The fifth phase of the Radiation Transfer Model Inter-Comparison exercise (RAMI-V): experiment description and results on actual canopy scenarios

Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

The Fifth Phase of the Radiation Transfer Model Intercomparison Exercise (RAMI-V): Experiment Description and Results on Actual Canopy Scenarios

Christian Lanconelli[1], Nadine Gobron[2*], Monica Robustelli[3], Jennifer Susan Adams[4], Kim Calders[5], Mathias Disney[6,7], Jean-Philippe Gastellu-Etchegorry[8], Adam Goodenough[9], Yves Govaerts[10], Robin J. Hogan[11,12], Huaguo Huang[13], Hideki Kobayashi[14], Andres Kuusk[15], Vincent Leroy[10], Niall Origo[16], Jianbo Qi[17], Sebastian Schunke[10], Martin van Leeuwen[18,19], Yingjie Wang[8], Donghui Xie[17], Yelu Zeng[20,21], and Feng Zhao[22]

[1]Uni Systems Italy, Milan, Italy. [2]European Commission, Joint Research Centre (JRC), Ispra, Italy. [3]Westpole SPA, Milan, Italy. [4]Remote Sensing Laboratories, Department of Geography, University of Zürich, Zurich, Switzerland. [5]Q-ForestLab, Department of Environment, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium. [6]Department of Geography, University College London, London, UK. [7]NERC National Centre for Earth Observation, UCL Gower Street, London WC1E 6BT, UK. [8]Univ Toulouse 3 Paul Sabatier, Univ Toulouse, CNES/IRD/CNRS/INRAE, CESBIO, Toulouse, France. [9]Digital Imaging and Remote Sensing Laboratory, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623, USA. [10]Rayference, 1000 Brussels, Belgium. [11]European Centre for Medium-range Weather Forecasts, Reading, UK. [12]Department of Meteorology, University of Reading, Reading, UK. [13]State Forestry and Grassland Administration Key Laboratory of Forest Resources and Environmental Management, Beijing Forestry University, Beijing 100083, China. [14]JAMSTEC—Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan. [15]University of Tartu, Tartu Observatory, Tõravere 61602, Estonia. [16]Climate and Earth Observation Group—National Physical Laboratory, Teddington, UK. [17]State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China. [18]Department of Geography, University at Buffalo, Buffalo, NY 14261, USA. [19]Van Leeuwen Imaging, Diedenweg 63-II, 6706CH Wageningen, Netherlands. [20]Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA. [21]State Key Laboratory of Remote Sensing Science, Jointly Sponsored by Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences and Beijing Normal University, Beijing 100101, China. [22]School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, Beijing 100191, China.

*Address correspondence to: nadine.gobron@ec.europa.eu

Citation: Lanconelli C, Gobron N, Robustelli M, Adams JS, Calders K, Disney M, Gastellu-Etchegorry JP, Goodenough A, Govaerts Y, Hogan RJ, et al. The Fifth Phase of the Radiation Transfer Model Intercomparison Exercise (RAMI-V): Experiment Description and Results on Actual Canopy Scenarios. *J. Remote Sens.* 2025;5:Article 0663. https://doi.org/10.34133/remotesensing.0663

Submitted 12 December 2024
Revised 13 May 2025
Accepted 18 May 2025
Published 29 July 2025

This paper presents the latest results of the radiation transfer model intercomparison (RAMI) of the realistic vegetation scenarios. RAMI-V included the same one-dimensional (1D) and 3D scenes of RAMI-IV phase and 2 new realistic ones, defined through a semiparametric (Savanna) and an empirical (Wytham Woods) approaches. The measurements to simulate were the bidirectional reflectance factor, directional–hemispherical reflectance, and bidirectional–hemispherical reflectance. In addition, the radiant flux transmission and absorption through and below the canopy and digital hemispherical photography were also proposed. The spectral bands were defined to mimic not only the ones of Copernicus optical missions, e.g., for the Sentinel-3 Ocean and Land Colour Imager (OLCI) and Sentinel-2 Multispectral Instrument (MSI), but also the Moderate Resolution Imaging Spectroradiometer (MODIS). New solar and viewing geometry configurations were adopted from realistic satellite overpasses for different seasons and geographical locations. The role of internal consistency checks was reinforced to provide more reliable feedback to the participants in the early stage of the experiment and reduce the role of outliers in the model-to-model comparison and the identification of a surrogate reference. Over 4 of the 8 scenarios proposed, a set of models agreed within 2% uncertainty thresholds for most of the virtual measurements defined in the experiment. Specifically, they were the birchstand both leaf-on (HET09) and leaf-off (HET15) versions, and the structured canopy models consisting of a citrus orchard (HET14) and a poplar forest (HET16). It is noteworthy that *less* was among the models designated to set a reference benchmark across all chosen instances. Besides, *dart*, *raytran*, and *wps* were contributing to the benchmark in most of the experiment proposed, especially referring to total BRF and DHR, and total absorption, while for the transmittance the results were more dispersed. *Dart*, *less*, *raytran*, and *wps* contributed by submitting 100%, 83.9%, 99.4%, and 86.2% of the experiment proposed, respectively. The proficiency testing of the models was performed by means of the z' metric defined in ISO-13528. A custom reference, based on a selection of models that showed the best agreement, as well as a reference based on robust statistic were adopted. Above the aforementioned selected scenes, and assuming a compliance threshold of 3% (5%) for bidirectional reflectance (albedo) measurements, *dart*, *less*, and *raytran* were in agreement in all (more than 95%) cases. The approach based on the robust statistic described in ISO-13528 confirmed its relevance in interlaboratory comparison exercises where the benchmark is not defined a priori, allowing us to obtain proficiency results equivalent to those defined against the customized references.

Introduction

The radiation transfer model intercomparison (RAMI) exercise (https://rami-benchmark.jrc.ec.europa.eu) was designed and implemented originally by the international radiative transfer (RT) community to benchmark the RT models used to simulate radiative measurements over plant canopy surfaces [1]. RAMI continued over the last 25 years to assess the model uncertainties and their compliance against the most updated requirements set by the Earth Observation scientific community [2]. Intercomparison exercises aims to support the modeling community in the validation and model physics development process, and to develop a community consensus on the best ways to simulate radiation transfer over different scenarios, which is important for the interpretation of remote sensing as well as in situ data.

RAMI was operated in successive phases, each one aiming at reassessing the capability, performance, and agreement of one-dimensional (1D) and 3D RT models, by expanding the set of experiments proposed and increasing their complexity. The first phase of RAMI (RAMI-1), issued in 1999, had the prime objective to document the variability between canopy reflectance model results under well-controlled experimental conditions [1]. The number of experiments was expanded to focus on the performance of RT models dealing with structurally complex 3D plant environments. RAMI-2 faced an increase in the number of participating models, and a better agreement between simulations for the structurally simple scenes inherited from RAMI-1 was observed, while the strong divergence of some 3D RT models over complex heterogeneous scenes was highlighted [3]. RAMI-3 [4] faced a further increase in the number of participants and experiments. The self-consistency (e.g., energy conservation) together with the absolute and relative performance of RT models were evaluated in detail [5]. It became possible to demonstrate, for the first time, a convergence of the whole set of submitted RT simulations and to document an agreement better than 1% between 6 of the participating 3D Monte Carlo RT models on heterogeneous and homogeneous abstract canopies, which allow to establish the bases for the RAMI On-line Model Checker (ROMC) (https://romc.jrc.ec.europa.eu/; [6]). This web tool enables model developers to autonomously verify the quality of their RTM against a reference for different scenarios, arising from the results of previous RAMI phases.

Widlowski et al. [7] proposed the use of ISO-13528 [8] in RAMI-IV to formalize the evaluation of the models' performances when the truth results are not known, by using a reliable conventional reference value. The pre-screening of data, the identification of reference solutions, and the choice of proficiency statistics were illustrated on simulation results from the RAMI-IV abstract canopy scenarios only. According to ISO-13528, proficiency testing is the evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons. After a series of initial consistency checks, this procedure involved (a) the definition of a tolerance criteria suitable for the determination of proficiency of RT models, (b) the definition of a surrogate reference solution against which the candidate models could be compared, and (c) the selection of appropriate evaluation metrics to quantify the performance of the models. During RAMI-IV phase, the complexity of the scenes was extended with actual scenarios, which were built considering realistic models of leaves and setting up trees including their wooden components. The simulation performed over the 6 actual realistic canopies of RAMI-IV showed much greater variance than those analyzed for the abstract canopy scenarios [9]. For these scenarios, the identification of a set of credible models, similar to that found for the abstract canopies during RAMI-3, failed because of the large spread among model results. Moreover, some RT models submitted simulation results for less than a quarter of the canopy architectures considered in RAMI-IV, preventing statistically significant conclusions for some of the experiments.

Whether the differences were caused by operator errors/choices or were intrinsic to the physical formulation or its implementation could not be determined. As a matter of fact, in RAMI-IV, some of the models that showed the largest deviation have never participated in previous phases, confirming the important role of open intercomparison exercises even in terms of model development. The evolution of the results from RAMI-1 to RAMI-IV showed that the repetition of a set of experiments in successive intercomparison rounds leads to a progressive improvement of most models, as developers gradually identify and remove model physics weaknesses, scenario implementation bugs, and generic software errors. It should also be considered that the range of complexity (and computational cost) can vary hugely between models: "Explicit" models aimed to perform reference calculations instantiate every object of a scene with a detailed realistic representation, whereas models to be used operationally parameterize the plant geometry in various ways to accomplish fast calculations. To give an idea, Stretton et al. [10] found that in urban areas the explicit *Dart* model was around 7 orders of magnitude more computationally expensive than the parametric *Spartacus* model. Intercomparison exercises are then important both to establish reference datasets from explicit models and to evaluate and improve the more approximate models.

RAMI-V maintained both abstract and actual scenarios and experiment definitions of RAMI-IV, while it introduced 2 additional actual scenes and modified the measurement configurations to adapt them to the European Union (EU) Copernicus program related to the passive remote sensing of land and vegetation in the solar spectrum. Specifically, Sentinel-2 Multispectral Imager (MSI) (https://sentiwiki.copernicus.eu/web/s2-mission; [11]) and Sentinel-3 Ocean and Land Color Imager (OLCI) (https://sentiwiki.copernicus.eu/web/s3-olci-instrument; [12]) bands and observation geometries have been considered to set up the virtual experiments. Additionally, similar information from the Moderate Resolution Imaging Spectroradiometer (MODIS) (https://modis.gsfc.nasa.gov/; [13]) has been considered to extent the study to middle-infrared spectral bands.

Experiment Description describes the experiment conceptualization in terms of (a) vegetation scenarios, (b) spectral properties associated to the canopy model primitives, (c) illumination characteristics, and (d) the measures to be performed for the actual canopy scenarios. Participation and Model Results provides a summary of the participant models and an overview of the submitted dataset for all actual scenes and measurements.

Model Intercomparison Analysis describes the analysis methodology, including the description of the review phase, established to support participants in identifying weird errors affecting their simulations with respect to the ensemble. It also describes (a) a set of internal model consistency checks, which was expanded with respect to the previous phase, (b) the model-to-model intercomparison, and (c) the process followed to determine a credible model ensemble, allowing us to approach a model-to-ensemble discussion (model proficiency), which was based here on the k' statistic, defined by ISO-13528.

Discussion and Concluding Remarks is dedicated to a discussion of the results of the experiment, with a focus to the new achievements with respect to RAMI-IV results.

Materials and Methods

In RAMI-V, each experiment is identified by the combination of a so-called testcase and a virtual measure to be simulated. The testcase consists of the combination of a vegetation scenario (ζ), the optical characteristics of scattering elements for a specific spectral band (λ), and an illumination geometry (Ω), which can be either direct or diffuse. The nomenclature of the experiment names and description has been slightly modified, compared to previous phases, to better separate the information regarding the intrinsic physical properties of the scene from the illuminations and the measures to be performed. All experiments were uniquely identified by a name formed by the combination of 4 well-distinguished tags as <scene>-<band>-<illumination>_<measure>, which are described in the following subsections. The full details of the scene and the files containing the scene definition in Rayshade [14], and Wavefront OBJ format for the new scenes only, can be found on the RAMI website.

Scene definitions

Among the 38 vegetation scenarios released in RAMI-V, only 8 of them describe complex actual canopies (Fig. 1), and this paper refers to their description and analysis only. Actual canopies are based on detailed inventories of both the structural and spectral properties of existing plantations and forest stands [7]. They were subdivided in (a) parametric scenes (6), (b) semi-empirical (1), and (c) empirical cases (1). The parametric ones are the same actual scenes of RAMI-IV listed in Table 1 and representative of 4 natural forests as follows: (a) a birch-stand deciduous leaf model, provided as a leaf-on (HET09), and (b) a leaf-off (HET15) version, to represent summer and winter conditions, (c) a summer (HET07), and (d) a winter (HET08) pinestand models [15–18]. Two crops were provided with structured arrangement of the trees. They were (e) a short-rotation poplar forest (HET16) [19] and (f) a citrus orchard (HET14) [20,21]. They maintained the definition given in RAMI-IV, where a hierarchical approach for the scene setup was followed, with the definition of the geometry models of the leafs, the definition of various tree models, and the distribution of the models in random or structured

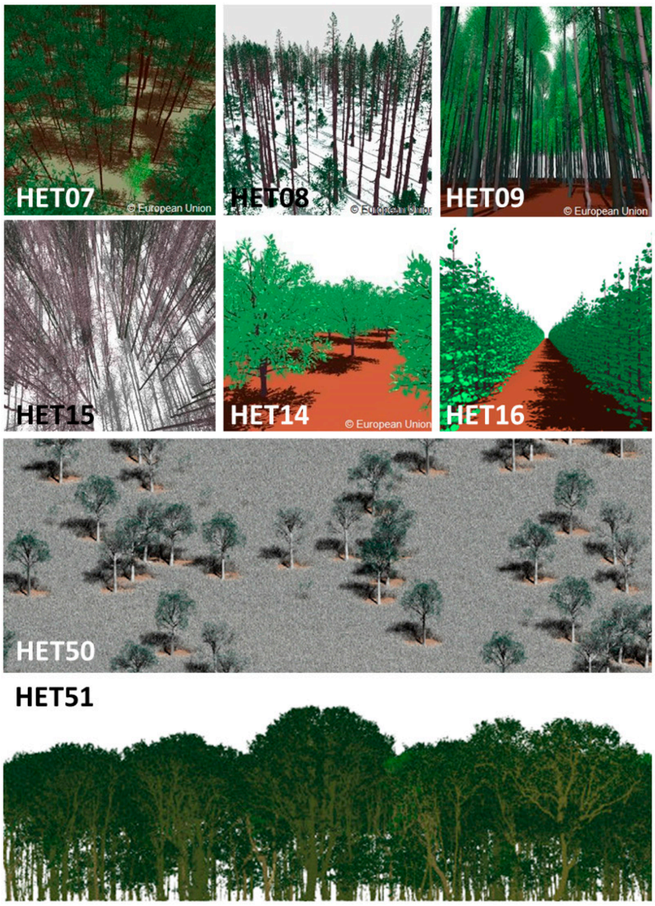


Fig. 1. Rendering of the 8 actual scenes issued with RAMI-V experiment. Scenes HET07 to HET14 were already described in RAMI-IV. Scenes HET50 (Savanna) and HET51 (modified Wytham Woods) were introduced in RAMI-V. The scene colors are purely illustrative, as the combination of RGB spectral properties was arbitrarily chosen to obtain good rendering effects.

Table 1. Overview of architectural characteristics for the 8 reconstructed RAMI-V actual canopy scenes. Note that the structural properties here may differ from those inventoried at the actual test sites. The density accounts for all plant objects (live trees, dead trees, understorey if defined) within 1 ha. The number of geometric primitives (triangles, ellipsoids, cylinders, discs) that described a given actual canopy scenario on the RAMI website is also indicated. More detailed information are available consulting the RAMI website.

RAMI-V id	Name	Plant density (ha^{-1})	Scene LAI (m^2/m^2)	Fractional cover (%)	Maximal height (m)	Primitives in the scene
HET09	Järsvelia Birch Stand	919	3.442	50.4	30.51	350,050,467
HET15	(Winter version)	919	-	-	30.51	138,251,607
HET07	Järsvelia Pine Stand	996	2.302	40.6	18.56	895,635,743
HET08	Ofenpass Pinestand	931	0.745	12.5	15.02	169,120,314
HET14	Wellington Citrus Orchard	991	2.691	39.2	4.12	89,618,249
HET16	Short-rotation forest	11,841	3.219	39.2	3.41	146,665,201
HET50	UCL Savanna	599 (trees)	1.119	–	11.31	4,718,130 (tri)
		2×10^5 (grass)	1.043	–	–	1,400,000 (cyl)
HET51	UCL Wytham Woods	528 (crowns files)	7.59	–	Around 30 m	6,124,335 (tri)
		558 (stems files)	–	–	–	1,715,905 (ccyl)

order across a flat square area with approximate dimensions of 1 ha.

All background surfaces were assumed lambertian in RAMI-V to slightly reduce the complexity of the scene implementation. For scenes common to RAMI-IV, the details are not reported here but can be obtained in [7] (see their figure 1 and tables 1 and 2) and the RAMI web pages. Nevertheless, the main information in terms of plant density, leaf area index (LAI), fractional cover (FCOVER), maximum height, and number of primitives in the 3D models are summarized in Table 1. (g) A savanna model (HET50) was adopted following [22]. It represented a 2-layer heterogeneous system including an over-storey (trees) and an under-storey (grass) layer. The canopy models were developed from detailed field measurements of structural and radiometric properties made at experimental plots with varying canopy cover around Skukuza (25.11°S, 31.42°E) and Pretoriuskop (25.17°S, 31.23°E), both sites belonging to long-term fire ecology experimental plots in the Kruger National Park, South Africa [23]. In RAMI-V, only the pre-fire model has been considered. The savanna plots are dominated by shallow rooted deciduous *Combretum* species, in particular *Combretum apiculatum* (Red Bushwillow), *Combretum hereroensis* (Russet Bushwillow), and *Combretum zeyheri* (Mixed Bushwillow), which cover the majority of the total biomass. Using the information from the field measurements, trees and grass models were generated to produce a 3D scene of 1-ha extent.

The structure for 3 merula trees with increasing size, 2 combretum models with leaf and 5 models without leaf, were developed using OnyxTREE software package (https://www.onyxtree.com/). They were parameterized using the detailed measurements of tree height, diameter at breast height (DBH), and crown size collected during the field campaign [22,24]. The trees were distributed randomly within a given plot, according to a predefined plant density, including stand and fall features, as determined from the field observations. Detailed information on the number of instances for each species can be obtained on RAMI-V pages. Table 1 shows the total number of trees, grass plants, and the corresponding LAI. The scene has a maximum height of 11.3 m corresponding to the highest merula model.

Due to large amount of under-storey grass cover (200,000 plants), cylinders of varying lengths and a fixed radius of 2.5 mm have been used to represent the grass objects to maintain efficiency of times' calculation. Given the huge number of instances, grassland contributes to the total LAI by ~93%.

(h) Finally, the Wytham Woods forest (HET51) is a 1-ha scene representing a real deciduous forest sampled by means of a combination of terrestrial laser scanner (TLS) and traditional census data, aimed to determine the species of each individual tree and allocate species-specific radiometric properties [25]. The original model was cropped to 1 ha typical of the other actual scenes to be adapted to RAMI-V needs—any primitive (triangle, cylinder, or disc) laying outside the new bounding-box has been omitted or commented from the original files. While the original Wytham Woods scene features a sloping terrain, all trees have also been shifted along the *z* axis to lay on a common flat surface. This guaranteed continuity on the boundaries and energy conservation in an infinite replication scheme to allow as many RT models as possible to ingest and process the scene without developing ad hoc plugins or exotic choices to guarantee energy conservation on scene boundaries. Each of the 7 tree species comes combined with a defined set of spectral properties associated to the the crown and the wooden parts, based on field spectroradiometer measurements. The 3D structure of the canopy was stored in a modified Wavefront OBJ format, able to ingest basic solid geometries such as cylinders, which allows to compress the representation with respect to a basic triangulation. A master object file hierarchically merges and clone 558 files containing the definition of the trunk and stems for each tree, and 528 files containing the definition of the crowns.

The leaves are represented with a pair of triangular meshes. The version of the scene used in RAMI-V is affected by an overestimation of the leaf area, with a resulting LAI over the cropped area doubled with respect to the original canopy value, which was around 3.8 m^2/m^2. The original repository was

Table 2. The RAMI-V bands defining the spectral dimension (λ) and their corresponding instrument origin. For OLCI and MSI, the central wavelengths are indicated in nm, while for MODIS the range covered by the specified band is given. The nearest RAMI-IV bands are reported for comparison. The word "all" refers to all measurements except *dhp* and *ftran_loc*. *refl* refers instead to all *brf* measurements and bihemispherical and directional–hemispherical reflectances. The bandwidth for OLCI bands was 10 nm except where indicated in parentheses.

OLCI Sentinel-3	λ_c (nm)	MSI Sentinel-2(A)	λ_c (nm)	MODIS Terra	λ range (nm)	RAMI-IV	λ_c (nm)	RAMI-V band name	Meas involved
O03	442.5	M01	443 (20)	MD3	459–479	B01		O03	All
O04	490	M02	490 (65)			B02	490	O04	All
O06	560	M03	560 (35)	MD4	545–565	B04	451	O06	All
O08	665	M04	665 (30)	MD1	620–670	B07	661	**O08**	All
O10	681.25 (7.5)					B08	674	O10	All
O11	708.75	M05	705 (14)			B10	705	O11	Refl
O12	753.75 (7.5)	M06	740 (14)			B13	753	O12	Refl
		M08	842 (105)					M08	Refl
O17	865 (20)	M8a	865 (21)	MD2	840–876	B15	872	**O17**	Refl
				MD5	1,230–1,250			MD5	Refl
		M11	1,610 (90)	MD6	1,630–1,650			M11	Refl
				MD7	2,105–2,155			MD7	Refl
		M12	2,202 (174)					M12	Refl

Table 3. Average zenith and azimuth angles (θ_s, ϕ_s) to be used for all measurements excluding *bhr* (which is performed under perfectly diffuse illumination only), *ftran_loc*, and *dhp*. They were obtained by averaging the sun position (θ_s, ϕ_s) for January–February, April–May, and July 2017 of the Sentinel-3 OLCI overpasses over each site associated to the specific vegetation scenario. They are averaged over the OLCI revisiting time period mostly spanning the reference months.

			Coordinates	Sun position		
Scene	Site	Country	latitude, longitude	January–February	April–May	July
HET07 HET09	Järvselja summer	Estonia	58.3°N, 27.3°E		56° 153°	41° 147°
HET08 HET15	Järvselja winter			76° 155°	56° 153°	
HET14	Wellington	South Africa	33.6°S, 18.9°E	42° 076°	60° 045°	67° 041°
HET16	Zerbolo	Italy	45.3°N, 8.9°E	71° 153°	36° 137°	34° 130°
HET50	Skukuza	South Africa	25.0°S, 31.5°E	37° 089°	50° 051°	60° 041°
HET51	Wytham Woods	UK	51.7°N, 1.3°W	75° 154°	46° 147°	35° 138°

updated to fix this bug, after the submission period of RAMI-V. Despite this issue, the scene was kept as valid, although its LAI touches the upper edge of biophysical realistic values. In future releases of the experiment, a fix will be considered. The understory has been removed, and a lambertian surface with a weighted average spectrum created from the combination of the spectrum of the original underground plants was provided.

In analogy with [7] (Table 2), we provided a summary of the new scene characteristics in Table 1. As usual, the detailed information and the relevant files containing structural and optical characterization of the scene were distributed through the experiment web pages.

The spectral bands

In RAMI-V, 13 spectral bands from 443 to 2,200 nm were selected among Sentinel-3 OLCI (8 bands), Sentinel-2(A) MSI (3), and MODIS (2) instruments (Table 2). The band selection criteria were based on the alignment between the 3 instruments' bands. They cover mostly, but not only, the spectral range used in the retrieval of biogeophysical parameters with remote sensing techniques.

Bands O08 and O17 common to all sensors are assumed as representative of red and near-infrared (NIR) bands, for a comparison of the results with previous RAMI phases. *brf* and *albedo* were performed in all the 13 bands, while flux measurements related to absorption and transmission through the canopy were simulated only in the photosynthetically active radiation (PAR) bands (the ones laying within 400 and 700 nm).

The spectral properties of the foliage, wooden components and surface, were calculated with the convolution of the extraterrestrial solar spectrum ($S_{0\lambda}$) with the spectral response of each sensor (R_{λ}) and the reflectance (transmission) spectrum associated to the primitives.

Both reflectance and transmission properties were defined for the foliage of each tree species belonging to a specific canopy model. For surface and wooden components, only the reflectance were defined, assuming them perfectly opaque to light ($T = 0$).

The scenes were different in terms of spectral complexity, with 1 single reflectance spectrum associated to all trees for the winter models HET08 and HET15 and crops HET14 and HET16, 2 reflectance spectra for HET07, 5 for the savanna (HET50), and 7 for HET09 and the Wytham Woods scenario (HET51).

Illumination and observation geometries

The illumination was either direct or perfectly diffuse (e.g., isotropic diffuse) for most of the experiments. By definition, for bidirectional and directional–hemispherical reflectance, the illumination was assumed to be perfectly direct (black sky), and for the bihemispherical reflectance, it was only diffuse (white sky).

To reinforce RAMI-V support to the validation of satellite-derived land essential climate variables (ECVs), the illumination and viewing geometries were adopted by averaging real Sentinel-3 OLCI configurations during January, April, and July 2017 over selected locations associated to each specific scene, as listed in Table 3. For the scenario associated to high latitudes (Järsveljia), the angular set was reduced considering only the representative angles on April–May and July for the summer version of the scene, and January-February and April–May for the winter version.

This approach was rather different with respect to previous RAMI phases, where the geometries were fixed to sun zenith (θ_s) of 20° or 50° and azimuth (ϕ_s) of 180° or 90° in some specific cases (heterogeneous structured canopies such as HET16), as we wanted to explore the RT model behaviors in realistic conditions related to Copernicus missions. In line with remote sensing conventions, the azimuth angles vary positively clockwise from geographical north (y axes).

Virtual measurement description

The types of measurement have not undergone substantial changes compared to the previous phases. The measurements are subdivided in top-of-canopy (TOC) bidirectional reflectance factor (BRF) [26,27] in the principal and orthogonal planes (indicated as *brfpp* and *brfop*) with a resolution of 2° over the range of θ_v from −76° (backward reflectance sector) to 76° (forward reflectance). The BRF had to be provided also in the azimuth ring (in steps of 2°) at a constant observation angle of $\theta_v = 37°$ (*brfazim*). For *brfpp* and *brfop* only, in line with previous phases, the BRF as originating from the single collided (*co_sgl*), multiple collided (*mlt*), as well as the un-collided (*uc_sgl*, e.g., collided only with the background surface) photons, was proposed.

In addition, the directional–hemispherical reflectance *dhr*(θ_s) and the bihemispherical reflectance (*bhr*) [26,27] were proposed. All *brf* and *dhr* had to be provided for the different positions of the Sun given in Table 3 and representative of different periods of the year. These measurements had to be performed over the whole set of 13 bands listed in Table 2.

The transmission (*trans_tot*), the foliage (*fabs_fol*), and total (*fabs_tot*) absorptions were foreseen over the RAMI-V bands laying within the PAR spectral region only (400 to 700 nm). Total absorption included all photons, except for those absorbed by the ground. In contrast, absorption by foliage included interactions with any scene elements identified as leaf (or needles for HET07 and HET08). The transmission simulations were also filtered to count only the single scattered and the un-collided rays, indicated as *ftran_coco* and *ftran_uc*, respectively. In principle, *ftran_tot* = *ftran_uc* + *ftran_coco* and a contribution from radiation bouncing back and forth between canopy elements and background.

The transmission of down-welling ($F\downarrow$) and up-welling ($F\uparrow$) flux through the canopy at 11 predetermined and equidistant levels along the z axis, dependent on the canopy height, has been maintained (*ftran_tot_vprof*). The highest level corresponded to the canopy height, and the lowest to $z = 0$. All fluxes were normalized to the incoming flux at the top of canopy ($F_{TOC}\downarrow$), which was set by any participant model to guarantee the convergence of the results to a stable value.

For *ftran* and *fabs* measurements , the illumination had to be considered isotropic diffuse, as well as direct, with the sun positions listed in Table 3. Finally, 2 measures relevant to in situ techniques for the determination of FAPAR and LAI were proposed: (a) the digital hemispherical photography (*dhp*) and (b) the local un-collided transmission at lower boundary from finite sized sun (*ftran_loc*). As the analysis of *dhp* did not highlight additional findings relative to RAMI-IV study, and because *ftran_loc* was submitted by *dart* team only, the results will not be discussed further.

As for the previous sections, any other details on the measurement implementation are given in the specific RAMI web pages.

Results

Model participation

Table 4 lists the 12 RT model names, relevant publication, and operator names that submitted results related to actual canopy experiments. Five of them were already involved in the previous RAMI-IV phase. Table 4 also provides an overview of the methods used by the different models to solve the RT problem and the main assumptions adopted to represent the scenarios. The most populated category was the Monte Carlo ray-tracing RT models, either operated in forward (FMC) or backward (BMC) modes. They generally explicitly describe the scenes in details by means of triangulations of surfaces and enclosed volumes, or can ingest additional primitives such as discs, ellipses, cylinders, and cones. Despite that *flies* uses an FMC solver, the representation of the scene is based on a simplification of the original trees. Abstract trees conforming to simple geometric representations of turbid enclosed shapes were created by using the individual tree model leaf area provided by RAMI and exactly positioned in the scene.

Other models such as *frt13*, *rapid* and *spartacus*, rely on different RT equation solutions, based on geometric and/or semi-analytical methods combined with simplified representation of the scenes. In *frt13*, the RT equation is solved by means of geometric considerations and analytical approach. *Spartacus*

Table 4. List of the 12 models, references, and operators contributing to the actual canopy experiments of RAMI-V phase. Previous RAMI-IV participations are indicated with ✓. Methods: FMC = forward Monte Carlo ray-tracing technique, BMC = backward/reverse Monte Carlo, BiMC = bidirectional Monte Carlo; RAMI-V scene representation assumptions: Exp = explicit with primitives; T = triangle meshes, P = more primitives, Tu = turbid medium; method to handle scene boundaries (e.g., ∞): C = cyclic boundary, R = scene replication; PAS stands for primitive-level anisotropic scattering; "—" indicates that the information was not provided.

Name	Operator	IV	Method	Scene	∞	PAS	References
dart	Wang Y.	✓	BiMC	Exp, T	—	✓	[44,45]
dirsig5	Goodenough A.		BMC	Exp, P	R	✓	[46]
eradiate	Schunke S.		FMC	Exp, T	R	✓	[47]
flies	Kobayashi H.	✓	FMC	Geo, Tu	—	No	[48]
frt13	Kuusk A.	✓	Hybrid	Geo, Tu	—	No	[15,49]
less	Qi J.		BMC+FMC	Exp, T+Tu	C	✓	[50]
librat	Origo N.	✓	BMC	Exp, P	R	—	[24,51]
rapid	Huang H.		Radiosity	Porous		No	[52,53]
raytran	Lanconelli C.	✓	FMC(Sprd)	Exp, P	C	✓	[54,55]
renderjay	Van Leeuwen M.		FMC	Exp, P	—	✓	[56]
spartacus	Hogan R.		DISORT	Sta, Tu	—	No	[57]
wps	Zhao F.		FMC(Sprd)	Exp, P	C	—	[58]

was designed to compute profiles of solar and thermal fluxes in vegetation canopies for use in weather and climate models. It solves the RT equation with the discrete ordinate method [28], ingesting a statistical description of the scene, such as the height-resolved variation in leaf and woody surface areas, with the main trunks described explicitly while positioned randomly and vertically across the scene, and branches and leaves distributed randomly in the horizontal domain.

The detailed contributions of each model to the various measurement type are summarized in Fig. 2. Only 2 teams, e.g., *dart* (100%) and *raytran* (99.4%), completed more than 99% of the total 3,596 experiments proposed for actual scenes in RAMI-V. Excluding *ftran_loc*, *raytran* submitted all the experiments considered in this work belonging to the context of *brf* and flux (absorption, transmission, and albedo). Overall, the teams of *less*, *raytran*, and *wps* submitted more than the 80% of the proposed experiments. *renderjay* reported only absorption (*fabs*) and transmission (*ftran*) measurements, while *spartacus* extended the submission to hemispherical reflectance measurements (*dhr* and *bhr*). The remaining models also submitted results for the *brf* measurements. Teams of *dart*, *dirsig5*, *less*, *raytran*, and *wps* contributed at least partially to all actual canopies. The remaining models excluded some scene from their analysis. In particular, *flies* did not submit results for savanna and Wytham Woods forest (HET50 and HET51), *rapid* submitted *brf* results for scenes HET07, HET14, HET50, and HET51, *spartacus* submitted flux results for the pinestands and the birchstands, and for HET14, *frt13* submitted *brf* results for pinestands and the summer version of the birchstand (HET09), *librat* submitted *brf* results for the new scenes only, *eradiate* submitted HET14, and *renderjay* focused on absorption and transmission over summer season forests (HET07 and HET09) and HET16.

As in previous phases, participants were not obligated to submit results for every proposed experiment. Instead, they had the freedom to select the subset of experiments that were relevant and suitable for their model. Therefore, not all comparisons could be conducted due to the absence of specific measurements, then the model-to-model comparisons were performed on a single scene basis.

The statistical information of density profiles of leaf and branch surface area was provided for most of the RAMI-V scenes, except HET50, HET51, and HET16 (for which the woody surface area was missing). The additional efforts required to define such statistical profiles from the explicit information likely penalized the participation, especially of the RT models not based on ray tracing. Future implementations of the intercomparison should include statistical version of all scenes to foster participation of model oriented to computation performance and operational deployment.

Selected examples of model results

Figure 3 shows a selection of results from BRF measure in the orthogonal and principal plane. Figure 3A shows an example for HET08 (winter pinestand). Qualitatively, a model dispersion similar to the one observed in RAMI-IV was observed, with only *frt13* and *wps* deviating considerably from the mean value. In this particular case, *frt13* was affected by a deviation from the bell shape characterizing all the other participants, suggesting some issue in the geometric implementation of the scene.

Figure 3B shows a case for HET14 (citrus orchard) in which the agreement among some models appeared better than in RAMI-IV with *flies* and *dirsig5* presenting an evident overestimation and underestimation, respectively. The figure also shows that *flies* probably suffered some geometrical issue due to either the canopy structure orientation or the illumination angle conventions adopted. For this particular scene, the trees are arranged in rows and the asymmetry of the *brfop* is expected, being the sun azimuth angle in RAMI-V not aligned, neither

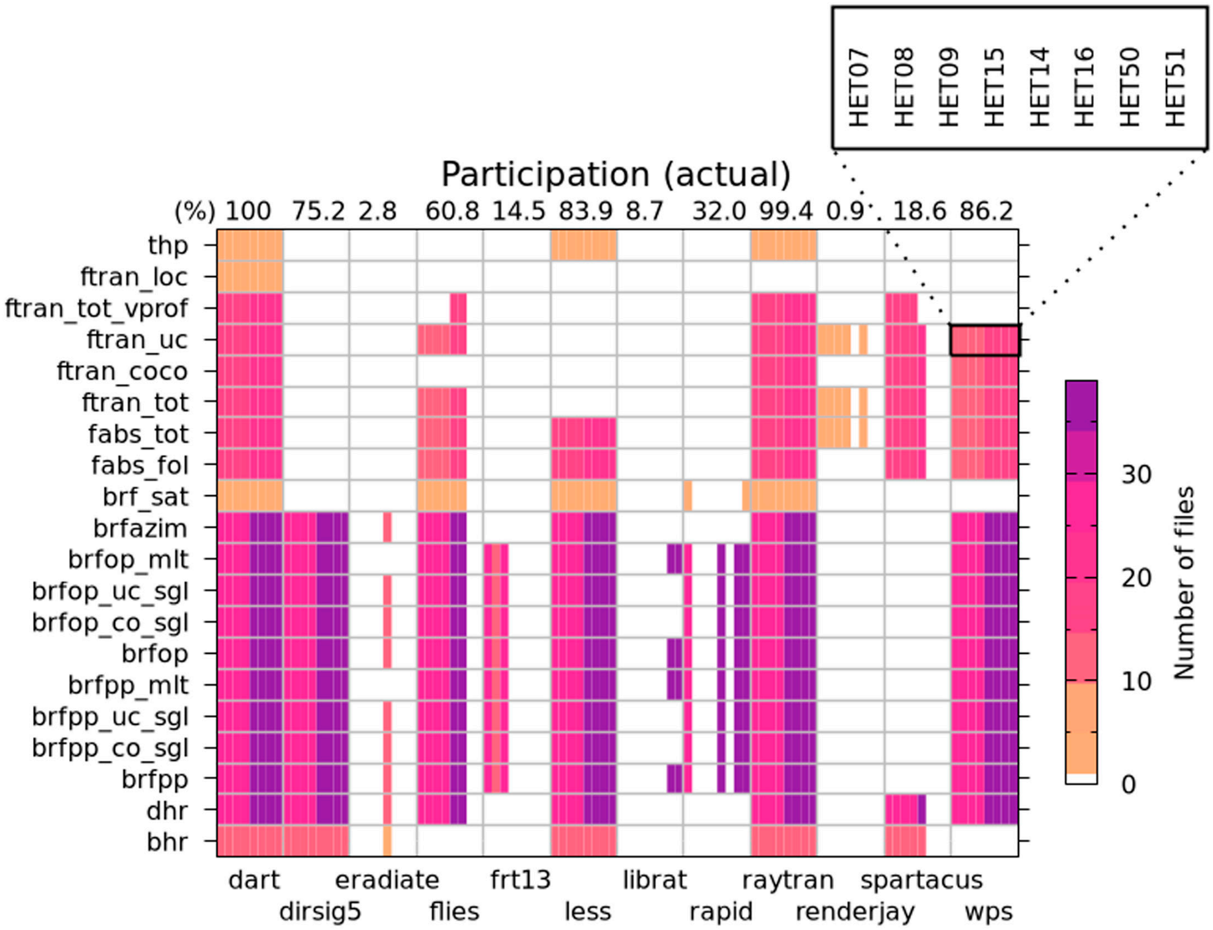


Fig. 2. Overview of the participation by scene and measurements performed by RT models contributing to the actual cases. Missing contributions are indicated by white cells. Each cell refers to 8 scenes ordered as in the scheme shown in the upper right. The total number of files expected per measure is as follows: 39 (26) for any *brf* and *dhr* measure, 13 for *bhr*, 15 for fluxes (*fabs* and *ftran*), 3 for *brf_sat*, 3 (2) for *ftran_loc*, and 1 for *thp*. They originate from the combination of bands × geometries proposed.

perpendicular, to the direction of the rows, as it was in RAMI-IV instead. This fact allowed us to address a certain number of inconsistency during the feedback phase, as we observed some models submitting results symmetric to the expected behavior.

Figure 3C shows an example for HET15 (winter birchstand). Excluding *flies* and *dirsig5*, the agreement in the red band (O10) appears excellent among the remaining 4 models (*dart*, *less*, *raytran*, and *wps*), and much better than that observed during RAMI-IV, where the difference between the minimum and maximum values was ~0.35, against the 0.1 seen here even including the worst cases.

Concerning *brfpp*, we included a pinestand summer case (HET07; Fig. 3D) which reports a better agreement with respect to RAMI-IV similar results, but the same odd features affecting *flies* in RAMI-IV, consisting in the lack of continuity of the bidirectional reflectance between −30° and 0° visual zenith angle range. All models reported a pronounced hotspot, and out of a few models, the others showed a good agreement, though not as better as the one observed in Fig. 3C for the HET15 scene, where the foliage component was partially missing. Figure 3E shows a comparison for the summer birchstand (HET09), where 5 of 7 models agreed within ±0.01, similarly to the results observed in RAMI-IV for the same forest, but for a larger set of models. Figure 3F refers to HET16 (a structured short-rotation forest), where 5 of 6 models showed an excellent agreement. A quantitative and comprehensive comparison of *brf* measurements is given in Model-to-model deviation ($\delta_{m\leftrightarrow c}$) and general model deviation (δ_m).

Figure 4 shows some results of *brfazim* measure for a uniform forest (HET09) and 2 structured canopies (HET14 and HET16). The figure shows, for different bands, the asymmetric patterns of BRF against the viewing azimuth angle ϕ_v, especially for HET14 and HET16 (e.g., structures aligned north-south, along the y axes). This asymmetry is due to the fact that in RAMI-V, conversely to RAMI-IV, the illumination azimuth angle was not aligned north-south (or east-west) with the structures of HET14 and HET16, but it represented realistic satellite observation made from Sentinel-3 OLCI. For the particular case given in Fig. 4, the solar azimuth angles were 147° (HET09), 76° (HET14), and 137° (HET16).

The forest scene HET09 features pseudo-randomly distributed trees, and considering that θ_s was 41° (not far from the θ_v of 37° required in brf_azim simulations), we observed that the maximum of the reflectance occurred around the hotspot region for all models ($\Delta\phi \sim 0°$), except *flies*. It looks like its hotspot was offset by ~60° in azimuth, suggesting a problem either in the setup of the scene or the illumination-viewing geometry, which was not fixed after the feedback phase.

The patterns observed in Fig. 4 for the structured canopies are more interesting. The absolute maxima of BRF are still

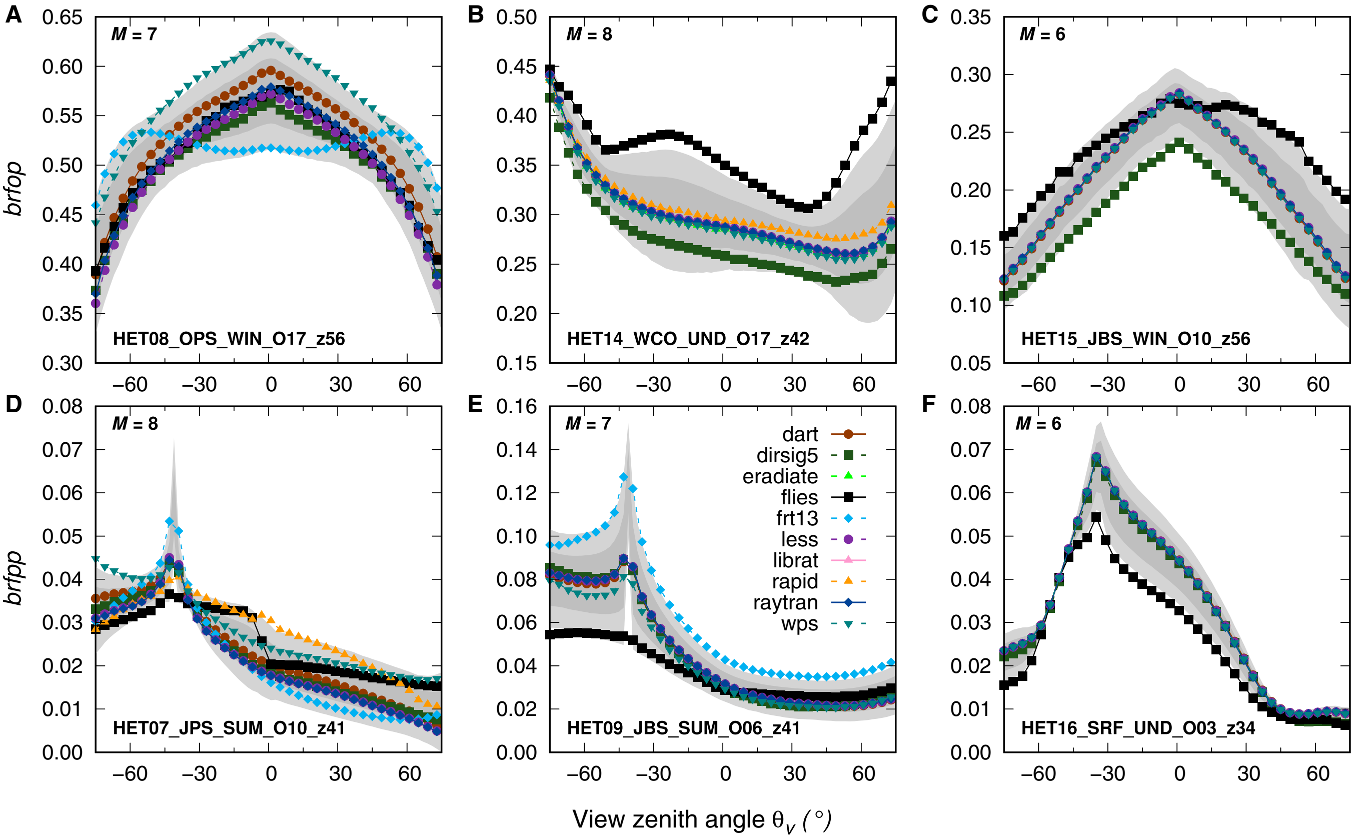


Fig. 3. Examples of model-simulated domain-level BRFs along the orthogonal plane (A: pinestand winter scene, B: citrusorchard, C: birchstand winter) and principal plane (D: pinestand, E: birchstand, and F: short-rotation poplar forest) for different spectral bands and for 6 actual canopy test cases. The specific cases are selected to mimic those of figure 3 of [7] for the most direct comparison. The dark gray bands show the standard deviation range (±1σ), while the light gray shows the outlier detection threshold ranges (see Preliminary screening, identification, and rejection of the outliers). *M* indicates the number of participants to the specific experiment.

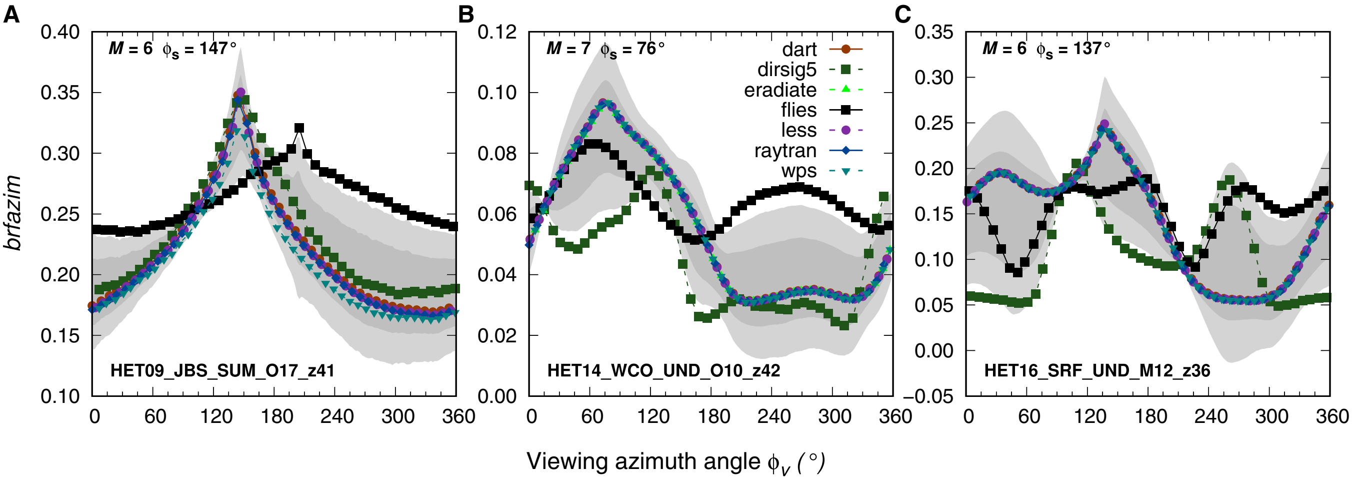


Fig. 4. Examples of model-simulated domain-level BRFs along an azimuth ring (*brfazim*) at $\theta_v = 37°$ for the birchstand forest (A), and the 2 structured actual canopies citrus orchard (B) and short-rotation forest (C). The results refer to different bands as indicated by the legend. *M* indicates the number of participants to the specific experiment. The azimuth angles are defined positive clockwise from the north. ϕ_s indicates the solar azimuth angle.

appreciable for $\phi_v \sim \phi_s$ in both cases, but local maximum/minimum was observed for intermediate angles by all models. Except *flies* and *dirsig5*, they agree rather well in representing the bidirectional reflectance along the proposed azimuth ring. For HET14, a local maximum can be observed around a ϕ_v of ~270° for most of the models. This appeared to be a privileged observation to collect the photons reflected by the soil without interaction with the vegetation. In the O10 band, the relatively higher reflectance of the soil (~0.2) with respect to the vegetation (~0.01) likely induced this effect. A symmetric peak at ϕ_v

~ 90° was masked by the enhanced backward reflectance of the vegetation, but still slightly appreciable, as a bulge around 120°.

Similar considerations can be done for HET16 among the agreeing ensemble of models, which were *dart*, *less*, and *raytran*. For *flies* and *dirsig5*, the results confirmed problematic interpretation of the experiment for structured canopies, which may arise from wrong assumption of the orientation of the structures or the convention used for the azimuth angles.

The bihemispherical reflectance results are shown in Fig. 5. A peculiarity of the *bhr* results is the lack of evident outliers as the standard deviation is always large enough to include all points within the acceptable thresholds defined by the Chauvenet criteria marked by light gray bands in the figures. The differences are evident between O11 and M11 NIR bands for HET07 (pinestand summer) and HET51 (broad-leaf forest) scenes, which present standard deviation of the order of 0.05 to 0.10, of the same order of magnitude of the physical quantity measured, producing a relative dispersion up to 100%. In the visible (VIS) spectral range, where the photons are mostly absorbed by dense canopies, the agreement among RT models appears better. For winter scenes (HET08 and HET15), the high reflectance in the VIS is comparable with that in the the NIR bands; the spread of model results is spectrally flat, confirming an impression that for higher reflectance at primitive level, the fact that more photons remain in the game longer induced larger uncertainties. Remarkably, for HET50 (savanna), the 4 contributing models agree rather well along the entire spectral range.

Concerning HET15 (citrus orchard), the 5 participants are grouped in 2 well-distinguished and spectrally coherent clusters, the lower being formed by *spartacus* and *dirsig5*, and the higher by *dart*, *less*, and *raytran*. Within each cluster, the agreement was excellent. It is then difficult to distinguish, from a statistical analysis, which is the most reliable result. In this case, the assessment of the model credibility based on their participation rate (Model participation) or the internal consistency check described in Section 4.2 supported the choice to discern the second in place of the first.

Concerning the directional–hemispherical reflectance, the participants were asked to submit results for 3 fixed sun angle configurations representative of January, April, and July average OLCI overpass conditions (Table 3), which were reduced to 2 for Järvselja scene as previously discussed.

Figure 6 shows the results for April–May geometry configurations listed in Table 3. The participation to *dhr* experiments was higher than that of *bhr*. In fact, *flies* and *wps* contributed to *dhr* (except over HET50 and HET51), while they did not submit results for *bhr*. *eradiate* submitted *dhr* for HET14 only. For HET15, the clustering discussed for *bhr* remains almost intact for *dart*, *less*, and *raytran* and was reinforced by *wps*, while the second cluster spreads out, with *spartacus* results deviating from *dirsig5* ones. Model *flies* overestimates the average behavior for HET09 (birchstand) and crops especially at the highest solar zenith angles. The good agreement observed for HET50 (savanna) was lost for MD5 and M11 RAMI-V bands.

Intercomparison approach

The proposed analysis of the results is consistent with the previous RAMI-IV phases. As such, it was based on (a) a set of consistency checks between the radiative quantities produced by a model, (b) a model-to-model deviation assessment, and (c) the proximity to a reference acting as a surrogate truth, originating from the ensemble of models agreeing within determined proficiency criteria. The ISO-13528 proficiency testing adopted by Widlowski et al. [9] provide rigorous guidelines to assess the quality of the results provided by a laboratory, considering the uncertainties arising from a combination of errors

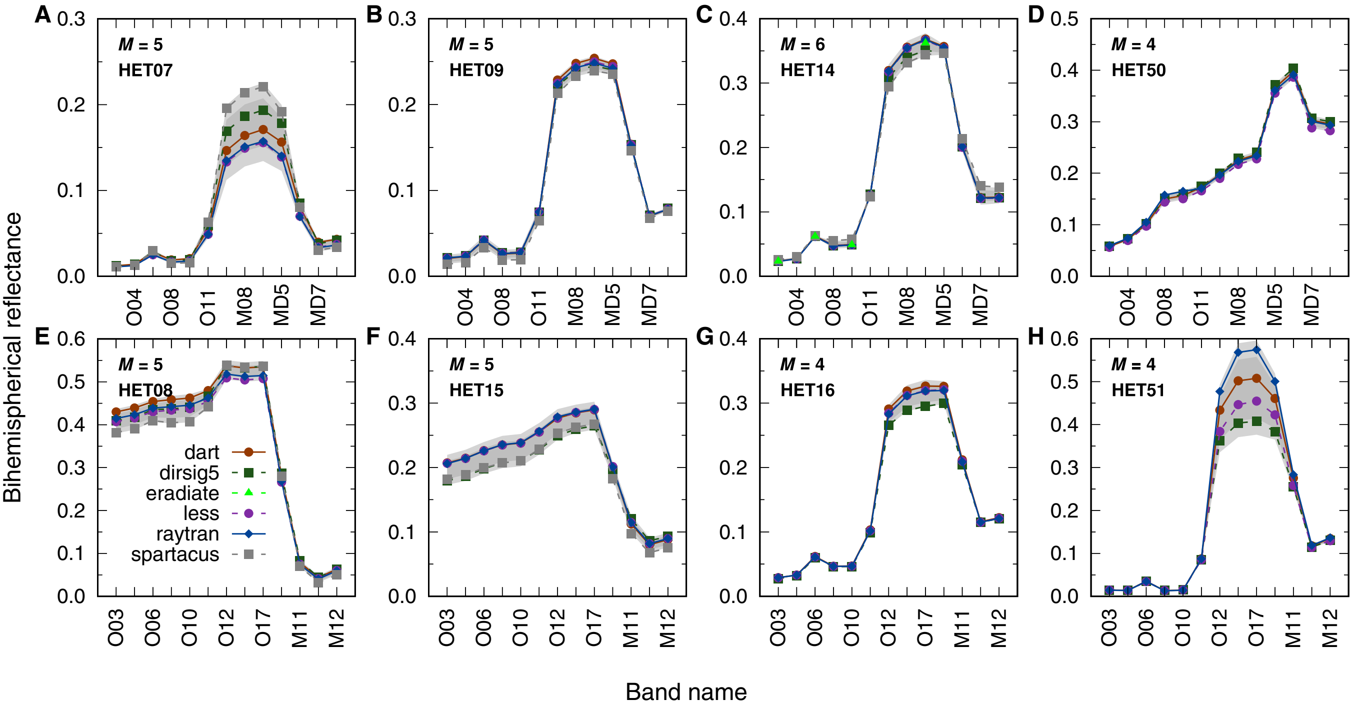


Fig. 5. Bihemispherical reflectance *bhr* results over RAMI-V actual scenes (A to H). The dark gray bands show the standard deviation range (± 1σ), while the light gray shows the outlier detection threshold ranges. M indicates the number of participants. *eradiate* submitted only 4 bands for HET14 case (C).

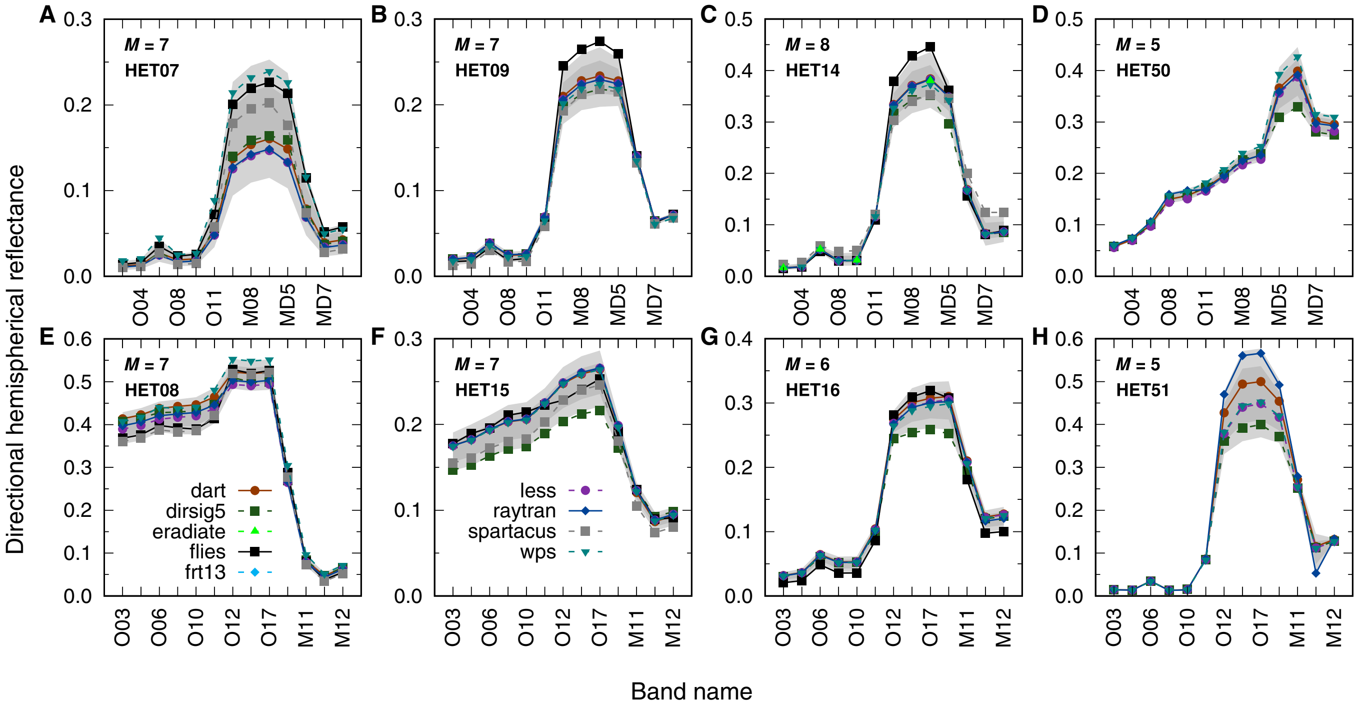


Fig. 6. Results of directional-hemispherical reflectance *dhr* for the April average sun configuration RAMI-V actual scenes (A to H). The dark gray bands show the standard deviation range (± 1σ), while the light gray shows the outlier detection threshold ranges. M indicates the number of participants. The legend was split across (E) and (F) for better rendering.

in (a) the scene representation assumptions, (b) the appropriateness of the implementation of the solution to resolve a 3D RT problem, and (c) the expertise of the operator performing the simulations.

To ensure the accuracy of the RAMI-V results, a thorough initial analysis was conducted. This analysis provided participants with feedback to identify and correct major errors in their experimental setup. Such errors could emerge at various stages of the experiment, such as during the scene format conversion, or the design of the assumptions made to fit the scene into particular code requirements, or merely from output files not meeting the specific format required by RAMI. Moreover, common issues are related to setting incorrect spectral properties or geometries, or to the misinterpretation of the virtual measurement concept.

Historically, RAMI adopted a rather strictly concept of blindness, preventing participant to have access to other model results. In RAMI-V, the blindness has been maintained with respect to the results of other models, but the outliers were identified and indicated to all participants in graphical form through the web interface. The decision to check and eventually fix the results, keeping results in business or withdrawn the result, was left to the participants. This approach has adopted to focus on the performance of RT codes in the optimum operative conditions, rather than assessing the overall capabilities of a specific laboratory consisting in the full process involving preparation of the dataset, simulation, and reformatting.

The following sections summarizes (a) the internal consistency checks, (b) the model-to-model comparison aimed to select a set of model to issue a candidate CRG benchmark, and finally (c) the model proficiency test based on k' metric to assess the model performance at least for the experiment where a surrogate reference was identified.

Internal consistency checks

The quality check and flagging of the data is a crucial step in any measure evaluation, and it was used to establish a first screening to flag a value as credible or suspicious. An internal consistency check is generally based on the comparison of an absolute or relative quantity, or of the combination of different virtual measures (typically differences or ratios), with respect to predetermined thresholds, which are defined from previous experiences or from a statistical analysis of the dataset under investigation. Here, we used the consistency checks to (a) identify suspicious behavior of the model submissions and isolate a credible population, and (b) to associate to each model a performance in terms of a metric based on the sum of the test value over a filtered set of submission (might be over all scenes, bands, geometries, or a combination of them), quantifying the level of confidence of the dataset submitted by a model.

The consistency checks implemented in this work are listed in Table 5. The following subsections describe each consistency checks and the results obtained for any model submissions, whenever all the virtual measurements required to perform the specific check were provided.

Energy conservation

This was a test common to all previous RAMI phases, consisting in the verification of energy conservation in terms of the quantity defined as $\Delta f_m = (1 + \alpha \times T - R - T) - A$, where A is the total energy absorbed by the canopy, R is the energy reflected at the TOC, and the term $(1 - \alpha) \times T$ represents the energy absorbed by the underlying surface under the assumption of Lambertian reflectance α and under a canopy transmission T. Alternatively, considering that FAPAR (e.g., the absorption) can be reliably obtained from the combination of the net fluxes at top and bottom of the canopies, the same equation

Table 5. List of the consistency checks applied in RAMI-V phase. PAR indicates the set of bands 003, 004, 006, 008, and 010. The last column specifies if a test was inherited from RAMI-IV phase: (y/n) stand for (yes/no). The + symbol indicates that an extended test with respect to RAMI-IV phase has been implemented.

Test	Test name	Required measurements	Models	Bands	RAMI-IV
r5cc1	Energy conservation	*dhr* (*bhr*), *ftran*, *fabs*	*dart*, *flies*, *raytran*, *spartacus*, *wps*	PAR	y+
r5cc2	BRF consistency	*brfpp(op)*, *co*, *uc*, *mlt*	*dart*, *dirsig5*, *flies*, *frt13*, *less*, *rapid*, *raytran*, *wps*.	All	y
	BRF versus Albedo	*brfpp(op)*, *dhr* (*bhr*)	*dart*, *dirsig5*, *eradiate*, *flies*, *less*, *raytran*,*wps*.	All	n
	Albedo versus TOC flux	*dhr* (*bhr*), *ftran_tot_vprof*	*dart*, *flies*, *raytran*, *spartacus*	PAR	n

can be seen as how reliable is the estimation of the absorption A from real flux measurements combination (Fig. 7).

Each term of the previous equation should be considered dependent on the canopy (ζ), wavelength (λ), and illumination angles (ω_i). The value of Δf_m should ideally be equal to zero for a perfect energy conservation and, accordingly to its definition, is positive when some photon loss among the components A, T, or R prevent the energy balance to be closed.

All the actual canopies in RAMI-V feature a lambertian background; hence, the value of α is an input parameter. The remaining information can be obtained from either *bhr* (under diffuse illumination conditions) or *dhr* (direct illumination) to define the reflectance term $\boldsymbol{R}$, *fabs*, and *ftran* to define the terms $\boldsymbol{A}$ and $\boldsymbol{T}$, respectively.

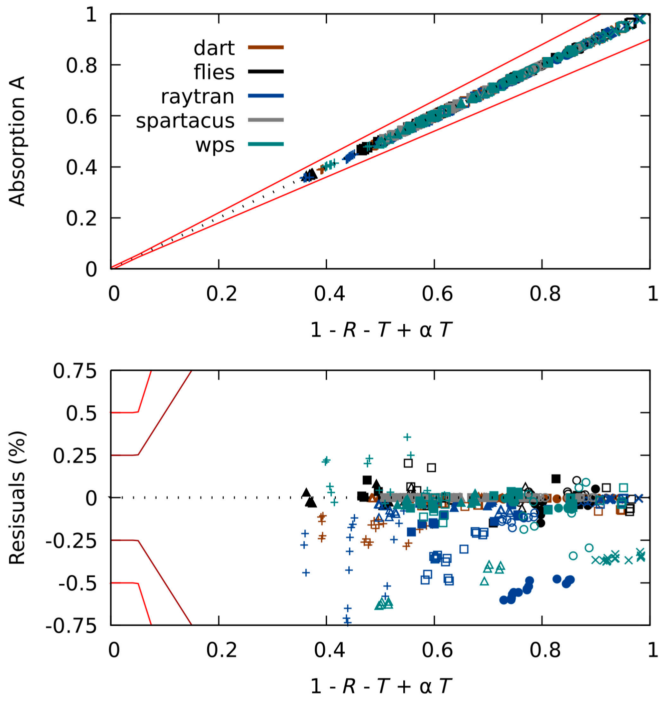


Fig. 7. Scatterplot of the absorption as obtained from *fabs_tot* versus the corresponding quantity evaluated through the net-flux combination. The lower panel shows the residual in percent. The red and brown lines correspond to the threshold and goal uncertainty requirement set by GCOS-244 for FAPAR [e.g., max(0.005, 10%) and max(0.0025, 5%) respectively]. The colors indicate different models as listed in Figs. 3 to 6, while the symbols identify the scenes as follows: HET07 (▲) and HET08 (△), HET09 (•) and HET15 (○), HET14 (■) and HET16 (□), HET50 (+) and HET51 (×).

In RAMI-V, the fabs and ftran measures have been required only for the bands laying in the PAR spectral region (400 to 700 nm); hence, this test could be applied only to 5 bands (Table 2). The models listed in Table 6 submitted all the required measures to perform it.

As RAMI-V proposed different scenes (ζ), spectral bands (λ), and illumination conditions (Ω_i), we summarized the result of the test for each model, in terms of the population statistics including the average bias Δf_m as done in RAMI-IV [9].

Table 6 summarizes some statistics of Δf_m distribution aggregated over scenes ζ, wavelength λ, and illumination Ω_i (including either diffuse or direct illumination conditions), including the median and interquartile ranges (IQRs) because of their robustness to outliers.

The average value for *dart* was triggered by a slight misbehavior of the model for savanna and short-rotation forest, and its excellent performance was highlighted by a median value lower than 0.005%. The best performance in terms of energy conservation was achieved by *spartacus*, although it should be observed that it did not submit results for any scenarios, which raised the major inconsistencies (HET16, HET50, and HET51). Also, the final results of *flies* showed very low mean and median difference values, confirming that the issues identified during the initial phase of the experiment were promptly fixed after the feedback phase.

A comparison with RAMI-IV in the NIR region cannot be performed as in RAMI-V absorption A and transmission T simulations were required only in the PAR region to support/validate in situ measurement protocols. Pertaining to VIS channels, Widlowski et al. [7] reported for actual canopies values of Δf_m up to 2% maximum average on band B18 (1,025 nm) for *Rayspread* (e.g., Raytran here), and from 5% to 40% for *Dart* (spectrally flat), while values for *Rgm* were not reported explicitly. The results obtained here for *dart* may reveal an excellent improvement in terms of energy conservation.

Overall, all models present a good behavior in terms of median/absolute maximum deviation being the worst values of 0.07%/0.64% acceptable in terms of proficiency in exploiting the RT model for practical applications, especially for sensitivity studies related to the assessment of FAPAR from flux measurements.

BRF consistency

This check consists of verifying that the sum of single-collided (sgl), uncollided (uc), and multiple-collided (mlt) BRFs coincides with the total BRF, as in the following equation

$$\Delta\rho = \left(\rho_{sgl}+\rho_{uc}+\rho_{mlt}\right) - \rho_{tot} \qquad (1)$$

In RAMI-V, all participant models, except *spartacus* and *renderjay*, submitted BRF results. Nevertheless, some of them did not submit all the filtered BRF and were excluded from this discussion. Specifically, *librat* submitted only total BRF and *eradiate* did not submit the ρ_{mlt} component. All 13 bands were involved in this test. On the initial round of submission, models *flies* and *dirsig5* have been warned about relevant inconsistencies and were requested to verify their simulations, which were partially fixed in the final results.

Table 7 shows the average, median, and maximum absolute difference $\Delta\rho$, along with its dispersion in terms of standard deviation and IQRs. For model *dirsig5*, the *brf* obtained from the sum of the filtered components was on average lower than the total *brf* for all scenes, with a median of the relative difference of ~5.4 ± 14.6%. In particular, HET14, HET16, and HET51 presented an *rms* of 0.06, 0.054, and 0.096, respectively. For *less*, a null difference was observed over all submissions, as a result, presumably, of the computation of 1 of the 4 components as the combination of the remaining 3 or, conversely, on the computation of total as the sum of the 3 components. Model *wps* exhibited the larger relative deviation on average (2.3×10^{-3}%), although it can be considered negligible for any practical application, and far below to the proficiency criteria set for the reflectance equal to the uncertainty *goal* for albedo, e.g., *max*(0.0015, 3%).

We also checked if the residuals were wavelength dependent by computing the mean absolute error (MAE) by band for each model and scene. For *dart*, *frt13*, *rapid*, and *raytran*, it was equally distributed across all bands for all scenes, with values of the order of 10^{-7} to 10^{-5}, indicating a deviation purely associated to rounding issue (RAMI requested to submit all values with the precision of $\pm 10^{-6}$). On the other hand, for *dirsig5*, *flies*, and *wps*, the MAE in VIS bands was negligible with respect to that observed in *NIR* bands, as might be expected for higher reflectance. Nevertheless, the values observed for *flies* were negligible (10^{-8}), *wps* presented some appreciable deviation especially for the winter scenarios (HET08 and HET15) with higher reflectances, while *dirsig5* remained problematic with a MAE variable between 0.1 and 0.8 for all scenes. The column |*Max*| in Table 7 reveals that there were some minor misalignment expressed by values of 0.65% and 0.17% for *raytran* and *wps*, respectively.

Total BRF versus albedo through the inversion of RPV model

A common approach to retrieve surface albedo from remote sensing techniques consists of fitting atmospherically corrected bidirectional reflectance with a bidirectional reflectance distribution function (BRDF) over a certain temporal window ranging from minutes to days, depending on the sensor observation strategy. It should guarantee the collection of a sufficient number of observations to perform the optimization of the adopted BRDF function's parameters. With this approach, the

Table 6. Summary of the energy conservation consistency test (r5cc1). All values, except *n* (number of testcase involved in the statistic), are given in percent (%).

RT model	Average	Median	|*Max*|	σ	IQR	*n*
dart	0.04	<0.005	0.29 (HET50)	0.07	0.03	140
flies	<0.005	0.01	0.20 (HET16)	0.06	0.05	70
raytran	0.19	0.10	0.91 (HET50)	0.22	0.31	140
wps	0.11	0.04	0.64 (HET08)	0.21	0.33	100
spartacus	<0.005	<0.005	<0.005 (HET08)	<0.005	<0.005	80

Table 7. BRF consistency check (r5cc2). Statistics of the relative difference (%) between the sum and the total BRF for each model as resulting from aggregated actual scenarios. As the results are submitted with 6 decimals, any term $<10^{-4}$% in the table indicates that the individual differences are predominantly null.

RT model	Average	Median	|*Max*|	σ	IQR	*n*
dart	1.7×10^{-5}	$<10^{-6}$	5.61×10^{-2}	2.1×10^{-3}	$<10^{-6}$	39,520
dirsig5	–10.9	–5.39	170	14.6	14.4	39,520
flies	$<10^{-6}$	$<10^{-6}$	4.8×10^{-4}	7.8×10^{-6}	$<10^{-6}$	27,664
frt13	8.0×10^{-6}	$<10^{-6}$	2.1×10^{-2}	2.3×10^{-3}	$<10^{-6}$	9,880
less	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	39,520
rapid	1.7×10^{-5}	$<10^{-6}$	1.0×10^{-2}	3.9×10^{-4}	$<10^{-6}$	21,736
raytran	-1.4×10^{-4}	$<10^{-6}$	6.6×10^{-1}	2.2×10^{-2}	$<10^{-6}$	39,520
wps	2.3×10^{-3}	$<10^{-6}$	1.7×10^{-1}	1.6×10^{-2}	6.6×10^{-3}	39,520

spectral directional–hemispherical and bihemispherical reflectances are obtained by single and double hemispherical integrations of the resulting model, respectively [29].

We used a similar approach here to verify the consistency between the total *brf* and the *dhr* or *bhr* results by means of the relative difference $\left(\alpha_{BRF} - \alpha\right) / \alpha$, where α_{BRF} represents the hemispherical integrals of the BRF function fitted on *brfpp* and *brfop* data. It might be a directional–hemispherical reflectance or a bihemispherical reflectance, and α represents the corresponding *dhr* or *bhr* submission, respectively.

Only 7 of 12 models were processed because *frt13*, *librat*, and *rapid* did not report albedo measures, *spartacus* did not report *brfs*', and *renderjay* did not report *brf* and albedo measurements.

Specifically, we fitted, for each scene and band, all the available bidirectional reflectances from 3 illumination angles (2 for Järvselja), and 76 × 2 (principal and orthogonal planes) viewing angles, with the 3-parameter semi-empirical Rahman Pinty and Verstraete (RPV) model [30]. This data aggregation was supported by the fact that RAMI-V scenes do not feature seasonal structural or optical property evolution.

The optimization produced a set of parameters for each combination of scene and band, which is described by ρ_0, the overall strength of the reflectance, k, the U-shaped or bell-shaped feature against θ_v (k parameter), and Θ, the asymmetry of the reflection feature against the azimuth position of the sun (Henyey–Greenstein scattering phase function). In the 3-parameter version of the RPV model, the hotspot characterization, of particular relevance for vegetation, is regulated by the ρ_0 parameter through a specific hotspot kernel function.

The fitting procedure was implemented using the Levenberg–Marquardt nonlinear least-squared methodology [31,32]. While the fitting algorithm provided also the estimation of the uncertainty associated to each parameter ($\delta\rho_0, \delta k, \delta\Theta$), we used the root mean square deviation *rms* metric to assess the capability of the model to represent the anisotropic features characterizing the dataset.

The *rms* increased generally with the reflectance and lay between 5% and 10% of the corresponding $dhr\left(\theta_s\right)$ reflectance values for most of the cases. Similar misbehavior features (rms variable between 0.04 and 0.2) were observed among all models affecting the fitting procedure for HET15 birchstand winter model and the poplar forest structured canopy (HET16), where the *rms* increased up to 20% of the corresponding measured *dhr*, the worst behavior corresponding to a cluster of cases for which the *rms* varied between 0.03 and 0.05 for reflectance values between 0.1 and 0.3.

Figure 8 illustrates 4 examples of the comparison between A_{BRF} and A, showing 2 models with a relatively higher dispersion (*dirsig5* and *flies*, upper panels) and 2 models with lower scatter of data (*dart* and *raytran*). The figure incorporates the values for all scene ζ (indicated by different symbols), band λ (which may be associated to the reflectance strengths), and season Ω_s (as the *dhr* was provided for 2 or 3 illumination angles).

The observed deviations depend on both the uncertainty of the RPV model to reproduce the BRF values (indicated by the color scale) and an actual inconsistency between the BRF and albedo simulations. It is not straightforward to distinguish their individual contribution to the overall deviation, although it is rather obvious that a poor performance of the RPV model to represent BRF (violet tones) likely induces large deviations, preventing the reliability of the consistency check, while when the optimization performance is relatively better the agreement is expected to improve.

For *dirsig5*, it is possible to identify some major deviation over HET14 around 0.2 and 0.4 (reference *dhr*), and a remarkable underestimation of the calculated albedo for some HET15 combinations, which is coherent with the poor relative performance indicated by the color scale. Focusing on *dirsig5*, A_{BRF} is practically unbiased on average (bias < 0.001), with an overall relative *bias* of 1.3% and an *rms* of 0.013 (or 7%). Somewhat slightly dispersed data affected *flies*, which showed a *bias* of 0.006 (2.7%) and *rms* of 0.019 (13%), mostly related to HET15 deviations appreciable in the scatterplot. Although even *dart* and *raytran* suffered similar RPV fitting quality issue for some scene, we do not appreciate any particular deviation for HET15, HET14, and HET16, indicating that the hemispherical integration compensated any over/underestimation of the *brf* field affecting the reflectance parametrization. Hence, the minor issues affecting *dirsig5* and *flies* over these scenes might indicate an actual problem of simulation consistency. These reinforce the findings of the examples shown in Figs. 3 and 4, where we already observed consistent *brf* deviations of *flies* and *dirsig5* from the other models, and lower values of the albedo (either *bhr* or *dhr*) in Figs. 5 and 6.

Figure S1 shows the consistency of BRF against albedo measurements in relative terms for all models. The heatmap reveals that *dart*, *dirsig5*, *less*, and *raytran* provided all possible combinations to complete the comparison. Model *eradiate* submitted only 4 bands and HET14 scene, *flies* missed the empirical scenarios and *bhr*, and *wps* did not submit *bhr*. The results for *bhr* are rather consistent across all models, with a slight broadband overestimation of the integrated values of albedo against the submitted albedo (<5 to 10%) for HET07, HET08 HET50, and HET51. Only *dirsig5* showed an underestimation in the *nir* bands for HET50.

The underestimation characterizes the behavior of all models for the remaining scenes (HET15, HET14, and HET16) over VIS and middle infrared bands (>1,600 nm), while in the *nir* the agreements lay within ±5%.

Moving to *dhr*, we observed on average a better agreement (less cells exceeding ±10%), with similar patterns with respect to scenes. The features for HET15 (birchstand winter) are inverted from blue (underestimation) to red (overestimation) tones, and with *flies* showing the worst relative agreement in January.

It would be rather speculative to associate this behavior to a specific cause with the poor fitting of RPV we observed for the birchsand winter model (HET15), while we could try to associate it to the fact that the RPV function was fitted on a dataset with θ_s of 76° and 56°, being the summer season missing in the requirements. This might offset the model representation to these specific conditions, and because for a bell-shaped anisotropy the reflectance descreased with θ_s, we miss the contribution of any lower reflectance terms in the bihemispherical integration, which causes the underestimation seen in *bhr* cells.

The other appreciable misalignment is for HET14 structured scene, which can be explained with similar argument as given for HET15, but reinforced by the fact that the role of the illumination azimuth angle ϕ_s in the description of the reflectance anisotropy is not handled by RPV model (neither by other common BRF functions).

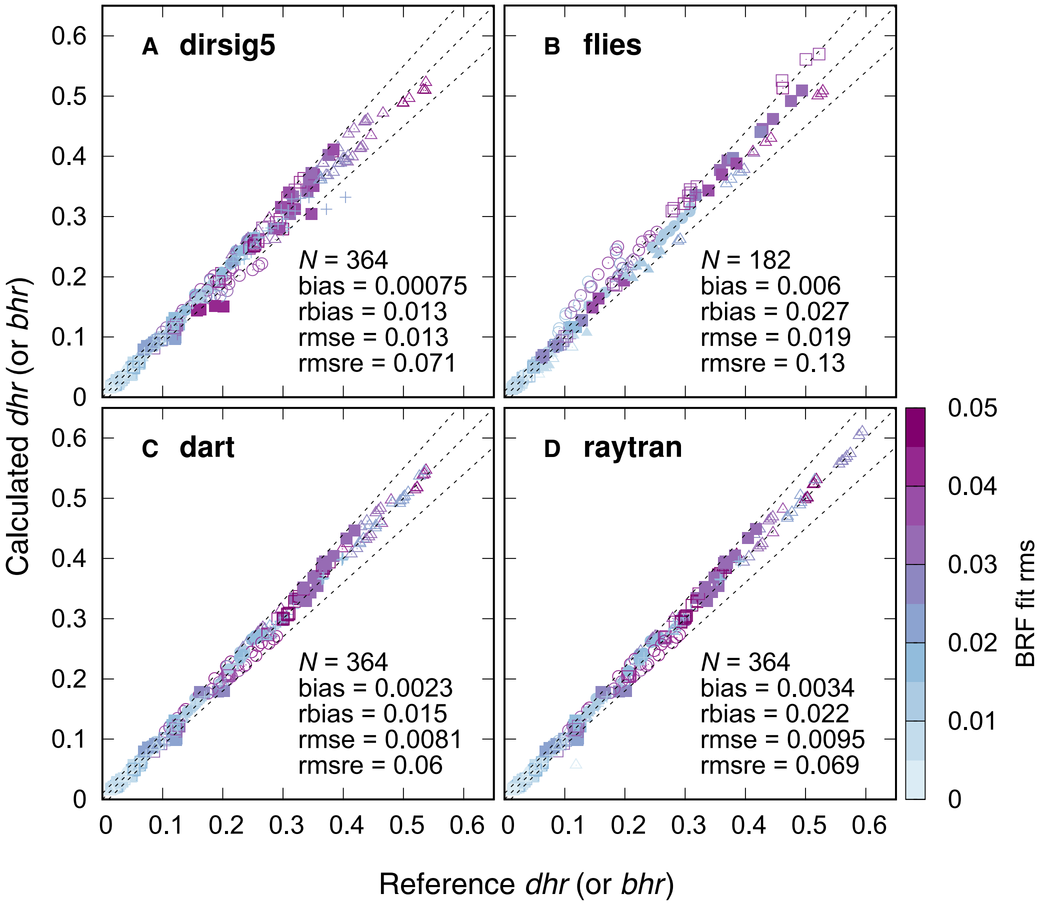


Fig. 8. Scatterplot of the *dhr* (or *bhr*) as calculated from BRF integrals versus the corresponding submissions for dirsig5 (A), flies (B), dart (C), and raytran (D) over all actual canopies: HET07 (▲) and HET08 (Δ), HET09 (●) and HET15 (○), HET14 (■) and HET16 (□), HET50 (+) and HET51 (×). Colors indicate the performance of the RPV optimization expressed in terms of *rms* of the modeled versus measured BRF.

In Table 8, a summary of the overall dispersion estimate expressed by means of some of the classical absolute (BIAS, RMSE) and relative (RBIAS, RMSRE) metrics describing scattered data (see appendix A) is given for each model. The BIAS (and RBIAS) should be taken with care as positive and negative terms cancel out. Absolute values are more representative for higher reflectance ranges, while the relative counterpart is offset toward lower reflectances. With this in mind, we observed that an overestimation of the albedo when computed by means of the integrals of the RPV funcion ($A_{BRF} > A$), of the order of $+2 \div 4 \times 10^{-3}$ (or $+1 \div 2\%$), was common to all models, with the exception of *flies*, which presented an absolute bias of 6×10^{-3} (or 3%). The dispersion of data expressed by RMSE (RMSRE) was of the order of 10^{-2} (or 5% to 10%) for all models, although slightly higher for *flies*.

Albedo versus fluxes at TOC

A flux-related consistency check was made by comparing the *bhr* (or *dhr*) measures with the ratio between the up-welling and downwelling terms at TOC as reported in ftran_tot_vprof measure. The profile of the transmission through the canopy has been submitted by *dart*, *flies*, and *raytran* for all the scenes, and by *spartacus* for 4 actual scenes. The cases with a difference higher than 10% have been flagged as suspicious and reported to the participants. Table 9 reports the median and IQR of the quantity $100 \times \left(\alpha - \alpha_{vprof}\right) / \alpha$, with values varying between −0.02% for *spartacus* and 0.01% for *flies*, confirming the excellent consistency of all the models with respect to this test.

Preliminary screening, identification, and rejection of the outliers

Accordingly to the Global Climate Observing System implementation plan [2] and the ECV requirement documentation, the uncertainties of the surface albedo should respect a threshold (goal) of 5% (3%) for values above 0.05, and to 0.0025 (0.0015) for the surface albedo below 0.05. The FAPAR uncertainty threshold (goal) was set to 10% (5%) for FAPAR values above 0.05 and 0.005 (0.0025) below 0.05. Considering this, any difference among participating models should be considered relevant whenever it is contributing effectively to broke these uncertainty requirements in absolute or relative terms.

The σ-clipping Chauvenet's outlier detection algorithm was adopted [33,34] to identify major discrepancies of the models. This method is particularly suitable to handle datasets with a low number of samples, and in RAMI, the number of independent measurements to be compared ranged between a minimum of 3 to a maximum of 10 depending on the individual experiment. With such a low number of measurements per experiment, the mean and standard deviation of all model's results—fundamental of the Chauvenet's algorithm—were

Table 8. Overall agreement between A_{BRF} versus A as aggregated over all scenes, bands, and illumination angles. The definitions of the metrics are given in Table S1.

Model	BIAS ×10^{-3}	RBIAS %	RMSE ×10^{-3}	RMSRE %	*N*
dart	2.3	1.5	8.1	6.0	364
dirsig5	<1.0	1.3	13.0	7.1	364
eradiate	3.6	0.9	11.0	12.0	16
flies	0.6	2.7	19.0	13.0	182
less	3.5	2.3	9.0	6.4	364
raytran	3.4	2.2	9.5	6.9	364
wps	3.8	2.5	8.2	5.7	260

Table 9. Median and IQR of the difference between *ftran_tot_vprof* and the corresponding *bhr* (or *dhr*). Notes (*): The values obtained for *flies* were filtered to consider only differences below 10%, and the remaining measurements were flagged as suspicious.

RT model	Median	IQR	*n*
dart	<0.005	<0.005	140
flies*	0.01	0.57	30
raytran	0.01	<0.005	140
spartacus	−0.02	0.05	80

considered more reliable than the median (and the IQR), which presents the risk to disproportionately favor the results of a specific model.

A set of summary tables reporting the average performance, in terms of fraction of outliers per experiment, has been distributed to each participant during the review phase to support them in identifying the critical cases. The tables were based on a green-light approach with (a) <10% (green), (b) <25% (yellow), and (c) ≥25% (red) thresholds. The fraction of spurious cases has been calculated on single brf-like experiments for *brf* measure (over $N_{tot} = 76$ values), while it was calculated by aggregating experiments by band ($N_{tot} = 13$or 5 for PAR fluxes) for the other one-value experiments. It was left up to the participant to withdraw, fix, or keep a flagged experiment in the final processing.

Figure S2 shows the mean fraction of outliers detected per experiment at an early stage of the experiment (date: 2022 February 18) and for the final frozen results (2023 September 21) for all measurements, except the filtered *brf* and the *ftran_tot_vprof*. For *raytran* and *renderjay*, the performances, in terms of outlier issues, improved substantially during the project. After the feedback, *raytran* operators were able to fix an issue that produced an overestimated absorption of photons and a general underestimation of the reflectance. Being in development phase, and primarily focused on abstract canopy experiments, *eradiate* was able to submit a full set of reflectances for HET14 in line with the model ensemble, while initially only *bhr* was submitted and identified as an outlier. For *dirsig5* and *flies*, Fig. S2 suggests a decline in performance. However, this is primarily due to other models converging toward a better alignment in their results after the initial review.

Focusing on the final results, Table S2 already indicates in a qualitative way the models that are in rather good agreement, per scene and measure. Excluding *eradiate* because of the low number of experiment submitted, *dart*, *less*, *raytran*, and *wps* were good candidates to verify the conformity across all measurements and scenes. *raytran* was still presenting outliers in the bidirectional reflectance for the Whytham Woods forest. Except over HET14, the flux measurement provided by *spartacus* was not rejected by the test. *flies* remained the model affected by a considerable set of outliers, except for the winter pinestand flux measures, while *dirsig5* showed acceptable agreement with the ensemble, in particular, for the boreal forests (HET07, HET08, and HET09) and HET51.

Model-to-model deviation ($\delta_{m\leftrightarrow c}$) and general model deviation (δ_m)

In Pinty et al. [1,3], the primary criterion to quantify the intermodel variability was a measure of distance between BRF fields generated under identical experiment conditions. It was indicated as local model deviation to quantify absolute relative distance of the bidirectional reflectance provided by a model m against all the other models $c \neq m$, for each observing angle θ_v, defined as $\delta_m(\theta_v) = \frac{2}{N}\sum_{\Omega_s}\sum_\lambda\sum_\zeta\sum_{c\neq m}\left|\frac{x_m-x_c}{x_m+x_c}\right|$, where each x_m and x_c were indexed values, $x_*(\zeta,\lambda,i,j)$. Because of the associative and symmetric properties of the sum, the quantity is representing the average of the normalized absolute difference $NAD(x_m,x_c)$ (Table S1) between x_m and x_c over the common experiments selected. As the addendum varies between the interval [0,2], the resulting metric will also be constrained in this interval, with lower values indicating minor deviation from the selected ensemble of experiments. A general deviation metric was then defined by just summing further on the remaining dimension (θ_v) to obtain a scalar value per model summarizing further the average behavior over any observation angle, $\delta_m = \frac{1}{N_v}\sum_{\Omega_v}\delta_m(\theta_v)$.

The concept of a model-to-model deviation $\delta_{m\leftrightarrow c}$ has been introduced in RAMI-3 [5] to focus on cross-model comparison, and is similar to the quantity $\delta_m(\theta_v)$ defined above, but summation was rearranged by removing the sum over any c model different from m, including the sum over observing angles in place of the sum over different scenarios. It appears as follows:

$$\delta_{m\leftrightarrow c}(\zeta) = \frac{2}{N_{m\leftrightarrow c}}\sum_{\lambda=1}^{N_\lambda}\sum_{i=1}^{N_{\Omega_s}}\sum_{j=1}^{N_{\Omega_v}}\left|\frac{x_c(\lambda,i,j)-x_m(\lambda,i,j)}{x_c(\lambda,i,j)+x_m(\lambda,i,j)}\right|. \qquad (2)$$

Equation 2 represents the $M(M-1)/2$terms of a symmetric matrix, where M is the number of models participating to a specific experiment. The sum is performed over bands λ (with $N_\lambda = 13$ or 5 for PAR flux measurements), illumination i (with $N_{\Omega_s} = 2$ or 3), and viewing j (with $N_{\Omega_v} = 76$) angles, the latter for the *brf* measure only. The number of common experiments $N_{m\leftrightarrow c}$ varied from 26 (39 when 3 illumination conditions) for the flux measurements to 1,976 (2,964) for the *brf*s measurements. By aggregating total *brfpp* and *brfop* to obtain a single summary indicator, these values should be multiplied by a factor of 2. We avoided aggregating $\delta_{m\leftrightarrow c}$ over ζ because of the heterogeneous agreement among the models over different scenarios.

The cluster of models producing the best $\delta_{m\leftrightarrow c}$ metrics was considered to define the benchmark values X_* for the computation of z', and eventually used to define new references to be ingested in the ROMC dataset [6].

Total *brf*

Figure 9 shows the results obtained for all total *brf* measures over different scenes, which are summarized in Table 10. Credible reference group (CRG) indicates the models for which δ_m is less than 2%.

Over the pinestand scenes (HET07 and HET08), we observed an agreement between *less* and *raytran* within <2% (0.7%) for the summer version and within <4% (2.1%) for the winter one, where *dirsig5* was also performing well against these models, with a maximum deviation of 3.9% against *raytran*.

For the birchstand leaf-on model (HET09), we observed *dart*, *less*, and *raytran* agreeing within 2% (1.8%). Within these models, *dirsig5* showed a maximum $\delta_{m\leftrightarrow c}$ of 4.5% against *dart*, while the worst agreement of *wps* was against *raytran* (5.6%). For the winter scene (HET15), *wps* joined the previous group of models with $\delta_{m\leftrightarrow c}$ lower than 2%. On cascade, *dirsig5* presented a $\delta_{m\leftrightarrow c}$ of 13.3% against *less*, while all the other models deviated by more than 20% from the core group.

Concerning citrus orchard (HET14) and birchstand forest (HET16), the group of models formed by *dart*, *less*, *raytran*, and *wps* was showing agreement as better as 0.8% and 1.5%, respectively, the worst combinations over the 2 scenarios. Further, *dirsig5* showed an agreement as better as 10% with this core group of models, while *rapid* and *flies*, over HET14, and *flies* over HET16, exceed this $\delta_{m\leftrightarrow c}$ threshold.

Finally, over savanna (HET50), we observed a very good agreement between *dart* and *raytran* (1.1%), with *less* agreeing within 4% with both models, and *dirsig5*, *librat*, and *rapid* within 10%. Over Wytham Woods forest model (HET51), the better agreement was observed between *less* and *wps* (0.7%), with on cascade *dirsig5* (4.4% at worst), *dart* (6.0%), *rapid* (10.2%), and *raytran* (15%).

Focusing on the tabular representation, we may assert that a good agreement ($\lesssim 2\%$) was observed for HET15 (birchstand winter) and HET16 (short-rotation forest) scenarios with 4 models suitable to form a group to define a benchmark (or CRG). For HET09 (birchstand summer) and HET14 (citrus orchard), the agreement remains of the same quality among the same models, but excluding *wps*. This might be related to minor issues of *wps* in representing the HET14 (2% to 4%) scenario, while the deviation from the CRG over HET09 (8% to 10%) appears induced by larger difficulty of the model to represent forest canopies, or the operator made some mistake or odd assumption while setting up the scene properties.

We had less strong arguments to support a reference group for the pinestand model, because only 2 models (*less* and *raytran*) over 8 (HET07) and 7 (HET08) participants lay within the 0% to 4% classes for either the summer or winter scene. To incorporate a third model (*dart*), a $\delta_{m\leftrightarrow c}$ of up to 12% should be accepted.

Somewhat more promising were the results for savanna (HET50) and Wytham forest (HET51), especially considering the fact that they are new scenes in RAMI. We registered an agreement within 6% over 3 models over HET50 (*dart*, *raytran*, and *less*) and within 4% among *less*, *wps*, and *dirsig5* for HET51.

Remarkably, *less* was the only model belonging to all scene- and measure-dependent CRGs. Contrarily, *flies* always lay outside the $\delta_{m\leftrightarrow c}$ limits considered in the table, by constantly presenting relative deviations above 20%, and further

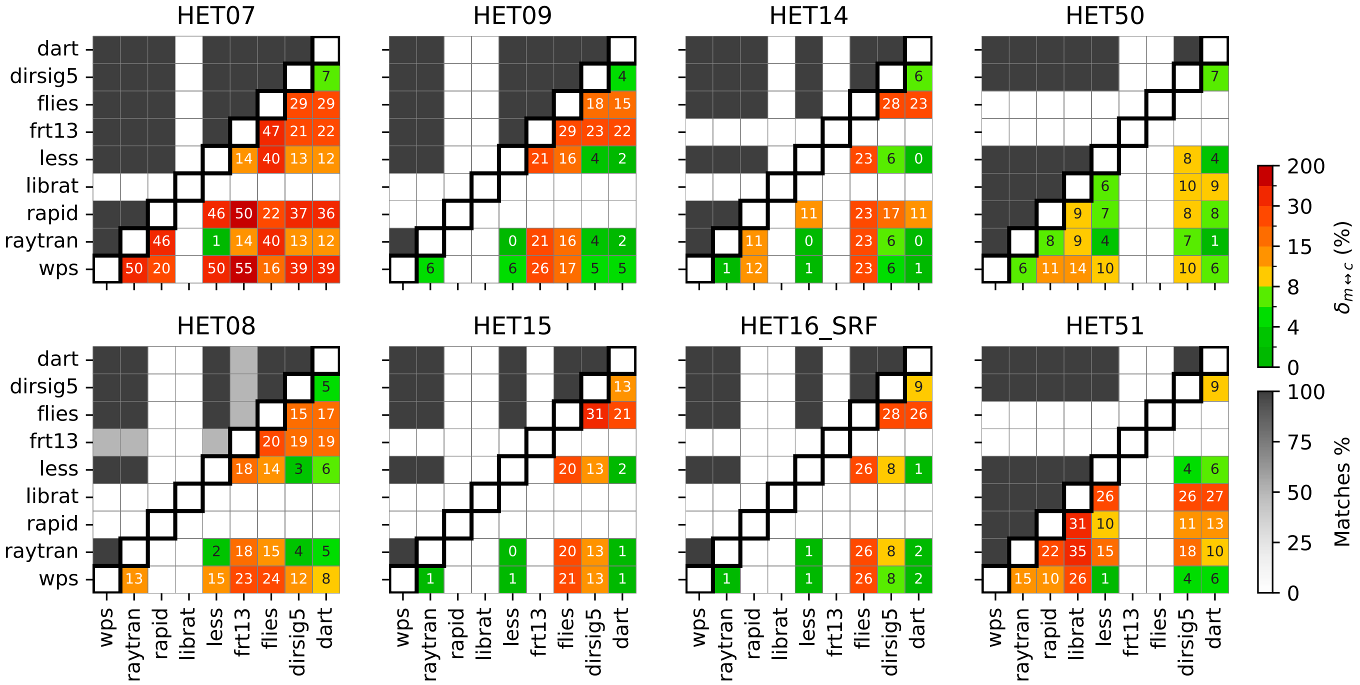


Fig. 9. Total *brfpp* and *brfop* model-to-model comparison calculated using Eq. 2. The lower part reports the $\delta_{m\leftrightarrow c}$ metric over existing matches. The upper part of the matrix reports the model-to-model matches (in % of the maximum number of comparisons) in gray tones. White cells indicate missing participation.

Table 10. Total *brfpp* and *brfop*: List of the models with different level of agreement in terms of Eq. 2 per scene. The level of agreement indicated in the column header identifies the $\delta_{m\leftrightarrow c}$ limits within which the specific model agrees with all the models listed on the left-hand side columns. The last column indicates the number of models belonging to the <2% group and the total participation. Within each cell, the models are always ordered alphabetically.

Scene	$\delta_{m\leftrightarrow c}$ <2% (CRG)	2–4%	4–6%	6–10%	10–20%	Tot models
HET07	less, raytran			dart, dirsig5	dart, dirsig5, frt13	2/8
HET08	–	dirsig5, less, raytran		dart	wps	0/7
HET09	dart, less, raytran		dirsig5, wps		flies	3/7
HET15	dart, less, raytran, wps				dirsig5	4/6
HET14	dart, less, raytran, wps			dirsig5	rapid	4/7
HET16	dart, less, raytran, wps			dirsig5		4/6
HET50	dart, raytran	less		dirsig5, librat, rapid	wps	2/7
HET51	less, wps		dirsig5	dart	rapid, raytran	2/7

verification of the ingestion of RAMI actual scenarios was then recommended.

Table 11 shows, for each scene, the statistics of the metric δ_m, which summarizes the agreement of each model c with all the other models taken together ($c \neq m$). The models are ordered by increasing δ_m, and the results grossly reflect the clustering given in Table 10. The ranges of δ_m also reflect the general agreement among the models over a specific scene, illustrating that over HET07, we obtained the lower agreement (22.4% to 38.2%), while over savanna we observed the most confident agreement among all models (5.8% to 9.5%).

We would also emphasize the specific case of HET07, where *dart* and *dirsig5* appear as the models with the less general deviation from all the model ensemble (Table 11), although the best model-to-model agreement was observed between *less* and *raytran* (Table 10). This reinforces the role of *dart* and *dirsig5* to issue a surrogate true value, despite their lower intermodel agreement (7%). However, we will not emphasize this thesis further, as we have already noted that the HET07 scenario showed the poorest overall agreement among all models. This likely indicates that the scene's input requires clearer definition, or that different models handle the RT process through ellipsoids representing needle-shaped leaves in such varied ways that achieving better alignment is currently challenging.

The difference in the vegetation reflectance between the VIS and NIR bands is used by many algorithms to retrieve the biogeophysical properties. As in RAMI-V, we considered mostly reflectance spectrum describing photosynthetically active vegetation; its range varies from a few percents to more than 0.5 after the red edge, which can influence the evaluation of $\delta_{m\leftrightarrow c}$ mainly because of the value of the denominator in Model-to-model deviation ($\delta_{m\leftrightarrow c}$) and general model deviation (δ_m). We verified how $\delta_{m\leftrightarrow c}$ varied as a function of the wavelength by calculating it over a selection of 3 VIS and 3 NIR bands. We aggregated the total *brf* in these bands over the 4 scenes for which we observed the best agreements. The results are shown in Fig. 10. As expected, being $\delta_{m\leftrightarrow c}$ a relative metric, its values over the VIS are larger, on average, than over the NIR. *frt13* and, on a lower extent *flies* and *rapid*, showed a considerable better agreement in the NIR, suggesting—excluding errors in the experimental setup—that their radiative schemes are sensitive to the strength of the reflectance. On the other hand, dirsig5 shows a slight worst behavior in the NIR. *eradiate*, which was not presented in the previous representation because it only submitted HET14 results, showed excellent agreement with *less*, *raytran*, and *wps* over both the VIS and NIR bands. Figure 10 shows also the comparison for filtered *brf*. It highlights some major inconsistencies that may be related to a wrong configuration of these experiments, especially for the multiple collided and uncollided components. The general impression is that the results obtained for the total *brf* are better, signal of the different approach adopted to obtain the filtered components.

Satellite geometries

Some of the participants submitted the total TOC *brf* for remote sensing actual geometries (*brf_sat*). Figure 11 shows $\delta_{m\leftrightarrow c}$ as aggregated over all bands pertaining, separately, to OLCI, MODIS, and MSI instruments. The results mimic and reinforce what have been observed for the *brf* in the principal and orthogonal planes presented in previous section. The agreement between 3D explicit models over HET07 was confirmed to be the worst, with only *less* and *raytran* presenting values of $\delta_{m\leftrightarrow c}$ below 2%. The agreement among these 3D RT models lay below 5% in terms of $\delta_{m\leftrightarrow c}$ for most of the instruments and scenarios, including the empirical HET50 (≤4%) and HET51. The results for OLCI and MODIS are rather similar for each RT model combinations, with a slightly worsening of the agreement highlighted by higher values of $\delta_{m\leftrightarrow c}$ for *flies*, except over HET15. RT models based on simplified representation of the canopy confirmed their higher values of $\delta_{m\leftrightarrow c}$. Because of the importance of such models in global applications, it is important to continue investigating the possible source of mismatch and to support the experiment with expanded definition of the statistical information needed to ingest the scenarios in such kind of models. This would allow us to distinguish between the problems related to the model physics itself from those related to the assumptions made while adapting the 3D scene to a statistical description suitable to be ingested in the simplified RT model.

Flux measurements

The same approach described in previous section was applied to flux measurements. Figure 12 shows the values of $\delta_{m\leftrightarrow c}$ for

Table 11. Ordered values of the general model deviation metric δ_m for total *brf* (*brfpp* and *brfop*) expressed in %

No.	HET07		HET08		HET09		HET15		HET14		HET16		HET50		HET51	
1	dart	22.4	raytran	9.3	less	8.1	less	7.2	dart	7.0	less	7.3	dart	5.8	wps	10.2
2	dirsig5	22.4	less	9.6	raytran	8.1	raytran	7.2	less	7.0	raytran	7.4	raytran	5.8	less	10.3
3	less	25.2	dirsig5	9.7	dart	8.4	wps	7.3	raytran	7.0	wps	7.5	less	6.5	dart	11.8
4	raytran	25.2	dart	10.1	dirsig5	9.8	dart	7.5	wps	7.2	dart	7.7	dirsig5	8.4	dirsig5	12.1
5	flies	31.9	wps	15.9	wps	10.7	dirsig5	16.7	dirsig5	11.4	dirsig5	12.2	rapid	8.6	rapid	16.2
6	frt13	31.9	flies	17.5	flies	18.4	flies	22.8	rapid	14.4	flies	26.2	wps	9.5	raytran	19.3
7	rapid	36.6	frt13	19.3	frt13	23.7			flies	23.9			librat	9.5	librat	28.3
8	wps	38.2														

dhr and *total absorption*. Concerning *dhr*, we summarized in Table 12(a) the results aimed to identify a set of models agreeing within similar discrete threshold classes presented in the total *brf* analysis, but limiting them to a maximum threshold of 8%.

Excluding HET08, the number of model combinations with a $\delta_{m\leftrightarrow c}$ metric lower than 2% varied between 2 (pinestand and empirical scenes) and 4, with the strongest agreement found over HET15, HET14, and HET16 (if we consider that *dart* maximum $\delta_{m\leftrightarrow c}$ against the CRG models was as better as 2.2%). The models that mostly contribute to this group were *dart*, *less*, *raytran*, and *wps*, as also confirmed by the results obtained for δ_m summarized in Table 13. As expected, the highest disagreement among models occurred over HET07, although a good agreement was confirmed between *less* and *raytran* hemispherical reflectances. Better matches were observed over HET15 (δ_m ranging between 3.8% and 10.4%) and HET50 (4.1% and 6.8%). This was likely related to the higher reflectance also in the VIS bands, contributing with lower values of the addendum fractions defining both $\delta_{m\leftrightarrow c}$ and δ_m. As *less* and *raytran* are consistent within 2% over 5 scenes, we considered them to form a custom reference for *dhr* and *bhr* in the next sections.

Similarly, Table 12(b) summarizes our findings for total absorption experiments. Although in the previous cases we always identified a single set of CRG group, here we encountered an additional issue. Over HET07, we observed 2 possible groups, each formed by 2 models, in the <2% class category, a first one formed by *less* and *raytran*, and a second formed by *dart* and *spartacus* (we excluded *renderjay* as it only submitted band O03 out of the 5 PAR bands pertaining to this comparison). In such cases, expressing a preference supported by the current RAMI-V results may be not trivial. On cascade, *wps* reinforces the first group with a $\delta_{m\leftrightarrow c}$ lower than 6%. Over HET08, 3 groups were identified with $\delta_{m\leftrightarrow c} < 2\%$; while extending the threshold up to 8%, *dart*, *wps*, and *flies* progressively increase the credit of the group formed by *less* and *raytran*. Table 12(b) shows that over birchstand scenarios, the structured canopies (HET14 and HET16) and Wytham Woods forest (HET51), the agreements among 4 (5) of 6 (7) models were good enough to consider issuing a benchmark reference value set for the absorption over the PAR spectral region. Remarkably, all the 4 participants to HET51 agreed within 2% for this measurement. Finally, over savanna (HET50), the best agreement was registered by *less* and *wps* (<4%), with the addition of *dart* by extending the threshold up to 6%. In line with the choice made for *dhr*, and considering that *less* and *raytran* agree within 2% over 7 of the 8 scenarios proposed, we selected them to calculate a surrogate custom reference for total absorption. It should be mentioned that the group formed by *dart* and *wps* could be a valid alternative over 6 of 8 scenes.

Table 12(c) summarizes the results obtained for the below canopy transmissions, aggregated over diffuse and direct illumination conditions. The model agreement over HET07 is confirmed to be problematic. Over HET08, *flies*, *renderjay*, and *spartacus* showed the best matches (<2%). The group formed by *dart*, *raytran*, and *wps* presented a good agreement over birchstand and structured canopy (HET14 and HET16) scenarios, with *renderjay* enforcing the group for a $\delta_{m\leftrightarrow c}$ threshold of 4% (8%) over the winter (summer) scene. The best match over savanna was among *dart* and *wps*, and no matches below 8% were observed over Wytham Woods. We decided to rely on the group formed by *dart*, *raytran*, and *wps* to define the custom reference over the 4 scenes from HET09 to HET16 as ordered in Table 12(c).

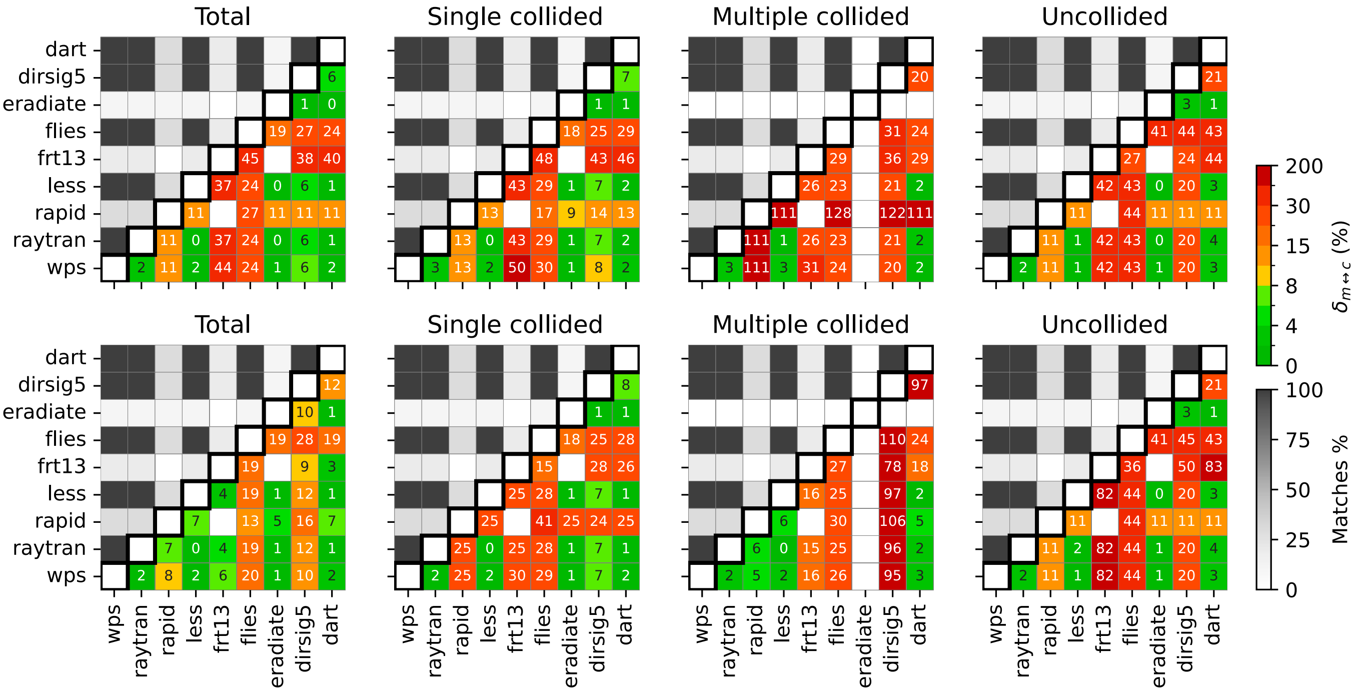


Fig. 10. As in Fig. 9 but with the aggregation performed over HET09, HET15, HET14, and HET16 for a set of consecutive VIS (O04, O06, O08)—first line of panels—and NIR (O12, M08, O17) bands, representative of the spectral ranges 490 to 665 nm and 750 to 865 nm, respectively. As *eradiate* submitted HET14 results, it replaced *librat* here.

The difference of the proficiency assessment made by using these custom references or the robust average methodology is discussed in the next session.

Determination of a surrogate reference through a custom selection (X^*) or a robust statistic (X_m^*)

Based on the findings of the previous section, we decided to use *dart*, *less*, and *raytran* results to issue a custom reference (X^*) for total *brf*—over HET09, HET15, HET14, and HET16. For flux measurements, the pair *less*–*raytran* was selected to form a custom reference for the joint hemispherical reflectance *bhr* and *dhr* (including HET07 in the list of scenes). Concerning *fabs_tot*, we continued to rely on *less*–*raytran*, as their $\delta_{m\leftrightarrow c}$ matched the 2% thresholds over all scenes except HET50 only. Nevertheless, an assessment of model proficiency assuming *dart*–*wps* as an alternative reference has been conducted and discussed. For *ftran_tot*, we relied on *dart*, *raytran*, and *wps* models, limiting the analysis over the same scenarios indicated for the total *brf* case.

Among these choices, we also approached the problem of proficiency testing by means of an alternative model-dependent reference (X_m^*), based on the same robust statistic described and adopted by Widlowski et al. [9] following ISO-13528 guidelines. The difference arising from the 2 approaches (e.g., custom versus robust statistic-based references) has been discussed.

For instance, the custom reference was set to the arithmetic average of the values provided by the models selected from the $\delta_{m\leftrightarrow c}$ analysis, and its uncertainty (u_{X^*}) to the reference group standard deviation. Concerning the robust approach, the reference ($u_{X_m^*}$)—*m*-index to indicate the model dependence reference—was derived by a combination of median value and an iterative outlier replacement, performed until the convergence of the model average lay within 10^{-6} (the RAMI result number format). The convergence was normally achieved in few iterations. Concerning its uncertainty, a conservative approach was adopted by defining $u_{X_m^*} = 1.25 s^*/\sqrt{N}$ and $s_m^* = 1.134\sqrt{\frac{1}{N-1}\sum\left(x_i^* - X_m^*\right)^2}$ [8].

Tables 14 and 15 report additional statistics originating from the comparison between the models belonging to the CRG, per scene and measure. Statistics such as the number of samples *N*, the relative mean bias error (rMBE), the root mean square error (RMSE), the signal-to-noise ratio (SNR), and the Pearson correlation coefficient (r^2) were also used in previous phases [9] (their figure 7 and appendix) to describe the uncertainty level of the surrogate reference identified for the filtered *brf* measurements over abstract canopies only. We included here *t-stat* and U_{95} because of their popularity in the characterization of the uncertainty associated to references obtained from the combination of 2 or more model results [35,36].

They were computed on a model-to-model basis, providing the statistics for all the combinations of models belonging to the CRG, being them [*dart*, *less*, and *raytran*] or [*dart*, *less*, and *wps*] for BRF and albedo, and absorption and transmission, respectively. To be conservative, the worst value of the metric among the 3 combination of models can be assumed to assess the credibility of the surrogate reference. In particular, the worst u_X among the combinations between the 3 models belonging here to CRG was assumed as representative of the benchmark uncertainty $u_X{}^*$.

Hence, by focusing on the values obtained by combining the 4 selected scenes [labeled as All in Tables 14 and 15 (a and b)], we obtained values of $u_X{}^*$ of 2.2×10^{-5}, 2.7×10^{-4}, and 5.9×10^{-4} for BRF, albedo, and fluxes, respectively. The analog statistics for each scene can be easily extracted form the same tables. The Pearson coefficient, not reported in the tables, confirmed an excellent correlation ($\gtrsim 0.9995$) on average over all the selected combinations.

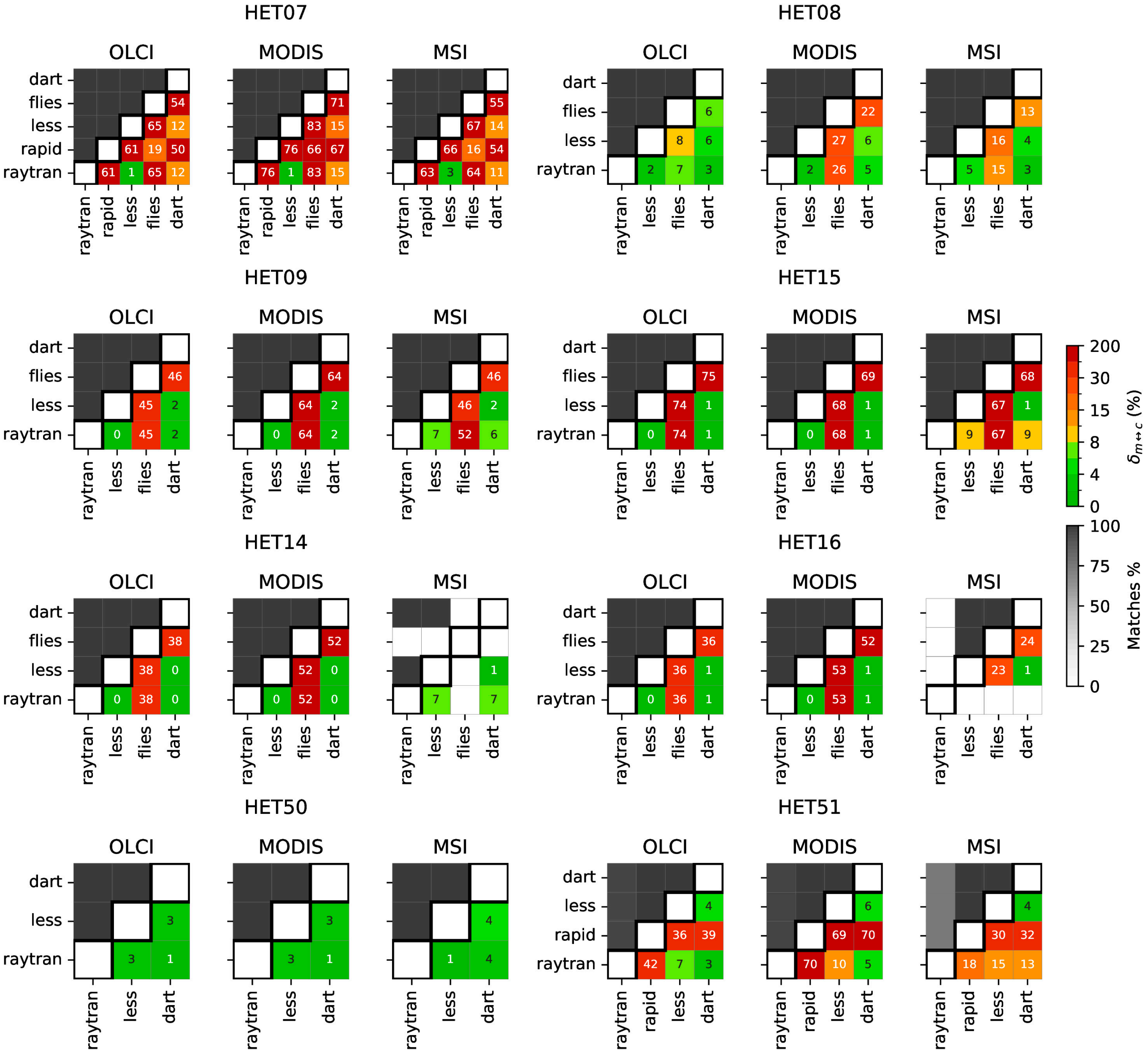


Fig. 11. Top-of-canopy *brf_sat* model-to-model comparison calculated using Eq. 2. The lower part reports the $\delta_{m\leftrightarrow c}$ metric over existing matches. The upper part of the matrix reports the model-to-model participation (%) in gray tones, and white cells indicate missing participation. The aggregation is performed per scene over all bands pertaining to OLCI, MODIS, and MSI.

The RMSE values observed here were almost an order of magnitude higher than those reported for abstract canopies in RAMI-IV (~10^{-4}), indicating a larger dispersion of the results, in line with the expectation for the more complex canopies under investigation. The SNRs varied between ~50 and ~500 for BRF, in line with the higher dispersion of the data observed for the abstract canopies in RAMI-IV, but assumed values similar to those observed for abstract canopies in RAMI-IV (>200, their figure 7) for the *raytran* against *less* combinations (R-L), especially over HET15 and HET14 for the total BRF.

Finally, when aggregating all scenes, the rMBE and σ_{MBE} values in absolute terms were −0.28% ± 0.21% for BRF (D-L case), 0.18% ± 0.10% for hemispherical reflectance, 0.45% ± 0.52% for the total absorption, and −1.0% ± 0.38% (R-W) for the total transmission. The values related to BRF were an order of magnitude higher than those observed for abstract canopies in RAMI-IV among *rayspread* and *librat*, but still indicated an average agreement as better as 0.3%, which is considerably lower than the proficiency criteria of 3% assumed for $\hat{\sigma}^2$. The values of the absolute uncertainty at 95% confidence level ($U_{95\%}$ All cases) were, on average, 7.1 × 10^{-3} (*brf*) (D-R case), 2.8 × 10^{-3} (*bhr* and *dhr*), 1.6 × 10^{-2} for *fabs_tot*, and 1.6 × 10^{-2} for textitftran_tot (D-R), with corresponding SNRs of 59, 137, 142, and 48, respectively.

Proficiency testing (by *z′* metric)

The aim of a proficiency assessment is to determine whether the deviations of a model from the reference (a) lay within an

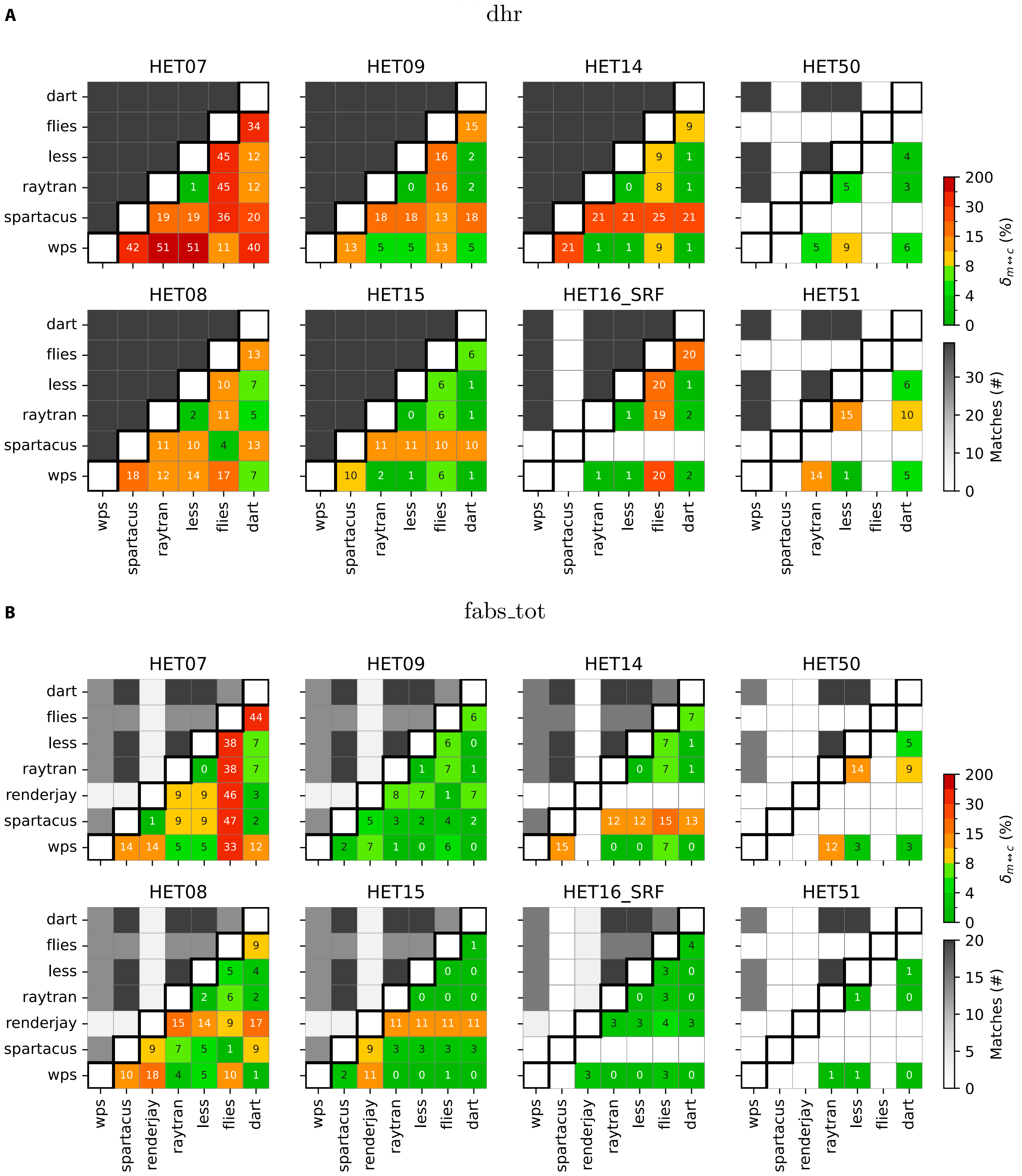


Fig. 12. Directional–hemispherical reflectance (A) and total absorption (B) model-to-model comparison calculated using Eq. 2. The upper part of each matrix reports the combined model-to-model participation in green tones. White cells indicate missing participation. The lower part reports the $\delta_{m\leftrightarrow c}$ aggregated over common illumination angles and relevant bands (all for *dhr* and PAR bands for absorption, but *renderjay* submitted only one band for *fabs_tot*). White cells indicate that the comparison was prevented due to missing matches.

acceptable confidence interval and/or (b) are explainable by the model uncertainties. To quantify the proficiency, metrics such as χ^2, t statistic, E_n, or z' can be adopted [8,35]. They all express a measure of the deviation of the model from a reference value, normalized against the combined model and reference uncertainties, or a required accuracy. The fitness for purpose of the

Table 12. As in Table 10 but for flux measurements (a) *dhr*, (b) *fabs_tot*, and (c) *ftran_tot*. Model *less* did not submit transmission, and *renderjay* generally does not appear in the table because it only submitted band 003. The indexes identify a group, or the agreement of a single model with a specific group.

Scene	<2% (CRG)	2–4%	4–6%	6–8%	Tot models
(a) Directional–hemispherical reflectance					
HET07	less, raytran	–	–	–	2/6
HET08	–	(less, raytran)^{1}, (flies, spartacus)^{2}	–	dart^{1}	0/6
HET09	dart, less, raytran	–	wps	–	3/6
HET15	dart, less, raytran, wps	–	–	flies	4/6
HET14	dart, less, raytran, wps	–	–	flies (<10%)	4/6
HET16	less, raytran, wps	dart (<2.2%)	–	–	3(+)/6
HET50	–	(dart, raytran)^{1}, (dart, less)^{2}	less	wps (<10%)	0/4
HET51	less, wps	–	dart	–	2/4
(b) Total absorption					
HET07	(less, raytran)^{1}, (dart, spartacus)^{2}	–	wps^{1}	–	2/7
HET08	(less, raytran)^{1} (flies, spartacus)^{2} (dart, wps)^{3}	dart^{1}	wps^{1}	flies^{1}	2/6
HET09	dart, less, raytran, wps	spartacus	–	flies	4/7
HET15	dart, less, raytran, wps, flies	spartacus	–	–	5/7
HET14	dart, less, raytran, wps	–	–	flies	4/6
HET16	dart, less, raytran, wps	flies	–	–	4/6
HET50	–	less, wps	dart		0/4
HET51	dart, less, raytran, wps				4/4
(c) Transmission					
HET07	–	(renderjay, spartacus)^{1}, (dart, spartacus)^{2}	dart	–	0/6
HET08	flies, renderjay, spartacus	(dart, raytran)	–	dart	3/5
HET09	dart, raytran, wps	(flies, renderjay)	–	spartacus	3/6
HET15	(dart, wps)^{1}, (dart, raytran)^{2}	dart, raytran, spartacus, wps	–	–	(2)/6
HET14	dart, raytran, wps	–	–	–	3/5
HET16	raytran, wps	dart	–	renderjay	0/6
HET50	–	–	dart, wps	–	0/3
HET51	–	–	–	–	0/3

model is then evaluated by calculating how many times it is compliant with predefined acceptability thresholds.

Coherently with previous RAMI phases, we selected the metric z' to discuss the proficiency of the model over selected cases. It is defined as

$$z'\left(m,\lambda,\zeta,\Omega_s,\Omega_v\right)=\frac{x_m^*\left(\lambda,\zeta,\Omega_s,\Omega_v\right)-X^*\left(\lambda,\zeta,\Omega_s,\Omega_v\right)}{\sqrt{\hat{\sigma}^2\left(\lambda,\zeta,\Omega_s,\Omega_v\right)+u_{X^*}^2\left(\lambda,\zeta,\Omega_s,\Omega_v\right)}} \tag{3}$$

where x_m^* is the virtual measure under investigation, X^* is the assumed reference value (derived either from a subset of models or from a robust statistic approach), $\hat{\sigma}$ is the standard uncertainty of the proficiency criteria, and u_{X^*} is the standard uncertainty associated to the reference value. We set $\hat{\sigma}$ either to (a) $f \cdot X^*$with $f = 0.03$for *brf*, (b) $max\left(0.0025,0.05X_R\right)/\sqrt{3}$ for albedo, or (c) $max\left(0.05,0.10X_A\right)/\sqrt{3}$ for absorption measurements, in line with the Global Climate Observing System (GCOS) recommendations and RAMI-IV assumptions [2,8,9].

For total *brf*, the fraction of cases for which the uncertainty of the standard reference value could be considered negligible with respect to the proficiency assessment ($u_X < 0.3 \cdot \hat{\sigma}$) was 49.3%. Being this value in line with the values reported by Widlowski et al. [9], we adopted the same prudent approach in the assessment of the proficiency, performing it by means of a z' metric (i.e., by including u_X in Eq. 3), rather than relying on a simple metric z, for which the denominator is defined by σ only. It is worth to recall that ISO-13528 associates proficiency compliance (later on indicated with a C) to any $|z'| < 2$ event and indicates that for any z' such that $2 \leq |z'| \leq 3$, a warning signal should be issued (W). For any $|z'| > 3$, ISO-13528 recommends to address an Action sign to the participant (A). We calculated the fraction of cases belonging to C-W-A classes for each model and scenario.

In this work, the reference values were obtained by (a) averaging *dart*, *less*, and *raytran* results for *brf* (or *less* and *raytran*

Table 13. Ordered values of the general model deviation metric δ_m for *dhr*, *fabs_tot*, and *ftran_tot* expressed in %

dhr	HET07		HET08		HET09		HET15		HET14		HET16		HET50		HET51	
1	dart	23.5	raytran	8.3	dart	8.1	dart	3.8	less	6.3	less	5.8	dart	4.1	wps	6.8
2	less	25.7	less	8.7	wps	8.3	wps	3.9	raytran	6.3	raytran	5.9	raytran	4.3	dart	7.0
3	raytran	25.7	dart	9.1	less	8.3	less	4.0	dart	6.5	wps	6.1	less	5.8	less	7.2
4	spartacus	27.2	flies	10.8	raytran	8.3	raytran	4.1	wps	6.6	dart	6.3	wps	6.8	raytran	13.0
5	flies	34.2	spartacus	11.3	flies	14.7	flies	6.9	flies	11.9	flies	19.7				
6	wps	39.2	wps	13.6	spartacus	16.0	spartacus	10.4	spartacus	21.8						
fabs_tot																
1	raytran	11.4	less	5.7	wps	2.7	wps	2.4	less	3.9	wps	1.4	wps	6.0	wps	0.5
2	less	11.5	raytran	6.0	dart	2.8	dart	2.4	raytran	3.9	less	1.4	dart	6.0	dart	0.6
3	dart	12.4	flies	6.7	less	2.8	less	2.4	dart	4.1	raytran	1.5	less	7.4	raytran	0.8
4	spartacus	13.6	spartacus	6.9	spartacus	2.9	raytran	2.4	wps	4.4	dart	1.7	raytran	11.9	less	0.9
5	renderjay	13.9	dart	7.1	raytran	3.3	flies	2.6	flies	8.4	renderjay	3.4				
6	wps	14.0	wps	8.2	flies	4.8	spartacus	3.7	spartacus	13.5	flies	3.7				
7	flies	41.2	renderjay	13.9	renderjay	5.7	renderjay	10.6								
ftran_tot																
1	dart	20.1	renderjay	5.5	spartacus	12.7	dart	10.4	raytran	10.6	raytran	7.3	dart	6.7	dart	18.4
2	raytran	20.5	dart	5.7	wps	13.4	spartacus	10.5	dart	10.8	wps	7.5	wps	8.8	raytran	22.0
3	spartacus	22.4	spartacus	5.8	raytran	13.4	raytran	10.6	wps	12.4	dart	8.9	raytran	10.6	wps	31.6
4	renderjay	22.7	raytran	6.4	dart	13.5	wps	10.9	flies	18.0	renderjay	9.6				
5	wps	23.0	flies	6.9	flies	22.5	flies	14.7	spartacus	35.3	flies	18.9				
6	flies	53.2	wps	9.9	renderjay	23.7	renderjay	37.5								

Table 14. Statistics of the comparison between the combination of models *c* and *m* belonging to the CRG, as computed on the aggregation of specific scenes as listed in the first column, for total *brf*. In the m-c column, the codes D-L, D-R, and L-R are shortcuts for *dart* against *less*, *dart* against *raytran*, and *less* against *raytran*.

Scene	m-c	meas	N	MBE	RMSE	r^2	rMBE	σ_{MBE}	t stat	SNR	U95	u_X
HET09	D-L	brfpp-|brfop-	3,952	3.9×10^{-4}	1.9×10^{-3}	0.999916	-9.3×10^{-3}	1.9×10^{-3}	13.2	59.2	5.2×10^{-3}	3.7×10^{-5}
	D-R	brfpp-|brfop-	3,952	5.4×10^{-4}	2.2×10^{-3}	0.999893	-8.9×10^{-3}	2.1×10^{-3}	16.2	52.8	5.9×10^{-3}	4.2×10^{-5}
	L-R	brfpp-|brfop-	3,952	1.5×10^{-4}	3.8×10^{-4}	0.999995	3.6×10^{-4}	3.5×10^{-4}	26.4	318.4	1.0×10^{-3}	6.9×10^{-6}
HET15	D-L	brfpp-|brfop-	3,952	-2.3×10^{-3}	2.8×10^{-3}	0.999938	-1.5×10^{-2}	1.6×10^{-3}	91.2	110.0	6.3×10^{-3}	3.2×10^{-5}
	D-R	brfpp-|brfop-	3,952	-2.2×10^{-3}	2.8×10^{-3}	0.999920	-1.5×10^{-2}	1.7×10^{-3}	83.9	106.3	6.3×10^{-3}	3.3×10^{-5}
	L-R	brfpp-|brfop-	3,952	1.1×10^{-4}	3.5×10^{-4}	0.999996	4.8×10^{-4}	3.3×10^{-4}	21.0	531.1	9.5×10^{-4}	6.6×10^{-6}
HET14	D-L	brfpp-|brfop-	5,928	-2.2×10^{-4}	7.4×10^{-4}	0.999996	1.1×10^{-3}	7.1×10^{-4}	24.4	224.8	2.0×10^{-3}	1.1×10^{-5}
	D-R	brfpp-|brfop-	5,928	-1.8×10^{-4}	7.5×10^{-4}	0.999994	1.2×10^{-3}	7.3×10^{-4}	19.0	218.7	2.0×10^{-3}	1.2×10^{-5}
	L-R	brfpp-|brfop-	5,928	4.5×10^{-5}	3.4×10^{-4}	0.999998	1.7×10^{-4}	3.3×10^{-4}	10.3	475.6	9.3×10^{-4}	5.4×10^{-6}
HET16	D-L	brfpp-|brfop-	5,928	1.3×10^{-3}	2.5×10^{-3}	0.999925	6.1×10^{-3}	2.2×10^{-3}	46.4	70.1	6.6×10^{-3}	3.5×10^{-5}
	D-R	brfpp-|brfop-	5,928	2.1×10^{-3}	3.6×10^{-3}	0.999865	1.2×10^{-2}	2.9×10^{-3}	54.0	51.7	9.1×10^{-3}	4.8×10^{-5}
	L-R	brfpp-|brfop-	5,928	7.5×10^{-4}	1.6×10^{-3}	0.999951	5.7×10^{-3}	1.4×10^{-3}	40.0	104.8	4.3×10^{-3}	2.3×10^{-5}
All	D-L	brfpp-|brfop-	19,760	-5.8×10^{-5}	2.1×10^{-3}	0.999882	-2.8×10^{-3}	2.1×10^{-3}	3.9	71.8	5.8×10^{-3}	1.9×10^{-5}
	D-R	brfpp-|brfop-	19,760	2.3×10^{-4}	2.5×10^{-3}	0.999830	-8.4×10^{-4}	2.5×10^{-3}	12.9	59.3	7.1×10^{-3}	2.3×10^{-5}
	L-R	brfpp-|brfop-	19,760	2.9×10^{-4}	9.4×10^{-4}	0.999980	1.9×10^{-3}	8.9×10^{-4}	45.7	168.4	2.5×10^{-3}	8.0×10^{-6}

Table 15. (a) As in Table 14 but for albedo measurements (*bhr* and *dhr* related experiments were aggregated). (b) As in Table 14 but for total absorption and (c) transmission. In the m-c column, the models are labeled as (L)ess, (R)aytran, (D)art, and (W)ps. "All" refers to the aggregation of the scenes indicated in each specific table.

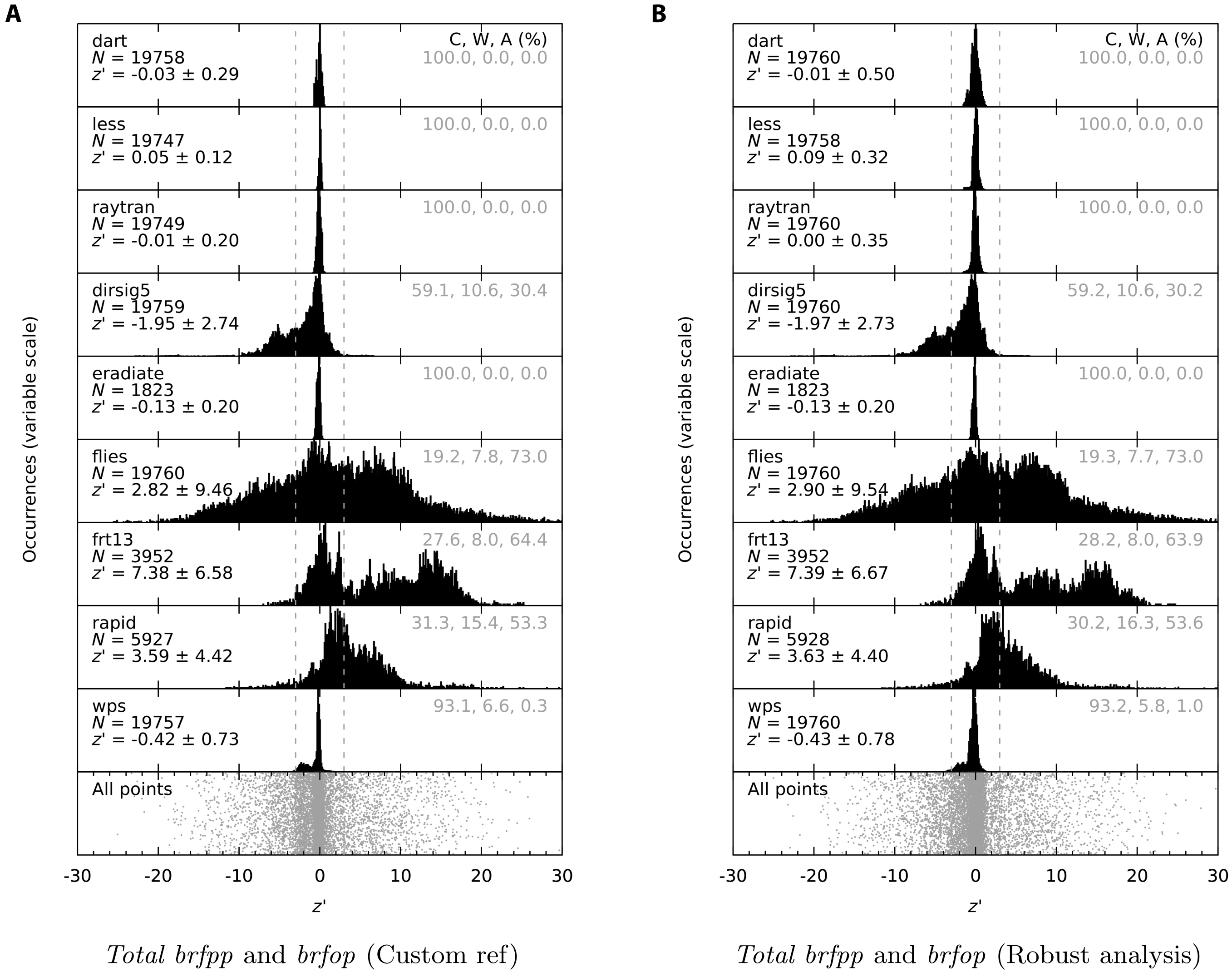
Scene	m-c	meas	N	MBE	RMSE	r^2	rMBE	σ_{MBE}	t stat	SNR	U95	u_X
(a) *bhr* and *dhr*												
HET07	L-R	bhr-|dhr-	39	-2.6×10^{-4}	7.8×10^{-4}	0.999973	1.4×10^{-3}	7.5×10^{-4}	2.2	86.9	2.1×10^{-3}	1.5×10^{-4}
HET09	L-R	bhr-|dhr-	39	3.7×10^{-4}	6.0×10^{-4}	0.999993	2.4×10^{-3}	4.8×10^{-4}	4.8	233.8	1.5×10^{-3}	9.6×10^{-5}
HET15	L-R	bhr-|dhr-	39	-2.2×10^{-4}	4.7×10^{-4}	0.999979	-1.1×10^{-3}	4.2×10^{-4}	3.3	441.4	1.2×10^{-3}	8.4×10^{-5}
HET14	L-R	bhr-|dhr-	52	-7.2×10^{-5}	4.0×10^{-4}	0.999996	6.4×10^{-5}	4.0×10^{-4}	1.3	404.5	1.1×10^{-3}	6.9×10^{-5}
HET16	L-R	bhr-|dhr-	52	9.7×10^{-4}	1.8×10^{-3}	0.999924	5.7×10^{-3}	1.6×10^{-3}	4.4	96.9	4.7×10^{-3}	2.7×10^{-4}
All	L-R	bhr-|dhr-	221	1.9×10^{-4}	1.0×10^{-3}	0.999958	1.8×10^{-3}	1.0×10^{-3}	2.8	137.1	2.8×10^{-3}	8.4×10^{-5}
(b) Total absorption												
HET07	L-R	fabstot-	15	-1.6×10^{-3}	1.8×10^{-3}	0.999920	-2.6×10^{-3}	7.1×10^{-4}	8.8	900.3	3.7×10^{-3}	2.3×10^{-4}
HET08	L-R	fabstot-	15	9.0×10^{-3}	9.0×10^{-3}	0.999993	1.6×10^{-2}	6.3×10^{-4}	55.0	932.2	1.8×10^{-2}	2.0×10^{-4}
HET09	L-R	fabstot-	15	5.0×10^{-3}	5.1×10^{-3}	0.999874	6.4×10^{-3}	7.3×10^{-4}	26.7	1,076.0	1.0×10^{-2}	2.3×10^{-4}
HET15	L-R	fabstot-	15	1.5×10^{-4}	2.7×10^{-4}	0.999992	1.9×10^{-4}	2.3×10^{-4}	2.4	3,401.7	6.9×10^{-4}	7.5×10^{-5}
HET14	L-R	fabstot-	20	2.1×10^{-4}	4.5×10^{-4}	0.999997	2.4×10^{-4}	4.0×10^{-4}	2.3	1,714.1	1.2×10^{-3}	1.1×10^{-4}
HET16	L-R	fabstot-	20	-3.7×10^{-4}	1.2×10^{-3}	0.999969	-5.1×10^{-4}	1.1×10^{-3}	1.5	627.6	3.1×10^{-3}	3.1×10^{-4}
HET51	L-R	fabstot-	20	1.1×10^{-2}	1.2×10^{-2}	0.999099	1.2×10^{-2}	4.3×10^{-3}	12.0	218.7	2.5×10^{-2}	1.2×10^{-3}
All	L-R	fabstot-	120	3.4×10^{-3}	6.2×10^{-3}	0.999389	4.5×10^{-3}	5.2×10^{-3}	7.3	142.1	1.6×10^{-2}	5.9×10^{-4}
(c) Total transmission												
All [a]	D-R	ftrantot-	70	6.2×10^{-4}	5.9×10^{-3}	0.999017	6.8×10^{-3}	6.0×10^{-3}	0.9	46.7	1.6×10^{-2}	8.9×10^{-4}
All	D-W	ftrantot-	50	7.8×10^{-4}	1.6×10^{-3}	0.999954	7.5×10^{-3}	1.4×10^{-3}	4.0	185.8	4.1×10^{-3}	2.4×10^{-4}
All	R-W	ftrantot-	50	-2.5×10^{-3}	4.6×10^{-3}	0.999695	-1.0×10^{-2}	3.8×10^{-3}	4.6	65.9	1.2×10^{-2}	6.8×10^{-4}

[a] For transmission only, the label "All" refers to HET09, HET15, HET14, and HET16.

for *hemispherical reflectances* and total absorption) to issue a custom benchmark, as well as by (b) applying a full robust analysis based on all model results as described by Widlowski et al. [9]. It is worth mentioning that the second approach implies model-dependent references, as the model under testing is left out from the computation of the robust average.

Figure 13A and B shows the histograms of z' obtained following these 2 approaches for the definition of X_*. The results were aggregated over the 4 scenes for which we were able to identify a surrogate reference as described in the previous section. For the models used to define the reference, the average and standard deviation of z' were -0.03 ± 0.29 (*dart*), 0.05 ± 0.12 (*less*), and 0.01 ± 0.20 (*raytran*) in case of custom reference approach. Similar values of the average z' were obtained from the robust analysis, with slightly large standard deviations $\sigma_{z'}$. For the remaining models, the dispersion was markedly higher, with *flies* and *frt13* presenting higher dispersion. Model *eradiate* submitted total *brf* over HET14 only, while it showed promising results. Figure 13 shows also the proficiency compliance report in terms of the fractions C, W, and A defined above and based on the module of z'. For the total brf, despite the approach chosen to define the reference value, 100% compliance was observed for *less*, *dart*, *raytran*, and *eradiate*. The higher dispersion of z′ observed for the other models triggers different percentages of warning or action signals. It is interesting to observe also the low dependence of z' statistics and C-W-A fractions by the method chosen to define the reference. The analysis based on the robust statistic described in ISO-13528 allows to easily characterize the proficiency of the various models, without necessarily going through a deep interpretation of the model-to-model comparison. The approach based on the robust mean will likely characterize any further evolution of the RAMI experiment, since the first evaluation period, in order to issue any useful warning and action signal to the participants, but still respecting its traditional blindness approach.

Figure S3A and B (provided in the digital annex) shows the histograms of z' for the aggregated *dhr* and *bhr*, and the total absorption, respectively, produced by means of the robust statistics analysis. Concerning hemispherical reflectance (total



Total brfpp and *brfop* (Custom ref)

Total brfpp and *brfop* (Robust analysis)

Fig. 13. Histograms of z' statistics for (A and B) total BRF simulations grouped over HET09, HET15, HET14, and HET16 scenes. Histograms are scaled vertically to the maximum value of occurrence. The number of cases *N*, and the average and standard deviation of z' are reported for each model. The labels on the right hand side report the compliant (C), warning (W), and action (A) z' fractions as defined in the text.

Table 16. Variation of model compliance based on z' for different reference choices. Note that *eradiate* submitted only 4 bands and HET14. All the values are expressed in %.

	dhr and *bhr*				*fabs_tot*		*ftran_tot*	
Model	Robust	L-R	D-W	D-L-R-W	Robust	L-R	Robust	D-R-W
dart	100	82.4	100	100	100	100	100	100
less	97.4	100	100	100	100	100	–	–
raytran	98.1	100	100	100	100	100	100	100
dirsig5	75.6	59.3	78.2	75	–	–	–	–
eradiate	100	100	100	100	–	–	–	–
flies	28.8	28.8	47.4	46.2	78.6	78.6	46.0	46.0
renderjay	–	–	–	–	81.8	81.8	42.9	42.9
spartacus	30.8	20.7	46.2	44.4	87.5	85.5	71.4	71.4
wps	80.1	76.9	100	100	100	100	100	100

absorption), the aggregation was performed over the 5 (7) scenes indicated in Table 15(a) [Table 15(b)].

For hemispherical reflectance (*fabs_tot*), the $\sigma_{z'}$ were 0.88 (0.27) and 0.89 (0.26) for *less* and *raytran*, respectively, following the robust methodology. *dart* presented a lower dispersion (0.53) but a higher bias (0.38) for the combined *dhr* and *bhr* case. Concerning both albedo and absorption, *wps* also showed good compliance achievements, while *eradiate* was not described further as it only submitted a few cases pertaining to only one band for the albedo measure.

To verify to what extent the statistics of z' and the compliance (C-W-A) fraction change depending on the method adopted to calculate the reference, we compared in Table 16 the successful occurrence for hemispherical reflectance, total absorption, and transmission. The results are representative of the aggregated analysis over the selected scenes that depend on the type of measurement, as summarized in Table 15. The compliance of dart, less, and raytran is always above 95% of the cases, except for *dart* when we defined the hemispherical reflectance reference based on *less* and *raytran* (82.4%). This is related to the deviation of *dart* results over HET07. The promising excellence proven by *eradiate* is limited to the 4 spectral bands results submitted for HET14 scenario only. For absorption and transmission, the robust analysis provided exactly the same results as the analysis based on the custom references, with a small difference affecting the compliance of *spartacus* (87.5 against 85.5).

Conclusion

Fourteen models participated to RAMI-V phase including 7 new comers, and 12 of them submitted results for the actual scenarios. *dart* and *raytran* completed both the full set of the proposed experiments, except that *raytran* did not complete the *ftran_loc*. Since only *dart* submitted results for *ftran_loc*, it was impossible to make a cross-comparison for this specific measurement.

Six of the 12 models submitted the full set of experiments concerning the bidirectional reflectance, allowing us to implement a thoughtful analysis, especially in terms of internal consistency checks and model-to-model comparison, with the aim to identify a suitable surrogate reference for each scenario and evaluate the proficiency of the whole set of models against it.

We implemented a preliminary screening based on a classical sigma clipping outlier detection method (e.g., the Chauvenet's method) and relied on the model-to-model comparison $\delta_{m\leftrightarrow c}$ metric, as well as on a robust statistics as defined in ISO-13528, to identify CRG scene-wise. Nobody associated an uncertainty estimation to the measurands, and we assumed the value of 3% as an artificial a priori uncertainty to perform the proficiency test by means of the z' metric. Implementing the internal consistency checks and the preliminary outlier detection scheme prevented trivial errors to induce large deviations and improved the overall quality of the dataset after the feedback phase of the project. It was left to the participants to fix their results with a re-submission or to keep them in the loop of the final analysis.

With respect to RAMI-IV, we also defined a new test aimed to verify the consistency between the total *brf* and the hemispherical reflectances (*bhr* and *dhr*). It was based on the fitting of a parametric BRDF, e.g., RPV, on the available set of *brfpp* and *brfop* submissions, and the verification of the correspondence between the hemispherical integrals of the function and the *bhr* (or *dhr*) submissions. For most of the models and scenes, we observed an agreement as better as ±0.05, while we were able to identify some major unconsistencies that may be used by the participants to verify their processing approach (Fig. S1). Notably, *raytran* (reflectance) and *renderjay* (flux measurements) outlier ratio performances improved from the initial to the final stage of the experiment (Fig. S2).

The overall analysis results proved that the level of agreement among the models depends on their characteristics and the specific scenario.

In particular, for total *brf*, we observed a model agreement within 2% (in terms of $\delta_{m\leftrightarrow c}$) between at least 2 models for most of the scenes, except for HET08. The most robust agreement was observed over birchstand scenes and structured scenes, with 4 models (*dart*, *less*, *raytran*, and *wps*) agreeing within 2%, except over HET09 where *wps* was agreeing only within 6% (Table 10 and Figs. 9 and 13).

Similar consistencies were observed for the hemispherical reflectances. Remarkably, *less* consistently belonged to the group of models agreeing within 2%, together with *raytran* and/or *dart* that completed it most of the time [Table 12(a) and Fig. 12], except over HET08 and HET50 where none of the models agrees within this threshold. Over pinestand and empirical scenes, the CRG was populated by only 2 models, likely because of the challenges associated with the detailed representation of new complex scenarios and the heterogeneity of model representation of needle leaves.

Concerning absorption [Table 12(b) and Fig. 12], we observed a considerable agreement among 4 models (namely, *dart*, *less*, *raytran*, and *wps*) over all scenes except pinestand and savanna. Nevertheless, over pinestands, many 2-model groups were identified, including the one formed by *less* and *raytran*. Notably, a good agreement over the new HET51 forest scenario has been achieved by the same group of models. Considering these results, *less* and *raytran* remain stronger candidates to issue a total absorption benchmark, as the persistent agreement over most of the scenario proposed (7 of 8) contributes to increase their robustness.

The results on the transmission were less encouraging, although *dart*, *raytran*, and *wps* (*less* did not submit transmission) were often supported by the best matches, although existing only over over HET09, HET14, HET15, and HET16 [Table 12(c)].

The various levels of agreements, expressed in terms of $\delta_{m\leftrightarrow c}$, vary between such different classes of RT models: on one hand, the ones having solver based on computationally expensive 3D ray tracing, and on the other hand, the ones oriented to large-scale operational computations (based on a combination of simple geometrical representation of the scene and RT equation analytical solutions). Some of the scenarios, such as savanna (HET50) and Wytham Woods forest (HET51), were new in RAMI-V, and the assumptions made by the different teams to represent them deserve some review. However, the lower agreement was observed over pinestand scenes, which was already proposed in previous phases. Nevertheless, while the agreement is still far to be comparable with the results obtained for abstract scenarios during RAMI-IV and RAMI-V [37], we were able to identify a set of models that could promisingly contribute to a new ROMC reference for actual scenarios.

These divergences can be explained by the models' assumptions made for each experiment and levels of details provided in the description of the actual scenarios. The fact that performances differ across the scenes raised that the community should explore further the establishment of a general 3D object format for representing the 3D world, addressing the inconsistencies in file formats. Specifically, while the Rayshade format is more flexible for representing volumetric shapes like cylinders, spheres, or ellipsoids, it may introduce deviations since modern models such as *dart*, *less*, and *eradiate* primarily support the OBJ format. This latter represents 3D geometry using vertices and faces, struggling precise representations of volumes. Developing a standardized format or advanced conversion tools could help bridge these differences and reduce model discrepancies. This recommendation of having a standard format for the representation of 3D scenarios was already highlighted in [38].

RAMI was confirmed to play a relevant role in the assessment of RT model accuracy performances by also considering the evolution in the definition of digital scenarios based on new field techniques, such as terrestrial laser scanning. Despite that the participation of models other than 3D Monte Carlo ray-tracing ones was encouraged, they still remain a subset. RAMI or other similar benchmarking activities may be improved on the participation side by spending additional efforts to provide the statistical information that summarize vertical profiles of the surface area density of leaves and branches. As the results obtained over coniferous still indicate a considerable dispersion even among 3D RT explicit models, additional investigations on the optimal representation of the needle leaves and twigs are deserved.

RAMI promotes a standardized approach to RT model verification, relying on extensive sets of experiments. To achieve this, RAMI conducted periodic benchmark rounds, known as phases, which occur every few years. Each phase builds upon the previous one, incorporating the same experiments to allow participants to refine and improve their models, with new experiments allowing the community to capitalize with space observation advancements. RAMI-V focused on Copernicus sensors with the aim to tackle the increased use of 3D RT in more concrete applications.

The integration of physically based models and machine learning/deep learning (ML/DL) models is a crucial aspect of advancing our understanding and predictions in various fields. Once RT models are validated, they can further support the training of ML/DL-based models by generating synthetic datasets to complement actual observations. Several studies have already demonstrated this potential to simulate VIS and multi/hyperspectral images [39] or by supporting the dynamic retrieval of olive tree properties using bayesian model and Sentinel-2 [40]. On the other hand, ML/DL techniques were already proficiently used to improve ray-tracing performances and reduce Monte Carlo noise of physically based RT models [41–43].

One additional step forward for RAMI would be related to the assessment of computational requirements. Moreover, none of the participants associated an uncertainty to their virtual measurements, which could help for the conformity test.

In conclusion, robust reference solutions are defined through RAMI results with the aim to be integrated into the ROMC, while RAMI phase proceeds with a new set of test cases. This methodology aligns with the ISO-13528 standard, which recommends ongoing proficiency testing. By providing access to the ROMC, model developers and users can independently assess the quality of a modeling tool at any time, rather than waiting for the next RAMI phase to become available. This approach enables continuous verification and improvement of modeling tools, fostering a culture of accountability and excellence in the field.

Acknowledgments

Funding: RAMI is supported by EU Copernicus program. The International Radiation Commission (IAMAS-IRC) endorses RAMI as a 3D RTM ongoing benchmarking activity (https://www.iamas.org/irc/3drt-three-dimensional-radiative-transfer/). The DART team (CESBIO, Toulouse, France) completed this work within the framework of the TOSCA program project funded by CNES (French Space Center). K.C. was funded by the European Union (ERC-2021-STG grant agreement no. 101039795). The views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. A.K. is supported by the

Estonian Research Council grant PRG1405. The Eradiate team (Rayference, Brussels, Belgium) performed this work within the framework of ESA contract 4000127201/19.

Author contributions: C.L., N.G., J.S.A., and M.D. conceived the idea and designed the experiment and submitted specific model results. C.L. and N.G. performed the intercomparison analysis and wrote the manuscript. M.R. designed and managed the web content and interface, implemented the routines for the initial screening of the data, contributed to tables and figures, and reviewed the draft. All the other authors submitted model results and contributed evenly to the paper by reviewing the initial manuscript.

Competing interests: The authors declare that they have no competing interests.

Data Availability

All RAMI-V scene description and descriptive files in Rayshade and OBJ formats can be found on RAMI website https://rami-benchmark.jrc.ec.europa.eu; *dart* can be obtained upon request from https://dart.omp.eu/#/, and the adapted RAMI-V scenes are available at https://dart.omp.eu/#/doc; *less* is available at https://lessrt.org/, with adapted RAMI-V scenes at https://lessrt.org/resources/3dscene/; the description and code of *eradiate* can be found at https://github.com/eradiate/; the source code and the RAMI-V scenes adapted for *renderjay* can be found at https://github.com/martinvanleeuwen/RenderJay.jl. The source code for *spartacus* may be downloaded from https://github.com/ecmwf/spartacus-surface, and the test cases in the package include the RAMI-V scenes used in this paper.

Supplementary Materials

Figs. S1 to S3
Table S1

References

1. Pinty B, Gobron N, Widlowski JL, Gerstl SAW, Verstraete MM, Antunes M, Bacour C, Gascon F, Gastellu JP, Goel N, et al. Radiation transfer model intercomparison (RAMI) exercise. *J Geophys Res Atmos*. 2001;106(D11):11937–11956.
2. GCOS. *The 2022 GCOS Implementation Plan*. Geneva (Switzerland): World Meteorological Organization; 2022.
3. Pinty B, Widlowski JL, Taberner M, Gobron N, Verstraete MM, Disney M, Gascon F, Gastellu JP, Jiang L, Kuusk A, et al. Radiation transfer model intercomparison (RAMI) exercise: Results from the second phase. *J Geophys Res Atmos*. 2004;109(D6):Article D06210.
4. Pinty B, Lavergne T, Dickinson RE, Widlowski JL, Gobron N, Verstraete MM. Simplifying the interaction of land surfaces with radiation for relating remote sensing products to climate models. *J Geophys Res Atmos*. 2006;111(D2): Article D02116.
5. Widlowski JL, Taberner M, Pinty B, Bruniquel-Pinel V, Disney M, Fernandes R, Gastellu-Etchegorry JP, Gobron N, Kuusk A, Lavergne T, et al. Third radiation transfer model Intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models. *J Geophys Res Atmos*. 2007;112(D9).
6. Widlowski JL, Robustelli M, Disney M, Gastellu-Etchegorry JP, Lavergne T, Lewis P, North PRJ, Pinty B, Thompson R, Verstraete MM. The RAMI on-line model checker (ROMC): A web-based benchmarking facility for canopy reflectance models. *Remote Sens Environ*. 2008;112(3):1144–1150.
7. Widlowski JL, Mio C, Disney M, Adams J, Andredakis I, Atzberger C, Brennan J, Busetto L, Chelle M, Ceccherini G, et al. The fourth phase of the radiative transfer model intercomparison (RAMI) exercise: Actual canopy scenarios and conformity testing. *Remote Sens Environ*. 2015;169: 418–437.
8. ISO-13528. Statistical methods for use in proficiency testing by interlaboratory comparison. Vol. 13528. 2015.
9. Widlowski JL, Pinty B, Lopatka M, Atzberger C, Buzica D, Chelle M, Disney M, Gastellu-Etchegorry JP, Gerboles M, Gobron N, et al. The fourth radiation transfer model intercomparison (RAMI-IV): Proficiency testing of canopy reflectance models with ISO-13528. *J Geophys Res Atmos*. 2013;118(13):6869–6890.
10. Stretton MA, Morrison W, Hogan RJ, Grimmond S. Evaluation of the SPARTACUS-urban radiation model for vertically resolved shortwave radiation in urban areas. *Bound-Layer Meteorol*. 2022;184(2):301–331.
11. Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, Hoersch B, Isola C, Laberinti P, Martimort P, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens Environ*. 2012;120:25–36.
12. Donlon C, Berruti B, Buongiorno A, Ferreira MH, Féménias P, Frerick J, Goryl P, Klein U, Laur H, Mavrocordatos C, et al. The global monitoring for environment and security (GMES) sentinel-3 mission. *Remote Sens Environ*. 2012;120:37–57.
13. Justice CO, Vermote E, Townshend JR, Defries R, Roy DP, Hall DK, Salomonson VV, Privette JL, Riggs G, Strahler A, et al. The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Trans Geosci Remote Sens*. 1998;36(4):1228–1249.
14. Kolb CE. Rayshade user's guide and reference manual. 1994.
15. Kuusk A, Nilson T, Paas M, Lang M, Kuusk J. Validation of the forest radiative transfer model FRT. *Remote Sens Environ*. 2008;112(1):51–58.
16. Kuusk A, Kuusk J, Lang M. A dataset for the validation of reflectance models. *Remote Sens Environ*. 2009;113(5):889–892.
17. Kuusk A, Nilson T, Kuusk J, Lang M. Reflectance spectra of RAMI forest stands in Estonia: Simulations and measurements. *Remote Sens Environ*. 2010;114(12):2962–2969.
18. Kuusk A, Lang M, Kuusk J. Database of optical and structural data for the validation of forest radiative transfer models. In: Kokhanovsky A, editor. *Light scattering reviews*. Berlin, Heidelberg: Springer; 2013. Vol. 7, p. 109–148.
19. Zenone T, Migliavacca M, Montagnani L, Seufert G, Valentini R. Carbon sequestration in short rotation forestry and traditional poplar plantation. Paper presented at: Proceedings of the Short Rotation Crops International Conference; 2008 Aug 18; Minneapolis, MN, USA.
20. Somers B, Delalieux S, Verstraeten WW, Coppin P. A conceptual framework for the simultaneous extraction of sub-pixel spatial extent and spectral characteristics of crops. *Photogramm Eng Remote Sens*. 2009;75(1):57–68.
21. Stuckens J, Somers B, Delalieux S, Verstraeten WW, Coppin P. The impact of common assumptions on canopy

radiative transfer simulations: A case study in citrus orchards. *J Quant Spectrosc Radiat Transf*. 2009;110(1–2):1–21.

22. Disney M, Kalogirou V, Lewis PE, Prieto-Blanco A, Hancock S, Pfeifer M. Simulating the impact of discrete-return lidar system and survey characteristics over young conifer and broadleaf forests. *Remote Sens Environ*. 2010;114(7):1546–1560.
23. Van Wilgen B, Govender N, Biggs H, Ntsala D, Funda X. Response of savanna fire regimes to changing fire-management policies in a large African national park. *Conserv Biol*. 2004;18(6):1533–1540.
24. Disney MI, Lewis PE, Bouvet M, Prieto-Blanco A, Hancock S. Quantifying surface reflectivity for spaceborne lidar via two independent methods. *IEEE Trans Geosci Remote Sens*. 2009;47(9):3262–3271.
25. Calders K, Origo N, Burt A, Disney M, Nightingale J, Raumonen P, Åkerblom M, Malhi Y, Lewis P. Realistic forest stand reconstruction from terrestrial LiDAR for radiative transfer modelling. *Remote Sens*. 2018;10(6):933.
26. Nicodemus FE. *Geometrical considerations and nomenclature for reflectance*. Washington (DC): US Department of Commerce, National Bureau of Standards; 1977. Vol. 160.
27. Schaepman-Strub G, Schaepman ME, Painter TH, Dangel S, Martonchik JV. Reflectance quantities in optical remote sensing—Definitions and case studies. *Remote Sens Environ*. 2006;103(1):27–42.
28. Stamnes K, Tsay SC, Wiscombe W, Jayaweera K. Numerically stable algorithm for discrete ordinate-method radiative transfer in multiple scattering and emitting layered media. *Appl Opt*. 1988;27(12):2502–2509.
29. Liang S, Fang H, Chen M, Shuey CJ, Walthall C, Daughtry C, Morisette J, Schaaf C, Strahler A. Validating MODIS land surface reflectance and albedo products: Methods and preliminary results. *Remote Sens Environ*. 2002;83(1–2):149–162.
30. Rahman H, Pinty B, Verstraete MM. Coupled surface-atmosphere reflectance (CSAR) model: 2. Semiempirical surface model usable with NOAA advanced very high resolution radiometer data. *J Geophys Res Atmos*. 1993;98(D11):20791–20801.
31. Moré JJ. The Levenberg-Marquardt algorithm: Implementation and theory. In: *Numerical analysis*. Berlin: Springer; 1978. p. 105–116.
32. Janert PK. *Gnuplot in action: Understanding data with graphs*. Shelter Island (NY): Manning Publications Co.; 2016.
33. Chauvenet W. *A manual of spherical and practical astronomy: Vol. II*. Philadelphia (PA): J.B. Lippincott Company; 1863.
34. Maples M, Reichart DE, Konz N, Berger TA, Trotter AS, Martin JR, Dutton DA, Paggen ML, Joyner RE, Salemi CP. Robust chauvenet outlier rejection. *Astrophys J Suppl Ser*. 2018;238:37.
35. Despotovic M, Nedic V, Despotovic D, Cvetanovic S. Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renew Sust Energ Rev*. 2016;56:246–260.
36. Gueymard CA. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. *Renew Sust Energ Rev*. 2014;39:1024–1034.
37. Gobron N, Lanconelli C, Gastellu-Etchegorry JP, et al. The 5th Phase of the RAdiative Transfer Model Inter-comparison (RAMI-V): Abstract Canopies for Copernicus Sensors Configurations. *J Remote Sens*. In Press.
38. Gobron N, Lanconelli C, Urraca Valle R, Govaerts Y. RAMI workshop—Radiative transfer modelling support to EO metrology and Cal/Val activities. Luxembourg: Publications Office of the European Union; 2023.
39. Lei T, Graefe J, Mayanja IK, Earles M, Bailey BN. Simulation of automatically annotated visible and multi-/hyperspectral images using the Helios 3D plant and radiative transfer modeling framework. *Plant Phenomics*. 2024;6:0189.
40. Abdelmoula H, Kallel A, Roujean JL, Gastellu-Etchegorry JP. Dynamic retrieval of olive tree properties using Bayesian model and Sentinel-2 images. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2021;14:9267–9286.
41. Gharbi M, Li TM, Aittala M, Lehtinen J, Durand F. Sample-based Monte Carlo denoising using a kernel-splatting network. *ACM Trans Graph*. 2019;38:1–12.
42. Munkberg J, Hasselgren J. Neural denoising with layer embeddings. *Comput Graph Forum*. 2020;39(4):1–12.
43. Muller T, McWilliams B, Rousselle F, Gross M, Novak J. Neural importance sampling. *ACM Trans Graph*. 2019;38:1–19.
44. Gastellu-Etchegorry JP, Demarez V, Pinel V, Zagolski F. Modeling radiative transfer in heterogeneous 3-D vegetation canopies. *Remote Sens Environ*. 1996;58(2):131–156.
45. Wang Y, Kallel A, Yang X, Regaieg O, Lauret N, Guilleux J, Chavanon E, Gastellu-Etchegorry JP. DART-Lux: An unbiased and rapid Monte Carlo radiative transfer method for simulating remote sensing images. *Remote Sens Environ*. 2022;274:Article 112973.
46. Goodenough AA, Brown SD. DIRSIG5: Next-generation remote sensing data and image simulation framework. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2017;10(11):4818–4833.
47. Leroy V, Nollet Y, Schunke S, Misk N, Marton N, Govaerts Y. Eradiate radiative transfer model. 2024. doi: 10.5281/zenodo.7224314. url: https://github.com/eradiate/eradiate.
48. Kobayashi H, Iwabuchi H. A coupled 1-D atmosphere and 3-D canopy radiative transfer model for canopy reflectance, light environment, and photosynthesis simulation in a heterogeneous landscape. *Remote Sens Environ*. 2008;112(1):173–185.
49. Kuusk A, Nilson T. A directional multispectral forest reflectance model. *Remote Sens Environ*. 2000;72(2):244–252.
50. Qi J, Xie D, Yin T, Yan G, Gastellu-Etchegorry JP, Li L, Zhang W, Mu X, Norford LK. LESS: LargE-Scale remote sensing data and image simulation framework over heterogeneous 3D scenes. *Remote Sens Environ*. 2019;221:695–706.
51. Lewis P. Three-dimensional plant modelling for remote sensing simulation studies using the botanical plant modelling system. *Agronomie*. 1999;19:185–210.
52. Huang H. Accelerated RAPID model using heterogeneous porous objects. *Remote Sens*. 2018;10(8):1264.
53. Huang H, Qi J, Li L. Enhanced branch simulation to improve RAPID in optical region using RAMI scenes. *J Remote Sens*. 2023;3:0039.
54. Govaerts YM. *A model of light scattering in three-dimensional plant canopies: A Monte Carlo ray tracing approach*. Brussels: Office for Official Publication of the European Communities; 1996.
55. Govaerts YM, Verstraete MM. Raytran: A Monte Carlo ray-tracing model to compute light scattering in three-dimensional heterogeneous media. *IEEE Trans Geosci Remote Sens*. 1998;36(2):493–505.
56. van Leeuwen M, Frye HA, Wilson AM. Understanding limits of species identification using simulated imaging spectroscopy. *Remote Sens Environ*. 2021;259:Article 112405.

57. Hogan RJ, Quaife T, Braghiere R. Fast matrix treatment of 3-D radiative transfer in vegetation canopies: SPARTACUS-vegetation 1.1. *Geosci Model Dev*. 2018;11(1): 339–350.

58. Zhao F, Li Y, Dai X, Verhoef W, Guo Y, Shang H, Gu X, Huang Y, Yu T, Huang J. Simulated impact of sensor field of view and distance on field measurements of bidirectional reflectance factors for row crops. *Remote Sens Environ*. 2015;156:129–142.