD estimates, per 0.1 AOD value bin, versus AERONET measurements. The suggested specification line is displayed with black and corresponds to the empirical uncertainty of MODIS land AOD retrievals (Remer et al., 2009), i.e., specs = $0.15*AOD_{550ref} + 0.05$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accurate aerosol retrievals, when high AOD events occur (Vermote and Kotchenova, 2008). Despite the general trend, Overland estimates AOD with uncertainties outside the suggested specifications in the whole range of values. Overall, iCOR and SMACG retrieve AOD with the lowest uncertainties (<0.12), while SIAC, AComp, MAJA and Sen2Cor follow with similar performances and uncertainty values <0.14 (See Table 3).

3.1.2. Water vapour validation over AERONET sites

The quality assessment of water vapour was performed by comparing the averaged retrievals over 9 km × 9 km study areas with the corresponding AERONET measurements. AComp, Planet_SR and SMACG did not provide WV, so they were excluded from this analysis. In Fig. 4 the black dashed line represents the 1:1 line, the ideal case of the perfect match between WV estimates and reference. The black solid lines indicate the uncertainty limits specifications, specs = $0.1*WV_{ref} + 0.2$ g/cm^2, according to the empirical uncertainty of Sentinel-2 WV retrievals as suggested by Sen2Cor developers (Pflug et al., 2020). The orange solid line represents the least-squares regression line for the two compared datasets.

Overall, WV retrieval is very accurate with strong correlations with the reference data, mostly >0.9, and with RMSE mainly <0.25 (Table 4). >90% of the results fall within the specifications for all the processors except LaSRC (89.5%), SIAC (85%) and FORCE (70%) whose results are more dispersed, mainly for WV values over 3 g/cm^2.

3.1.3. Surface reflectance validation using AERONET-derived reference

The APU analysis based on the AERONET-derived SR (Section 2.3.4) was implemented per site and using only LaSRC's or combination of LaSRC's and each processor's quality layer. For reasons of space, the cases presented here are limited to the ones using the common quality layer and including all the 79 common study sites. Similar to AOD and WV results, the complete analysis can be found in CEOS Cal/Val portal. In Fig. 5 APU bar plots highlight the performances per Sentinel-2 band and processor. Band 10 is omitted, as it is a cirrus band and does not contain surface information. MAJA and Sen2Cor are also excluded from the analysis, after their developers' team request, due to the differences in the SR computation between the reference and MAJA, i.e., RTM, adjacency effects correction, etc. iCOR and SMACG were not included in this analysis either, because of inconsistencies in the file types, filenames and/or folder structure that did not allow the processing of the submitted results. The ATCOR outcomes for B02, B03, B04 and B08 are also missing, as they had not been provided when this analysis was performed.

FORCE and Planet retrieved surface reflectance within the specifications across all bands, apart from B01. Overall, the uncertainty for B01 is relatively larger due to the strong impact of aerosols that challenge an accurate atmospheric correction at this wavelength. Considering all the processors, accuracy mainly remains within the specifications showing that the SR retrievals are not biased for most of the bands. Moreover, the APU results are improving from visible (V: B01, B02, B03, B04) and red-edge/near infra-red (RE/NIR: B05, B06, B07, B08, B8a) to the short wave infra-red (SWIR: B11, B12), since the scattering by aerosols and gaseous molecules decreases as the wavelength increases. LaSRC and SIAC exhibit similar performances with FORCE and Planet, but besides B01, the uncertainty slightly exceeds the specifications also for B02. AComp's surface reflectance retrievals are getting closer to specifications from VNIR to SWIR spectral range, apart from B05, B08 and B12 for which uncertainty is rather beyond the suggested requirements. ATCOR took part in the analysis with smaller number of bands, but it succeeded in estimating accurate surface reflectance for the participating bands except for B01 and B8a. The estimates of Overland are beyond the uncertainty specifications across all bands.

In this analysis, the processors' performances could likely be biased by the differences between the AC approaches and the approach followed to compute the simulated SR reference, i.e., RTM and adjacency effects correction (Table 1). Nevertheless, regarding the differences amongst various RTMs, several studies reported that the relative differences between 6SV and MODTRAN are varying in the range of around 4–11% for the visible part of the spectrum, while remaining rather constant between 6SV and libRadtran, with errors around 3%–4%. In general, the shorter the wavelength is, the greater the discrepancies are amongst the various RTMs (Kotchenova et al., 2008; Callieco and Dell'Acqua, 2011; Vicent et al., 2020). Concerning the adjacency effects, in the previous ACIX implementation, a twofold analysis was performed with and without including their correction and no big discrepancies were observed, either due to land cover homogeneity or the insignificant number of pixels affected at this certain 9 km × 9 km image subset.

3.1.4. Surface reflectance validation with RadCalNet measurements

The quality of the estimated Sentinel-2 surface reflectance was quantitatively assessed based on the available RadCalNet measurements. 44 and 40 Sentinel-2 scenes were available with valid RadCalNet observations over La Crau and Gobabeb stations respectively. An overview of the processors' performance regarding the relative differences in all the bands is presented in Fig. 6. Considering that the expected Sentinel-2 absolute calibration accuracy is ±3%, propagating this figure into surface reflectance, it means up to 6% uncertainty in the blue and 3% in the SWIR (after subtracting the atmospheric path reflectance). If we also add quadratically the 3% calibration uncertainties of ROSAS stations, we could get biases up to between 4.5% to 6.5% depending on the spectral bands, with higher values in the short wavelengths.

Generally, the processors perform well given the aforementioned calibration uncertainties of Sentinel-2 sensors and RadCalNet instruments, with the exception of ATCOR on La Crau, and Overland on Gobabeb. Most of the processors underestimated the surface reflectance measured at La Crau and Gobabeb for all the bands apart from B11. ATCOR and LaSRC follow relatively similar trends with the rest, but with a limited performance in the blue bands (B01-B02). Overland is the only processor to overestimate surface reflectances in all bands for both sites, and although it has comparatively low biases in La Crau, it does not succeed in estimating accurately the brighter reflectances in Gobabeb.

The consistent performance of the processors is also demonstrated by the low values of the standard deviation in the relative differences between estimated and on-site surface reflectance (Fig. 6). A small variation in the SRs is expected over time in both RadCalNet sites, which are originally calibration sites with little seasonal changes. Therefore, the low values of standard deviation indeed confirm the smooth time series and the ability of a processor to correct for the variations due to atmospheric conditions. With the exception of ATCOR, iCOR and Overland, all the processors have very similar standard deviation values.

Table 3
The overall averaged Accuracy (A), Precision (P) and Uncertainty (U) of AOD estimates versus the AERONET measurements. The last row refers to the percent of AOD estimates with Uncertainty scores within the suggested specifications. The best performance per metric is highlighted in bold.

	AComp	ATCOR	FORCE	iCOR	LaSRC	MAJA	Overland	Sen2Cor	SIAC	SMACG
A	0.040	0.026	0.053	**0.001**	0.018	0.033	0.116	0.019	0.014	0.027
P	0.126	0.174	0.187	0.116	0.165	0.131	0.170	0.135	0.123	**0.114**
U	0.132	0.176	0.194	0.116	0.166	0.135	0.206	0.136	0.124	**0.117**
U in the specs (%)	41.73	58.27	59.36	62.45	64.54	63.84	39.94	65.84	69.02	**78.39**

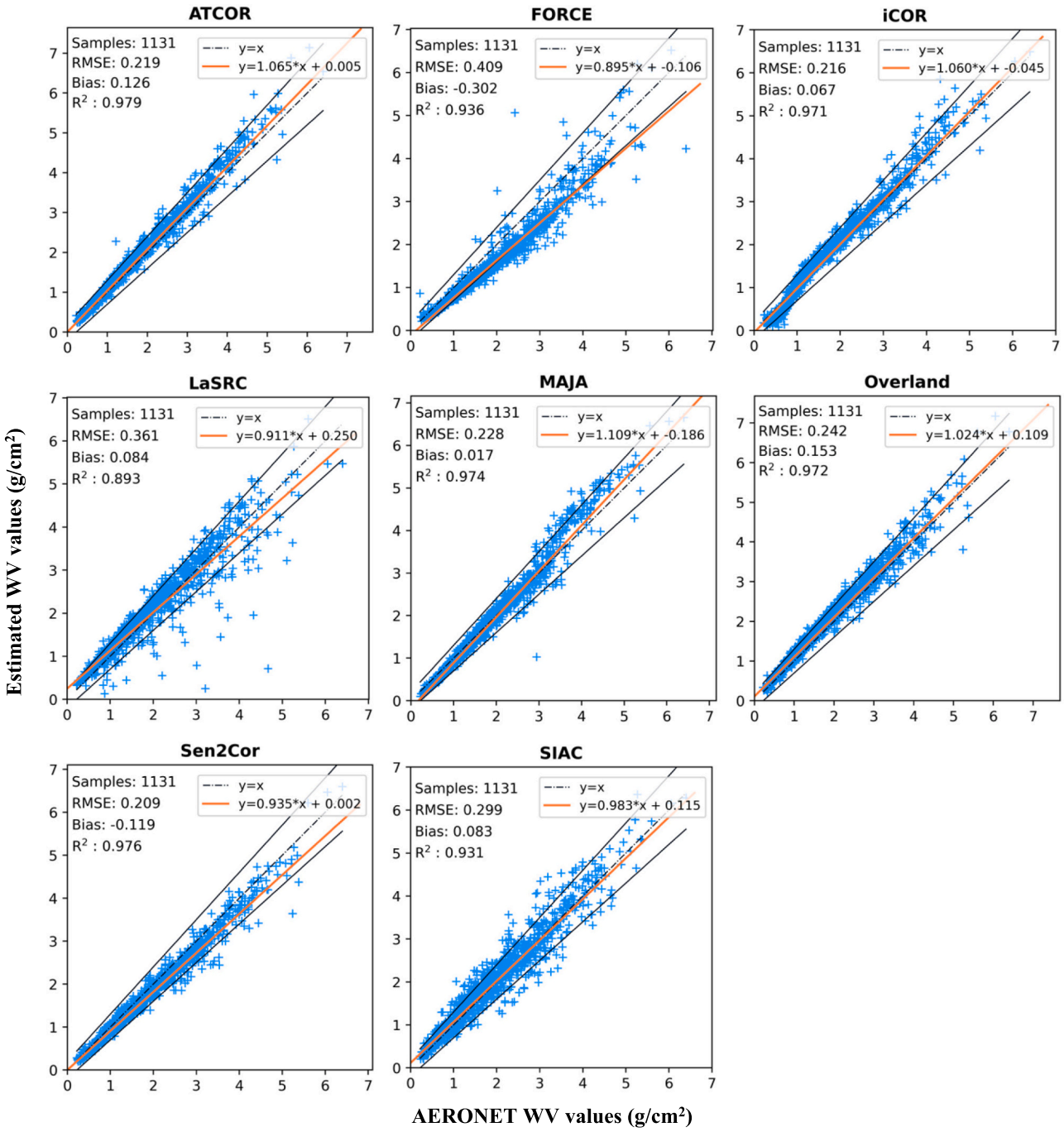


Fig. 4. Scatterplots of WV retrieval over WV reference from AERONET. The 1:1 agreement line is indicated with a dashed line, while the black solid lines represent the uncertainty limits specifications, specs = 0.1*WVref +0.2. The orange solid line is the least-squares regression line for the estimated datasets and AERONET reference measurements.

Table 4
RMSE, Bias and R^2 for the WV estimates of all the processors. The best performance per metric is highlighted in bold.

	ATCOR	FORCE	iCOR	LaSRC	MAJA	Overland	Sen2Cor	SIAC
RMSE	0.219	0.409	0.216	0.361	0.228	0.242	**0.209**	0.299
Bias	0.126	−0.302	0.067	0.084	**0.017**	0.153	−0.119	0.083
R^2	**0.979**	0.936	0.971	0.893	0.974	0.972	0.976	0.931

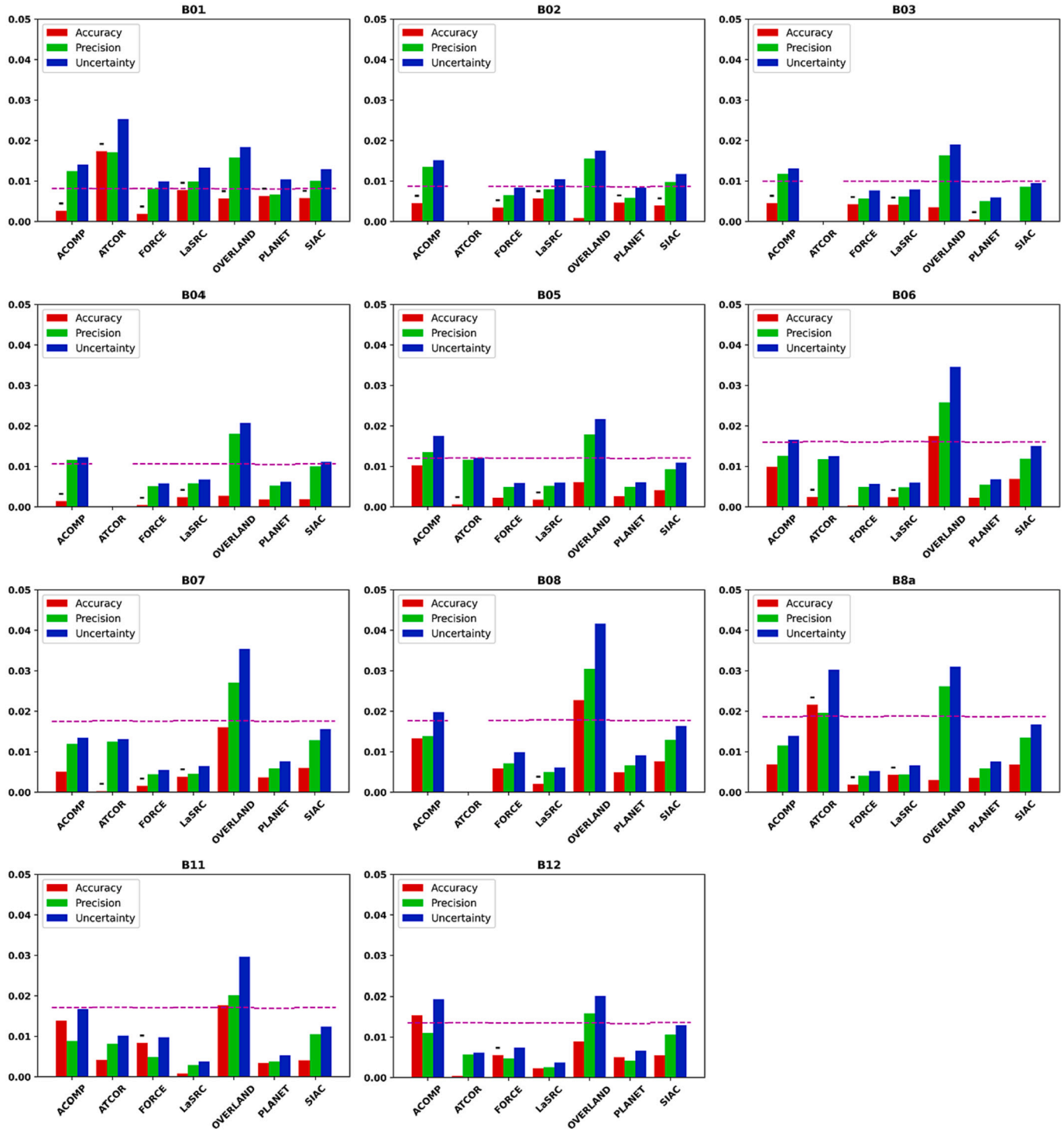


Fig. 5. Accuracy (Red), Precision (Green), Uncertainty (Blue) surface reflectance analysis results over all ACIX-II Land sites. The specification ($0.05\rho + 0.005$) is displayed with the magenta dashed line and is calculated using the average reference surface reflectance for each processor and band. The minus sign (−) over the bars indicates the negative values of the metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.5. Noise estimation over short-term time series

The noise criterion described in Section 2.3.4 was computed over the sites for which at least 20 acquisitions sensed by the same Sentinel-2A/-2B orbit were available. Having the same relative orbit, hence, the same viewing geometry, is essential for this atmospheric correction validation criterion, because noise in short-term time series can also be due to different observation and illumination geometries. The BRDF effect correction, appropriate to eliminate these differences, was requested to be omitted when it was part of the nominal AC processing chain for results' consistency purposes. From the total of 79 common sites, 9 sites fulfilled the criterion of minimum 20 scenes, which were split in two broad categories: arid/urban and vegetated sites.

The noise estimations for the 10 m Sentinel-2 bands are presented in Fig. 7 per processor and land cover type, i.e., arid/urban and vegetated.

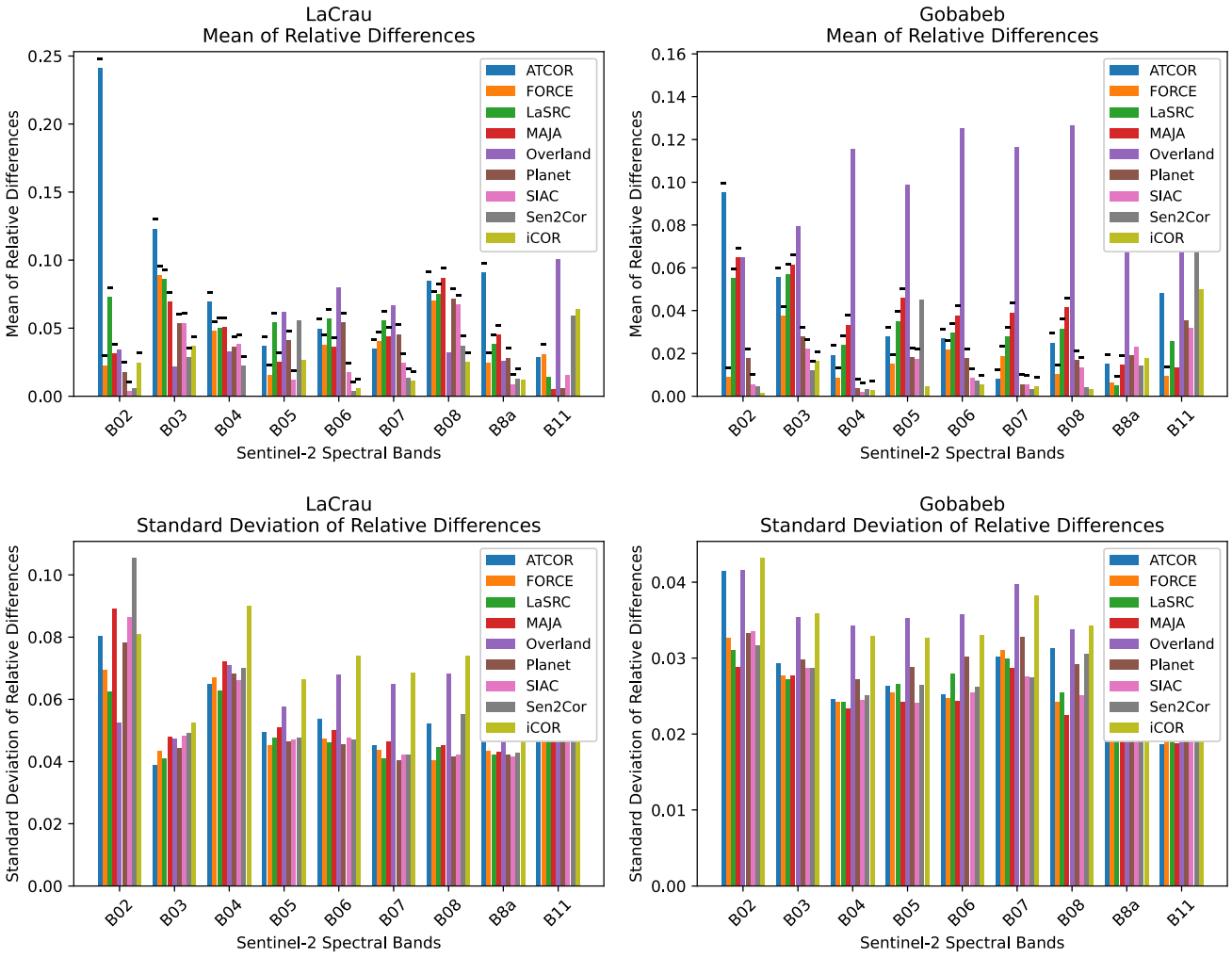


Fig. 6. Mean and standard deviation of the relative differences between reference (RadCalNet) and estimated SRs per band and processor for LaCrau (France) and Gobabeb (Namibia) site. The minus sign (−) over the bars refers to the negative values of the bias.

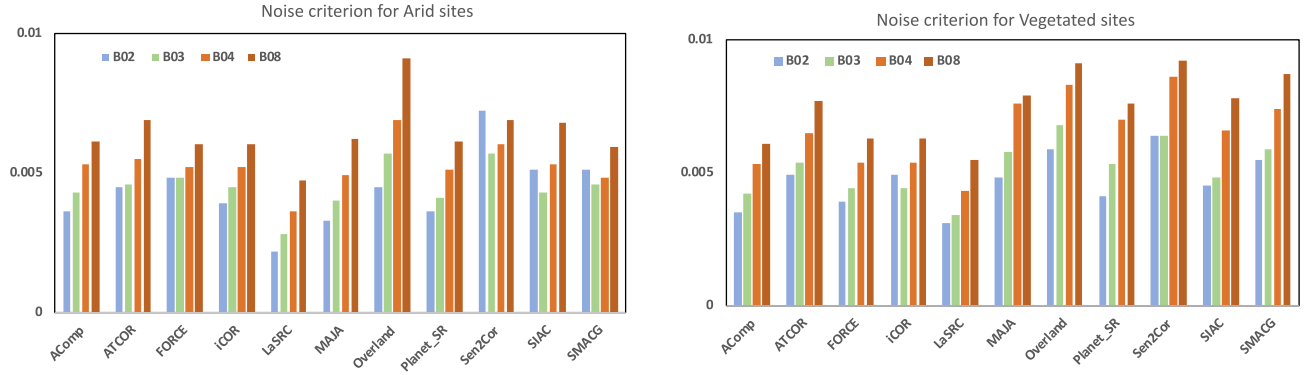


Fig. 7. Noise criterion over short-term surface reflectance time series for 4 arid/urban sites, i.e., Burjassot, Mezaira, Mongu_Inn, Tunis_Carthage and 5 vegetated sites, i.e., Evora, Granada, Kanpur, Lumbini, Murcia. The results are associated with the 10 m Sentinel-2 bands and presented by processor.

Overall, the processors produce surface reflectances with lower variability over arid than over vegetated sites, due to the seasonal changes of vegetation that may contribute to surface reflectance variance. LaSRC has achieved the lowest noise values in the SR time series, but without great discrepancies from most of the processors. This may have occurred because the metric was calculated over a rather small study area (9 km × 9 km) involving 81 samples and utilizing common quality masks to identify the quality approved pixels.

3.2. Landsat 8 OLI

Nine developer teams participated in the exercise and implemented their AC processors on about 1250 Landsat 8 scenes acquired over 110 AERONET sites around the world (Fig. 1). Eight of the processors were also part of the Sentinel-2 processing, while EMBAC is applicable only to Landsat data. Towards a clear performance comparison, only the common sites and dates amongst all processors were involved, similar to

Sentinel-2 data analysis, and the sites were reduced to 62 in total. For the complete analysis of the entire dataset as submitted by each processor, the results can be found on ACIX-II Land web subpage (https://calvalportal.ceos.org/acix-ii-results).

3.2.1. AOD validation with AERONET measurements

The quality assessment of the AOD retrieval was performed by comparing to the corresponding AERONET measurements. The scatterplots in Fig. 8 show that the results are in better agreement with the reference in low aerosol values, while they are more scattered for values >0.3. Overall, iCOR and LaSRC outperform with the lowest RMSEs and highest correlations while SIAC and AComp follow with slightly inferior performance. ATCOR and FORCE underestimate AOD mainly because of the fallback solution applied when not enough scene pixels were detected as dark reference pixels (DDV and dark water) (Table 5). Overall, SIAC, iCOR, ATCOR and LaSRC estimated >90% of the AOD events inside the suggested uncertainty specifications whereas FORCE (89%) and AComp (87%) follow with very similar results (Table 6).

The good performance at low to medium (<0.2) aerosol values is also demonstrated in the APU analysis in Fig. 9. All the processors' retrievals have uncertainty inside or close to suggested specifications, apart from Overland and EMBAC that overestimate AOD through the whole range.

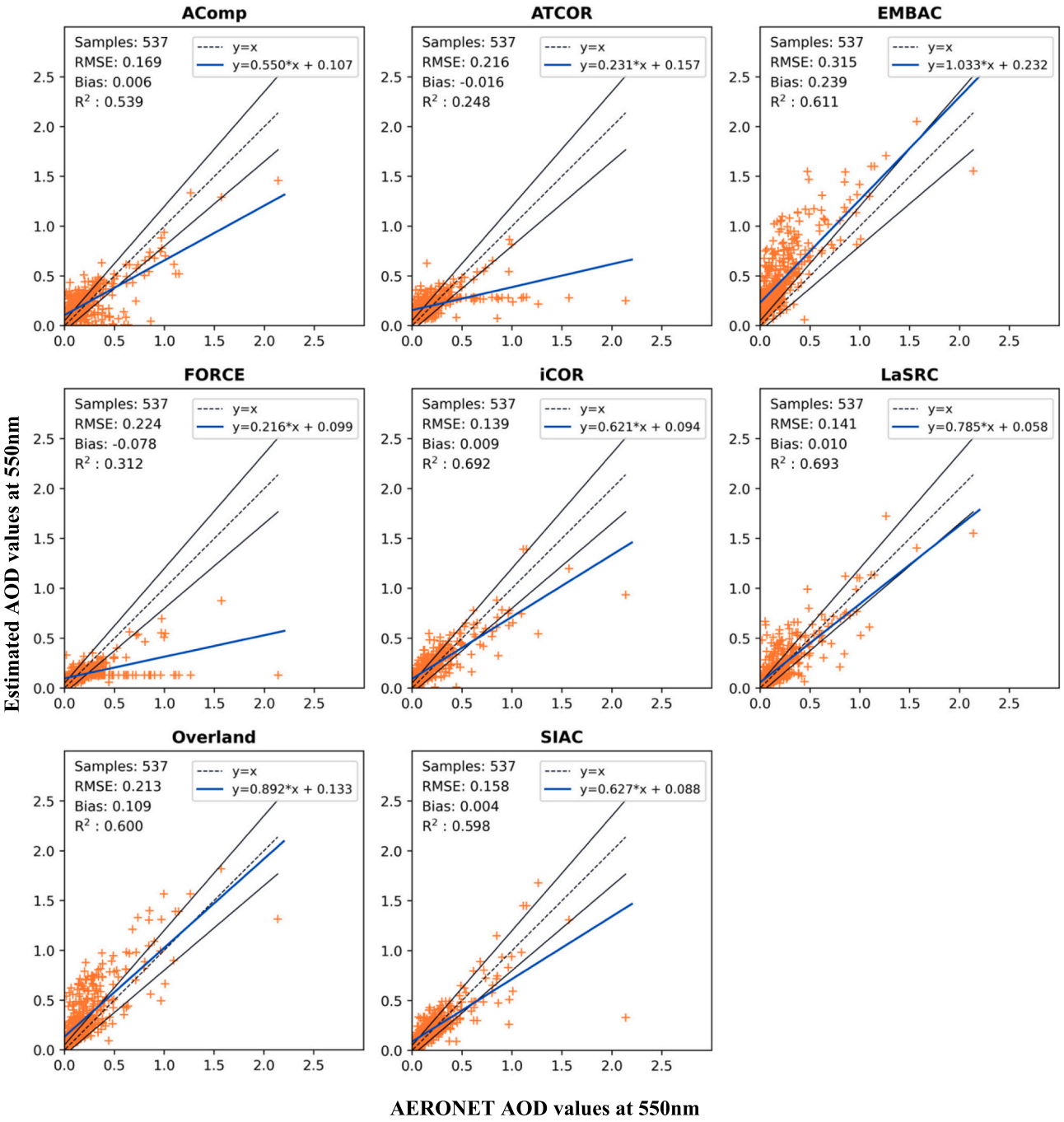


Fig. 8. Scatterplots of AOD estimates based on Landsat 8 data versus AERONET measurements. The 1:1 agreement line is indicated with the black dashed line, while the black solid lines represent the uncertainty specifications, specs = $0.15*AOD_{550ref} + 0.05$. The solid blue line is the least-squares regression line for the estimated datasets and AERONET reference measurements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
RMSE, Bias and R^2 for the AOD estimates of all the processors. The best performance per metric is highlighted in bold.

	AComp	ATCOR	EMBAC	FORCE	iCOR	LaSRC	Overland	SIAC
RMSE	0.169	0.216	0.315	0.224	**0.139**	0.141	0.213	0.158
Bias	0.006	−0.016	0.239	−0.078	0.009	0.01	0.109	**0.004**
R^2	0.539	0.248	0.611	0.312	0.692	**0.693**	0.600	0.598

Table 6
The overall averaged Accuracy (A), Precision (P) and Uncertainty (U) of AOD estimates versus the AERONET measurements. The last row refers to the percent of AOD estimates with Uncertainty scores within the suggested specifications. The best performance per metric is highlighted in bold.

	AComp	ATCOR	EMBAC	FORCE	iCOR	LaSRC	Overland	SIAC
A	0.006	0.016	0.239	0.078	0.009	0.010	0.109	**0.004**
P	0.169	0.216	0.205	0.211	**0.139**	0.141	0.183	0.158
U	0.169	0.216	0.315	0.224	**0.139**	0.141	0.213	0.158
U in the specs (%)	87.15	90.69	57.54	89.20	94.03	90.32	78.03	**95.16**

3.2.2. *Surface reflectance validation using AERONET-derived reference*

The APU analysis over the 62 common sites with valid results for all the processors is presented in Fig. 11. The SR reference and the metrics were computed according to the description given in Section 2.3.2. Each subplot corresponds to the results per Landsat 8 band (Fig. 10).

The overall analysis demonstrates that the APU results are improving from visible (Bands 1–4) and NIR (Band 5) to SWIR (Band 6, Band 7) remaining mainly within the suggested specifications ($0.05\rho + 0.005$). Apart from the Band 1 (Ultra Blue), where the aerosol scattering is stronger, FORCE and Planet succeeded in estimating the SR with uncertainties inside the specifications across all bands. Similar performance was found for SIAC, LaSRC, iCOR and EMBAC with the quality of the results to be deteriorated only for Band 1 and Band 2. AComp's performance follows the general trend and the uncertainty is improved from VNIR to SWIR. In terms of Accuracy, all processors' results remain within the specifications producing unbiased SRs.

3.2.3. *Surface reflectance validation with RadCalNet measurements*

The quality assessment of the computed SRs was performed based on the RadCalNet measurements over La Crau and Gobabeb stations. In total, 5 and 14 Landsat scenes matched the valid ground data of each site respectively. ATCOR is excluded from the analysis over Gobabeb, due to

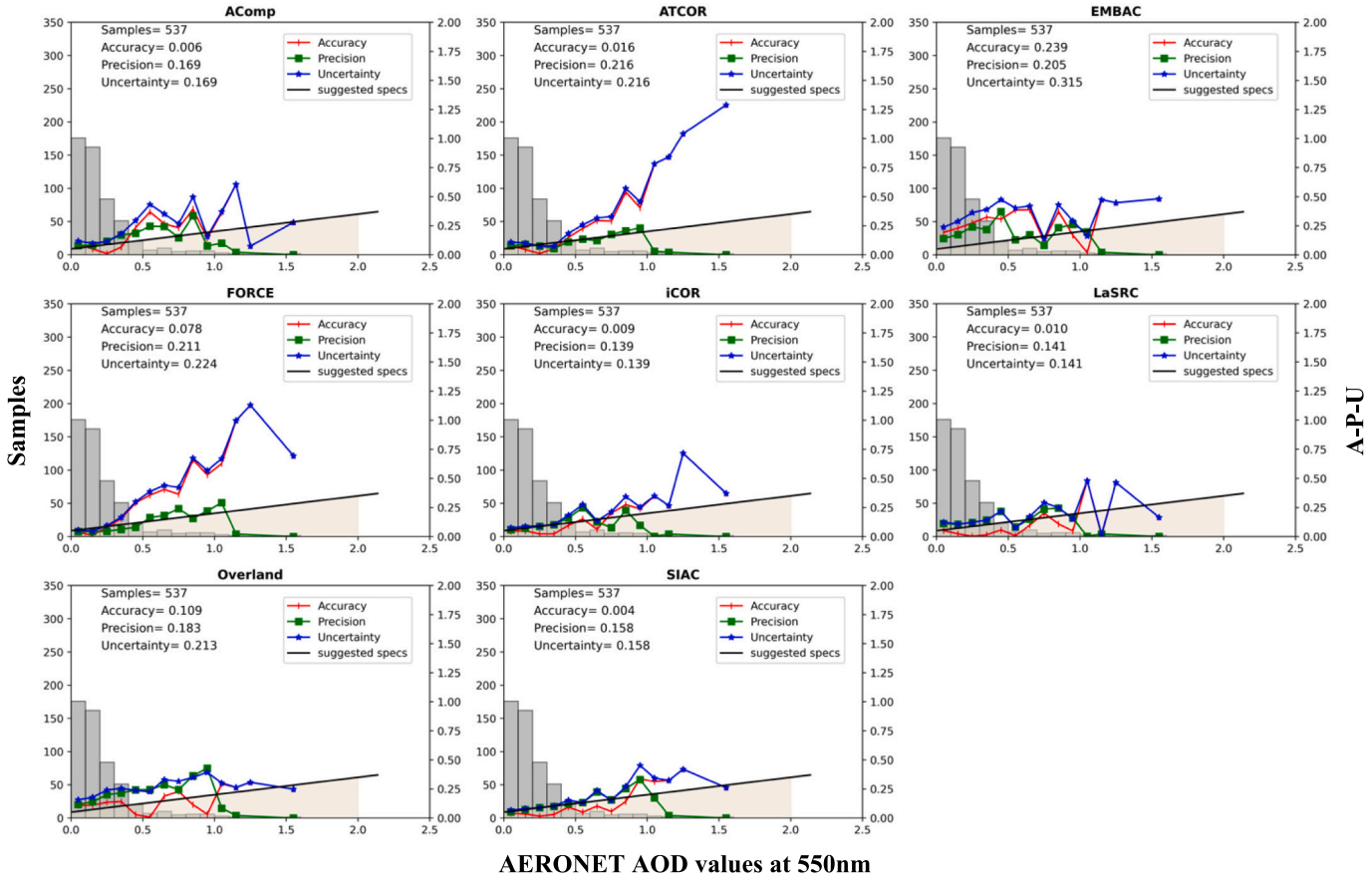


Fig. 9. Accuracy (Red), Precision (Green) and Uncertainty (Blue) plots of AOD estimates, per 0.1 AOD value bin, versus AERONET measurements. The suggested specification line is displayed with black and corresponds to the empirical uncertainty of MODIS land AOD retrievals (Remer et al., 2009), i.e., specs = $0.05 + 0.15\tau$, for the corresponding AOD value τ. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

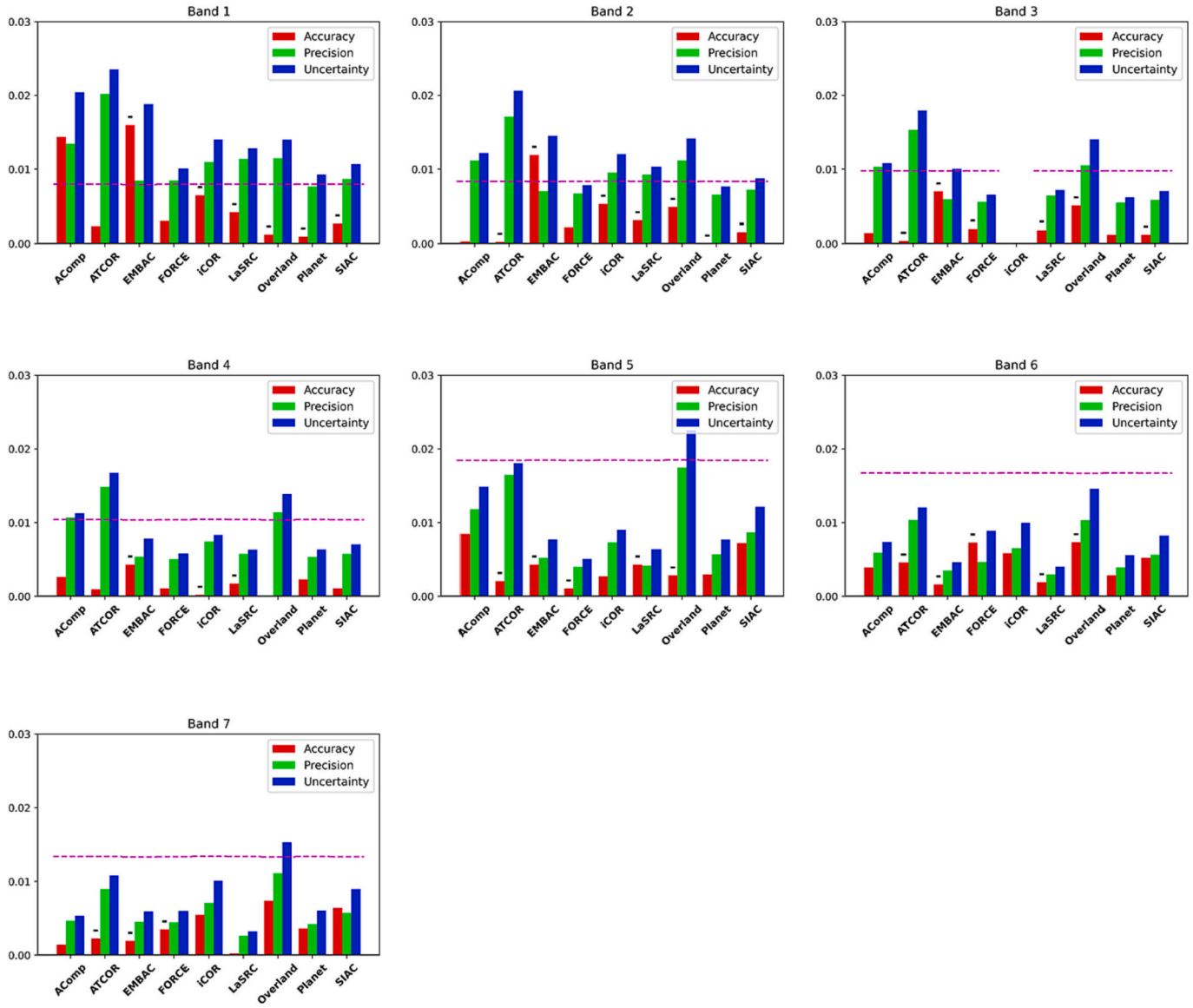


Fig. 10. Accuracy (Red), Precision (Green), Uncertainty (Blue) surface reflectance analysis results over all 110 AERONET sites. The specification ($0.05\rho + 0.005$) is displayed with the magenta dashed line and is calculated using the average reference surface reflectance for each processor and band. The minus sign (−) over the bars refers to the negative values of the metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

incorrectly projected outcomes. iCOR did not provide Band 3, so it is also missing from the analysis.

Although the statistical sample is not adequately large, plotting the cross-comparison results in Fig. 11 exhibits the diverse performance of the processors for the six Landsat 8 bands. The minus sign (−) over the bars refers to the negative values of the Relative bias. Overall, the processors underestimate SRs for both sites. Nevertheless, large discrepancies were observed for EMBAC and LaSRC when producing SRs in the VNIR bands. Moreover, the processors produce consistent SRs with low standard deviation values for all the bands, that is mostly <0.01 in Gobabeb and <0.015 in La Crau.

4. Discussion

The summary of processors' performance in terms of SR uncertainty *U*, as computed with the AERONET-derived SR reference (Section 2.3.2), is presented in the annotated heatmap of Fig. 12. The results here concern only the seven processors implemented on both Sentinel-2 and Landsat 8 data, i.e., AComp, ATCOR, FORCE, LaSRC, Overland, Planet_SR and SIAC, and only for the similar bands of MSI and OLI. Overall, the performances are improved for SWIR and NIR followed by the visible bands, while for Landsat 8 relatively lower uncertainties are observed than for Sentinel-2. LaSRC, FORCE and Planet_SR outperform in this assessment with the lowest uncertainty values overall. It is noted that FORCE did perform less accurately in the AOD retrieval due mainly to the AOD fallback values, when DDV pixels could not be detected. However, as the same RTM and internal model constants are used in the inversion process to derive AOD, and in the forward calculation to obtain SR, this inaccuracy self-corrects and produces a reliable estimate of SR nonetheless (Frantz et al., 2016). In addition, FORCE's performance was much improved, when the AOD fallback cases were excluded, achieving the correlation and RMSE scores of the best performers. SIAC, AComp and ATCOR yield SR with relatively low uncertainties between 0.005 and 0.015 for the red, NIR and SWIR bands. The variances amongst the approaches, i.e., RTM and adjacency effects correction (Table 1), for implementing AC and computing the reference

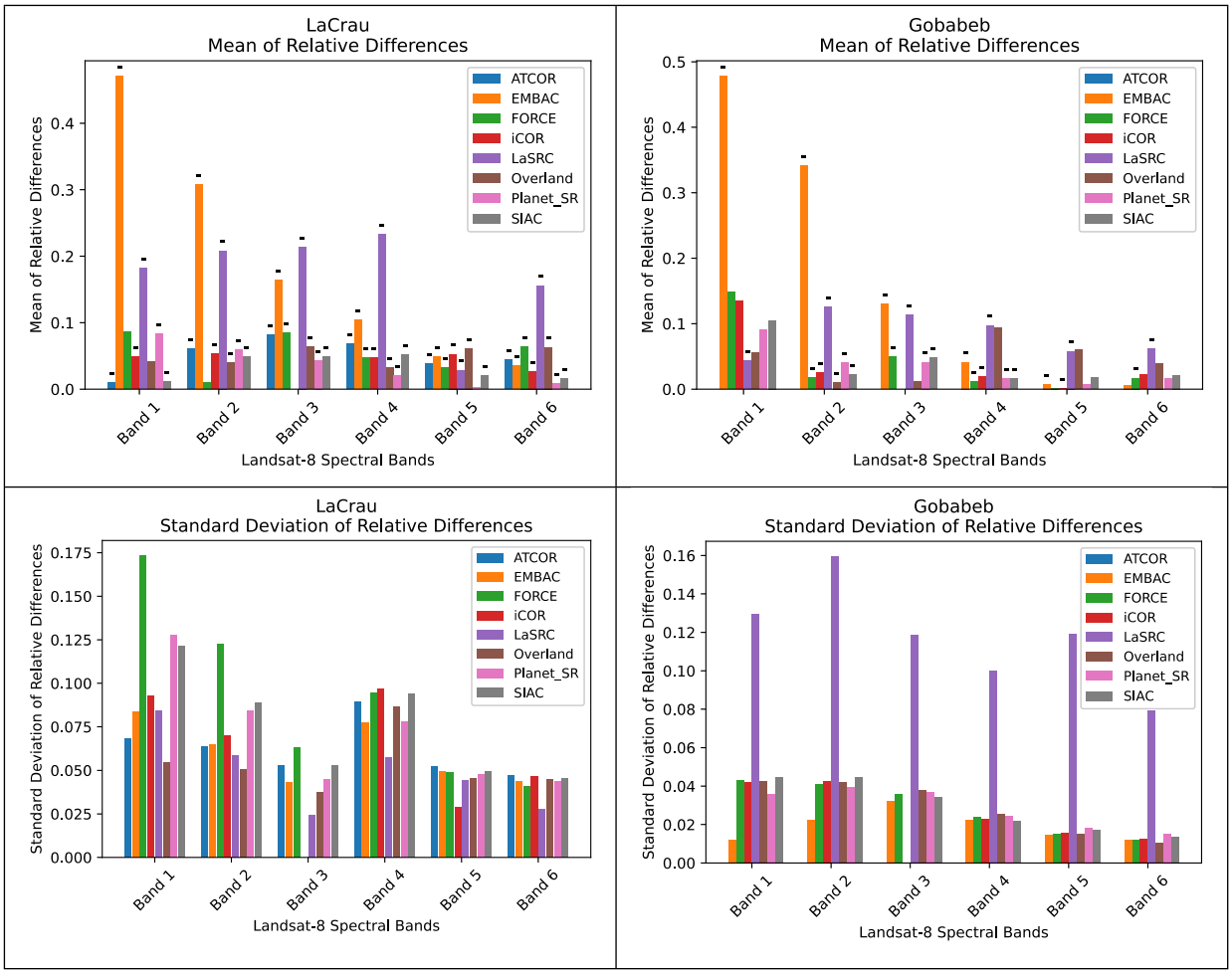


Fig. 11. Mean and standard deviation of the relative differences between the reference (RadCalNet) and estimated SRs per band and processor for LaCrau (France) and Gobabeb (Namibia) sites. The minus sign (−) over the bars refers to the negative values of the metric.

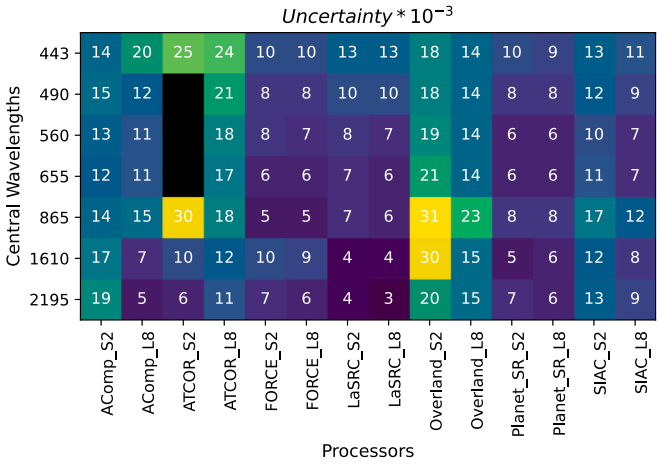


Fig. 12. The estimated surface reflectance uncertainty (U*10^{-3}) for Sentinel-2 MSI (S2) and Landsat 8 OLI (L8) common wavelengths as calculated using AERONET-derived reference reflectances (Section 2.3.2).

data may affect the results in this analysis. However, several RTM comparison studies have demonstrated that the differences between 6SV and MODTRAN do not exceed 11% for the visible part of the spectrum, from 0.4 to 0.6 μm, while the differences in general decrease as the wavelength increases (Kotchenova et al., 2008; Callieco and Dell'Acqua, 2011; Vicent et al., 2020). Nevertheless, it is recommended to involve different RTMs in the future SR simulations and/or to intercompare the various models and examine their discrepancies and uncertainties. To this end, ESA and the European Commission are currently organizing a benchmark exercise, the RAdiation transfer Model Intercomparison for Atmosphere (RAMI4ATM), for the inter-comparison of coupled surface–atmosphere RTMs over a variety of atmospheric scenarios.

Regarding the validation based on the RadCalNet ground-based measurements, the processors' performance is summarized in Fig. 13, for both sensors and sites. FORCE has the best agreement with the reference measurements, as it yields the lowest differences of standard deviations for almost all Sentinel-2 bands. Planet_SR and SIAC are the ones following with comparable performance over La Crau, while MAJA performs similarly well with FORCE over Gobabeb. In comparison to Sentinel-2, significantly fewer ground measurements were available for Landsat 8 results validation, with LaSRC having the best match with the reference over La Crau, contrary to Gobabeb where it mainly underestimated SRs. EMBAC estimated the most accurate SRs over Gobabeb with standard deviations <0.01 for all the related bands.

Although small discrepancies are observed amongst the processors, given that our findings are based on a limited number of samples, the results should be treated with considerable caution. In addition, the network is designed for the radiometric calibration of satellite data, so the sites are mainly spatially homogeneous and stable with limited atmospheric and cloud variations. La Crau is more suitable for SR validation than Gobabeb due to the low vegetation in the area, but it still represents a rather ideal case scenario. However, a new RadCalNet site is currently being established by CNES in Lamasquère (FR-Lam) over an

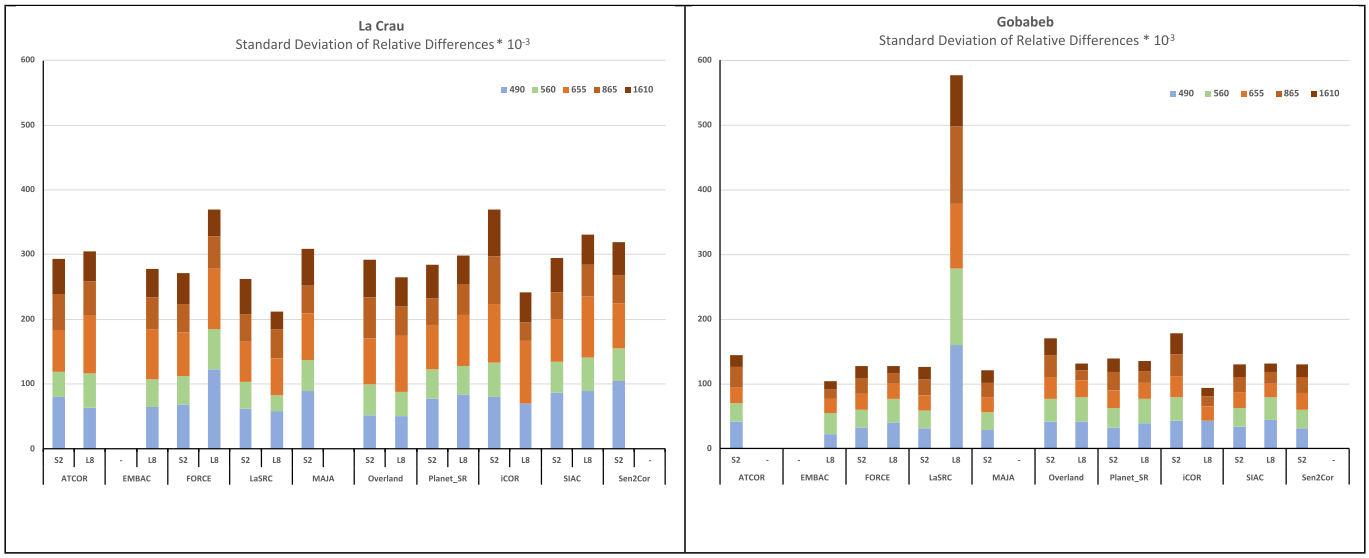


Fig. 13. The standard deviation of the relative differences between the reference (RadCalNet) and estimated SRs per band and processor for LaCrau (France) and Gobabeb (Namibia) sites.

agricultural area (Lamasquere project details). The observations in this case could support SR validation studies, as the site is spatially heterogeneous, covered by different crops and located next to woods, so representative for considering adjacency effects.

Moreover, the short-term surface reflectance analysis (noise metric) showed consistent results across 4 arid/urban sites and 5 vegetated ones. LaSRC provided the smoothest, without big variations, SRs, with similar results coming from the rest of the processors. The greatest differences amongst them were 0.004 for B02 (blue band) in the arid/urban cases and 0.005 for B04 (red band) in the vegetated ones. It is possible that these relatively uniform outcomes were caused by the fact that only 81 pixels were involved in this assessment and were also the commonly quality approved ones by all processors.

Overall, the proper validation of SR products and the assessment of AC processors in our study was limited by the absence of a global network of ground-based SR observations over land. Continuous measurements over diverse surface and atmospheric conditions are needed to provide the reference for the robust validation of AC performance. Hyper-spectral and multi-angular observations are also essential to fit all optical bands on all satellite missions and to correct for the effects of the BRDF. Currently, there are some efforts for the establishment of such networks, e.g., HYPERNETS (Goyens et al., 2021), and to add new sites in RadCalNet, in order to provide reference observations for future SR validation activities (Niro et al., 2021).

5. Conclusion

The Atmospheric Correction Intercomparison eXercise (ACIX) is organised by ESA and NASA in the frame of CEOS WGCV. In this second implementation over land 12 atmospheric correction processors participated for Sentinel-2 or Landsat 8 data. The AOD and WV estimates were assessed over a wide variety of AERONET sites distributed globally, providing a good indication of the retrieval capability of the processors. Most of them succeeded in providing AOD with high accuracy (U < 0.08) for aerosol loads lower than 0.2, and demonstrated very good performances for WV estimation with RMSE <0.25 g/cm^2. Regarding SR quality assessment, the analysis based on AERONET-derived reflectances demonstrated lower uncertainties for SWIR and NIR with increasing values at the short wavelengths. Although the computation of AERONET-derived SR is a valuable data source for assessing AC performance, it may introduce biases in the analysis when dealing with processors based on different approaches. Regarding the RTMs differences, a more thorough examination involving other models is suggested for similar future SR validation activities. Good performances with standard deviations of the relative differences mostly <0.05 were witnessed in the analysis based on the RadCalNet measurements. Nevertheless, the available observations in this case were limited over stable surface and atmospheric conditions. Networks of ground-based stations distributed globally and dedicated to SR validation are expected in the near future to provide valuable data for robust quality assessment of optical missions over land.

CRediT authorship contribution statement

Georgia Doxani: Conceptualization, Writing – original draft, Software, Formal analysis, Visualization, Validation. **Eric F. Vermote:** Conceptualization, Writing – review & editing, Methodology, Software, Formal analysis, Visualization, Validation. **Jean-Claude Roger:** Conceptualization, Writing – review & editing, Methodology, Software, Formal analysis. **Sergii Skakun:** Conceptualization, Writing – review & editing, Methodology, Software, Formal analysis. **Ferran Gascon:** Conceptualization, Writing – review & editing. **Alan Collison:** Conceptualization, Writing – review & editing, Methodology, Software. **Liesbeth De Keukelaere:** Conceptualization, Writing – review & editing, Methodology, Software. **Camille Desjardins:** Methodology, Software. **David Frantz:** Conceptualization, Writing – review & editing, Methodology, Software. **Olivier Hagolle:** Conceptualization, Writing – review & editing, Methodology, Software. **Minsu Kim:** Methodology, Software. **Jérôme Louis:** Methodology, Software. **Fabio Pacifici:** Conceptualization, Writing – review & editing, Methodology, Software. **Bringfried Pflug:** Conceptualization, Writing – review & editing, Methodology, Software. **Hervé Poilvé:** Conceptualization, Writing – review & editing, Methodology, Software. **Didier Ramon:** Methodology, Software. **Rudolf Richter:** Methodology, Software. **Feng Yin:** Conceptualization, Writing – review & editing, Methodology, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The input data involved in this study are open and freely available from ESA/Copernicus (Sentinel-2) and USGS (Landsat 8). The results can be accessed at the CEOS Cal/Val portal.

Acknowledgment

We thank all the PI(s) and Co-I(s) and their staff for establishing and maintaining the over than 120 AERONET sites used in this investigation. We acknowledge also CNES, ESA and NPL for maintaining the RadCalNet stations involved in the study and special thanks to Aimé Meygret and his team in CNES for processing the corresponding measurements. In addition, we thank Bryan Keary (University College Cork, Ireland) for proof-reading the paper. F.Y. (UCL and NERC) was supported for this work by the National Centre for Earth Observation (Grant agreement or award number: 525861). K.M. (KBR) was supported by USGS under the Contract 140G0121D0001.

References

Bouvet, M., Thome, K., Berthelot, B., Bialek, A., Czapla-Myers, J., Fox, N.P., Goryl, P., Henry, P., Ma, L., Marcq, S., Meygret, A., 2019. RadCalNet: A radiometric calibration network for Earth observing imagers operating in the visible to shortwave infrared spectral range. Remote Sens. 11 (20), 2401. https://doi.org/10.3390/rs11202401.

Callieco, F., Dell'Acqua, F., 2011. A comparison between two radiative transfer models for atmospheric correction over a wide range of wavelengths. Int. J. Remote Sens. 32 (5), 1357–1370. https://doi.org/10.1080/01431160903547999.

Claverie, M., Vermote, E.F., Franch, B., Masek, J.G., 2015. Evaluation of the Landsat-5 TM and Landsat-7 ETM+ surface reflectance products. Remote Sens. Environ. 169, 390–403. https://doi.org/10.1016/j.rse.2015.08.030.

De Keukelaere, L., Sterckx, S., Adriaensen, S., Knaeps, E., Reusen, I., Giardino, C., Bresciani, M., Hunter, P., Neil, C., Van der Zande, D., Vaiciute, D., 2018. Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. Eur. J. Remote Sens. 51 (1), 525–542. https://doi.org/10.1080/22797254.2018.1457937.

Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., Mangin, A., Pahlevan, N., Pflug, B., Vanhellemont, Q., 2018. Atmospheric correction inter-comparison exercise. Remote Sens. 10, 352. https://doi.org/10.3390/rs10020352.

Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L., 2018. Analysis ready data: enabling analysis of the Landsat archive. Remote Sens. 10 (9), 1363. https://doi.org/10.3390/rs10091363.

Frantz, D., 2019. FORCE—Landsat+ Sentinel-2 analysis ready data and beyond. Remote Sens. 11 (9), 1124. https://doi.org/10.3390/rs11091124.

Frantz, D., Röder, A., Stellmes, M., Hill, J., 2016. An operational radiometric Landsat preprocessing framework for large-area time series applications. IEEE Trans. Geosci. Remote Sens. 54 (7), 3928–3943. https://doi.org/10.1109/TGRS.2016.2530856.

Giles, D., Sinyuk, A., Sorokin, M., Schafer, J., Smirnov, A., Slutsker, I., Eck, T., Holben, B., Lewis, J., Campbell, J., Welton, E., Korkin, S., Lyapustin, A., 2019. 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 optical depth (AOD) measurements. Atmos. Meas. Tech. 12, 169–209. https://doi.org/10.5194/amt-12-169-2019.

Goyens, C., De Vis, P., Hunt, S.E., 2021. Automated Generation of Hyperspectral Fiducial Reference Measurements of Water and Land Surface Reflectance for the Hypernets Networks. In: IEEE International Geoscience and Remote Sensing Symposium IGARSS, 7920-7923. https://doi.org/10.1109/IGARSS47720.2021.9553738.

Holben, B.N., Eck, T.F., Slutsker, I.A., Tanre, D., Buis, J.P., Setzer, A., Vermote, E., Reagan, J.A., Kaufman, Y.J., Nakajima, T., Lavenu, F., 1998. AERONET—A federated instrument network and data archive for aerosol characterization. Remote Sens. Environ. 66 (1), 1–16. https://doi.org/10.1016/S0034-4257(98)00031-5.

Kotchenova, S.Y., Vermote, E.F., Matarrese, R., Klemm Jr., F.J., 2006. Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. Part I: Path radiance. Appl. Opt. 45 (26), 6762–6774. https://doi.org/10.1364/AO.45.006762. Optical Society of America.

Kotchenova, S.Y., Vermote, E.F., Levy, R., Lyapustin, A., 2008. Radiative transfer codes for atmospheric correction and aerosol retrieval: intercomparison study. Appl. Opt. 47, 2215–2226. https://doi.org/10.1364/AO.47.002215.

Li, Y., Chen, J., Ma, Q., Zhang, H.K., Liu, J., 2018. Evaluation of sentinel-2A surface reflectance derived using Sen2Cor in North America. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 11 (6), 1997–2021. https://doi.org/10.1109/JSTARS.2018.2835823.

Liang, S., Wang, J., 2020. Chapter 4 - Atmospheric correction of optical imagery. In: Advanced Remote Sensing, second ed. Academic Press, pp. 131–156. https://doi.org/10.1016/B978-0-12-815826-5.00004-0.

Marcq, S., Meygret, A., Bouvet, M., Fox, N.P., Greenwell, C., Scott, B., Berthelot, B., Besson, B., Guilleminot, N., Damiri, B., 2018. New Radcalnet Site at Gobabeb, Namibia: Installation of the Instrumentation and First Satellite Calibration Results. In: Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July. https://doi.org/10.1109/IGARSS.2018.8517716.

Meygret, A., Santer, R.P., Berthelot, B., 13 September 2011. ROSAS: a robotic station for atmosphere and surface characterization dedicated to on-orbit calibration. In: Proceedings of the Earth Observing Systems XVI, San Diego, CA, USA. https://doi.org/10.1117/12.892759.

Niro, F., Goryl, P., Dransfeld, S., Boccia, V., Gascon, F., Adams, J., Themann, B., Scifoni, S., Doxani, G., 2021. European Space Agency (ESA) calibration/validation strategy for optical land-imaging satellites and pathway towards interoperability. Remote Sens. 13, 3003. https://doi.org/10.3390/rs13153003.

Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M., Giardino, C., Hunter, P., Simis, S., Spyrakos, E., Tyler, A., 2021. ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. Remote Sens. Environ. 258. https://doi.org/10.1016/j.rse.2021.112366.

Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Quang, C., Gascon, F., Boccia, V., 20 September 2020. Next updates of atmospheric correction processor Sen2Cor. Proc. SPIE 11533, Image and Signal Processing for Remote Sensing XXVI, 1153304. https://doi.org/10.1117/12.2574035.

Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., Ying, Q., 2020. Landsat analysis ready data for global land cover and land cover change mapping. Remote Sens. 12, 426. https://doi.org/10.3390/rs12030426.

Remer, L.A., Tanre, D., Kaufman, Y.J., Levy, R., Mattoo, S., 2009. Algorithm for remote sensing of tropospheric aerosol from MODIS for collection 005: Revision 2 Products: 04_L2, ATML2, 08_D3, 08_E3, 08_M3. https://modis.gsfc.nasa.gov/data/atbd/atbd_mod02.pdf (accessed 22 October 2022).

Richter, R., 1998. Correction of satellite imagery over mountainous terrain. Appl. Opt. 37, 4004–4015. https://doi.org/10.1364/AO.37.004004.

Roger, J.-C., Vermote, E., Skakun, S., Murphy, E., Dubovik, O., Kalecinski, N., Korgo, B., Holben, B., 2022. Aerosol models from the AERONET database: application to surface reflectance validation. Atmos. Meas. Tech. 15, 1123–1144. https://doi.org/10.5194/amt-15-1123-2022.

Rouquié, B., Hagolle, O., Bréon, F.-M., Boucher, O., Desjardins, C., Rémy, S., 2017. Using copernicus atmosphere monitoring service products to constrain the aerosol type in the atmospheric correction processor MAJA. Remote Sens. *9*, 1230. https://doi.org/10.3390/rs9121230.

Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O., López-Puigdollers, D., Louis, J., Lubej, M., Mateo-García, G., Osman, J., Peressutti, D., Pflug, B., Puc, J., Richter, R., Roger, J.-C., Scaramuzza, P., Vermote, E., Vesel, N., Zupanc, A., Žust, L., 2022. Cloud Mask Intercomparison eXercise (CMIX): an evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. Remote Sens. Environ. 274, 112990. https://doi.org/10.1016/j.rse.2022.112990.

Vermote, E.F., Kotchenova, S., 2008. Atmospheric correction for the monitoring of land surfaces. J. Geophys. Res.-Atmos. 113, D23S90. https://doi.org/10.1029/2007JD009662.

Vermote, E., Justice, C.O., Bréon, F.-M., 2009. Towards a generalized approach for correction of the BRDF effect in MODIS directional reflectances. IEEE Trans. Geosci. Remote Sens. 47, 898–908. https://doi.org/10.1109/TGRS.2008.2005977.

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens. Environ. 185, 46–56. https://doi.org/10.1016/j.rse.2016.04.008.

Vicent, J., Verrelst, J., Sabater, N., Alonso, L., Rivera-Caicedo, J.P., Martino, L., Muñoz-Marí, J., Moreno, J., 2020. Comparative analysis of atmospheric radiative transfer models using the Atmospheric Look-up table Generator (ALG) toolbox (version 2.0). Geosci. Model Dev. 13, 1945–1957. https://doi.org/10.5194/gmd-13-1945-2020.

Yin, F., Lewis, P., Gomez-Dans, J., Wu, Q., 2019. A sensor-invariant atmospheric correction method: application to Sentinel-2/MSI and Landsat 8/OLI. https://doi.org/10.31223/osf.io/ps957 [Preprint].

Web references

Hagolle, Olivier, Huc, Mireille, Desjardins, Camille, Auer, Stefan, Richter, Rudolf, 2017. MAJA Algorithm Theoretical Basis Document (1.0). Zenodo. https://doi.org/10.5281/zenodo.1209633 (accessed 28/06/2022).

Sentinel-2 L-1C User guide. https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/product-types/level-1c (accessed 09/09/2022).

https://www.usgs.gov/landsat-missions/landsat-level-1-processing-details (accessed 09/09/2022).

Lamasquere project details. https://osr.cesbio.cnrs.fr/les-2-sites-flux-icos/lamasquere/ (accessed 20/01/2022).

Level-2A Algorithm Theoretical Basis Document. https://sentinels.copernicus.eu/documents/247904/4363007/Sentinel-2-Level-2A-Algorithm-Theoretical-Basis-Document-ATBD.pdf/fe5bacb4-7d4c-9212-8606-6591384390c3 (accessed 25/02/2022).

RAdiation Transfer Model Intercomparison for Atmosphere (RAMI4ATM) Web Portal. https://rami-benchmark.jrc.ec.europa.eu/_www/RAMI4ATM.php (accessed 20/01/2022).