Contents lists available at ScienceDirect

Remote Sensing of Environment

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

# An inter-comparison exercise of Sentinel-2 radiometric validations assessed by independent expert groups

Nicolas Lamquin[a,*], Emma Woolliams[b], Véronique Bruniquel[a], Ferran Gascon[c], Javier Gorroño[b], Yves Govaerts[d], Vincent Leroy[d], Vincent Lonjou[e], Bahjat Alhammoud[f], Julia A. Barsi[g], Jeffrey S. Czapla-Myers[h], Joel McCorkel[g], Dennis Helder[i], Bruno Lafrance[j], Sebastien Clerc[a], Brent N. Holben[g]

[a] ACRI-ST, 260 Route du Pin Montard, 06904 Sophia Antipolis, France
[b] National Physical Laboratory (NPL), Hampton Road, Teddington, Middlesex, TW11 0LW, United Kingdom
[c] ESA/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Italy
[d] Rayference, Avenue Paul Deschanel, 247 B3, 1030 Brussels, Belgium
[e] Centre National d'Etudes Spatiales (CNES), Centre spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[f] ARGANS Ltd., Chamberlain House, 1 Research Way, Plymouth PL6 8BU, United Kingdom
[g] NASA Goddard Space Flight Center (GSFC), 8800 Greenbelt Rd, Greenbelt, MD 20771, United States of America
[h] College of Optical Sciences, The University of Arizona, 1630 E. University Blvd, P.O. Box 210094, Tucson, AZ, United States of America
[i] South Dakota State University (SDSU), Brookings, SD 57007, United States of America
[j] CS Systèmes d'Information, Parc de la Grande Plaine - 5, Rue Brindejonc des Moulinais - BP 15872, 31506 Toulouse Cedex 05, France

ARTICLE INFO

*Keywords:*
Sentinel-2
Radiometric validation
Vicarious methodologies
Inter-comparison
Expert groups

ABSTRACT

Copernicus is the European Union's Earth Observation and Monitoring programme, delivering free access to operational and historical environmental data to support applications in a wide range of societal benefit areas. To allow meaningful long-term environmental monitoring and robust decision-making, it is essential to ensure that satellite-retrieved products are of high quality and consistency. This paper describes the outputs of an international workshop on the radiometric calibration validation of the Copernicus Sentinel-2A and Sentinel-2B Multi-Spectral Instrument. A wide range of vicarious methodologies have been applied independently and then compared per type of target. All methods agree on the good radiometric performance of both Sentinel-2A and Sentinel-2B with respect to the mission requirements as well as on evidence of a slight bias between the two instruments. Comparisons of all these results are discussed to highlight the benefits and advantages of the methods as well as to propose potential improvements either for the methods themselves and/or for the comparison exercise.

## 1. Introduction

Copernicus is the European Union's Earth Observation and Monitoring program, delivering free access to operational and historical environmental data to support applications and to feed services for a wide range of societal benefit areas. The space component of Copernicus is principally provided by the operational services of the Sentinel satellites as well as by third party missions providing complementary observational capacity. To be of real use in environmental monitoring and decision-making, it is essential to ensure that satellite-retrieved products are of high quality and consistency and that they are interoperable: that is, data from different satellite sensors, including those from different agencies, can be meaningfully combined.

Since early 2017, the Sentinel-2 mission has provided continuous monitoring of terrestrial surfaces and coastal waters at a global scale completing a five-days revisit frequency using two identical Sentinel-2 platforms launched respectively on June 23rd 2015 and March 7th 2017. Both platforms carry an instance of the Multi-Spectral Instrument (MSI) acquiring measurements in 13 channels in the visible (VIS), near-infrared (NIR) and shortwave infrared (SWIR) spectral domains at spatial resolutions of 10 m, 20 m or 60 m (Gascon et al., 2017) over a total swath width of 290 km (Table 1 for details).

L1C products are Top-of-Atmosphere (TOA) reflectance images delivered to the public after geometric and radiometric calibration

[*] Corresponding author.
*E-mail address:* nicolas.lamquin@acri-st.fr (N. Lamquin).

https://doi.org/10.1016/j.rse.2019.111369
Received 3 January 2019; Received in revised form 4 August 2019; Accepted 10 August 2019
Available online 31 August 2019

**Table 1**
MSI channels, central wavelength, bandwidth, and spatial resolution.

| Spectral domain | Band | Central Wavelength (nm) | Bandwidth (nm) | Resolution (m) |
|---|---|---|---|---|
| VIS | B01 | 443 | 20 | 60 |
| | B02 | 490 | 65 | 10 |
| | B03 | 560 | 35 | 10 |
| | B04 | 665 | 30 | 10 |
| | B05 | 705 | 15 | 20 |
| | B06 | 740 | 15 | 20 |
| NIR | B07 | 783 | 20 | 20 |
| | B08 | 842 | 115 | 10 |
| | B8a | 865 | 20 | 20 |
| | B09 | 945 | 20 | 60 |
| SWIR | B10 | 1375 | 30 | 60 |
| | B11 | 1610 | 90 | 20 |
| | B12 | 2190 | 180 | 20 |

(Gascon et al., 2017). The radiometric calibration validation described in this paper is based on these L1C products for both Sentinel-2A (S2A) and Sentinel-2B (S2B) MSI instruments. Products are available from the Sentinels Scientific Data Hub and are formatted per tile (ortho-images in UTM/WGS84 projection) of about 100 × 100 km$^2$. Because Sentinel-2A was launched before Sentinel-2B, there are more data underpinning the statistical analysis of this sensor, and therefore the results presented here are considered more robust for Sentinel-2A.

The Sentinel-2 MSI instruments are routinely calibrated by the Sentinel-2 Mission Performance Centre (S2-MPC). The calibration process uses a combination of onboard and vicarious methods to adjust the inflight instrument calibration model and, in the processing chain, to account for instrument changes since pre-flight calibration.

The nominal radiometric calibration of Sentinel-2 MSI sensors is based on images acquired over ocean at night (for the dark signal calibration) and on-board sun-diffuser images (to perform the absolute radiometric calibration and the pixel equalization). The principles of radiometric calibration are fully described in Gascon et al. (2017). The dark signal coefficients are assessed by averaging counts over dark acquisition lines. The dark signal is particularly stable for MSI sensors. For VNIR bands, variations are usually smaller than one digital count. For SWIR bands, the dark signal is generally stable and also very small, though it may be as high as five digital counts for a few pixels.

The absolute and relative gain coefficients are estimated from sun-diffuser acquisitions on a monthly basis. The method relies on the comparison of the bright and uniform image of the sun-diffuser acquisition to a simulation of the reflected radiance. The simulation includes the solar spectral irradiance from Thuillier et al. (2003), convolved with the Sentinel-2 spectral band response functions, and with a fine calculation of the Earth-to-Sun distance based on Orekit flight dynamics library (Maisonobe and Pommier-Maurussane, 2010). The reflectance of the diffuser panel is modelled by a pre-launch characterisation of its Bidirectional Reflectance Distribution Function (BRDF). For each spectral band, the acquisition image of the diffuser is compared to the modelled image so as to provide an image of the ratio between measurement and simulation over the complete field-of-view (FOV). An average over the whole image of the ratio gives the absolute gain coefficient for the considered band. At the same time, the comparison of the equalized sun-diffuser image to the simulation allows the detection of changes in the relative gain coefficients, per pixel, per detector. The monthly calibration leads to an update of the dark signal coefficients and of the absolute and relative gain coefficients for the Level-1 processing chain.

The Sentinel-2 mission requirements aim to achieve a radiometric uncertainty at TOA of no > 3% goal, 5% threshold. The threshold requirement represents the limit above which the observation quality becomes unbeneficial for its application. The goal requirement is the ideal requirement which it is not necessary to exceed. Inter-band consistency should be within 3% (ESA Sentinel-2 Team, 2007). For some applications, particularly climate applications, lower uncertainties are desirable (e.g. GCOS Observation Requirements: albedo 5% absolute, 1% stability, https://gcos.wmo.int/en/essential-climate-variables/requirements, see also Wielicki et al., 2013).

In addition to the formal radiometric calibration activities carried out by S2-MPC, radiometric validation must be performed independently to ensure that the mission requirements are met, to assess the level of operational bias between sensors, and to validate the estimated uncertainty.

The Committee on Earth Observation Satellites (CEOS) Working Group on Calibration and Validation (WGCV) has defined validation as “the process of assessing, by independent means, the quality of the data products derived from satellite instrument measurements” (http://ceos.org/ourwork/workinggroups/wgcv/). There are many different methods used for satellite optical sensor radiometric validation, as described in the next section.

Since the aim of satellite radiometric validation is to assess the quality of the operational calibration, as the operational demands become increasingly challenging (requiring lower uncertainties), validation methods must likewise become more accurate. A valuable way to assess validation methods is to perform comparisons between different methodologies, and especially between independent analyses by different experts.

In the frame of the S2RadVal project (http://s2radval.acri.fr/) of the ESA Scientific Exploitation of Operational Missions programme (SEOM, http://seom.esa.int/), a workshop was organised at ESRIN in January 2018 involving international scientific experts in calibration and validation in order to compare their assessments of the Sentinel-2 radiometric validation. This paper, co-authored by all participants, presents the output of the workshop.

The principles of the different radiometric validation methods are presented in the following section, along with a description of the targets used for the validation. Comparisons of the implementation of each method by different expert groups are then presented, showing discrepancies between the independent implementations. The different methods are finally compared together with a view to discuss their limitations and potential improvement. Interestingly, the dispersion of all the individual results tends to cancel out by using each method as transfer target methodogy to estimate biases between S2A and S2B, without a temporal coregistration between the two MSI instruments.

## 2. Targets and methods

The most commonly-used methods for radiometric validation are summarised in Table 2. These methods rely on a comparison of sensor measurements of radiance or reflectance over natural (and in rare cases artificial) targets with an independent estimate of the TOA radiance or reflectance of that target. In some cases (e.g. Rayleigh scattering) the independent estimate is from a simulated signal of the same target in the same configuration (e.g. spectral band, geometry) from a physical model of the target, in other cases it is from an independent observation by a different satellite sensor under the same configuration (or by the same sensor at different times). Generally, two satellite instruments have different spectral response functions, and these should be accounted for in any sensor comparison (Teillet et al., 2007; Chander et al., 2013).

Many of the methods use intermediate cases – neither fully simulated nor fully measured. Here measurements (from other sensors, from other spectral bands of the same sensor or from ground observations), along with measurements or models of, e.g. atmospheric conditions, are combined with, or used to parameterise, an accurate physical model of the TOA signal, considering radiative transfer modelling (RTM) and properties of the surface.

In our analysis, all methods have been treated statistically; that is multiple comparisons are performed for each method and the results

**Table 2**
Link between natural targets, their properties, the radiometric validation, and their spectral range for use in validation of EO optical sensors.

| Target | Properties | Radiometric validation methods | Spectral range |
|---|---|---|---|
| Instrumented calibration sites (in situ) | Characterized on site. Usually bright, homogeneous to allow scaling | Calibration relative to ground observations and modelled atmospheric radiative transfer | Spectral range of the in situ measurements, often VNIR* and sometimes SWIR* |
| Pseudo-invariant calibration sites (desert) | Bright, invariant, homogeneous | Radiometric trends, absolute calibration. Cross-sensor radiometric comparisons | VNIR-SWIR |
| Wide homogeneous snowy areas (Arctic/Antarctic) | Bright, invariant, homogeneous | Radiometric trending | VNIR |
| Oceans – Rayleigh scattering | Dark, homogeneous | Absolute calibration relative to well-understood physical processes | VIS |
| Oceans – Sunglint | Bright, white (spectrally homogeneous) | Interband calibration using well-understood physical processes | VNIR-SWIR |
| Deep convective clouds | Bright, white (spectrally homogeneous) | Radiometric trends, interband calibration | VNIR |
| Moon | Bright, invariant (modelled phase/ libration variability), No attenuation | Absolute calibration | VNIR-SWIR |

* VNIR = VIS-NIR, SWIR = Short-Wave Infrared.

averaged. There are always some biases due to configurational differences, seasonal effects and site-specific noise (e.g. from atmospheric variability) that can, along with instrument noise, be partially averaged out by combining data from different comparisons.

The participants of the S2RadVal workshop submitted results exploiting deep convective clouds (DCC), oceanic targets, pseudo-invariant calibration sites (PICS), and ocean and land sites equipped with in situ measurement facilities. Each target is detailed below.

### 2.1. Deep convective clouds: interband calibration validation

Deep Convective Clouds (DCCs) are very vertically-extended (about ground to tropical tropopause) and opaque (optical thickness of the order of 100) clouds. They exhibit bright and almost white spectra from the visible to the near-infrared, with nearly-isotropic reflectivity when illumination and viewing angles do not exceed 30°. Cloud brightness is driven by their optical thickness with reflectance levels ranging from about 0.7 to about 1 or more as observed from radiometers with spatial resolutions coarser than 1 km. Three-dimensional structural/optical effects do cause small-scale heterogeneous cloud reflectance (see Lamquin et al. (2017) for details) which are seen for the spatial scales of S2-MSI. As a second order effect, the microphysical composition of the clouds (ice crystals properties on the top) drives a slight variability of the spectral shape.

DCCs are used to monitor the temporal degradation and the interband calibration of a sensor in the VNIR relatively to a reference band. The reference band is assumed to be calibrated well enough to capture the level of brightness. These targets have been used for EO sensor calibration for about two decades (Vermote and Kaufman, 1995; Hu et al., 2004).

Two independent DCC methods have been implemented and compared. One method was developed by CNES and was initially applied to PARASOL as reported in Fougnie and Bach (2009). Revel et al. (2019) adapted this method to high spatial resolution imagery and applied it to Sentinel-2 data. The other method, developed by ACRI-ST in the framework of S2RadVal, is further detailed in Lamquin et al. (2017).

### 2.2. Open ocean clear waters ("Rayleigh" method): absolute calibration validation

Open ocean clear waters are used as targets in the so-called "Rayleigh" method. This approach is well described in Fougnie and Henry, n.d. (QA4EO community-specific guidelines) and originates from Vermote et al. (1992) applied to SPOT and Hagolle et al. (1999) applied to POLDER.

In the absence of clouds and wind (no roughness of the sea causing potential sunglint) and by selecting the most isolated oceanic locations with smallest aerosol loadings, the ocean provides a stable target. The TOA signal is composed of the signature of pure oligotrophic water (the signal of which is modelled through the use of climatologies and radiative transfer look-up-figs) along with the signature of the overlying clear atmosphere where most of the signal in the visible comes from Rayleigh scattering. Rayleigh Scattering can contribute to up to 90% of the signal in the blue spectral region.

The Rayleigh method is an almost pure "absolute" method as the observed signal can be straightforwardly compared with the modelled signal. Because the Rayleigh scattering signal decreases with longer wavelengths, this method is most accurate at the shortest wavelengths (blue), which are often the most challenging for other methods. Results are presented only from the blue (443 nm) to the deep red (740 nm).

Two independent Rayleigh methodologies have been implemented and compared. The first methodology has been implemented by CNES. A description of the methodology can be found in Fougnie et al. (2010) and its application to S2-MSI data is reported in Revel et al. (2019). The second methodology has been applied using the DIMITRI (Database for Imaging Multi-spectral Instruments and Tools for Radiometric Intercomparison) package, developed and maintained by ESA/ESTEC, ARGANS and MAGELLIUM (https://dimitri.argans.co.uk). The method is described in Barker et al. (2014) and applied on Sentinel-2 by ARGANS following Alhammoud et al. (2018).

### 2.3. Coastal waters: low radiance absolute calibration validation

Coastal waters provide a valuable target for the verification of the S2-MSI low radiance calibration. Coastal waters often contain suspended particles, which increase the water reflectance compared to open waters. Aerosols can also contribute to the TOA signal as coastal areas can also be affected by aerosols emitted over nearby land. The AERONET-Ocean Colour (OC) network (Zibordi et al., 2009) can supply sea conditions and optical properties, including the aerosol optical thickness, in several bands matching some of the S2-MSI ones. These measured properties are used to parameterise a model to provide simulated S2-MSI observations over the AERONET-OC stations.

The proposed verification method therefore consists of simulating Sentinel-2/MSI observations acquired over AERONET-OC stations and comparing observations against simulation results, accounting for their respective uncertainties. The following AERONET-OC stations were selected for the methodology over coastal waters: AAOT, COVE SEAPRISM, Gustav Dalen Tower, GOT Seaprism, Galata, Gloria, Helsinki Lighthouse, LISCO, MVCO, Pålgrunden (Philipson et al., 2016), Thornton, USC, WaveCIS. The 6SV Radiative Transfer Model (Kotchenova et al., 2006) has been selected to perform numerical simulations over coastal waters. This RTM accounts for polarisation, which cannot be neglected at short wavelengths, and, as it is a well-documented and well-structured code, it was straightforward to modify it for the purpose of this study.

S2-MSI observations acquired over AERONET-OC stations were extracted from the corresponding tiles in bands B01 to B08 (443 nm to 842 nm), and where a coincident AERONET-OC measurement existed, and meteorological conditions met criteria minimising the simulated surface reflectance uncertainties, the S2-MSI observations were simulated using the 6SV RTM. The 6SV RTM has been modified to maximise the use of AERONET-OC data, in particular concerning water-leaving reflectance and the spectral variations of the aerosol optical thickness (AOT).

The Guide to the Expression of Uncertainty in Measurement (GUM) recommendations (http://www.bipm.org/en/publications/guides/gum.html) were used to determine the global uncertainty of simulated MSI reflectances. Uncertainties were added in quadrature without covariance as there is no reason to believe that the errors associated with the different ground observations, are common for water targets. Processing more samples should reduce the contribution of non-systematic uncertainties. Once enough S2-MSI data have been processed, the overall biases between simulations and observations are provided in each spectral band with the corresponding uncertainties.

### 2.4. Land calibration equipped sites: absolute calibration validation

Ground-based in situ measurements of surface reflectance and atmospheric conditions, combined with an RTM, can provide TOA values for comparison with satellite sensor observations. The traditional reflectance-based approach uses a well-calibrated reference panel and portable radiometer to determine the surface reflectance for a specific region of interest. The surface reflectance is determined using the ratio of surface to reference panel measurements, with metrological traceability to a laboratory calibration of the reference panel. Another approach is to use absolutely-calibrated radiometers to measure the radiance of the surface, combining this with a RTM calculation of downwelling irradiance to convert surface radiance into surface reflectance factor. Atmospheric measurements are typically made using solar radiometers or sun photometers. Land calibration equipped sites are typically chosen based on the requirements of the sensors to be calibrated. The typical moderate-resolution sensor used for Earth observation in the solar-reflective regime benefits from a site that is large, spatially homogenous, spectrally flat, high in altitude, has a high surface reflectance, and is as temporally invariant as possible. Desert regions provide the majority of sites that fit these criteria.

For this paper, calibration validations over instrumented land targets are derived from two groups of sites: one group concerns the RadCalNet sites (Gobabeb in Namibia, with results from CNES and NPL, and the Railroad Valley site in the USA, with results provided by University of Arizona and NPL), the other concerns acquisitions during the Salar de Uyuni (Bolivia) campaign with results provided by NASA.

#### 2.4.1. RadCalNet stations

RadCalNet is a CEOS WGCV initiative to provide a coordinated network of instrumented land-based test sites. Each site provides nadir-view ground reflectance at 30-minute intervals from 9 am to 3 pm local standard time at 10 nm intervals from 400 nm to 2500 nm along with the necessary atmospheric measurements (surface pressure, columnar water vapour, columnar ozone, aerosol optical depth and the Angstrom coefficient). RadCalNet sites owners perform a quality check on the data prior to uploading the data for RadCalNet processing by NASA/GSFC. The RadCalNet processing propagates the ground reflectance to TOA through a standard MODTRAN® analysis (MODerate resolution atmospheric TRANsmission, see below for details).

The Gobabeb site is operated as an ESA-CNES site, with ESA supported by NPL. It has an automatic ground-based station comprising a CIMEL photometer on a 10 m mast (Marcq et al., 2018). This makes measurements for multiple viewing geometries, which are used to parameterise a site bidirectional reflectance distribution function (BRDF) model and to determine downwelling irradiance. The official RadCalNet product takes nadir-viewing measurements from this model, NPL performed a comparison of S2-MSI to the official RadCalNet product. The CNES analysis compared the S2-MSI reflectance to that simulated by the full model, as described in Revel et al. (2019).

The Railroad Valley site is operated and maintained by the University of Arizona and has been used as a vicarious calibration test site for over 20 years and as an instrumented site for 10 years. The Radiometric Calibration Test Site (RadCaTS) uses four custom-built ground-viewing radiometers (GVRs) developed at the University of Arizona (Anderson and Czapla-Myers, 2013; Anderson et al., 2013) which collect data every 2 min during cloud-free daylight in eight spectral bands. Atmospheric measurements are made using a Cimel CE318-T photometer following the AERONET protocol. The GVRs are positioned throughout the RadCaTS 1 km$^2$ region of interest to minimise uncertainties due to spatial sampling (Czapla-Myers et al., 2007; Czapla-Myers et al., 2008).

The surface reflectance data collected by the multispectral GVRs are converted to hyperspectral (1 nm) reflectance by fitting the average surface reflectance determined in each of the eight GVR channels using a library of ~700 hyperspectral data sets collected at Railroad Valley from 2000 to 2018 using ASD spectrometers. The hyperspectral surface reflectance and atmospheric data are used as input into MODTRAN®, and the TOA reflectance is determined for a given sensor overpass in the sensor geometry and using the band relative spectral responses (Czapla-Myers et al., 2015). Sentinel-2A and -2B observe RadCaTS with two off-nadir view angles (~6° and ~11° at 18:44 UTC and 18:33 UTC, respectively). The University of Arizona used the full RadCaTS data to compare with S2-MSI while NPL used the official RadCalNet product derived from this.

For both RadCalNet sites, the NPL analysis involved, for each overpass that had both RadCalNet and Sentinel-2 data, extracting the relevant data from the Sentinel-2 tile, interpolating the RadCalNet data (temporally) and convolving the RadCalNet data with the Sentinel-2 spectral response functions (SRFs) before comparing band-integrated TOA reflectances. The Sentinel-2 L1C products were screened to discard data if the complete 100 × 100 km$^2$ tile had > 60% cloud or > 20% invalid data, or if any cloud or invalid mask intersects the region of interest (ROI). Further screening was done on the trend over each site, based on statistical analysis of the site (rejecting outliers in terms of mean, standard deviation, skewness and kurtosis). Finally, images were visually screened. The RadCalNet datasets used for the comparison are output files version 02.00 for the Gobabeb site and version 02.03 for the Railroad Valley downloaded on the 22nd of May 2018.

#### 2.4.2. Salar de Uyuni campaign

A validation campaign was held in June 2017 for the Geostationary Operational Environmental Satellite-16 Advanced Baseline Imager (GOES-16 ABI) and this campaign also provided match ups to Sentinel-2. The reflectance-based approach of vicarious calibration was used with in situ measurements of ground reflectance relative to a laboratory-calibrated reference reflectance panel along with measurements of atmospheric parameters (Thome, 2001).

Reflectance measurements were made using an ASD FieldSpec4 from 350 nm–2500 nm, to characterise four 1 km transects, oriented to match the GOES-16 azimuth, that were separated by 333 m. Reference measurements of a Spectralon target with NIST-traceable BRDF measurements were collected at the beginning, end, and every 250 m of each transect. Most measurements were taken in the GOES-16 viewing geometry, but nadir-view measurements were made to match the S2A overpass on 18 June 2017.

Six AERONET sun photometers were deployed for the duration of this campaign (Holben et al., 1998) and designated SDU 1 through SDU 6. SDU 1 was positioned near the perimeter of the test site; 3 units, SDU2–4, were place at a radius of approximately 20 km around the test site, and the remaining two were placed at the perimeter of the salt flat. The average of the four sun photometers located near the test site

(SDU1–4) are used to specify the atmospheric properties in the radiative transfer model.

The MODTRAN® 5 version 2 radiative transfer code was used to account for the Earth atmosphere by propagating the measured ground reflectance to the top of the atmosphere (Berk et al., 2005). Combining radiative transfer uncertainty with measurement repeatability and reference panel characterisation uncertainty, the total SI-traced uncertainty of the predicted TOA reflectance is on the order of 2% for the VNIR channels and 2.2% for the SWIR channels with exception of the cirrus band (B10 1.3 μm) which exhibits 15%–20% uncertainty though results of this channel are not presented.

### 2.5. *Pseudo-invariant calibration sites: absolute calibration, cross-sensor calibration*

Pseudo-invariant calibration sites (PICS) are locations that are radiometrically very temporally stable and spatially homogeneous. Desert regions are ideal for PICS as they receive little rainfall, have less frequent cloud cover, and have stable and homogeneous surface features. PICS are used to monitor the temporal stability over a sensor's lifetime. They have been used to transfer absolute calibration from one point in time to the full lifetime of the mission (Markham and Helder, 2012). Recently work has been undertaken to model the PICS as absolute calibration sources (Mishra et al., 2014 and Bouvet, 2014) using well-calibrated satellite data along with surface BRDF and RTM models. These modelling techniques would allow for absolute calibration of instruments without on-board calibration devices or when the on-board calibrators are no longer stable.

Additionally, the stability of individual PICS regions allows for direct comparison between two satellite sensors that may not see the site at exactly the same time. The sites are not equipped with instrumentation at the surface, so assumptions must be made that the atmospheric conditions and surface reflectances haven't changed significantly between acquisitions.

Four methodologies using pseudo-invariant calibration sites (PICS) are provided for this paper: one from NASA/GSFC, one from the Sentinel-2 Mission Performance Center (S2-MPC) (ARGANS/DIMITRI), one from CNES, and one from South Dakota State University along with the United States Geological Survey (USGS).

NASA/GSFC performs cross-calibration with Landsat-8 Operational Land Imager (OLI). L8-OLI and S2-MSI acquire coincident images (within 20 min) of specific PICS regions every 80 days, when their orbits match. By correcting the TOA reflectance for the differences in their spectral responses and solar zenith angle effects, the MSI and OLI reflectances can be compared directly, for spectral bands which OLI and MSI have in common. The PICS regions in use here are 20 × 20 km$^2$ areas in the Sahara Desert, defined in Lachérade et al. (2013). Sentinel-2A and Landsat-8 acquire Libya-4 and Algeria-3 on the same dates and Sentinel-2B and Landsat-8 acquire Egypt-1 and Algeria-5 on the same dates. The methodology is detailed in Barsi et al. (2018). The TOA reflectances are extracted from the S2-MSI L1C and the L8-OLI L1TP products. The reflectances are corrected for solar zenith angle (if necessary) and region-specific spectral band differences. The comparison metric is the ratio between the region-average reflectances.

In the framework of the S2-MPC activities, ARGANS performs the validation and monitoring of the S2-MSI L1C–product radiometry using the Desert-PICS method implemented in DIMITRI following Bouvet (2014). This method simulates TOA reflectance in the VNIR spectral range over the PICS using a physical radiative transfer model calibrated from 4 years of MERIS observations. Sub-regions of 20 × 20 km$^2$ were selected over each CEOS-PICS test-site as reported in Alhammoud et al. (2018) and compared to S2A-MSI (July 2015–April 2018) and S2B-MSI (March 2017–April 2018).

CNES routinely monitors S2-MSI cross-calibration with respect to other reference sensors: Landsat-8, MERIS, MODIS, SPOT-5 and Sentinel-3 OLCI. The methodology is described in Cabot et al. (2000) and Lachérade et al. (2013), the latest results are reported in Revel et al. (2019). Only results from cross-calibration with Landsat-8 OLI and MERIS are presented here for comparisons with the two other methodologies: NASA/GSFC for OLI and ARGANS/DIMITRI for MERIS.

South Dakota State University and USGS EROS performed a comparison to the Absolute PICS (APICS) which uses the MODIS Terra instrument as a calibrated radiometer to develop an absolute model for the Libya 4 PICS (Vaidya et al., 2014). The model combined the radiometric accuracy and viewing angles available from MODIS Terra with the hyperspectral information available from the EO-1 Hyperion to develop a simple model predicting the nadir upwelling radiance from Libya 4. The model can be used any time Libya 4 is viewed and has been validated with a broad range of sensors to have an accuracy of 3% or better. Recently the model has been upgraded to provide more accurate results at the shortest wavelengths for Sentinel-2 (Kaewmanee and Helder, 2017).

## 3. Comparison results

In this section all methodologies presented above are compared per target following the same order as presented. For the sake of comparability all independent results are computed as ratios between S2-MSI TOA reflectance $\rho_{\mathrm{MSI}}(\lambda)$ and the reference TOA reflectance $\rho_{\mathrm{REF}}(\lambda)$ (measured and/or simulated) per waveband as:

$$g(\lambda) = \frac{\rho_{\mathrm{MSI}}(\lambda)}{\rho_{\mathrm{REF}}(\lambda)} F_{\mathrm{SBAF}}$$

where $F_{\mathrm{SBAF}}$ is a spectral band adjustment factor (Teillet et al., 2007; Chander et al., 2013) that accounts for any mismatch in the spectral response function between the reference and S2-MSI (not always needed).

Interband ratios $g'(\lambda)$ are computed relative to the reference band B04 ($\lambda_{\mathrm{ref}}$ = 665 nm) so that

$$g'(\lambda) = \frac{g(\lambda)}{g(\lambda_{\mathrm{ref}})}.$$

In all figures dashed lines indicate the 3% (goal) and 5% (threshold) radiometric requirements. Full uncertainty analysis is not always performed. “Error bars” usually represent the dispersion of individual results, in other cases the signification if explained.

For methodologies involving low radiance levels, this ratio is between two small numbers and therefore less meaningful. Instead, the comparison is done for absolute radiance differences and compared with the absolute radiance equivalent to a 3% or 5% reflectance uncertainty for typical land surface radiance levels.

### 3.1. *Deep convective clouds*

Two independent DCC methods were implemented to determine the interband ratio $g'(\lambda)$, relative to the reference band B04. The CNES method uses a fixed cloud parameterisation while the ACRI method was repeated for different cloud model parameterisations (macrophysical and microphysical descriptions of the cloud, see Lamquin et al. (2017) for details). Both methods use statistics of comparisons over a very large number of carefully selected cloudy pixels. Vertical bars for the CNES method (with results shown in red) represent the standard deviation around the mean. Results from the ACRI method (shown in blue) are displayed for three cases: one mean and two extremals. Each different cloud parameterisation (3 × 3 combinations) led to different results and the minimal and maximal values are shown in dashed lines, with the mean comparison in plain line. Similar dispersion, per type of model, has been obtained from this methodology but is not displayed here for legibility. Further discussion in expressing the total uncertainty of this methodology can be found in Lamquin et al. (2017). Results are presented for S2A (left) and S2B (right) on Fig. 1 in the VNIR where

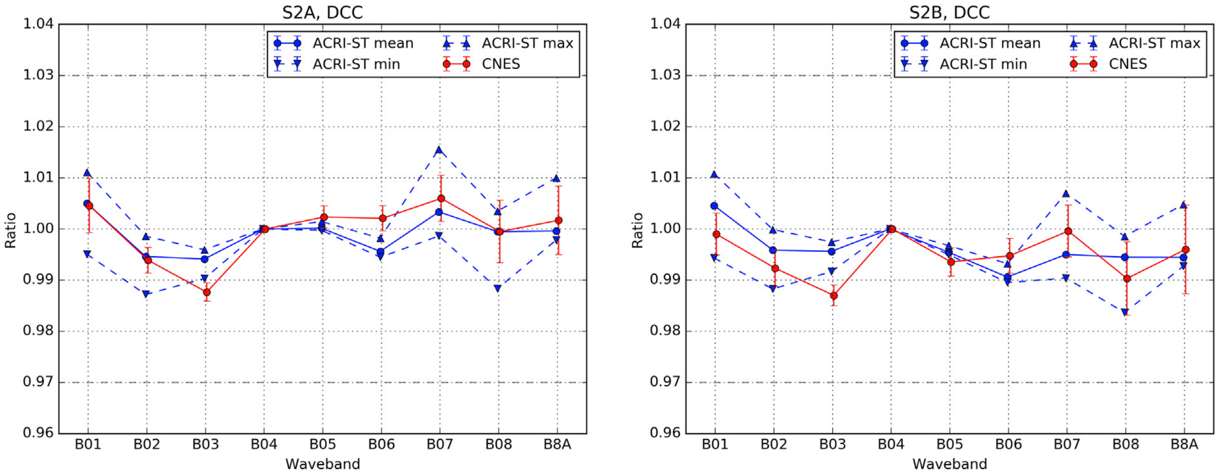

**Fig. 1.** Interband ratios $g'(\lambda)$ from DCC methods of CNES (red) and ACRI-ST (blue) with band 4 (B04) as reference. S2A (left) and S2B (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

both methods are applicable.

All results agree within the interband requirement of 3% and even within 2%. Except for B03 (S2A and S2B), B05 (S2B), and B06 (S2A), the CNES results are well within the boundaries found by the extremal simulations of the ACRI result. Differences between the two methods at B03 may be due to a different handling of ozone absorption, combined with a different temporal/regional sampling, since this spectral band is the most sensitive to ozone. Indeed, a change of 50 DU, the order of magnitude of values in the Tropics, changes the ozone transmission to about 2%. These results provide evidence that DCCs can be used to verify interband calibration with an accuracy of about 1%–2%. Both methods show that Sentinel-2 interband requirements are met both for S2A and S2B in the VNIR. Comparisons between S2A and S2B results show consistent behaviour for both instances of the MSI. We observe similarities between S2A and S2B interband ratios but without being able to determine if these come from a bias in our models or from a true instrumental effect.

### 3.2. *Open ocean clear waters (Rayleigh)*

In the CNES implementation of the Rayleigh method, the Sentinel-2 products are split into 1 km$^2$ boxes. The large size of the Sentinel-2 swath, and the scale of the oceanic sites (between 1000 km and 2500 km) mean that there are hundreds of thousands of independent measurements for comparison. The combined (averaged) results are then considered statistically significant. The L1C cloud mask and a threshold on the B8A reflectance is applied to remove cloudy measurements and a filter on the viewing angle is used to keep only the data that are not polluted by sun glint. Applying these masks and filters results in the use of acquisitions in the northern hemisphere from December to February and southern hemisphere from June to August. With an acquisition local time of 10:30 am, and a near-nadir viewing geometry, the east side of the Sentinel-2 swath contains more rejected data. The results are not perfectly stable as a function of the position in the field of view (about a 3% variation between left and right in Fig. 8 of Revel et al., 2019) and variations between sites can also be found. We consider these discrepancies as insignificant (averaged out; which is also the statement in Revel et al., 2019) because of the statistical heterogeneity of the field-of-view coverage as well as due to the inherent limitations of the method (notably uncertainties in the modelling).

In the ARGANS/DIMITRI implementation, due to the large size of Sentinel-2 images at full spatial resolution (10 m), the radiometric and geometric measurements of the L1C products are re-sampled onto the band B01 grid (60 m spatial resolution); and then extracted over ROIs of about 10 × 10 km$^2$. An automatic cloud screening in DIMITRI is applied, followed by visual inspection and manual screening of individual images. The TOA-reflectance, the solar and sensor geometry, cloud-mask and auxiliary variables are stored for each pixel. A threshold on the B8A reflectance is applied to remove measurements contaminated by glint and high aerosol load, selecting the most relevant measurements but reducing the number of measurements that can be averaged. The average and the standard deviation of the TOA-reflectance, solar and sensor geometry and the percentage of the cloud-cover over the ROI are computed. Then, the Rayleigh method is applied. Finally, each acquisition is associated to a set of gain coefficients by band (B01 to B04).

Results of the comparison between the simulated Rayleigh scattering signal and the S2-MSI measured signal are shown in Fig. 2 for the ratio $g(\lambda)$; no spectral band correction is needed here. Error bars shown for both methods are the standard deviation of the independent results over different oligotrophic regions used for assessing the ratios. Except for the ARGANS results for B01, all absolute ratios are within the goal uncertainty of 3%. The dispersion of the results is higher for spectral bands B01 to B03 where reflectance is the highest. This higher dispersion may be related to some imperfection in the estimated marine reflectance climatology as well as in the estimation of the Rayleigh and aerosols contributions to TOA signal. Results from both CNES and ARGANS seem to demonstrate a slight absolute bias between the two sensors, with S2A measured reflectance being higher than S2B.

The difference between the ARGANS and CNES results may partly be due to the different regions of interest used by the two groups (CNES results are based on larger sites). The two groups also use different hypothesis for the marine reflectance: CNES analysis is based on a climatology derived from SEAWIFS and MODIS data while ARGANS/DIMITRI uses a marine model following Morel and Maritorena (2001).

### 3.3. *Coastal waters*

Table 3 and Fig. 3 present the results of the coastal water radiance comparisons. In the red part of the spectrum, radiances are very low over these sites and therefore comparing ratios leads to a lack of precision. The relatively higher complexity of the observed scene, compared to open oceans, adds even more uncertainty. We therefore uniformly compare the results in absolute radiances rather than as relative signals. To do so, the requirement of 5% relative uncertainty is converted to a requirement in the absolute values based on TOA reflectance computed with 6SV for a reference vegetated surface as described by Gobron et al. (1996) with a Leaf Area Index (LAI) value of 3, using a US76 standard atmosphere, an aerosol optical thickness of 0.2 at 0.55 μm, nadir viewing conditions and a sun azimuth angle of 45° (see

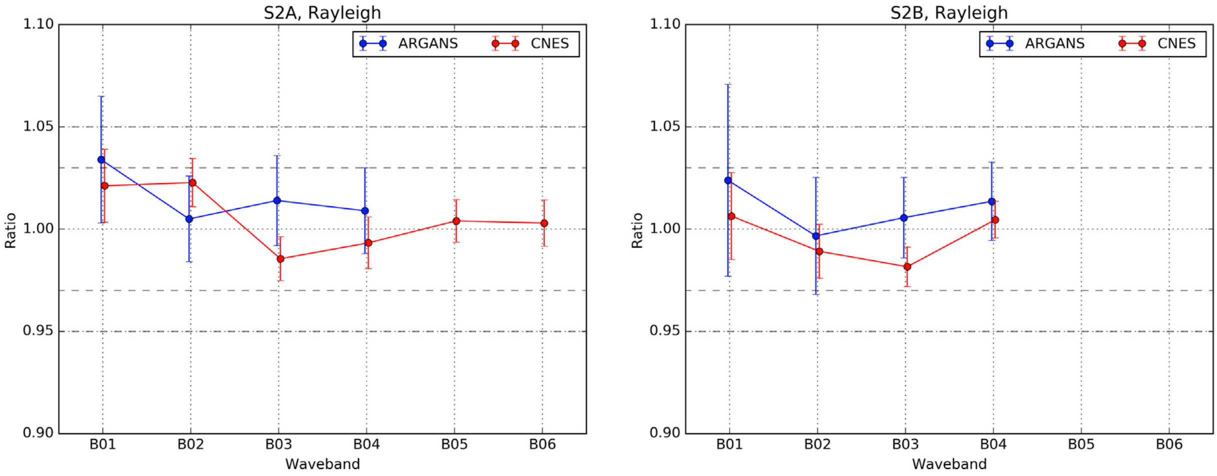

**Fig. 2.** Absolute ratios from Rayleigh methods of S2-MPC, ARGANS/DIMITRI (blue) and CNES (red). S2A (left) and S2B (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
Absolute radiometric requirements in MSI band B01 to B08 for S2A and S2B.

| Band | S2A | | S2B | |
|---|---|---|---|---|
| | TOA BRF | ABS. REQ. | TOA BRF | ABS. REQ. |
| B01 | 0.1298 | 0.00649 | 0.1311 | 0.00656 |
| B02 | 0.0990 | 0.00495 | 0.1010 | 0.00505 |
| B03 | 0.0914 | 0.00457 | 0.0918 | 0.00459 |
| B04 | 0.0666 | 0.00333 | 0.0666 | 0.00333 |
| B05 | 0.0811 | 0.00405 | 0.0809 | 0.00404 |
| B06 | 0.2374 | 0.01187 | 0.2305 | 0.01153 |
| B07 | 0.3381 | 0.01690 | 0.3364 | 0.01682 |
| B08 | 0.3245 | 0.01622 | 0.3243 | 0.01622 |

Table 3). Resulting absolute requirements are almost equal for S2A and S2B.

Fig. 3 displays the results obtained with this method, only results for S2A are provided as there were insufficient S2B match ups to get a statistically valid comparison. The horizontal-axis is labelled with S2-MSI spectral band names and the vertical-axis with the absolute bias between the simulated TOA reflectance values and corresponding MSI observations. The orange dots and error bars represent the mean and standard deviation on the bias. The green error bars represent the estimated uncertainty on the mean value and the blue brackets the requirements taken from Table 3. As the 5% MSI requirements have been formulated for a standard green vegetated surface observed from space, the corresponding absolute requirement values range from 0.003 to 0.006 in the visible spectral region and from 0.011 to 0.017 in the NIR one. In the visible spectral region, the mean bias is close to the upper bound of the requirement interval while in the NIR region it is closer to the centre. The large standard deviation on the bias in the S2-MSI visible bands results from the low number of successfully processed observations (about 40). The estimated method uncertainty (green error bars) is at the order of magnitude of the 5% requirement in the visible region. Further improvement of this method is needed to conclude on meeting the 3% goal requirement.

### 3.4. Land calibration sites

#### 3.4.1. RadCalNet stations results

For this analysis NPL provided analysis from the RadCalNet database, CNES and University of Arizona provided results from their own

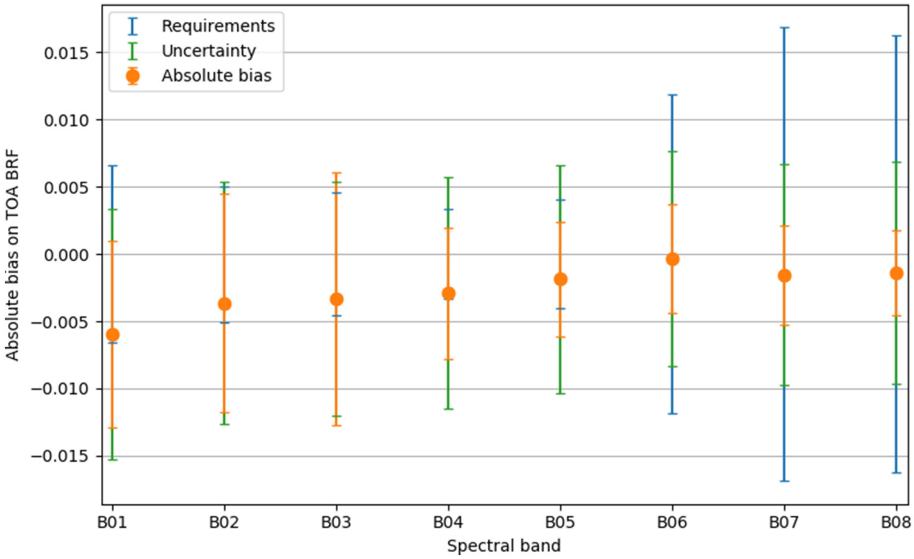

**Fig. 3.** Results for the low-radiance method applied to coastal water targets. For absolute bias, the mean and standard deviation are provided for the entire statistical series, alongside the uncertainty on simulation results and the absolute requirements.

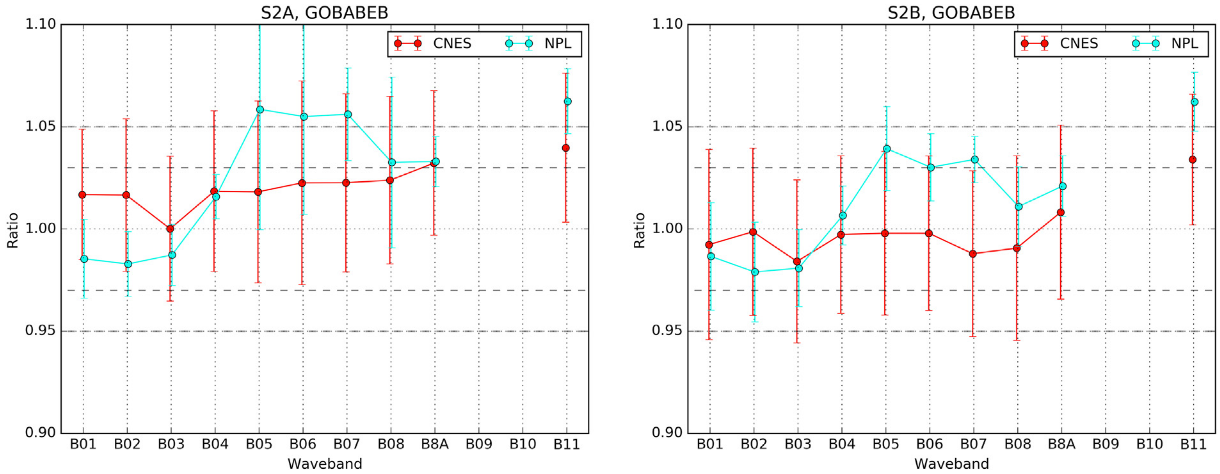

**Fig. 4.** Absolute ratios from in situ comparison on GOBABEB (CNES in red, NPL in blue). S2A (left) and S2B (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

datasets. RadCalNet datasets include associated figures of uncertainty along with the TOA reflectance measurements. For the NPL analysis, the Sentinel-2 uncertainty over the ROI has been obtained as in Gorroño et al. (2018). However, for now, results below are presented without including the RadCalNet uncertainty. This is because a full analysis of the uncertainty associated with the band-integrated and temporally interpolated values has not yet been performed. Further work is planned to consider the error correlation structure of the RadCalNet data and its impact over the accumulation of individual comparisons. In this paper, only the standard deviation of the individually determined Sentinel-2 to RadCalNet ratios is given, providing insight on the dispersion of the single measurements (Fig. 4).

a. Gobabeb results

Over Gobabeb there is a significant difference between the NPL (RadCalNet) results and the CNES results. The discrepancies for the lower wavelength bands — B01, B02, B03 — can be associated to discrepancies between the propagation of the surface reflectance to TOA. The propagation is set with 6SV for CNES results whereas the RadCalNet TOA propagation is calculated using MODTRAN®. In addition, different aerosol models and radiative transfer models have been applied.

For the narrow bands near the vegetation red-edge (B05, B06, and B07), the major discrepancy has been found to be due to the RadCalNet spectral bands. The propagation of the surface reflectance values to RadCalNet TOA reflectance values uses a MODTRAN® spectral configuration with a 10 nm sampling but 20 nm Full Width Half Maximum (FWHM). The RadcalNet TOA reflectance in the proximity of absorption areas is “smooth” due to the FWHM bands effect. When the signal is convolved by the narrow Sentinel-2 bands response functions and compared to the L1C data it results in a positive bias (not shown). Further work is on-going to determine its exact impact and potential mitigation although a preliminary assessment has estimated this effect at the 2%–3% level at the Gobabeb site. Other, more minor, effects can be associated to the closer temporal match of the CNES analysis, compared to interpolating the 30 min RadCalNet intervals, and the RadCalNet nadir view.

b. Railroad Valley

NPL compared Sentinel-2 to the official RadCalNet product using the same protocol as for Gobabeb, Namibia. The University of Arizona performed analysis using the full RadCaTS data. Results from both methodologies are shown in Fig. 5. Here NPL results are only provided for S2A, as there were insufficient match ups for S2B.

The two analyses agree better over Railroad Valley than the NPL (RadCalNet) analysis did with the CNES analysis over Gobabeb, especially in the blue spectral region. This is in part because, at Railroad Valley, there are more observations due to the S2 overpass from two different orbits. Also, results in bands B01, B02, and B03 are closer than for Gobabeb because both results on RRV use MODTRAN® for the propagation to TOA. The S2A B05, B06, and B07 bands do show a discrepancy around 2% - 3% between RadCaTS and RadCalNet methods, which are similar to the comparison to CNES analyses over Gobabeb. This is also likely to be due to the bias introduced by the FWHM of the RadCalNet TOA reflectance values when applied to the Sentinel-2 mission. Absorption peaks as well as narrow spectral bands introduce a systematic effect that cannot be reduced with the addition of further observations.

### 3.4.2. *Salar de Uyuni campaign (Bolivia)*

Fig. 6 shows, for the campaign in Bolivia, the ratios between the measured S2A-MSI reflectance values and the reference reflectances determined by propagating the measured ground reflectances to TOA using MODTRAN®. The results in Channels B01-B08 are spectrally flat and show that the reference reflectance is about 3% brighter than Sentinel-2A measurements. The reference reflectances are also brighter for Channels B11 and B12, but the disagreement is slightly larger at the 5% level. The result for the strong water vapour absorption channel (B10) is not shown due to large uncertainties of the reflectance-based validation technique in this spectral region. The result for the other water vapour absorption channel (B09, less absorbing) is kept for information though it seems not reliable as well.

### 3.5. *Pseudo-invariant calibration sites (PICS) results*

From NASA/GSFC, the Sentinel-2A comparison with Landsat-8 is based on the average of 18 coincident image pairs acquired between July 2015 and January 2018; the Sentinel-2B comparison is based on 8 coincident image pairs acquired between August 2017 and January 2018. The S2A-MSI agrees with L8-OLI to within 1% across all the common bands. The S2B-MSI data indicate a bias relative to L8-OLI of about 3% for B01 and B02 and about 1% for the other common bands (except for B11).

CNES results are based on > 1500 (S2A) or 500 (S2B) matchups with L8-OLI. Among the 19 desert PICS available in the database, about half have been used in this analysis because the observing and solar geometries from Sentinel-2 and Landsat-8 were sufficiently closely matched. CNES found that S2A-MSI agrees with L8-OLI to within 2%

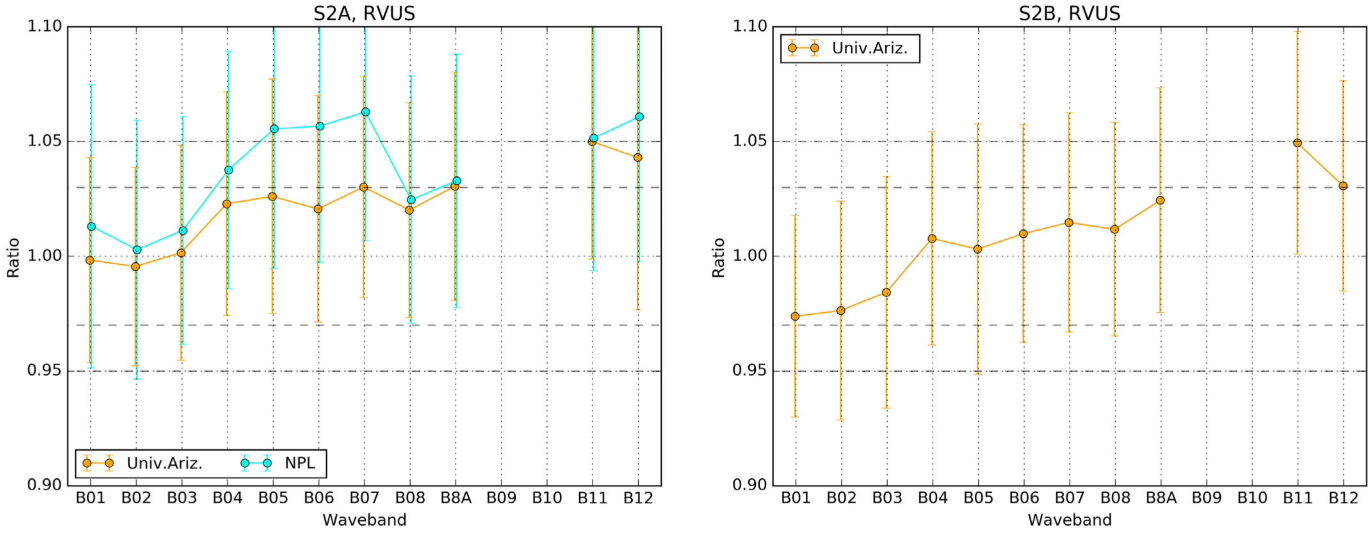

**Fig. 5.** Absolute ratios from in situ comparison on Railroad Valley (U. Ariz. orange, NPL blue). S2A (left) and S2B (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

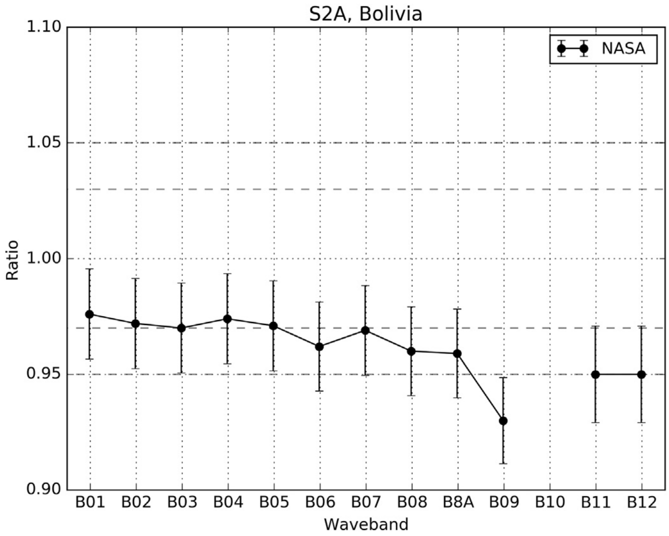

**Fig. 6.** Absolute ratios from comparison between the Salar de Uyuni campaign (Bolivia) and S2A.

except for B12 where the biggest difference is obtained with about 3%. S2B-MSI/L8-OLI agreement is less good, with cross-calibration ratio between 0% and −4% depending on the spectral band. The results indicate a possible mean bias of about 1% to 2% between the two sensors. In general terms, CNES and NASA/GSFC results are in good agreement within a few percent. In particular, they both report an indication for a small S2B-MSI/L8-OLI bias of the order of 1% to 2%. The small differences between the CNES and NASA/GSFC results can be explained by their respective methodologies. Basically, CNES uses a larger data set. The only criterion for data selection is based on geometric conditions, assuming a temporal stability of a site's reflectance. On the other hand, NASA/GSFC uses a very restricted data set because a temporal coincidence criterion has been added. Consequently, while less numerous, the NASA/GSFC matchups are of higher quality. Finally, the spectral band-adjustment correction, applied to compensate the sensors' spectral response differences, is not the same in the two methodologies.

Because there are a very large number of MERIS PICS acquisitions available (~22,000), many matchups have been found between MERIS and S2A (4700) and between MERIS and S2B (1500). A very good agreement is noticed between MERIS and S2A, typically within 1%. S2B agreement with MERIS is less good, with an indication for a bias between the two sensors between −1% and −2%. While this bias only appears in ARGANS analysis in bands B01 to B04, the same trend is observed in the ARGANS results as in the CNES results: the S2B/MERIS calibration ratios are systematically lower than the S2A/MERIS ones.

APICS results from the University of South Dakota and USGS, also shown in Fig. 7, are based on data collected from August 2015 to October 2017 for S2A and July 2017 to September 2017 for S2B. All results are within the 3% accuracy of the method and, except for the red band in S2A and the deep blue band in S2B, are within 1% of the calibration for each instrument.

## 4. Discussion

In Section 3 we presented the results of five methods to validate the radiometric calibration of the MSI sensors on the S2A and S2B platforms, with multiple implementations of these different methods by different sensor expert groups. Section 3 showed differences between different implementations of the same method for each sensor. Fig. 8 summarises all "absolute radiometric" results for S2A (i.e. results of methods that can validate the radiometric gain of each band independently) onto a single figure. This shows that all results validate the S2 threshold requirement (agreement between S2A MSI and validation approaches is better than 5%) and many results validate the S2 goal requirement (3%). However, the spread of results also shows that these methods need to be developed and compared further to provide consistent validation for sensors below the 3% level.

As in previous applications of these methods, a full uncertainty analysis has not been performed for all the methods presented here, and in particular, not for the combination of multiple comparisons (mean comparison). But with a growing need to provide validation and comparison of sensors at better than 3%, more rigorous uncertainty analysis is also required. Such an uncertainty analysis would require the uncertainty associated with the averaged observation in a selected ROI (Gorroño et al., 2018), as well as uncertainties associated with the validation data, with the comparison process (e.g. for temporal, spectral and spatial mismatch) and an understanding of the error correlation structure to enable the uncertainty associated with the mean of multiple comparisons to be correctly determined.

Of the absolute radiometric methods shown in Fig. 8, the Rayleigh methods show the best consistency between the two implementations, with discrepancies between the two implementations of 2%–3% and

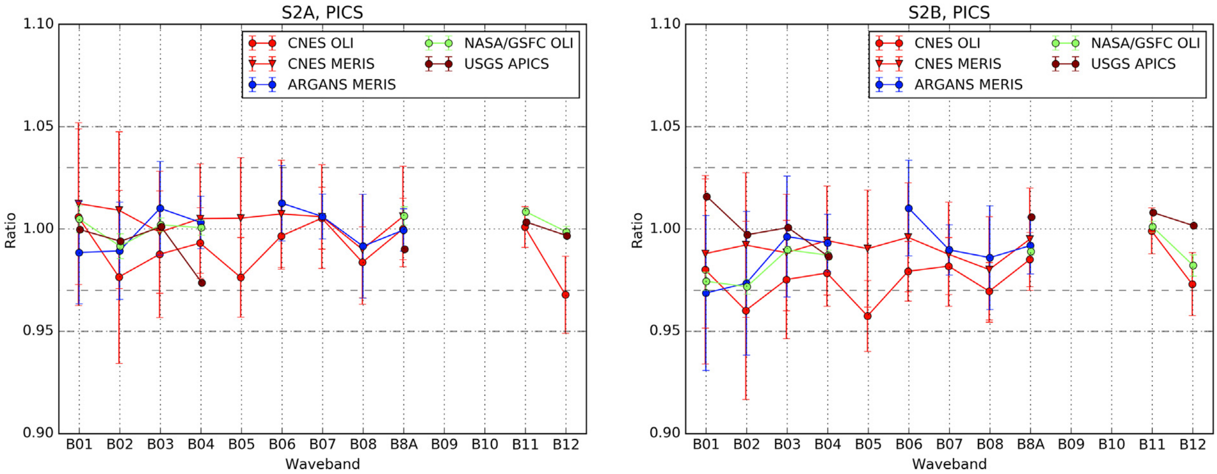

**Fig. 7.** Absolute ratios from all PICS methods. S2A (left) and S2B (right).

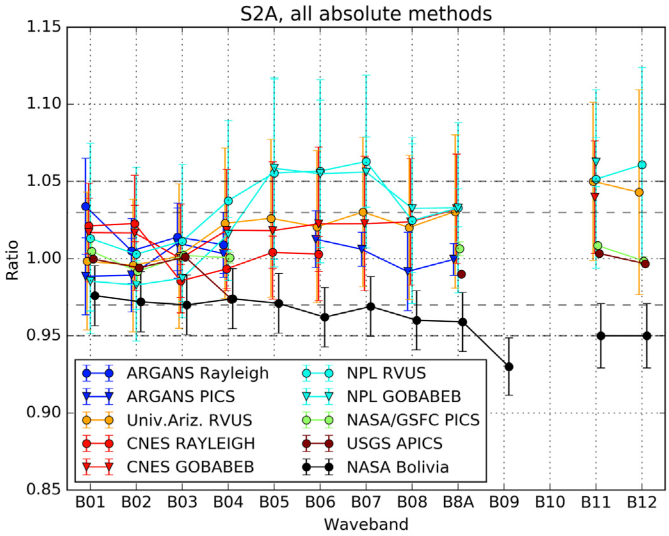

**Fig. 8.** Results from all absolute methodologies combined and compared. Error bars provide a standard deviation of different comparisons, and not a complete uncertainty analysis.

low dispersion of results for different sites (error bars). The approach is "statistical", it can rapidly collect statistics for a large number of matchups between reference sites and the S2 sensors, and can provide results over the entire FOV quite rapidly, although the number of matchups is reduced on the Eastern part of the FOV, where pixels are affected by Sun glint (see Revel et al., 2019). A deeper analysis across the FOV is necessary to disentangle differences due to modelling from differences due to sampling strategies. A comparison of methodologies, over a common test framework, would be useful to that regard.

PICS methods show similar results with dispersions and discrepancies about 3%. As with the Rayleigh method, the PICS methods have a reasonably large number of matchups. A clear advantage, compared to the Rayleigh method, is the full spectral coverage of the solar spectrum. On the other hand, PICS targets are much more heterogeneous and require a more accurate a priori knowledge of the target BRDF.

The least consistent methods are those based on matchups with in situ measurements. Here the spread of results for a single method is about 3%–5% and differences between methods are sometimes as large as 5%. The characterisation of the target BRDF, the spectral sampling of the measurements, as well as the choice of the RTM used to propagate the measurements to TOA are the strongest contributors in the discrepancies between the methodologies and should be further investigated to improve these techniques. In particular, the RadCalNet data used as provided on the RadCalNet portal ("NPL" approaches in Fig. 8), suffer from spectral resolution issues, and, partly as a result of these results, the RadCalNet Work Group is now planning on changing the bandwidth from 20 nm to 10 nm FWHM.

The main advantage of these in situ matchup methods, compared to the other methods, is that they have the potential for providing the best characterisation of the site, with well-calibrated, SI-traceable instrumentation providing the most reliable inputs. One clear drawback is the small number of matchups for low-earth orbit sensors (limited by the number of overpasses and the cloudiness of the sites), and the fact that the targets usually cover only a small part of the instrument FOV. As for the other methods, the aggregation of many individual comparisons is necessary to obtain statistical convergence. The single result obtained from the Salar de Uyuni campaign is representative of any individual result from the other implementations of this method. A single result can indeed appear different from the others, but an accumulation of similar analyses may provide convergence toward more similar results.

Fig. 8 does not provide data for the low-signal coastal water method, discussed in Section 3.3, as this comparison was performed in terms of absolute rather than relative reflectance. The method showed good agreement between S2A and S2B and the modelled water-leaving reflectance. The method provides valuable information on the full dynamic range of the MSI instruments.

As well as the absolute gain comparisons described above, the methods can also be used to validate interband consistency, which is also evaluated by the DCC method. As expected, DCCs, with the whitest targets, highest SNR and simplest assumptions, provide the best results for the interband ratios. The two independent implementations gave very consistent results (differences < 0.5%). These results (Fig. 1) show interband consistencies (compared to B04) of well under 1% (with the possible exception of B03, where one result showed interband differences of just over 1%).

Fig. 9 presents a similar interband analysis for the absolute methods, presenting the results shown previously translated into interband ratios, relative to the reference band B04. The comparison with the results obtained from the DCC methods provides evidence of interband biases in the other methods.

As expected, the variability of the results underlines the limitations inherent of all methodologies (and not necessarily specific of all methods). Interestingly, the single result from the Salar de Uyuni campaign (Bolivia) provides the closest similarity to the one from the DCC comparisons. The "NPL" results, which use the official RadCalNet product, show significant discrepancies in the bands that are most

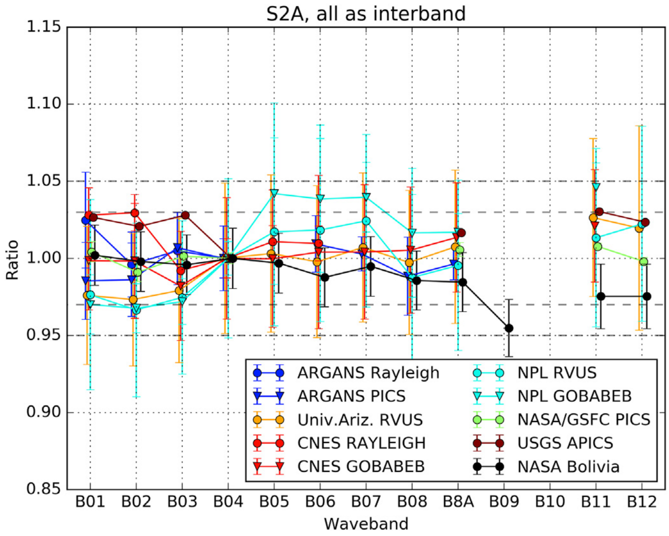

**Fig. 9.** Results from all absolute methodologies combined and compared as interband ratios to the reference band B04.

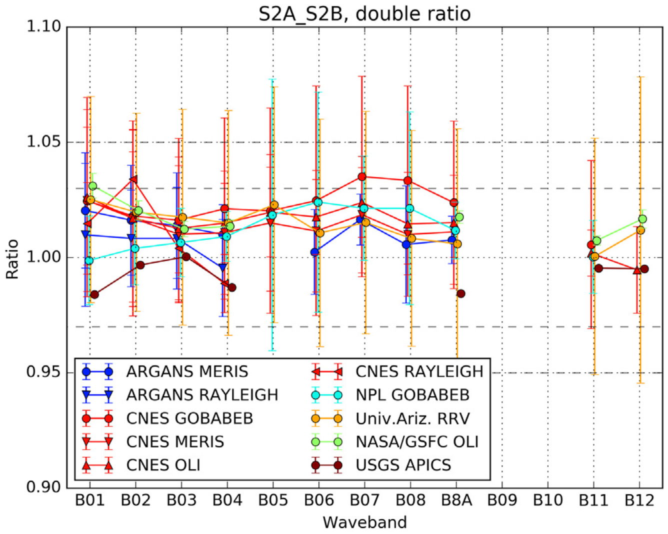

**Fig. 10.** Double ratios g(S2A)/g(S2B) for each calibration validation method.

sensitive to the RadCalNet bandwidth, as discussed above, and the methods based on the physical Rayleigh model, show very good interband consistency.

Combining the understanding that has been gained by comparing the different methods, Table 4 provides a summary of the benefits and drawbacks of each method, as well as potential improvements regarding either the method itself and/or the comparison exercise performed in this paper.

Bringing together different approaches in this way has enabled us to understand the capabilities and limits of the different methods. For a single sensor, the methods can provide evidence that the sensor meets the specifications, but would need improving, and more rigorous uncertainty analysis, to test the gain at a lower uncertainty level. However, many of the sources of uncertainty lead to systematic effects that can cancel out when applying the methodologies as transfer targets between the two MSI sensors. Indeed, absolute calibration validations, independently for S2A and S2B, allow an estimation of the inter-sensor calibration between S2A and S2B, an approach referred to as "double-ratio" (as in e.g. Revel et al., 2019). As S2A and S2B cannot be co-registered there is no opportunity to intercompare the two sensors directly. By comparing to an external reference, indirect comparisons can be made as an interesting alternative.

Fig. 10 presents results from such indirect comparisons for all the appropriate methods described earlier. Results are presented as ratios of the individual calibration ratios from any absolute calibration method $g_A(\lambda)/g_B(\lambda)$.

All results, except the APICS method, exhibit values higher than unity in the VNIR (bands $<$ B09). This means that S2A reflectances are brighter than those of S2B with a bias of about 1%–2%, in agreement with the results of Revel et al. (2019). The SWIR bands do not show this difference.

The observed difference between S2A and S2B could be due to a diffuser BRDF effect (the differences are similar to the uncertainties associated with the BRDF measurement) or possibly the baffle straylight. Although the two instruments were manufactured in the same way, there could be some differences in how the straylight goes through. Pre-flight modelling suggests, for example, straylight errors during diffuser acquisition could be up to 0.7%, so it would be difficult to explain a 1.5% difference between the sensors that are nominally identical. The S2 measurement uncertainties are generally smaller for the VIS than for the SWIR. However, the 1.5% difference could point to a bias during the VIS characterisation. This bias does not seem to affect the SWIR BRDF characterisation which has been done separately.

The double-ratio differences for the RadCalNet Gobabeb results present a global value around the 1.5% but this value fluctuates in the 0%–2.5% range depending on the bands. It is plausible that such results could come from the effects of the spectral resolution and aerosol modelling discrepancies as discussed in the Gobabeb section, which may affect not only the absolute level but also the relative one, if the atmospheric conditions are changing between the two different Sentinel overpasses.

**Table 4**
Summary of benefits and drawbacks of all methods/targets applied in this paper, potential improvement to be considered.

| Method/target | Benefits | Drawbacks | To be considered for improvement |
|---|---|---|---|
| Deep convective clouds | Bright signal, lowest complexity for modelling. Best interband. High statistics. | Only interband or long-term monitoring with a reference. Currently VNIR. | Extension to an absolute method covering VNIR and SWIR is desirable. |
| Rayleigh | High statistics covering all the FOV. Low complexity of the target. | Contamination by Sun glint reduces statistics at some parts of the FOV. Not full wavelength coverage. Need of water reflectance a priori knowledge. | Comparisons needed independently for different ROIs and within the FOV to disentangle systematic model effects from sampling effects. |
| Coastal waters | Good knowledge of the surface and atmospheric properties using ground fiducial measurements (SI-traceable). | Higher complexity of the target than for Rayleigh, needs RTM finely tuned. Low statistics. | More statistics shall provide better results as statistical convergence is very slow for small targets. |
| PICS | High statistics, low complexity for model. | Needs a comparison to other satellite database (not ground truth) which relies on the uncertainty of this database (including potential calibration residual bias). | One absolute reference BRDF model would be welcome. |
| Land calibration equipped sites | Best knowledge of the surface and atmospheric properties from fiducial ground measurements (SI-traceable). | Very small statistics and very partial FOV coverage. | Improvements in RTM and BRDF models, and, in some cases, spectral sampling. |

## 5. Conclusions

The ESA S2RadVal workshop provided a great opportunity to gather independent Sentinel-2 experts to compare radiometric validation methods. Such an exercise is to be repeated and encouraged for any optical remote sensing mission in the solar spectrum range.

All the methods presented here investigated the use of natural targets to monitor interband, absolute, and cross-sensor radiometric calibrations as well as validating for low radiances. The combined methods provide evidence of the excellent radiometric performance of the Sentinel-2 mission (both S2A and S2B sensors) with overall results always achieving threshold requirements (5%) and sometimes achieving goal requirements (3%). The interband validation (1%–2%) is better than the goal requirement.

The relatively large discrepancies between the presented methodologies are representative of the difficulty to reach below the 3% goal accuracy for validation at this stage. We have discussed possible approaches to improve these vicarious methods. In the meantime, the methods can only be used for validation, rather than for calibrating an instrument aiming for sub 3% absolute accuracy. This supports the continuing need for on-board calibration devices, particularly as the vicarious methods rely on the statistical collation of many months of observation to obtain meaningful comparisons, and therefore these methods cannot produce accurate results on the same monthly-based timescales of the on-board systems.

Despite S2B having only recently been launched, and despite the fact the two satellites never have direct matchups, a comparison of S2A and S2B is possible through the double-ratios estimation of intersensor calibration. This provides overall agreement on a bias of about 1–2% between S2A and S2B in the VNIR and none in the SWIR, S2A radiometry being slightly brighter than S2B.

## Acknowledgments

This work, part of the S2RadVal project, was supported by the European Space Agency under Grant 4000116454/16/I-Sbo. Science Systems and Applications, Inc. work was performed under NASA contract NNG15HQ01C. The University of Arizona was funded by NASA Research Grants NNX16AH44G and NNX16AL25G, and USGS Cooperative Agreement G14AC00371. NPL received funding from the MetEOC-3 project from the EMPIR programme co-financed by participating states and from the European Union's Horizon 2020 Research and Innovation Framework Programme.

We would like to thank the following people involved in the production and distribution of AERONET-OC data: Giuseppe Zibordi (Venise, Gustav_Dalen_Tower, Galata_Platform, Gloria, Helsinki_Lighthouse sites), Sam Ahmed and Alex Gilerson (LISCO site), Hui Feng and Heidi M. Sosik (MVCO site), Dimitry Van der Zande (Thornton_C-power site), Burton Jones and Curtiss Davis (USC_SEAPRISM site), as well as Alan Weidemann, Bill Gibson, and Robert Arnone (WaveCIS_Site_CSI_6 site).

Finally, we are grateful to three anonymous reviewers for providing very constructive comments that led to better organization and consolidation of this paper.

## References

Alhammoud, B., Jackson, J., Clerc, S., Arias, M., Bouzinac, C., Gascon, F., Cadau, E.G., Iannone, R., 2018. Sentinel-2 level-1 radiometry validation using vicarious methods from DIMITRI database. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium (IGARSS 2018), Valencia, Spain.

Anderson, N., Czapla-Myers, J., Leisso, N., Biggar, S., Burkhart, C., Kingston, R., Thome, K., 2013. Design and calibration of field deployable ground-viewing radiometers. Appl. Opt. 52 (2), 231–240.

Anderson, N.J., Czapla-Myers, J.S., 2013. Ground viewing radiometer characterization, implementation and calibration applications: A summary after two years of field deployment. In: Proc. SPIE 8866, 88660N–88660N-10.

Barker, K., Marrable, D., Hedley, J., Mazeran, C., 2014. DIMITRI Algorithm Theoretical Basis Document. Rayleigh Scattering Methodology for Vicarious Calibration MO-SCI-ARG-TN-004b v 1.0.

Barsi, J.A., Alhammoud, B., Czapla-Myers, J., Gascon, F., Haque, M.O., Kaewmanee, M., Leigh, L., Markham, B.L., 2018. Sentinel-2A MSI and Landsat-8 OLI radiometric cross comparison over desert sites. European Journal of Remote Sensing 51 (1), 822–837. https://doi.org/10.1080/22797254.2018.1507613.

Berk, A., Anderson, G.P., Acharya, P.K., Bernstein, L.S., Muratov, L., Lee, J., Fox, M., Adler-Golden, S.M., Chetwynd, J.H., Hoke, M.L., Lockwood, R.B., Gardner, J.A., Cooley, T.W., Borel, C.C., Lewis, P.E., 1 June 2005. MODTRAN 5: A reformulated atmospheric band model with auxiliary species and practical multiple scattering options: Update. In: Proc. SPIE 5806, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, https://doi.org/10.1117/12.606026. https://doi.org/10.1117/12.606026.

Bouvet, M., 2014. Radiometric comparison of multispectral imagers over a pseudo-invariant calibration site using a reference radiometric model. Remote Sens. Environ. 140, 141–154.

Cabot, F., Hagolle, O., Henry, P., 2000. Relative and multitemporal calibration of AVHRR, SeaWiFS, and VEGETATION using POLDER characterization of desert sites. In: Proc. Int. Geosci. Remote Sens. Symp., Honolulu, HI. 5. pp. 2188–2190.

Chander, G., Mishra, N.G., Helder, D.L., Aaron, D.B., Angal, A., Choi, T., Xiong, X., Doelling, D.R., 2013. Applications of spectral band adjustment factors (SBAF) for cross-calibration. IEEE Trans. Geosci. Remote Sens. 51 (3), 1267–1281. https://doi.org/10.1109/TGRS.2012.2228007.

Czapla-Myers, J., McCorkel, J., Anderson, N., Thome, K., Biggar, S., Helder, D., Aaron, D., Leigh, L., Mishra, N., 2015. The ground-based absolute radiometric calibration of Landsat 8 OLI. Remote Sens. 7 (1), 600–626.

Czapla-Myers, J.S., Thome, K.J., Buchanan, J.H., 2007. Implication of spatial uniformity on vicarious calibration using automated test sites. Proc. SPIE 6677, 66770U–10.

Czapla-Myers, J.S., Thome, K.J., Cocilovo, B.R., McCorkel, J.T., Buchanan, J.H., 2008. Temporal, spectral, and spatial study of the automated vicarious calibration test site at Railroad Valley, Nevada. In: Proc. SPIE 7081, 70810I-9.

ESA Sentinel-2 Team, 2007. GMES Sentinel-2 Mission Requirements Document. EOP-SM/1163/MR-dr v2.0. https://earth.esa.int/pub/ESA_DOC/GMES_Sentinel2_MRD_issue_2.0_update.pdf.

Fougnie, B., Bach, C., 2009. Monitoring of radiometric sensitivity changes of space sensors using deep convective clouds: operational application to PARASOL. IEEE Trans. Geosci. Remote Sens. 47 (3) MARCH.

Fougnie, B.Henry, P. Absolute Calibration Using Rayleigh, QA4EO-WGCV-IVO-CLP-007. http://qa4eo.org/documentation.html.

Fougnie, B., Llido, J., Gross-Colzy, L., Henry, P., Blumstein, D., 2010. Climatology of oceanic zones suitable for in-flight calibration of space sensors. In: Proceedings Earth Observing Systems VX. SPIE.

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Fernandez, V., 2017. Copernicus sentinel-2A calibration and products validation status. Remote Sens. 9, 584. https://doi.org/10.3390/rs9060584.

Gobron, N., Pinty, B., Verstraete, M.M., Govaerts, Y.M., 1996. A semi-discrete model for the scattering of light by vegetation. J. Geophys. Res. 102, 9431–9446.

Gorroño, J., Hunt, S., Scanlon, T., et al., 2018. Providing uncertainty estimates of the Sentinel-2 top-of-atmosphere measurements for radiometric validation activities. European Journal of Remote Sensing 51 (1), 650–666. https://doi.org/10.1080/22797254.2018.1471739.

Hagolle, O., Goloub, P., Deschamps, P.-Y., Cosnefroy, H., Briottet, X., Bailleul, T., Nicolas, J.-M., Parol, F., Lafrance, B., Herman, M., 1999. Results of POLDER in-flight calibration. IEEE Trans. Geosci. Remote Sens. 37 (3), 1550–1566.

Holben, B.N., Eck, T.F., Slutsker, I., Tanre, D., Buis, J.P., Setzer, A., Vermote, E., Reagan, J.A., Kaufman, Y., Nakajima, T., Lavenu, F., Jankowiak, I., Smirnov, A., 1998. AERONET - a federated instrument network and data archive for aerosol characterization. Rem. Sens. Environ. 66, 1–16.

Hu, Y.B., Wielicki, B.A., Yang, P., Stackhouse Jr., P.W., Lin, B., Young, D.F., 2004. Application of deep convective cloud albedo observations to satellite-based study of terrestrial atmosphere: monitoring the stability of spaceborne measurements and assessing absorption anomaly. IEEE Trans. Geosci. Remote Sens. 42 (11), 2594–2599.

Kaewmanee, M., Helder, D., (2017), Refined Absolute PICS Calibration Model Over Libya-4 Using Sentinel2A and Landsat 8 Collection-1 Data for Validation, PECORA 20, Sioux Falls, SD Nov 13–16, 2017.

Kotchenova, S. Y., Vermote, E. F., Matarrese, R., and Klemm, 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: 6762–74.

Lachérade, S., Fougnie, B., Henry, P., Gamet, P., 2013. Cross calibration over desert sites: description, methodology and operational implementation. IEEE Trans. Geosci. and Remote Sensing 51 (3).

Lamquin, N., Bruniquel, V., Gascon, F., 2017. Sentinel-2 L1C radiometric validation using deep convective clouds observations. European Journal of Remote Sensing 51 (1), 11–27. https://doi.org/10.1080/22797254.2017.1395713.

Maisonobe, L., Pommier-Maurussane, V., 2010. Orekit: An Open-Source Library for Operational Flight Dynamics Applications, 4th ICATT, May 2010.

Marcq, S., Meygret, A., Bouvet, M., Fox, N., 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, Submitted to IGARSS 2018.

Markham, B.L., Helder, D.L., 2012. Forty-year calibrated record of earth-reflected radiance from Landsat: a review. Remote Sens. Environ. 122, 30–40.

Mishra, N., Helder, D.L., Angal, A., Choi, J., Xiong, X., 2014. Absolute calibration of optical satellite sensors using Libya 4 Pseudo invariant calibration site. Remote Sens. 6, 1327–1346.

Morel, A., Maritorena, S., 2001. Bio-optical properties of oceanic waters: a reappraisal. J.

of Geophys. Res. 106, 7763–7780.

Philipson, P., Kratzer, S., Ben Mustapha, S., Strömbeck, N., Stelzer, K., 2016. Satellite-based water quality monitoring in Lake Vänern, Sweden. Int. J. Remote Sens. 37 (16), 3938–3960. https://doi.org/10.1080/01431161.2016.1204480.

Revel, C., Lonjou, V., Marcq, S., Desjardins, C., Fougnie, B., Coppolani-Delle Luche, C., Guillemot, N., Lacamp, A.-S., Lourme, E., Miquel, C., Lenot, X., 2019. Sentinel-2A and 2B absolute calibration monitoring. European Journal of Remote Sensing 52 (1), 122–137. https://doi.org/10.1080/22797254.2018.1562311.

Teillet, P.M., Fedosejevs, G., Thome, K.J., Barker, J.L., 2007. Impacts of spectral band difference effects on radiometric cross-calibration between satellite sensors in the solar-reflective spectral domain. Remote Sens. Environ. 110, 393–409. https://doi.org/10.1016/j.rse.2007.03.003.

Thome, K.J., 2001. Absolute radiometric calibration of Landsat 7 ETM+ using the reflectance-based method. Remote Sens. Environ. 78 (1–2), 27–38.

Thuillier, G., Hersé, M., Labs, D., Foujols, T., Peetermans, W., Gillotay, D., Simon, P.C., Mandel, H., 2003. The solar spectral irradiance from 200 to 2400 nm measured by the SOLSPEC spectrometer from the ATLAS and EUREKA missions. Sol. Phys. 214, 1–22.

Vaidya, A., Helder, D., Mishra, N., 2014. Absolute Radiometric Calibration Using Pseudo Invariant Calibration Sites, JACIE 2014, Louisville Kentucky, March 26–28.

Vermote, E., Kaufman, Y.J., 1995. Absolute calibration of AVHRR visible and near infrared channels using ocean and cloud views. Int. J. Remote Sens. 16 (13), 2317–2340.

Vermote, E., Santer, R., Deschamps, P.-Y., Herman, M., 1992. In-flight calibration of large field of view sensors at short wavelengths using Rayleigh scattering. Int. J. Remote Sensing 13, 3409–3429.

Wielicki, B.A., et al., 2013. Achieving climate change absolute accuracy in orbit. Bull. Amer. Meteor. Soc. 94, 1519–1539. https://doi.org/10.1175/BAMS-D-12-00149.1.

Zibordi, G., Holben, B., Slutsker, I., Giles, G., D'Alimonte, D., Melin, F., Berthon, J.F., Vandemark, D., Feng, H., Schuster, G., Fabbri, B.E., Kaitala, S., Seppala, J., 2009. AERONET-OC: a network for the validation of ocean color primary products. J. Atmos. and Oceanic Technology. 26, 1634–1651. https://doi.org/10.1175/2009JTECHO654.1.