

1 Machine learning-driven characterization and prescription of 2 aerosol optical properties for atmospheric models

3 Nilton Évora do Rosário¹, Karla M. Longo², Pedro H. Toso¹, Saulo R. Freitas², Marcia A.
4 Yamasoe³, Luiz Flávio Rodrigues², Otavio Medeiros², Haroldo Campos Velho², Isilda da Cunha
5 Menezes⁴, Ana Isabel Miranda⁴

6 ¹Departamento de Ciências Ambientais, Universidade Federal de São Paulo, Diadema, SP Brazil

7 ²Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, Brazil

8 ³Departamento de Ciências Atmosféricas, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São
9 Paulo, Cidade Universitária, São Paulo, SP, Brazil

10 ⁴Center for Environmental and Marine Studies (CESAM), Department of Environment and Planning, University of Aveiro,
11 Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

12 Correspondence to: Nilton do Rosário (nrosario@unifesp.br)

13

14

Abstract

15 Accurate modeling of aerosol optical properties is critical to simulate aerosol radiative effects.
16 However, uncertainties regarding the simulation of aerosol-intensive optical properties are still
17 significant. Therefore, the use of observations to constrain aerosol optical properties in models has
18 been indicated as an option. Also, explicit computations of optical properties are still too costly for
19 operational models, which makes observation-based prescriptions a convenient solution. We
20 developed an observation-based prescription of aerosol optical properties driven by
21 machine-learning techniques that can be applied in models. The Iberian Peninsula (IP) was taken as
22 the reference domain, and the aerosol products from the AERONET sites across the IP were the main
23 dataset. First, clustering was applied to define the typical aerosol optical regimes affecting the IP
24 atmosphere. Five typical regimes were identified. Two of them were dominated by coarse mode,
25 which was associated with Saharan dust. One was found to be close to pure dust, while the other
26 indicated a mixed scenario of dust and pollution. Two of the non-dust regimes, strongly and
27 moderately absorbing, were found to be associated with smoke. The remaining non-dust regime,
28 with no clear association, occurs mostly in the eastern portion of the IP. Afterward, using
29 aerosol-type columnar mass density from MERRA-2, a model was trained as a predictor of the
30 optical regimes using the Random Forest method. The model was tested under distinct aerosol
31 scenarios. Predictions' accuracy ranged from 60 to 75%, depending on the regime, while presenting
32 an average accuracy of 70%.

33 **Keywords:** Aerosol Optical Properties, AERONET, MERRA-2, Machine-Learning, Random Forest

34

35

36

37 1. Introduction

38 The importance of aerosols in the Earth's climate system is undisputed. Aerosol particles
39 participate directly in the planetary energy budgets via the scattering and absorption of
40 terrestrial and solar radiation (Kim and Ramanathan 2008; IPCC, 2021; Li et al., 2022).
41 However, this participation is permeated by high complexity due to the variety of aerosol
42 particles sizes and composition, which cause significant uncertainty (Spencer et al. 2019;
43 IPCC, 2021; Li et al., 2022). The uncertainties and challenges in accurately representing
44 aerosol particles' processes in climate, weather, and environmental models arise from
45 various limitations. For instance, when it comes to aerosol direct interaction with radiation,
46 the current global aerosol monitoring system does not provide a comprehensive
47 spatial-temporal characterization of spectral complex refractive index and size distribution
48 of the aerosol particles, critical information to characterize the particle absorption and
49 scattering (Samset et al. 2018; Li et al., 2022). This lack of observational data contributes
50 significantly to uncertainty in aerosol modeling and, therefore, to the uncertainty of the
51 aerosol radiative forcing.

52 The difficulty of traditional libraries of aerosol optical and microphysical properties
53 (Shettle and Fenn, 1979; Koepke et al., 1997; Hess et al., 1998) to describe geographical
54 variation aerosol properties, for instance, soil dust mineralogy (Adebisi et al., 2023), has
55 been central in the aerosol optical properties uncertainty debate. Another critical aspect is
56 the characterization of the state of the mixture of the aerosol particles in the model's
57 aerosol modules (Samset et al. 2018; Sand et al., 2021). Given the complex dynamic of
58 aerosol particle emission, transport, and removal in the atmosphere, numerical modelling of
59 the state of the mixture and the resultant complex refractive index and size distribution is
60 widely recognized as one of the most important sources of uncertainty in addressing
61 aerosol particles' radiative forcing (Sand et al., 2021). According to Sand et al. (2021),
62 aerosol absorption is poorly constrained, and the current climate models present a large
63 range in the quantification of the main absorbing aerosol species (black carbon (BC),
64 organic aerosols (OA), and mineral dust). Brown et al. (2021) findings indicate that
65 biomass-burning aerosols in most climate models are too absorbing, mainly due to
66 treatments of aerosol mixing state. They found the internal mixing assumptions used in
67 climate models to overestimate Black Carbon(BC) absorption when compared to the
68 observations. Saharan dust, a critical component of the global aerosol system, has been
69 found to absorb less solar radiation than models estimate (Adebisi et al., 2023), and the
70 primary cause pointed out is the models' overestimation of the dust imaginary refractive
71 index. Absorption is not the only issue facing aerosol particle representation in climate
72 models; the relative contribution of fine and coarse mode particles is also a challenge. For
73 instance, Adebisi et al. (2023) also found models underestimating large dust particles when
74 representing North African dust plumes.

75 Observation-constrained models have been recommended to mitigate models' current
76 difficulty in fully simulating aerosol properties and processes accurately (Samset et al.
77 2018; Proske et al., 2024). In addition to the uncertainty aspects, explicit simulation of
78 aerosol composition and microphysical properties, followed by explicit computation of
79 intensive optical properties, is still too expensive computationally for operational models,

80 which also makes observational-based prescriptions a convenient solution. Zhong et al.
81 (2022) used relationships from an ensemble of aerosol models and satellite observations to
82 identify the primary source of uncertainty in aerosol modelling results in biomass burning
83 regions. Their study pointed out the incorrect simulations of lifetimes and the
84 underestimation of mass extinction coefficients as the main reasons for their difficulty in
85 matching observed aerosol optical depth (AOD). As the largest, time and device-consistent
86 observational network, capable of constraining multiple aerosol intensive microphysical
87 and optical properties, the AERosol RObotic NETwork (AERONET) has been used worldwide
88 to constrain models and satellite algorithms (Omar et al., 2005; Li et al., 2010; Levy et al.,
89 2010; Rosario et al., 2013; Russel et al., 2014; Chen et al., 2023). Chen et al. (2023)
90 developed an aerosol optical module with observation-constrained Black Carbon properties
91 to improve aerosol absorption simulation. Their sensitivity simulations show a reduction of
92 18%–69% in the biases of aerosol single-scattering co-albedo when compared with global
93 observations from AERONET. Li et al. (2010) used AERONET retrievals to evaluate and
94 improve the performance of a GCM aerosol optical module. They found their GCM to
95 simulate flatter Aerosol Optical Depth (AOD) spectral dependence, indicating an Angstrom
96 Exponent (AE) biased to low values, which suggests that the aerosol sizes simulated were
97 too large. After adjusting the aerosol's size based on AERONET retrievals, the agreement
98 between simulated and observed AOD improved for all aerosol regimes, but especially for
99 smoke and dust scenarios. Rosario et al. (2013) used a set of spectral optical models
100 developed from AERONET sky retrievals over distinct biomes combined with the concept of
101 anisotropic areas of influence of the AERONET sites (Hoelzemann et al., 2009) to constrain
102 smoke aerosol radiative effect modelling during South American biomass burning. By doing
103 so, they were able to capture the effect of the regional variability of smoke optical
104 properties (absorption and size-related) on the surface solar irradiance related to the
105 biomes' distinct nature of smoke.

106 Global and regional cluster analysis of AERONET long-term retrievals of aerosol properties
107 has proved valuable to classify observations in terms of aerosol optical regimes, providing
108 means to qualitative constraints on aerosol properties (Omar et al, 2005; Levy et al., 2007;
109 Russell et al., 2014; Li et al., 2019; Fan et al., 2020; Zhou et al., 2023). In these studies, the
110 number of identified typical aerosol optical regimes varied from 4 to 10, numbers that were
111 expected to likely represent either global or regional major aerosol scenarios, according to
112 each study's focus. In their study, Zhou et al. (2023) found that regional aerosol regime
113 classifications performed better than global classifications when applied to simulate AOD
114 during pollution episodes and in different seasons in Beijing, China. They found larger
115 differences between the strong and moderately absorbing aerosol regimes, namely dust and
116 smoke regimes, when comparing global and regional clustering results. This is a
117 consequence of the differences between China's regional dust and smoke aerosol particles'
118 physical and chemical characteristics and those of global dust and smoke mean features.
119 Another aspect highlighted by Zhou et al. (2023) is that smoke and dust-dominated optical
120 regimes are more frequent globally than in China. Their result suggests that regional
121 classification better captures typical aerosol optical regimes influencing a specific domain
122 and, therefore, with the potential to improve observation-constrained simulations of
123 aerosol radiative forcing.

124 Focusing on the Iberian Peninsula (IP), this study sought to characterize the typical aerosol
125 optical regimes driving the variability of aerosol-intensive properties over the peninsula,
126 aiming to constrain aerosol optical properties prescription in atmospheric models using a
127 novel machine-learning approach. IP is a region affected by a highly dynamic and complex
128 set of aerosols mixing, including natural and anthropogenic particles (Cachorro et al., 2016;
129 Gomez-Amo et al., 2017). Natural sources include marine aerosols from the Atlantic Ocean
130 and Mediterranean Sea, mineral dust from North Africa, and, eventually, wildfire emissions.
131 Major anthropogenic sources are urban-industrial, particularly in more densely populated
132 regions, and biomass burning driven by human activities, especially in the north and central
133 Portugal and eastern and northern Spain. Regional column-integrated optical properties are
134 highly sensitive to the mixing of this diversity of aerosol types, in particular to dust and
135 smoke mixing (Gomez-Amo et al., 2017).

136 The manuscript is organized as follows: Section 2 includes a brief overview of the Iberian
137 Peninsula, focusing on the main atmospheric circulation features and major aerosol particle
138 sources affecting the region, followed by the description of the dataset and methods
139 adopted to identify, characterize, and prescribe the identified aerosol typical regimes.
140 Results and discussions are presented in Section 3. First, the identified aerosol optical
141 regimes and their major features are described and contextualized. Subsequently, the
142 results of the novel machine-learning approach to prescribing the optical regimes are
143 discussed and evaluated. Finally, the main findings of our study are highlighted in the
144 conclusion section.

145

146 **2. Study Region, Data and Methods**

147

148 **2.1 Study region**

149 The Iberian Peninsula (**Figure 1**), comprising Spain and Portugal, exhibits diverse climate
150 conditions due to its complex topography and proximity to the Atlantic Ocean, the
151 Mediterranean Sea, and North Africa. The wind circulation over the peninsula is shaped by
152 its location between the Atlantic Ocean and the Mediterranean Sea, diverse topography, and
153 interactions between regional and global atmospheric patterns, leading to complex wind
154 circulations that significantly influence the region's climate. This results in distinct climate
155 zones, from arid deserts to lush green forests. The Mediterranean climate spans most of
156 Spain, including the eastern and southern coastal regions and central Portugal, featuring hot
157 and dry summers, especially inland. Winters are mild, rarely dropping below 10°C in
158 coastal areas. Most precipitation, often rain, occurs in autumn and winter, leading to dry
159 summers that increase wildfire risks. Wildfires regularly occur in the IP region, fueled by
160 extreme weather conditions, abnormal high temperature records combined with strong, dry
161 winds (Asfaw et al., 2022; Ermitão et al., 2023). Under these scenarios, the entire region can
162 be affected by smoke plumes that often shape the entire region's optical properties (Elias et
163 al, 2004; Gomez-Amo et al., 2017). But wildfires are more frequent in the north and central
164 region of Portugal and the north and eastern portion of Spain (Ermitão et al., 2023; Alvares
165 et al., 2024). Oceanic climate is typical in northern coastal regions of Spain, such as Galicia,

166 Asturias, and the Basque Country, and parts of northern Portugal. The Atlantic Ocean
167 influences mild temperatures year-round, with minimal seasonal variation and abundant,
168 evenly distributed rainfall. Annual precipitation can exceed 1,000 mm, with frequent cloud
169 cover and high humidity, especially in winter. The Continental climate of the central plateau
170 (Meseta Central) and the Ebro Valley feature extreme temperature variations, with hot
171 summers, highs often above 35°C, and winters below freezing. The central regions have less
172 precipitation than the coastal areas, with a semi-arid climate in some parts. Most rainfall
173 occurs in spring and autumn. Arid and Semi-Arid Climates are found in Southeastern Spain,
174 especially in Murcia and Almería, and parts of the Ebro Valley. These areas receive very low
175 rainfall, often less than 300 mm annually, leading to desert-like conditions like those in the
176 Tabernas Desert. Summers are extremely hot, while winters are mild. Southern Spain,
177 especially the Andalusian region, can be affected by hot and dry winds from the Sahara,
178 causing heat waves and dust storms.

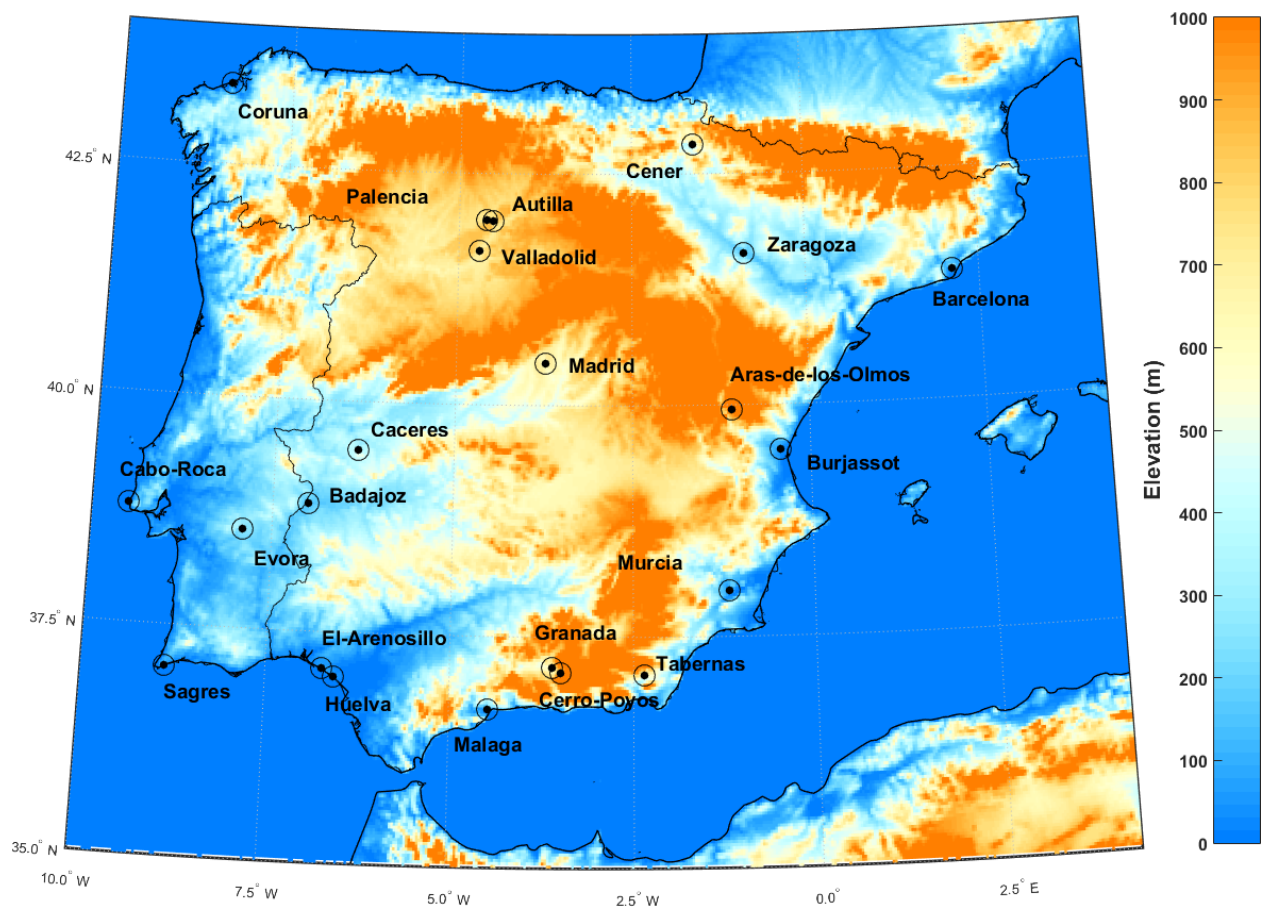
179 The occurrence of Saharan dust events on the Iberian Peninsula usually peaks in March and
180 June, with a marked minimum in April and lowest occurrence in winter according to
181 Cachorro et al. (2016). Depending on the synoptic conditions and circulation patterns, dust
182 transport can affect the entire peninsula (Toledano et al., 2007). The prevailing westerlies,
183 blowing from west to east, are the dominant wind pattern over the Iberian Peninsula. These
184 winds are most prominent in the mid-latitudes, including the Iberian Peninsula. More
185 pronounced in the northern region during autumn and winter, these winds bring moist air
186 from the Atlantic, increasing precipitation in Galicia, the Basque Country, and northern
187 Portugal. While they also affect central and southern areas, their impact is moderated by the
188 peninsula's topography and other wind systems. The northeast trade winds affect the
189 southern and western coasts of Portugal and southwestern Spain, creating a mild and dry
190 climate, especially in summer. In contrast, Mediterranean winds affect the eastern and
191 southeastern coasts. Additionally, the Iberian Thermal Low, resulting from intense heating
192 of the Iberian interior, creates a low-pressure area that draws air from the Atlantic and
193 Mediterranean shores, leading to converging wind patterns. This circulation pattern
194 enhances sea breeze penetration and moderates coastal temperatures. Southern Spain is
195 influenced by the Sahara winds, as said, these dry winds often carry dust, increasing the
196 temperature and reducing air quality. Calima is a type of wind that occurs when Saharan
197 dust reaches the peninsula, especially in summer, causing hazy skies, a reddish tint, and low
198 visibility. These winds are linked to high-pressure systems over North Africa and
199 low-pressure systems over the western Mediterranean.

200 The wind circulation over the Iberian Peninsula is a dynamic and complex system shaped by
201 global atmospheric patterns, regional geography, and local topography. The interaction of
202 prevailing westerlies, trade winds, Mediterranean breezes, and local wind systems creates a
203 diverse wind regime that affects the peninsula's climate. Understanding these patterns is
204 essential for weather prediction, agriculture management, and tackling environmental
205 challenges. According to Cachorro et al. (2016), these complex and contrasting influences of
206 air masses from the Atlantic Ocean, Mediterranean Sea, European continent, and North
207 Africa lead to a large spatio-temporal variability in aerosol properties, types, and mixing
208 processes over the Iberian Peninsula. This makes the peninsula a challenging region for
209 online modeling of aerosol microphysical properties and mixing state, therefore an

210 interesting region to evaluate observation-based approaches, such as those based on
211 climatological aerosol intensive optical properties from AERONET (Li et al., 2019; Fan et al.,
212 2020; Zhou et al., 2023). The computation of optical properties for radiative transfer
213 computations is usually based on a mass-weighted average of individual species at each grid
214 point. This assumption of external mixing may not always be accurate, leading to significant
215 uncertainties, such as excessive absorption by smoke aerosols and inaccuracies in dust size
216 fractions. Observation-based approaches, such as those provided by AERONET retrieval
217 climatology, attribute intensive optical properties to an effective aerosol based on actual
218 observations. This method aims to reduce the uncertainties arising from the explicit
219 simulation of these properties in climate models. The Iberian Peninsula, influenced by a
220 variety of aerosol types, including dust, smoke, urban-industrial emissions, and marine
221 aerosols, presents an interesting region to test this hypothesis.

222

223



224

225 **Figure 1:** AERONET sites locations displayed on top of the Iberian Peninsula topography.

226

227

228 2.2 AERONET aerosol inversion product

229 AERONET is a global ground-based network of sun photometers mainly aimed at
230 characterizing columnar aerosol particle properties (Holben et al., 1998). From the direct
231 Sun attenuation measurements, AERONET algorithms derive spectral Aerosol Optical Depth
232 (AOD_{λ}) at the wavelengths 0.34, 0.38, 0.44, 0.50, 0.67, 0.87, 0.94, and 1.02 μm . The interval
233 between direct sun measurements is typically 15 minutes, but only cloud-free conditions
234 are considered for aerosol retrievals. From the spectral dependency of AOD at these
235 wavelengths, AERONET provides Angstrom Exponent (AE), a parameter sensitive to the
236 aerosol particle size distribution (Eck et al., 1999). AERONET also provides several other
237 intensive properties that depend not on the amount but on the nature of the aerosol, related
238 to particle size, shape, and composition, from sky radiance measurements up to nine times a
239 day at the wavelengths 0.44, 0.67, 0.87, and 1.02 μm (Sinyuk et al., 2020). These intensive
240 properties include microphysical parameters, such as refractive indices ($n+ik$) and volume
241 size distribution, and also optical parameters like Single Scattering Albedo (SSA),
242 asymmetry parameter (ASY), Lidar Ratio (LR), Linear Depolarization Ratio (LDR), Angstrom
243 Exponent, among others (Holben et al., 1998; Dubovik et al. 2002). This set of aerosol
244 intensive properties is expected to capture most of the important aspects that differentiate
245 the distinct aerosols' optical regimes that affect the study region. For instance, the
246 imaginary part of the complex refractive index (k) and single scattering albedo (SSA) are
247 properties that separate highly absorbing aerosol regimes from moderate and low
248 absorbing regimes. Angstrom Exponent (AE) and Asymmetry Parameter (ASY) are
249 properties that help separate aerosol regimes characterized by distinct size distributions.
250 LR is highly sensitive to size and composition-related information, for instance, the real part
251 of the complex refractive index. Meanwhile, LDR has high sensitivity to particle morphology,
252 and it is widely used to separate dust particles from other aerosol types. Given the
253 dependency of these intensive properties on the aerosol type (size and composition) and
254 mixed state, it is possible to characterize the aerosol scenarios over a specific AERONET site
255 in terms of their nature and sources (Eck et al., 1999; Dubovik et al., 2002). For instance, the
256 LR is the ratio of the extinction coefficient to the backscatter coefficient and is crucial for
257 identifying different aerosol types. It reflects how light scattering varies with particle size
258 relative to the light wavelength. Small particles, like smoke, have a high LR, while large
259 particles, like sea salt, have a low LR. Therefore, with a well-distributed regional network of
260 AERONET's sun photometers, as that covering the Iberian Peninsula, one can characterize
261 the spatial dynamics of aerosol types and mixture state influencing the regional aerosol
262 regimes. Regarding the time period for the current study, it extends from 2003 to 2023.
263 However, due to calibration and other operational aspects, some AERONET sites present
264 different time ranges within this period.

265 Three key aspects of aerosol nature have been widely used to link aerosol regimes with
266 particle emission sources. These aspects are absorption efficiency, size distribution, and
267 shape (Dubovik et al., 2002). For instance, combustion-based sources, including biomass
268 and fossil fuel burning, produce aerosol dominated by fine mode particles, and absorption
269 ranges from moderate to strong, depending on the nature of biomass burning, fossil fuel,
270 and ageing processes. In contrast, natural sources, such as deserts and marine
271 environments, produce aerosols dominated by coarse-mode particles. Marine aerosol

272 particles are characterized by very low absorption, while dust aerosol can exhibit high
273 absorption, mainly in the UV and VIS bands (Smirnov et al., 2002; Dubovik et al., 2002).
274 Furthermore, the irregular shape of dust particles is a key factor that differentiates them
275 from other aerosol types. This distinctive feature is captured by AERONET retrievals of the
276 LDR (Shin et al., 2018). Source attribution provides valuable insights into the typical
277 intensive optical properties affecting the atmospheric column of a site resulting from
278 complex aerosol state mixtures. This understanding is crucial as it addresses a major
279 challenge that current aerosol modules in CMIP6 climate models face (Zhao et al., 2022).
280 Reproducing climatological aerosol-intensive properties scenarios over specific regions has
281 been a major goal of atmospheric models. In addition to evaluating aerosol modules in
282 atmospheric models, AERONET's optical properties in typical regimes, which can be
283 expressed as spectral aerosol optical models (Omar et al., 2005; Levy et al., 2007; Rosario et
284 al., 2013; Zhou et al., 2023), are valuable for simulating aerosol direct radiative effects in
285 environmental models (Rosario et al., 2013; Li et al., 2019). This approach is especially
286 beneficial when high computational capacity is unavailable and explicit aerosol modules are
287 not feasible.

288 With more than 25 years of operating a vast network of Cimel Electronique Sun-sky
289 radiometers across the world, AERONET has provided highly accurate, ground-truth
290 measurements of aerosol optical depth and other properties (Giles et al., 2019). It has been
291 widely used as the main reference to evaluate and validate satellites (Gupta et al., 2018) and
292 model products (Gloß et al., 2021). The two most critical intensive optical properties to
293 estimating aerosol radiative forcing retrieved by AERONET, single scattering albedo (SSA)
294 and asymmetry parameter (ASY), are related, respectively, to absorption and size of the
295 aerosol. Their accuracies are aerosol loading dependent (Dubovik et al., 2002). For AOD >
296 0.4 at 440 nm (or > 0.2 at longer λ), SSA uncertainty $\approx \pm 0.03$, for lower AOD, uncertainty can
297 be ± 0.05 – 0.07 or larger. Regarding ASY, uncertainty is about ± 0.02 – 0.05 when AOD is high
298 (≥ 0.4 at 440 nm, ≥ 0.2 at longer wavelengths) but can be significantly larger at low AOD.

299 Aiming to identify a representative set of typical aerosol regimes that affect the Iberian
300 Peninsula, we applied cluster analysis methods (described in Sec. 2.4) to the AERONET sky
301 radiance retrievals dataset from 2003 to 2023, taking advantage of the extensive coverage
302 of AERONET sites across the region. Within that period, a total of 4395 retrievals from
303 AERONET Level 2.0 retrievals products were obtained and applied in the clustering process.
304 **Table 1** presents a set of intensive properties provided by AERONET that was used to
305 identify typical aerosol scenarios in the Iberian Peninsula atmospheric column. The
306 variables displayed cover all three previously mentioned aspects, absorption efficiency, size
307 distribution, and shape, which are expected to characterize the distinct nature of aerosol
308 types and mixture anticipated in the study region.

309

310

311

312

313 **Table 1:** List of AERONET inversions products (variables) used in clustering process
 314 followed by their abbreviation as defined by AERONET.

Variables	Abbreviation
Refractive Index - Real Part	RI _{Real} (440), RI _{Real} (670), RI _{Real} (870), RI _{Real} (1020)
Refractive Index - Imaginary part	RI _{Imag} (440), RI _{Imag} (670), RI _{Imag} (870), RI _{Imag} (1020)
Single Scattering Albedo	SSA(440), SSA(670), SSA(870), SSA(1020)
Asymmetry Parameter	ASY(440), SSA(670), SSA(870), SSA(1020)
Linear Depolarization ratio	LDR(440), LDR(670), LDR(870), LDR(1020)
Lidar Ratio	LR(440), LR(670), LR(870), LR(1020)
Fine and Coarse modes Volume median radius	VMR-F,VMR-C
Standard deviation from volume median radius, for Fine and Coarse modes	STD-F, STD-C
Fine and Coarse modes Effective radius	Reff-F, Reff-C

315

316 We selected only AERONET sites that operated for at least two years and that have sky
 317 radiance inversion available with the highest quality level 2.0. Some selected sites are still
 318 operational, while others have been discontinued. **Figure 1** illustrates the geographical
 319 distribution of the chosen sites. Our selection encompasses various landscapes of the
 320 Iberian Peninsula, from coastal plains regions (Coruña, Sagres, Burjassot) to highland
 321 plateaus in the interior (Madrid, Valladolid, Aras-de-los-Olmos) and lowland valleys
 322 (Zaragoza, Murcia). Regarding external air mass influence, sites in the southern border of IP
 323 are typically the first to experience the transport of dusty air masses from North Africa, with
 324 locations such as El- Arenosillo, Huelva, Malaga, and Sagres affected. The eastern sites
 325 (Barcelona, Burjassot, and Murcia) are expected to be strongly influenced by the
 326 Mediterranean air masses. Western and northern sites (Cabo da Roca, Coruna, Sagres) are
 327 directly under the influence of air mass from the Atlantic Ocean. Additionally, the
 328 Portuguese countryside (Evora) and eastern Spanish sites (Badajoz, Caceres) are located in
 329 regions that very often experience biomass burning during the dry season (Ermitão et al.,
 330 2023; Silva et al., 2023; Hammed e tal., 2024; Alvares et al., 2024).

331

332 2.3 Merra-2 Aerosol Diagnostic Product

333 The MERRA-2 (Modern-Era Retrospective Analysis for Research and Applications, Version
 334 2) Aerosol Diagnostic Product (ADP) is a comprehensive dataset provided by NASA that
 335 offers global information about atmospheric aerosols (Gelaro et al., 2017; Buchart et al.,

336 2017). MERRA-2 combines observational data with numerical models (reanalysis project)
337 to create a detailed long-term record of atmospheric dynamics and composition from 1980
338 to the present. Among other variables, the MERRA-2 ADP product offers a long-term view
339 of aerosol mass distribution by types and the related optical properties (Buchart et al.,
340 2017). Its extended temporal coverage allows analysis of aerosol trends, such as those
341 related to changes in atmospheric composition due to human activity and the impact on
342 climate. Key features of the MERRA-2 ADP include aerosol microphysical and optical
343 properties such as optical depth, mass concentration, and size distribution. These
344 properties are crucial for understanding aerosol loading and composition in the
345 atmosphere and their role in the Earth's radiation budget and climate system. A key aspect
346 of MERRA-2 ADP for this study is that it provides aerosol-type column mass density, our
347 target variable as a predictor of aerosol optical model regime. The MERRA-2 ADP includes
348 diagnostics for the aerosol types considered in most chemistry transport models: Dust (DT),
349 Black-Carbon (BC), Organic Carbon (OC), Sea-Salt (SS), and Sulfate (SF). The aerosol-type
350 diagnostics variables cover mass concentration at specific levels and are integrated into the
351 entire atmospheric column, which are applied to estimate columnar optical properties, such
352 as extinction, scattering, and absorption optical depths, at multiple wavelengths. For this
353 study, the 550 nm wavelength was used as a reference. Optical properties are a function of
354 aerosol species, particle size, and relative humidity. To convert from the simulated aerosol
355 masses to optical quantities such as aerosol optical depth, MERRA-2 uses Optics look-up
356 tables (LUTs) derived from Mie calculations using parameters from the Optical Properties of
357 Aerosols and Clouds (OPAC; Hess et al., 1998) , as described in Chin et al. (2002) and
358 Colarco et al. (2010), except for dust-type aerosol, which is based on Colarco et al. (2014).
359 Therefore, these optical properties are by-products of running the MERRA-2 reanalysis
360 system and made available to the community via MERRA-2 ADP. Further details on this can
361 be found in Buchard et al. (2017). From these extensive aerosol-driven optical properties, it
362 is possible to derive several MERRA-2 ADP intensive optical properties, such as Single
363 Scattering Albedo (SSA).

364

365 Given that the aerosol optical properties retrieved from each AERONET site are influenced
366 by mixtures of different aerosol types, it is reasonable to assume that the impact of each
367 aerosol type on the column's intensive optical properties is primarily determined by its
368 concentration. Based on this premise, we propose a machine-learning approach that utilizes
369 the aerosol-type column mass density predicted by chemistry transport models to prescribe
370 the most accurate possible spatial distribution of the aerosol spectral optical model
371 developed through cluster analysis of AERONET data. A description of the method
372 presented in this study, exploring MERRA-2 products, can be found in subsection 2.5.

373

374 **2.4 Optical models development: Cluster Analysis**

375 Cluster analysis has been extensively used to develop aerosol optical models based on
376 AERONET sky inversion products (Omar et al., 2005; Levy et al., 2007; Russel et al., 2014).
377 The underlying principle is that AERONET instantaneous retrievals can be grouped into a
378 certain number of clusters, each representing different categories of aerosol regimes. These

379 studies have explored mainly the K-means clustering method, one of the most popular
380 unsupervised machine learning algorithms for partitioning a dataset into a pre-defined
381 number of clusters. However, specifying the number of clusters in advance poses a
382 significant challenge for the K-means method. Fortunately, there are techniques available
383 that reduce the subjectivity involved in this pre-definition. In our study, we adopted the
384 Elbow method (Shi et al., 2021), a widely used method for determining the optimal number
385 of clusters (k) in a K-Means clustering algorithm. It examines the relationship between a
386 range of number of clusters and the Within-Cluster Sum of Squares (WCSS), which
387 measures the variance within each cluster (**Eq. 1**)

388

389

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

390

391

392 Where k is the number of cluster C_i is the set of point in a cluster μ_i is the centroid of cluster
393 i , and $\|x - \mu_i\|^2$ is the squared Euclidean distance. The WCSS measures the compactness of
394 the clustering, and one wants it to be as small as possible. If clusters are tight and
395 well-separated, WCSS will be small, because points are close to their centroids. If clusters
396 are loose or overlapping, WCSS will be large. The Elbow method is based on a plot of
397 WCSS(k) against the number of clusters(k). As k increases, WCSS always decreases; more
398 clusters mean tighter groups. The goal of the method is to identify the point (k) where the
399 rate of decrease in WCSS sharply slows down, indicating that adding more clusters yields
400 diminishing returns. We ran our clustering algorithm with k varying from 2 to 10 clusters.
401 For each k , we calculated the total WCSS. The k results against WCSS were displayed in a
402 plot, and the optimal number of clusters was defined based on the location (k) of the bend
403 (elbow) in the plot from which the change in WCSS slows down. This bend in the curve
404 indicates the point where adding more clusters no longer gives meaningful improvement.

405 The Elbow method has been widely used because of its straightforward approach to
406 estimating the most appropriate number of clusters. However, we recognize that it still
407 carries a certain degree of subjectivity, as it relies on visual interpretation. To reduce this
408 subjectivity, we combined the Elbow and Stability methods to evaluate the optimal number
409 of clusters that best represent the major aerosol regimes affecting the study region.
410 Although more rigorous methods are available in the literature, defining the number of
411 clusters remains a challenge, and different approaches often lead to distinct solutions
412 (Krishnaveni et al., 2023). Nevertheless, despite the limitations of the Elbow method, the
413 number of clusters identified in our study seems to provide a coherent characterization of
414 optical regimes affecting the Iberian Peninsula.

415 The K-means clustering analysis was performed using Scikit-learn, a Python open source
416 machine learning library that supports supervised and unsupervised learning (Abraham et
417 al. 2014). All our analyses were performed using the Scientific PYTHON Development
418 EnviRonment (Spyder), an open-source IDE for the Python language. As part of the
419 preprocessing step, the values of the features listed in Table 1 were normalized using the

420 *StandardScaler()* utility from *sklearn.preprocessing* package, which computes the mean and
421 standard deviation. Then the scaled dataset was obtained subsequently by subtracting all
422 values of the features from their respective mean and normalized by the corresponding
423 standard deviation. With this scaling process we expected to avoid the dominance of
424 features with large numeric ranges.

425 To evaluate the robustness of the clustering results, a set of sensitivity analysis using the
426 Scikit-learn library was performed. These tests were meant to test numerical stability,
427 sampling robustness, clusters reproducibility, separation quality and presence of
428 transitional regimes. A summary with all the results of these performed tests can be seen in
429 the supplementary material. First, a jitter robustness test was conducted by adding
430 gaussian noises taking different uncertainty estimates (5%, 10%, 20%) to the standardized
431 feature matrix and re-running the clustering algorithm multiple times (200 runs). The
432 resulting partitions were compared with the original clustering using the Adjusted Rand
433 Index (ARI), which assesses the sensitivity of the clustering structure to perturbations in
434 feature space (mimicking features uncertainties). Second, a bootstrap robustness analysis
435 was implemented by repeatedly resampling the dataset with replacement and recomputing
436 the clustering solution for each resampled dataset. The ARI between the original and
437 bootstrap-derived clusterings was calculated to evaluate the dependence of the clustering
438 structure on specific samples and to test its reproducibility under data variability. In
439 general, the mean ARI results for both jitter and bootstrap robustness analysis (Table S3 in
440 the supplement) presented values above 0.80, which indicate high similarity between the
441 originally predicted clusters and the perturbed, which support the clustering structure
442 robustness to random noises.

443 Cluster-level stability was quantified using a cluster retention score, defined as the average
444 probability that members of an original cluster remain grouped together across
445 perturbation runs also applying gaussian noises references (5%, 10%, 20%) to the
446 standardized feature matrix and re-running the clustering algorithm multiple times (200
447 runs). This metric allows identification of highly stable regimes as well as potentially
448 transitional or mixed clusters. With most of the cluster retention scores above 0.95 (Table
449 S4 in the supplement), results show that the obtained cluster regimes are highly stable. The
450 exception is cluster 3, which presented a retention score of 0.88, which can be considered
451 stable but somewhat overlapping. Finally, a consensus (co-assignment) matrix was
452 constructed by computing, for every pair of samples, the proportion of runs in which the
453 two samples were assigned to the same cluster. From this matrix, mean intra-cluster and
454 inter-cluster co-assignment probabilities were derived to quantify internal cohesion and
455 external separation (Table S5 in the supplement). The results of high intra-cluster
456 probability (>0.85) and low inter-cluster probability (<0.10) show that in general members
457 of the same cluster are consistently grouped together, which corroborates the quality of the
458 clustering results. This test was also conducted by adding gaussian noises considering
459 different uncertainty estimates (5%, 10%, 20%).

460 Together, these diagnostics provide a comprehensive evaluation of clustering stability,
461 reproducibility, and structural robustness.

462

463 2.5 Optical models spatial prescription: Random Forest Technique

464 Once the optimal number of clusters is defined, which corresponds to the expected number
465 of major optical properties regimes to influence the study region, and the clustering process
466 is performed, each cluster is characterized by a set of AERONET instantaneous retrievals of
467 optical and microphysical properties that are expected to express an optical property
468 regime. Also, each AERONET instantaneous retrieval is tagged with the cluster number that
469 it belongs to. By averaging the instantaneous properties of each cluster, we set the reference
470 values that represent the mentioned major aerosol optical properties regimes.

471 We propose a machine-learning approach that utilizes the well-known random forests
472 supervised algorithm (Breiman, 2001) to spatially represent the aerosol optical models
473 defined by the cluster analysis for each AERONET site (described in section 2.4). The
474 implemented method was tested using exclusively aerosol column mass density fields from
475 MERRA-2 (**Table 2**) to establish the spatial distribution of the optical regime defined by the
476 cluster's average. This approach is also suitable for chemistry transport models.

477 MERRA-2 time series of column mass density for each aerosol type (DT, BC, OC, SS, SF) over
478 each AERONET site were collocated with the network inversion products used to derive the
479 clusters representing the distinct aerosol regimes over the Iberian Peninsula (described in
480 section 2.4). Only MERRA-2 column mass density fields were used in the Random Forest
481 training process; optical properties fields from the reanalysis system were not used at this
482 stage. Merra-2 column mass densities are available with a frequency of 1 hour, while
483 AERONET optical properties instantaneous retrievals are provided at irregular times, due to
484 its dependence on cloud cover and AOD criteria (AOD at 440 nm > 0.4). So, for each
485 AERONET retrieval, our script searches for the MERRA-2 closest hour to synchronize the
486 two datasets. Therefore, the collocation between AERONET retrievals and MERRA-2, for a
487 spatial matching we considered the nearest neighbor by taking the MERRA-2 grid cell that
488 contains the AERONET station location, and for temporal collocation MERRA-2 hourly
489 aerosol diagnostics were matched to the closest AERONET observation time. Each
490 AERONET instantaneous aerosol microphysical and optical properties inversion retrieval
491 (Sinyuk et al., 2020) was connected to the corresponding cluster to which it belongs.
492 Likewise, as mentioned, each instantaneous aerosol microphysical and optical properties
493 inversion retrieval was also connected to the closest in-time combination of MERRA-2 data
494 of aerosol-type column mass density (DT, BC, OC, SS, SF). With this, we built a time series of
495 4309 points of spatial and temporal collocated MERRA-2 aerosol types of column mass
496 density with the developed clusters occurrences over each AERONET site, which was used
497 in a training process aiming to predict the suitable cluster given a specific combination of
498 aerosol types column mass density predicted by MERRA-2. This time series was randomly
499 divided into training and testing subsets, with 70% (3016 collocated combinations) of the
500 data used for model training and 30% (1293 collocated combinations) reserved for
501 independent evaluation. This was done using the `train_test_split` utility from the Scikit-learn
502 library (Abraham et al. 2014).

503 The algorithm uses training data to learn the relationship between the combination of
504 aerosol-types' columnar mass density and the target, which are the previously developed

505 clusters from AERONET aerosol-intensive properties. The training was done using the
506 Random Forest Classification algorithm (RandomForestClassifier) from Scikit-Learn .

507 We used stratified k-fold cross-validation integrated within a RandomizedSearchCV
508 hyperparameter optimization process as a strategy. Our hyperparameter optimization was
509 performed using RandomizedSearchCV with five-fold cross-validation. The search space
510 included the number of trees (n_estimators) sampled uniformly between 50 and 500 and
511 the maximum tree depth (max_depth) sampled between 1 and 20. A total of five random
512 hyperparameter combinations were evaluated, and the best-performing model was refitted
513 on the full training dataset. The random search methodology was used to find parameter
514 combinations inside the parameter space without the processing demands of grid search
515 and with the stratified k-fold cross-validation we search to ensure that each fold has
516 approximately the same class proportions as the full dataset, which allows a fair evaluation,
517 since every validation set includes samples from all classes. This also contributes to the
518 meaning of the performance metrics in relation to the minority classes, for example strong
519 absorbing aerosol regimes. The stratified k-fold also favours a more stable training, given
520 that in every training split the less frequent aerosol property regimes are also seen, which
521 helps to reduce variance in model performance across folds. This strategy contributes to
522 improving the hyperparameter tuning, once the RandomizedSearchCV won't select
523 parameters based on misleading folds. So, by preserving class distribution in every fold and
524 preventing biased results, the strategy based on stratified k-fold cross-validation helps to
525 handle class imbalance, which in turn improves the model reliability and generalization.
526 Class imbalance is typical in atmospheric aerosol characterization, where extreme but
527 radiatively important aerosol regimes, like intense smoke episodes, are rare compared to
528 more common background conditions. Therefore, with the strategy described, we also
529 aimed to address the issues of class imbalance of aerosol regime classification in our study.

530 Cross-validated performance indicators were used to select the final configuration in order
531 to reduce overfitting and ensure consistent performance across aerosol regimes. To evaluate
532 model performance across aerosol regimes, and not to rely only on the overall accuracy, the
533 performance metrics were computed individually for each regime

534 The confusion matrix utility from the Scikit-learn library (confusion_matrix) was used to
535 visualize the performance of the models by comparing true labels with predicted labels. It
536 allowed us to evaluate performance for each class individually and to support the
537 interpretation of the per-class metrics calculated, namely Accuracy, Precision and Recall,
538 and F1 score. While the confusion matrix provides the context that explains the model
539 performance, the per-class metrics provide numerical performance values for each class.

540 Accuracy represents the number of correctly classified data instances over the total; it
541 checks the predictions against the actual values in the test set and returns the percentage of
542 times the model got right. Precision and recall are two critical metrics for evaluating the
543 performance of a classification model. Precision is the proportion of true positives among
544 all the predicted positive cases (true and false), meaning it measures the accuracy of
545 positive predictions (**Eq. 2**). Recall is the proportion of true positives among all actual
546 positive cases (true and false), meaning it measures the model's ability to identify positive
547 cases (**Eq. 3**). The F1 score, the harmonic mean of a model's precision and recall, takes both

548 precision and recall and provides a more balanced measure of a model's performance (Eq.
 549 4). The F1 score is set to be a value between 0 and 1, indicating, respectively, poor precision
 550 and recall and high precision and recall, which is ideal.

551
$$\text{Precision} = \text{True positive} / (\text{True positive} + \text{False positive}) - (2)$$

552
$$\text{Recall} = \text{True positive} / (\text{True positive} + \text{False negative}) - (3)$$

553
$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Recall} + \text{Precision}) - (4)$$

554

555 **Table 2:** Predictor variables from Merra-2 (aerosol-type column mass density) used in the
 556 machine learning process to prescribe the aerosol optical regime (optical model).

Variables	Abbreviation	Unity	Spatial resolution
Dust column mass density	DUCMASS	kg/m ²	0.5° × 0.625°
Black carbon column mass density	BCCMASS	kg/m ²	0.5° × 0.625°
Organic carbon column mass density	OCCMASS	kg/m ²	0.5° × 0.625°
SO ₂ column mass density	SO2CMASS	kg/m ²	0.5° × 0.625°
SO ₄ column mass density	SO4CMASS	kg/m ²	0.5° × 0.625°
Sea salt column mass density	SSCMASS	kg/m ²	0.5° × 0.625°

557 To identify which aerosol types most influence aerosol regime prescription and to
 558 understand whether meaningful types are driving predictions, we used the
 559 *best_rfc.feature_importances_*, an utility in scikit-learn's RandomForestClassifier, to calculate
 560 the scores indicating the importance of each aerosol types in the training dataset. The
 561 importance of an aerosol type in making predictions was based on how much it reduces
 562 impurity across all trees. Each decision tree in the forest splits data using different features,
 563 and each split reduces impurity. The reduction in impurity is attributed to the feature used
 564 at that split. An aerosol type importance is based on the frequency that it is used to split
 565 nodes.

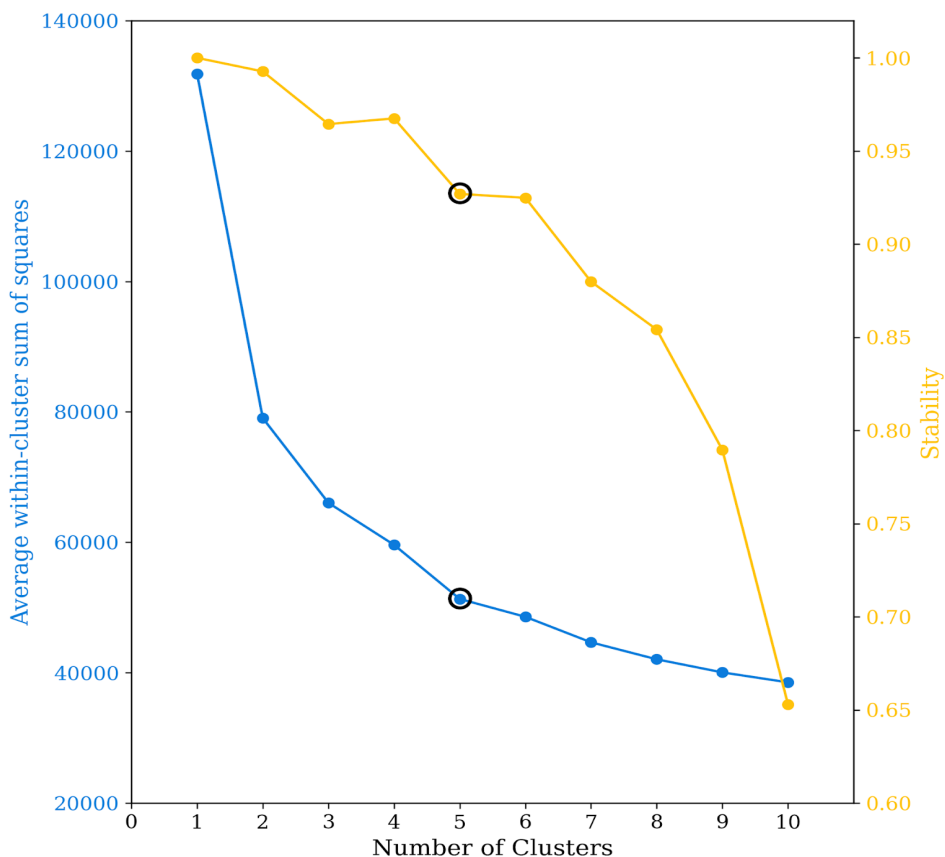
566 3. Results

567 The results section is divided into three subsections. The first one presents the results of
 568 identifying the typical aerosol optical regimes affecting the Iberian Peninsula using cluster
 569 analysis. The second subsection discusses the results and the performance of spatial
 570 prescription of these typical aerosol regimes by applying machine learning (Random Forest)
 571 to the columnar density of MERRA-2 aerosol components. Finally, case studies applying the
 572 method developed are presented and discussed.

573

574 3.1 Cluster Analysis: Optical models development

575 The number of clusters (k) selected to characterize the typical optical aerosol regimes over
576 the Iberian Peninsula was defined based on the Elbow method (**Figure 2**), which indicated
577 five clusters were the optimal number to capture the aerosol regime variability. We also
578 evaluated from the Elbow method that there is a sharp bending at $k=2$, which we associated
579 with a clustering separation between aerosol regimes strongly dominated by coarse mode,
580 dust regimes, and regimes dominated by fine mode, non-dust regimes. However, to cover
581 more specific regimes within these two macro-regimes (dust regimes vs non-dust regimes),
582 a higher k is required, and $k=5$ is revealed to be the second sharpest bending. Cluster
583 stability as a function of the number of clusters was also evaluated as a way of evaluating
584 whether the clusters obtained are meaningful and not just artifacts of randomness or noise.
585 High stability suggests clusters represent real structure in the data, not just random
586 fluctuations. The stability for $k=5$ is above the 90% threshold, similar to $k=6$, a number
587 after which stability sharply decreases. Therefore, combining the Elbow method and
588 stability reinforced $k=5$ as an optimal cluster number to capture the typical aerosol
589 scenarios over the Iberian Peninsula, reducing the subjectivity usually associated with the
590 K-means clustering method.

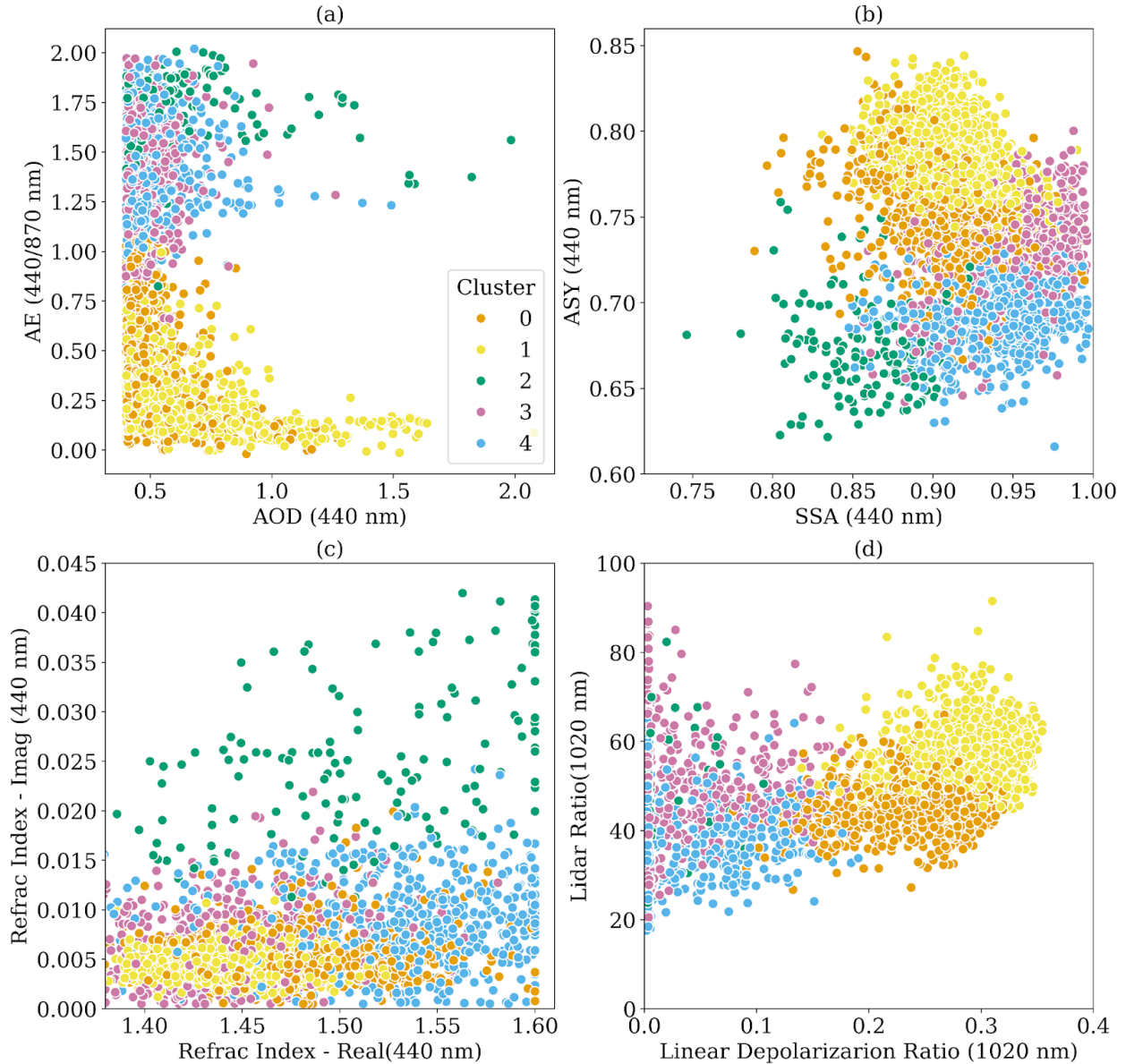


591

592 **Figure 2:** Average of sum of squares within-cluster and cluster stability as function of the
593 number of clusters.

594 We applied the cluster analysis once we defined the optimal number of clusters. As
595 described, the result for clustering process robustness was assessed using the fraction of
596 points changing cluster membership relative to the original classification by perturbing the
597 input data by ± 1 standard deviation (SD) and rerun the k-means clustering. The fraction of
598 points changing cluster assignment was 15.3% for the +1SD perturbation and 17.6% for the
599 -1SD perturbation, yielding a mean sensitivity of 16.5%. Only 0.5% of points changed
600 cluster membership under both perturbations, indicating that the observed sensitivity is
601 largely confined to boundary points, while the cluster cores remain robust.

602 **Figure 3** presents a combination of graphics used for aerosol properties analysis,
603 highlighting the obtained clusters' behavior and distinction. The first graphic (Fig. 3a)
604 represents the Aerosol Optical Depth (AOD) as a function of Angstrom Exponent (AE),
605 which allows us to relate aerosol loading variability with aerosol regimes dominated either
606 by coarse or fine mode (Eck et al., 1999). This analysis shows that two of the clusters (C0
607 and C1) are regimes dominated by coarse mode particles ($AE < 1.0$), while the remaining
608 three (C2, C3, and C4) are regimes under the stronger influence of fine mode particles ($AE >$
609 1.0). The second plot displays the asymmetric parameter against the single scattering
610 albedo at 440 nm. This plot aims to elucidate the clusters' distinctions related to particle
611 absorption efficiency and the asymmetry between hemispherical forward and backward
612 scattering. Aerosol regimes dominated by coarse particles tend to exhibit more significant
613 forward scattering and, consequently, higher asymmetry parameter values. In contrast,
614 lower asymmetry parameter values are expected in fine mode regimes (Eck et al., 1999;
615 Dubovik et al., 2002). This pattern is evident in the graphic; clusters C0 and C1 present
616 higher asymmetry parameter values. It is also possible to identify the distinction between
617 the non-dust regimes C2, C3, and C4. C2 presents the lowest asymmetry parameter values,
618 while it is the most absorbing of the clusters, according to its single scattering albedo
619 values. Small and highly absorbing particles are commonly associated with urban pollution
620 or fresh smoke plumes from biomass burning (Dubovik et al., 2002; Omar et al., 2005; Levy
621 et al. 2010; Martins et al., 2009). The C3 cluster differs significantly from C2 by presenting
622 higher asymmetry parameter values, an indication of a shift to larger particle sizes. C3 has
623 higher single-scattering albedo values, indicating a less absorbing aerosol regime. SSA alone
624 did not help to differentiate the two clusters dominated by coarse mode particles (C0 and
625 C1). C0 asymmetry parameter values tend to be lower than those of C1, suggesting that the
626 former could be a dusty mixture not as close to a pure dust scenario as C1. The traditional
627 plot of Lidar Ratio (LR) against Linear Depolarization Ratio (LDR) (Kanitz et al. 2013,
628 Illingworth et al., 2015) confirms this hypothesis (Fig. 3d). Pure dust regimes of aerosol, due
629 to their high level of non-spherical particles, produce higher LDR (Groß et al., 2011). The C1
630 cluster presents higher values of LDR than C0, indicating that C1 is closer to a pure dust
631 regime. The C0, while a dust regime, is likely to represent a mixed scenario given its LDR
632 values consistent with dust and smoke mixing (Kanitz et al. 2013). LDR values below 15%,
633 which is the case of the clusters C2, C3, and C4, are typically associated with fresh/aged
634 smoke, urban-industrial pollution, and marine particles scenarios. The analysis of the real
635 part versus the imaginary part of the complex refractive index (Fig. 3c) emphasizes C2 as
636 the aerosol regime with the largest absorption and highlights that the real part of the
637 complex refractive index is the main aspect differentiating C3 and C4.



638

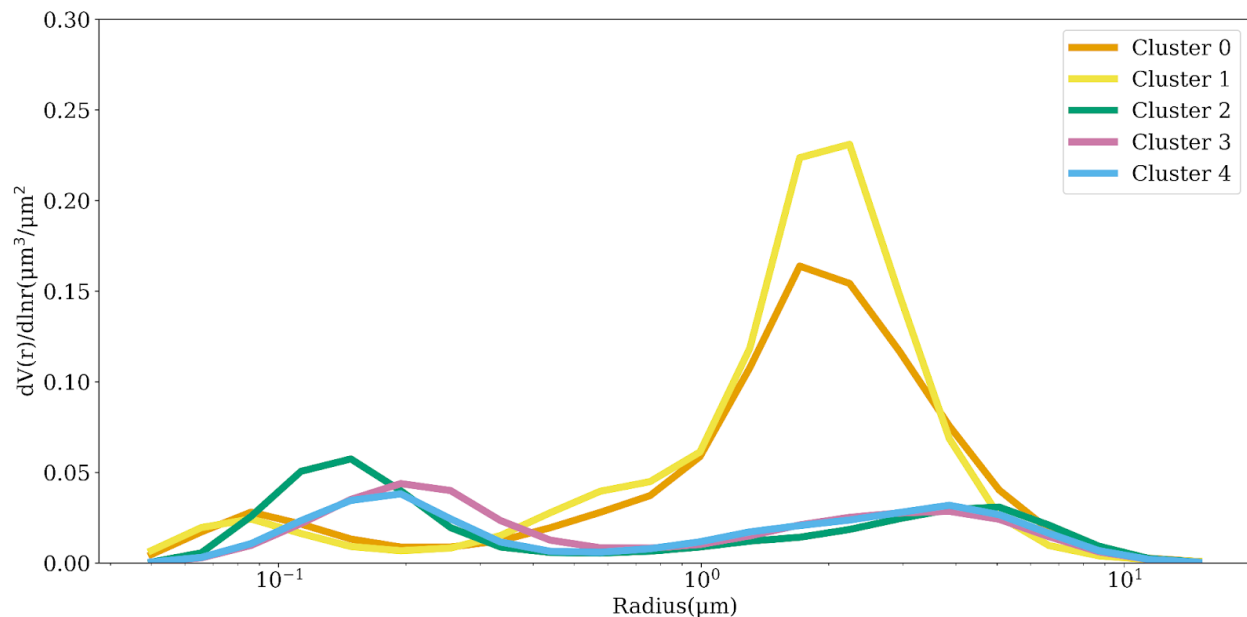
639 **Figure 3:** Scatterplot of the clusters elements as function of different parameters: (a) Extinction
 640 Angstrom Exponent (AE) as function of Aerosol Optical Depth (AOD) at 440 nm; (b) Asymmetry
 641 Parameter (ASY) as function of Single Scattering Albedo (SSA) at 440 nm; (c) Lidar Ratio as a function
 642 of Linear Depolarization Ratio at 1020 nm; (d) Refractive index at 440 nm: Imaginary part as function
 643 of Real part.

644

645 **Figures 4 and 5** present the clusters averages for selected features: size distribution,
 646 complex refractive index, single scattering albedo, and asymmetry parameter. A more
 647 detailed summary of the mean behavior of the clusters is presented in **Table 3**. The average
 648 size distribution of the clusters confirms that aerosol regimes affecting the Iberian
 649 Peninsula vary between two scenarios dominated by coarse mode (C0, C1), named here as
 650 dust regimes, and three scenarios when coarse mode is not dominant, here considered as

651 non-dust regimes. There are differences between the dust regimes: C1 is associated with a
 652 higher coarse particle loading than C0. Among the non-dust regimes (C2, C3, and C4), the
 653 main difference is seen between C2 and the other two. C2 is characterized by a larger fine
 654 particle loading. Between C3 and C4, one can observe a larger radius spread for C3
 655 regarding the contribution of the fine mode, which indicates a potential growth of particles
 656 via processes such as water uptake, aging, and coagulation, or that the aerosol regime
 657 mixture includes sources that naturally produce slightly larger fine particles. These features
 658 usually indicate more aged, more hygroscopic, or more humidified aerosol compared to
 659 freshly emitted, dry fine particles.

660



661

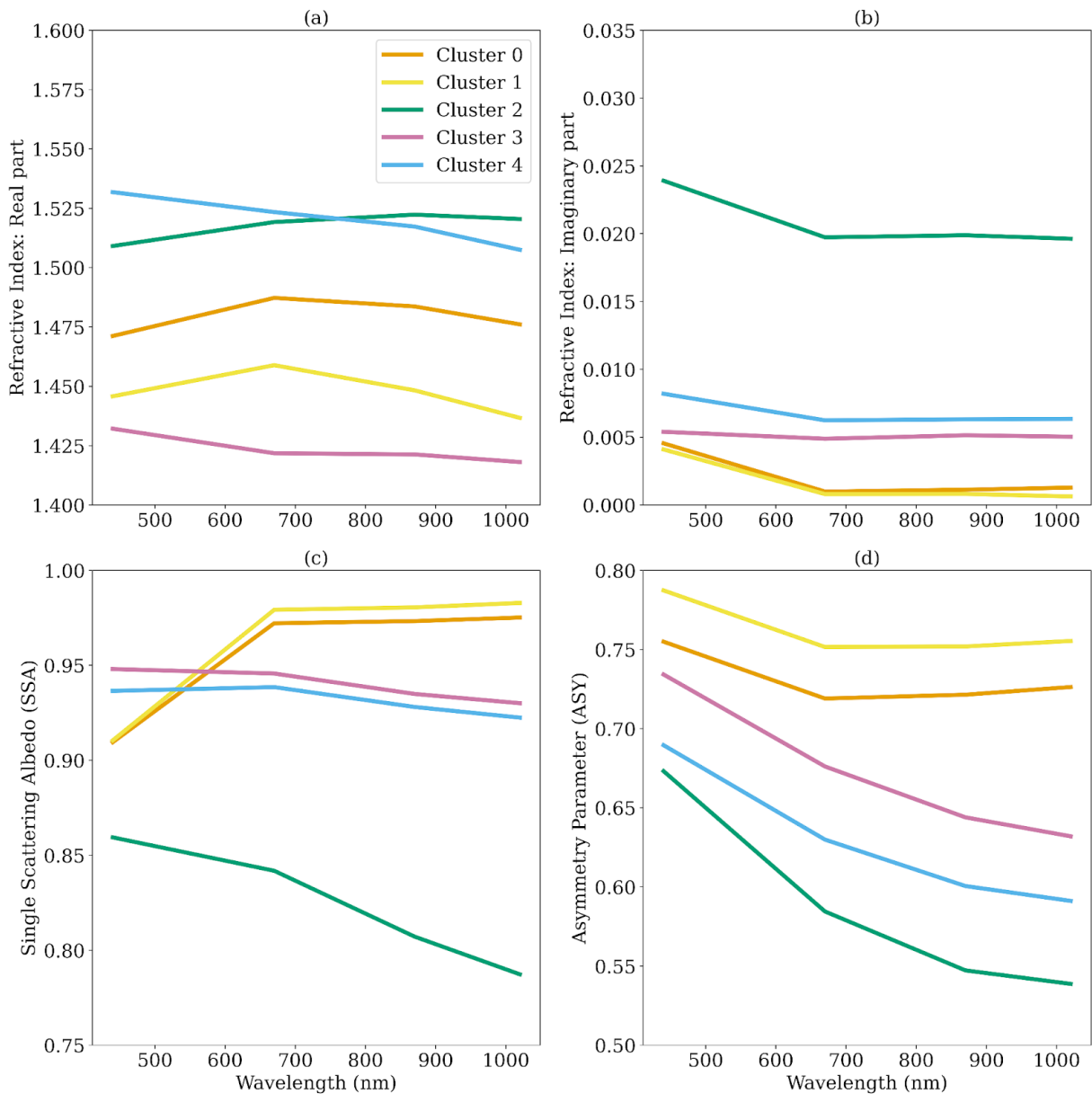
662 **Figure 4:** Clusters mean volume particle size distribution as a function of radius. These size
 663 distributions correspond to the average of the instantaneous size distributions retrieved by
 664 AERONET from each identified cluster. The numeric values of each cluster size distribution can
 665 be found in Table S2 in the supplement.

666

667 Clusters C2 and C4 have close values for the real part of the refractive index, but cluster C2
 668 has a much larger imaginary part, justifying its lowest SSA (**Figure 5**). The C2 strong
 669 absorption combined with its smaller particles suggests that it is likely associated with
 670 fresh smoke (Reid et al., 1998; Reid et al., 2005). The average of the real part of the complex
 671 refractive index corroborates the difference between the C3 and C4 aerosol regimes.
 672 According to Moise et al. (2015), a variation as such observed between C3 and C4 (1.4 to
 673 1.5) could produce an increment of 12 % in estimating the direct aerosol radiative forcing
 674 over the solar spectrum wavelength range. Zhao et al. (2019) also showed that the direct
 675 aerosol radiative forcing is estimated to vary by 40 % when the real part of the complex
 676 index values varies between 1.36 and 1.56. The reasons for the differences observed
 677 between the real parts of C3 and C4 remain unclear. However, the spatial distribution of the
 678 clusters (see Fig. 6) indicates that C3 is more prevalent in the eastern region of the Iberian

679 Peninsula, which is the wettest area and more exposed to air masses from the
 680 Mediterranean and Eastern Europe. Additionally, the low values of the real part of the
 681 complex refractive index for C3 align with aerosol regimes that have a strong contribution
 682 from sulfate particles. The spectral dependency of the single scattering albedo corroborates
 683 our attribution of the C0 and C1 to a dust regime. Dust particles are characterized by strong
 684 absorption in the UV spectrum (Dubovik et al., 2002), which decreases as the wavelength
 685 increases, a feature present in both C0 and C1. Also, consistent with dust-dominated
 686 regimes, C0 and C1 have the largest mean asymmetry parameters at all wavelengths.

687



688

689 **Figure 5:** Clusters average of complex refractive index, (a) Real and (b) Imaginary parts, (c)
 690 single scattering albedo and (d) asymmetry parameter.

691 The analysis above and the summary provided by **Table 3** provide several specific
692 characteristics that help us to contextualize the clusters. To enhance this understanding, we
693 add the spatial (**Figure 6**) and seasonal (**Figure 7**) distribution of the clusters into our
694 analysis. C0 and C1 aerosol regimes are dominated by dust, where C1 is the closest regime
695 to what we could call a pure dust scenario. Both aerosol regimes, C0 and C1, affect
696 practically the entire Peninsula (**Figure 6**) and all year round, but it is more frequent in the
697 southern part of the Peninsula, an expected feature considering that the dust particles are
698 mainly transported from North Africa (Cachorro et al., 2016; Gómez-Amo et al., 2017). The
699 C2 cluster is the most absorbing regime and is characterized by the smallest fine mode
700 particles (**Table 3**). Its spatial distribution (**Figure 6**), with more frequent occurrence along
701 the belt spanning from Evora, in Portugal, to Caceres, in Spain, a region known for high
702 recurrence of biomass burning, reinforces our hypothesis. Additionally, the seasonal
703 distribution of C2 in this region coincides with the peak of the biomass burning season. The
704 C3 aerosol regime also occurs over all AERONET sites throughout the year, but it is
705 dominant in the eastern and northeastern portions of the Iberian Peninsula. Among
706 non-dust regimes, its unique feature is its very low real part of the refractive index. C4, as
707 C3, is weakly absorbing according to its single scattering albedo. C4 is present across the
708 entire Peninsula, but its occurrence increases in the central and northern portions, which
709 are more prone to biomass burning. An important feature of C4 is that its occurrence
710 increases during the summer and the beginning of autumn (**Figure 7**) in the central region
711 of the Iberian Peninsula, from Évora (Portugal) to Madrid (Spain), when the region's
712 biomass burning season is underway. These aspects led us to hypothesize that C4 is an
713 aerosol regime under the strong influence of smoke aerosol particles.

714

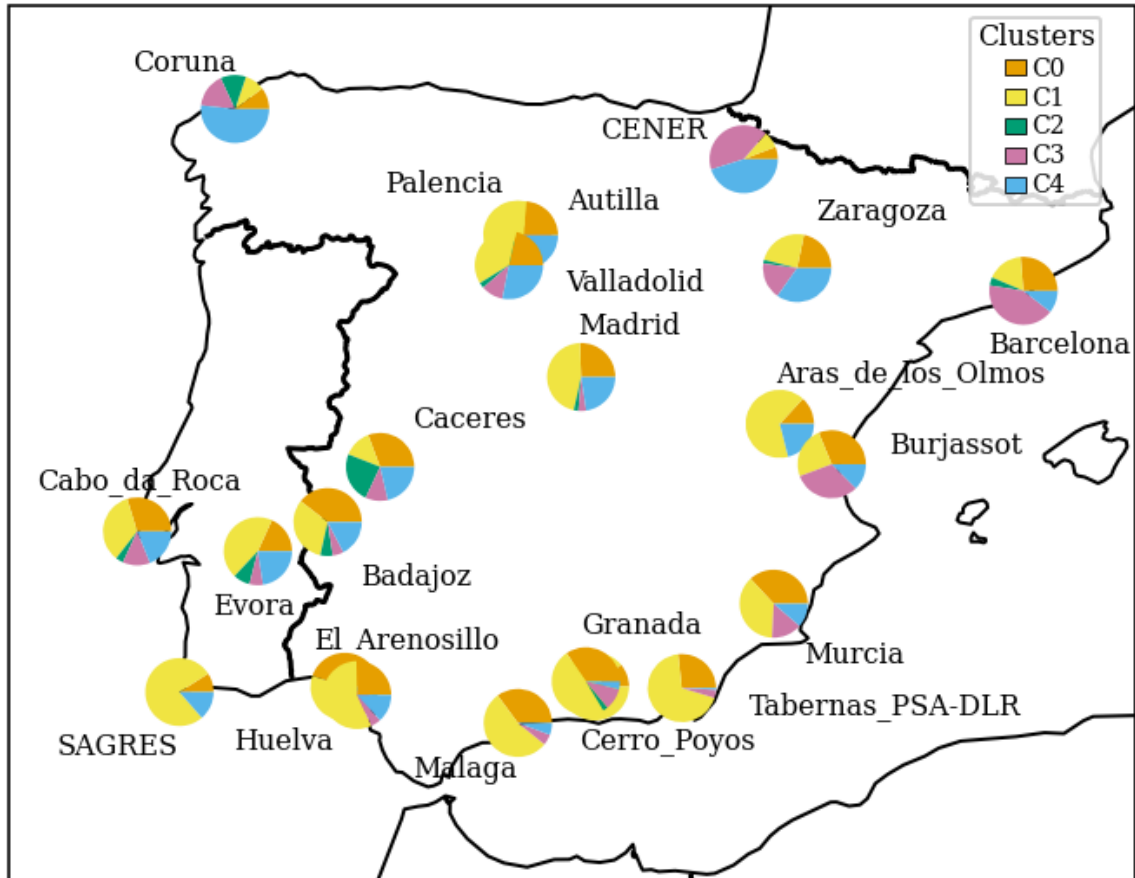
715 **Table 3:** Summary of the clusters based on the average of optical and microphysical properties. A
716 detailed description of the clusters can be found in Tables S1 and S2 in the supplement. The
717 values in the brackets correspond to standard deviation.

Properties	Cluster0 (Polluted dust)	Cluster1 (Pure dust)	Cluster2 (Strongly absorbing smoke)	Cluster3 (Urban-Indust rial Pollution)	Cluster4 (Moderately absorbing smoke)
Number of records	1308	1665	153	660	609
Percentage (%)	29.76	37.88	3.48	15.01	13.74
Ref_Idx_Real (440 nm)	1.47(0.04)	1.44(0.03)	1.51(0.07)	1.43(0.06)	1.52(0.05)
Ref_Idx_Img (440 nm)	0.005(0.002)	0.004(0.001)	0.025(0.009)	0.006(0.004)	0.009(0.004)
VMR-F	0.14(0.03)	0.14(0.03)	0.16(0.02)	0.21(0.04)	0.18(0.04)
STD - F	0.61(0.09)	0.67(0.07)	0.42(0.06)	0.47(0.06)	0.41(0.05)
REff-F	0.12(0.02)	0.12(0.02)	0.14(0.02)	0.18(0.03)	0.17(0.03)
REff-C	1.68(0.16)	1.61(0.13)	2.44(0.43)	2.31(0.38)	2.25(0.49)
VMR-C	2.02(0.23)	1.88(0.17)	3.10(0.45)	2.82(0.42)	2.82(0.57)
STD-C	0.60(0.52)	0.54(0.04)	0.68(0.06)	0.63(0.05)	0.67(0.05)
AOD (440 nm)	0.50(0.11)	0.58(0.21)	0.64(0.29)	0.48(0.09)	0.51(0.13)
SSA (440 nm)	0.91(0.03)	0.91(0.02)	0.86(0.03)	0.95(0.03)	0.94(0.03)

ASY (440 nm)	0.76(0.02)	0.79(0.19)	0.67(0.03)	0.73(0.03)	0.69(0.02)
AE(440/870 nm)	0.40(0.25)	0.24(0.14)	1.67(0.20)	1.43(0.26)	1.47(0.25)
LR(1020 nm)	64(9)	70(8)	89(16)	77(17)	61(15)
LDPR(440 nm)	0.17(0.04)	0.21(0.04)	0.01(0.03)	0.03(0.04)	0.03(0.05)

718

719

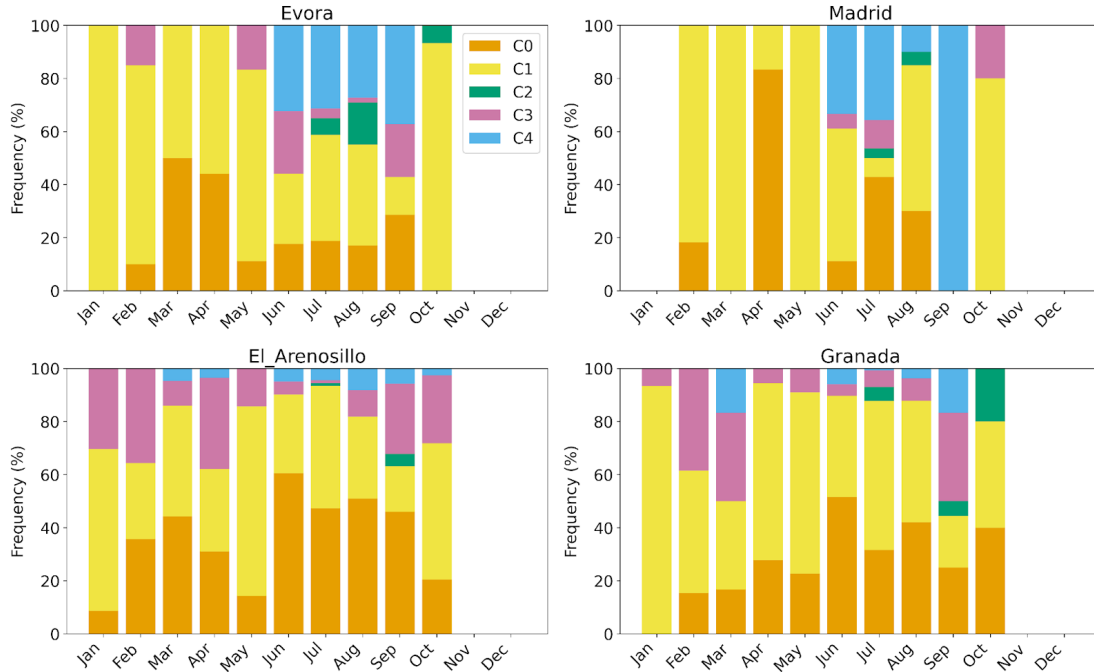


720

721 **Figure 6:** Proportions of the occurrence of the clusters of aerosol regimes at the AERONET
722 sites across the Iberian Peninsula.

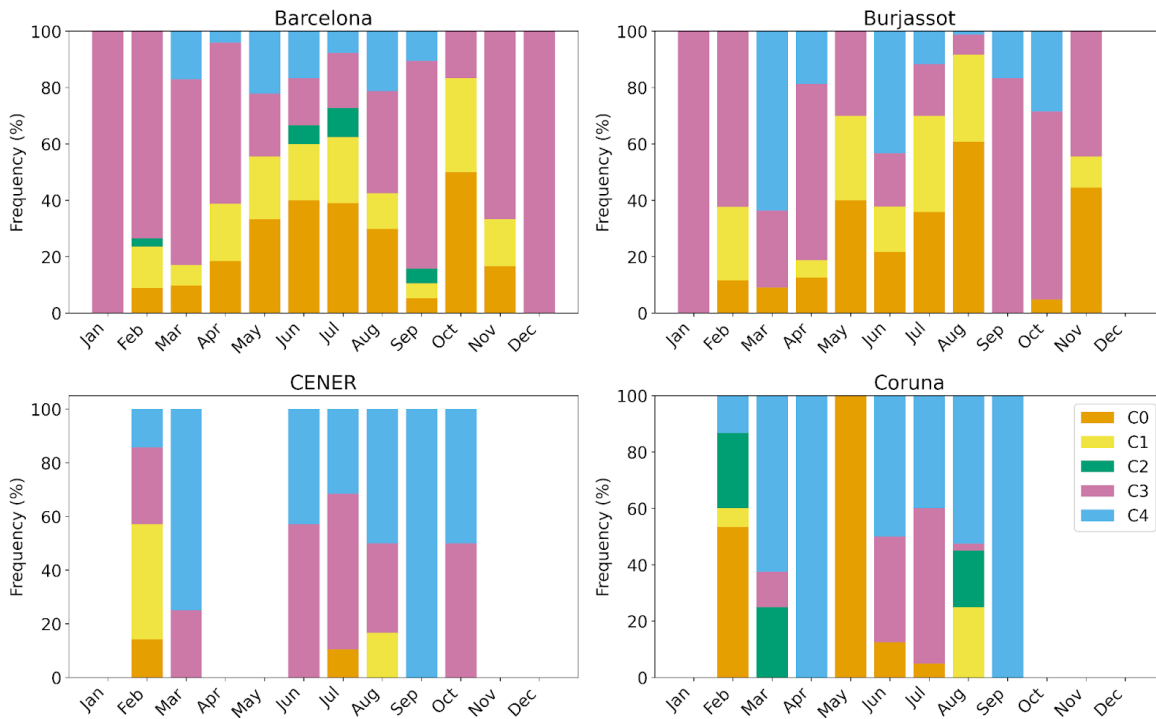
723

724 **Figures 7 and 8** provide a perspective view on the seasonal occurrence of each cluster
725 based on sites that represent different regions of the Iberian Peninsula.



726

727 **Figure 7:** Clusters relative monthly occurrence over the AERONET sites representative of the
 728 Iberian Peninsula western lowlands (Evora), highlands plateau (Madrid) and southeast
 729 lowlands (El Arenosillo, Granada).



730

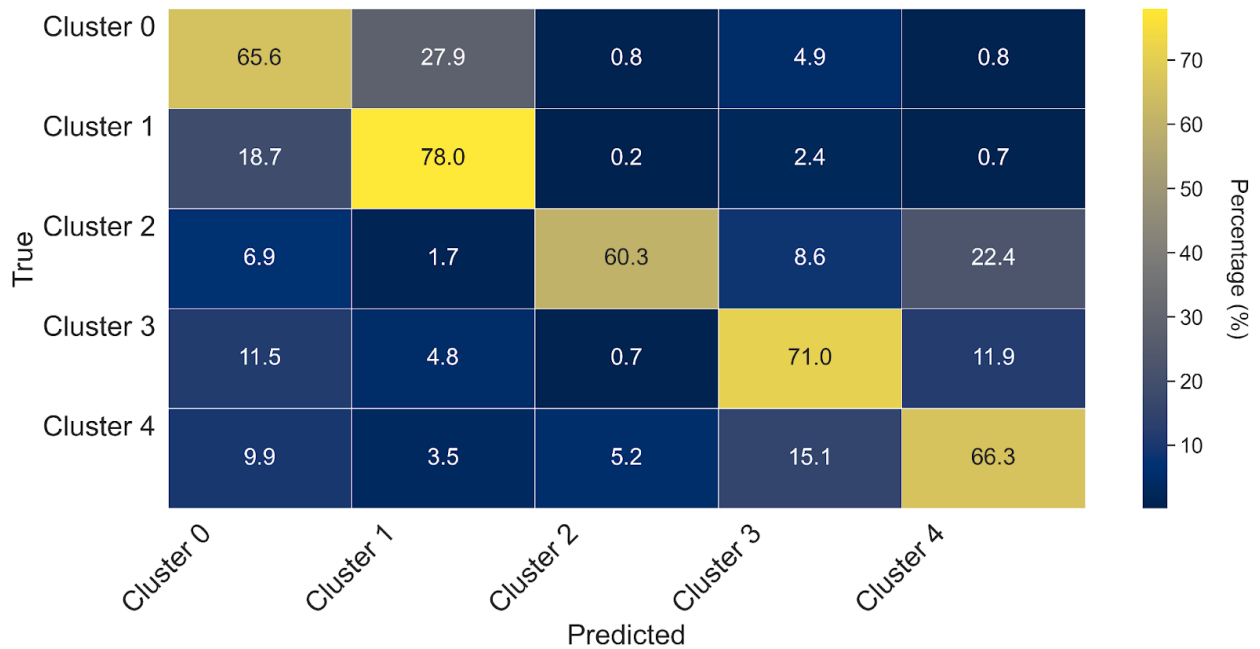
731 **Figure 8:** Clusters relative monthly occurrence over the AERONET sites representative of the
 732 following Iberian Peninsula regions: Eastern Coast (Barcelona, Burjassot) and Northern
 733 (Coruna, CENER).

734 3.2 Random Forest Classifier: Performance and Optical models spatial dynamic

735 The Random Forest training of MERRA-2 aerosol-type column mass density as predictors of
736 aerosol optical regime covered 70% of the AERONET sky inversions used in this study,
737 combining datasets from all sites. The testing dataset, constituted by the remaining 30%, was
738 used to evaluate the model's performance. The best parameters obtained from the optimization
739 using RandomizedSearchCV were the number of decision trees of 477 ($n_estimators = 477$)
740 and the maximum depth of trees of 19 ($max_depth=19$). There are several metrics for
741 assessing machine learning performance. **Figure 9** presents the one used in this study, the
742 Normalized Confusion Matrix (NCM), which expresses the percentage of correct and
743 incorrect predictions (where the classifier got confused). In the matrix, the rows represent
744 the true labels, and the columns represent the predicted ones. The values along the diagonal
745 indicate the percentage of times where the predicted matches the true label. The other cells
746 reflect instances where the classifier mislabeled an observation; the column tells us what
747 the classifier predicted, and the row tells us the correct label.

748 For all clusters, the classifier's correct predictions surpassed the incorrect predictions, with
749 a maximum frequency of correct predictions close to 80% obtained for C1. The minimum
750 percentage of correct prediction, about 60%, was obtained for C2, the highest absorbing
751 cluster. Regarding dust regime clusters, despite the struggle to predict C0, it is possible to
752 see that, in this case, the classifier's main confusion is with the C1, which is also a cluster
753 related to an aerosol scenario dominated by coarse mode particles (dust regime), as with
754 C0. The classifier's confusion in this case is between the two dust-regime models; therefore,
755 the induced error in optical properties prescription would be lower than that if the
756 confusion was between a dust and a non-dust regime, especially like C2, which is
757 substantially different from any of the dust regimes. Rarely does the classifier take either
758 C0 or C1 as C2, C3, and C4, a case where substantial error in the optical properties
759 prescription would be expected. By combining C0 and C1 results in the NCM, the
760 percentage of correct predictions achieved by the classifier indicating dust regime is higher
761 than 95%. Similarly, the classifier rarely takes C3 and C4 as C0, C1, and C2. Given that C3
762 and C4 are also close in terms of their optical properties, especially concerning absorption,
763 some degree of confusion among them is expected. Nevertheless, these aspects of the
764 confusion matrix among close clusters are important to identify where the model needs
765 extra training, for instance, considering longer time series when available and adding new
766 and relevant predictors, such as Brown Carbon, an important aerosol component not
767 available in the current MERRA-2 aerosol reanalysis products. C2, the least frequent and the
768 one representing the most absorbing aerosol regime over the Iberian Peninsula, is rarely
769 mislabeled as C0 or C1, but often mislabeled as C3 or C4. Still, the score percentage is
770 around 60%.

771



772

773 **Figure 9:** Normalized confusion matrix of the Random Forest classifier applied to the
 774 prediction of the clusters that describe the typical aerosol optical regime based on MERRA-2
 775 aerosol components column mass density.

776 To provide further insight into the model performance, we also examined other metrics
 777 commonly used to evaluate Random Forest training. Precision, Recall, and F1 score were
 778 calculated for both scenarios, the trained model applied to the test and to the train dataset
 779 (**Table 4**). The results indicate that the model generalizes well, without significant
 780 overfitting. Even for Cluster 2, which has a small number of occurrences, the model was able
 781 to maintain high precision and score. The general accuracy did not drop critically for the
 782 test data (0.70) when compared with the train dataset (0.88), another indicator of the
 783 model's ability to generalize. The trained model applied to the test dataset achieved a
 784 general accuracy of 70 %, meaning it correctly predicted the aerosol regime in three out of
 785 four cases. For all clusters, all metrics adopted were higher than 0.60, with precision and
 786 recall values exceeding 0.75 in some cases. The precision metric indicates how often the
 787 positive predictions are correct. The model precision varied within the specific optical
 788 regimes (ex., non-dust) and among optical regimes (dust, non-dust). It showed higher
 789 precision in identifying C1 than C0, the two dust regimes. Among the non-dust regime
 790 clusters, the lowest precision obtained was 0.62 for the prediction of C2; nevertheless, this
 791 precision is still a promising outcome considering the limited number of samples of this
 792 cluster available for the training process. Given its strong absorption nature, mislabeling the
 793 C2 aerosol regime would translate into high error in optical properties prescription;
 794 therefore, as mentioned, extra training is required to improve the model prediction for C2
 795 occurrence.

796

797 **Table 4.** Performance metrics values of the trained model applied to the test and the train
 798 (within parenthesis) dataset to predict aerosol optical regimes based on aerosol-type
 799 column mass density.

Clusters	Precision	Recall	F1-Score	Support (N)
0	0.62(0.89)	0.62(0.78)	0.62(0.83)	361(909)
1	0.68(0.86)	0.70(0.94)	0.69(0.90)	517(1130)
2	0.62(0.92)	0.60(0.76)	0.61(0.83)	47(106)
3	0.76(0.93)	0.73(0.94)	0.74(0.94)	206(429)
4	0.68(0.91)	0.69(0.93)	0.69(0.92)	162(442)

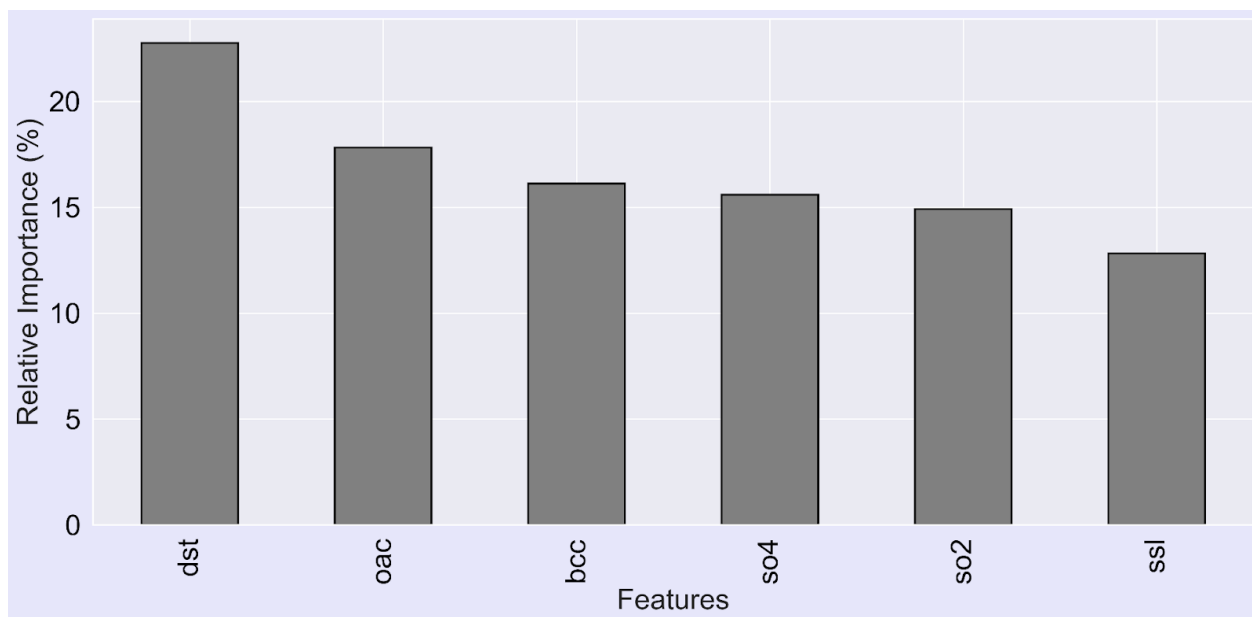
800

801 **Figure 10** illustrates the relative importance of the predictor variables for the grids
 802 consisting of the AERONET sites, highlighting the influence of each aerosol-type column
 803 mass density on the model's decision-making. The results indicate that the presence of dust
 804 over the Iberian Peninsula is the primary factor affecting the aerosol optical properties in
 805 this region. This finding aligns with actual conditions, as the transport of Saharan dust to
 806 the peninsula is the main driver of aerosol optical properties variability in the area. Dust is
 807 followed by organic carbon, sea salt, and sulfate aerosol types. Organic carbon relevance is
 808 associated with biomass burning, a critical aerosol source during the dry season.
 809 Interestingly, black carbon column mass density did not rank among the top predictors.
 810 Despite the expectation that black carbon might serve as a significant indicator of the
 811 aerosol optical regime due to its association with smoke-influenced aerosols. There is
 812 considerable uncertainty in black carbon simulations in atmospheric chemistry models,
 813 including reanalyses such as MERRA-2, which may hinder its effectiveness in predicting the
 814 aerosol regime observed at AERONET monitoring sites.

815 We also managed to calculate the relative importance of the predictors from Table 1 in the
 816 cluster prediction; the result is presented in the supplement (Fig. S2). Consistency can be
 817 observed between the score scale from Fig. S2 and that derived from MERRA-2 with respect
 818 to aerosol types ((Figure 10). In Figure 10, dust (dst) mass variability emerges as the most
 819 influential factor in determining which cluster should be applied. In the Fig. S2 figure, which
 820 presents the importance of the optical parameters from Table 1 for clustering, the scores
 821 appear well distributed, with a maximum value close to 0.1. Nevertheless, it is evident that
 822 higher wavelengths (near-infrared at 870 and 1020 nm) and specific optical
 823 parameters—namely the Asymmetry Parameter and the Linear Depolarization
 824 Ratio—exhibit the greatest importance, as they are most effective in distinguishing dust
 825 from other aerosol types.

826

827



828

829 **Figure 10:** Relative importance of the predictor variables, i. e. the degree of influence of each
 830 aerosol-type column mass density on the model decision-making. dst - Dust , oac - Organic
 831 Carbon, ssl - Sea-Salt, so4- Sulfate so2 - Sulfur dioxide(precursor of so4), bcc - Black Carbon.

832

833 3.3 Application: Case studies

834 From the testing dataset, we selected some case studies that significantly impacted local
 835 populations, garnered media attention, and represented different aerosol scenarios in the
 836 Iberian Peninsula. This selection provides a visual (qualitative) demonstration of the
 837 model's predicting capability (Table 5).

838

839 Table 5: List of case studies of aerosols high loading events over Iberian Peninsula selected
 840 to highlight as examples of the classifier trained model application.

Case study	Date	Nature (Reference link)
#01	June 27, 2023	Smoke1
#02	October 16, 2017	Dust and Smoke2
#03	August 11, 2016	Smoke3
#04	March 17, 2022	Dust4

841 1-<https://earthobservatory.nasa.gov/images/151507/canadian-smoke-reaches-europe>

842 2-<https://atmosphere.copernicus.eu/saharan-dust-and-smoke-over-france-and-uk>

843 3-<https://earthobservatory.nasa.gov/images/88552/fires-rage-in-portugal>

844 4- <https://earthobservatory.nasa.gov/images/149645/dusty-storm-clouds-over-europe>

845

846 We set our trained model to prescribe the spatial distribution of aerosol optical regimes
847 (clusters) that best fit various scenarios based on MERRA-2 aerosol-type column mass
848 density. The results for all cases studied are presented in Figure 11. To minimize
849 uncertainties associated with the estimates of aerosol absorptivity, AERONET SSA retrievals
850 are limited to cases where the AOD at 440 nm exceeds 0.4 (Dubovik and King, 2000; Holben
851 et al., 2006). Therefore, the discussion of the optical regime prescriptions was focused on
852 areas where AOD was above this threshold. The uncertainty in retrieved SSA is ~ 0.03 at
853 AOD at 440 nm = 0.4 and decreases at higher AOD levels (Sinyuk et al., 2020).

854 For our regional analysis, we used the MERRA-2 AOD field as a reference, since AERONET
855 provides local AOD at specific sites. Figure S1 in the supplement exhibits as example the
856 seasonal cycle aerosol AOD at 550 nm over the Iberian Peninsula. Given that the
857 prescription is done based on a map of the combination of aerosol types column density
858 from models, in this case MERRA-2, the only way to filter areas across the Iberian Peninsula
859 where AOD at 440 nm > 0.4 is to use the AOD field from MERRA-2.

860 Case#01 occurred from June 1 to 25, 2023, coinciding with large-scale wildfire events in
861 Quebec, Canada. A substantial portion of smoke from these wildfires crossed the Atlantic
862 Ocean and reached Western Europe, especially the Iberian Peninsula, resulting in darkened
863 skies in the affected countries. Our trained model predicted that the most suitable aerosol
864 optical regime for the areas impacted by the smoke (Portugal, Western, and Northern Spain)
865 is C4, which corroborates our previous discussion associating the C4 optical regime with
866 regional smoke.

867 Case#02 features an emblematic event on October 16, 2017, marked by a simultaneous
868 massive wildfire in central and northern Portugal and a strong dust transport from North
869 Africa via the south of Portugal. The path connecting the smoke and dust produced a strong
870 northward transport affecting the United Kingdom, influenced by the synoptic conditions
871 associated with the ex-hurricane Ophelia, located just north of the Iberian Peninsula
872 (Osborne et al., 2019). The optical regime prescription identified the C4 cluster as the
873 appropriate regime from central Portugal northward to the UK. Meanwhile, the area
874 affected by dust, spanning from North Africa to southern and central Portugal, was
875 characterized by a mix of C0 and C1, the clusters associated with dust regimes. As the dust
876 plume arrived in Portugal, the model indicated a gradual transition from C1, indicative of
877 pure dust, to C0, which represents conditions of dust mixed with smoke (Gómez-Amo et al.,
878 2017). The random distribution of C2 within the larger C4 regions likely reflects the model's
879 response to the specific conditions dictated by the aerosol-type column mass densities. This
880 could suggest patches of high-absorbing aerosol-type within a less-absorbing large-scale
881 smoke plume, although there is insufficient evidence to draw definitive conclusions.

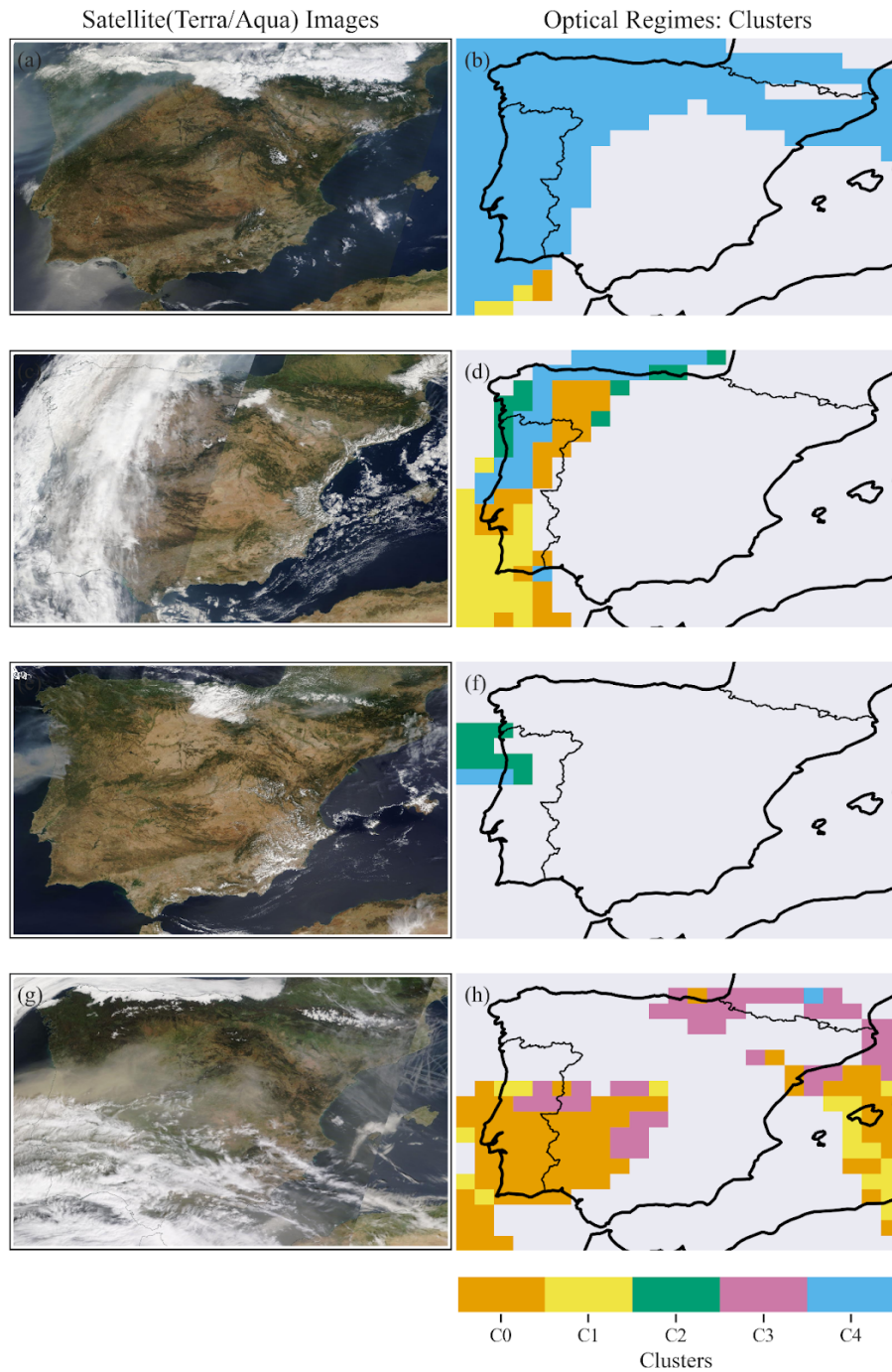
882 Case#03, dated August 16, 2016, involved strong wildfire emissions in northern Portugal.
883 Most of the smoke was transported toward the Atlantic Ocean, while the remainder of the
884 peninsula experienced low aerosol loading conditions. Consistent with fresh smoke aerosol

885 scenarios, the model prescribed the C2 optical regime, the highest absorbing cluster. In
886 strong biomass burning events, especially at the early stage of the emission process, the
887 ratio of elemental carbon to organic carbon is usually high, which has been shown to explain
888 the high absorption features of fresh smoke plume (Schwink et al, 2024). Additionally,
889 previous studies have also shown that Brown Carbon(BrC) absorption is strongest in fresh
890 smoke plumes and decreases with atmospheric processing (Saleh, R., et al., 2014; Pokhrel,
891 et al., 2017).

892 Case#04 pertains to an extreme Saharan dust transport that affected most of the Iberian
893 Peninsula on March 15-17, 2022. During this event, the 24-hour average concentration of
894 PM_{2.5} reached as high as 700 $\mu\text{g m}^{-3}$ in parts of Spain (Rodriguez and López-Darias, 2024).
895 The pollution episode was dominated by dust, and indeed, the model prescribed the optical
896 regimes C0 and C1, which indicate pure dust and dusty conditions for most of the Iberian
897 Peninsula. This demonstrates our approach's capability to differentiate specific scenarios
898 within dust regimes. For non-dust regimes such as C2, a highly absorbing regime, we would
899 not expect to see widespread prescriptions, as we hypothesize that it is associated with
900 fresh, high-absorbing pollution plumes.

901 **Figure 6**, depicting the occurrence of each cluster across the Iberian Peninsula,
902 corroborates our hypothesis by indicating that the C2 regime is mainly present in specific
903 areas where aerosol loading increases are primarily attributed to biomass burning, such as
904 the western lowlands of the Iberian Peninsula (Evora, Badajoz, and Caceres) and in the
905 Galicia region (Coruna). The C3 optical regime was not linked to large-scale dust transport
906 or smoke plumes across the Iberian Peninsula, suggesting it might be associated with high
907 levels of local or regional pollution. **Figure 6** shows that the C3 regime is commonly
908 observed throughout the year in the eastern portion of the Iberian Peninsula. The results of
909 these case studies, combined with performance evaluations, highlight the capability and
910 potential of this machine-learning approach, which uses clustering and random forest
911 classification to prescribe optical models from aerosol-type columnar mass density to
912 calculate aerosol particles' direct radiative effect in atmospheric models. By constraining
913 modelling with observational aerosol optical data, we can help mitigate the known
914 uncertainties related to aerosol direct radiative forcing.

915



916

917 **Figure 11:** Case studies of distinct aerosol scenarios over the Iberian Peninsula selected to test
 918 our machine-learning based approach to predict the best optical property regime: (a, b)
 919 Case#01 on June 27, 2023; (c, d) Case#02 on October 16, 2017; (e, f) Case#03 on August 11,
 920 2016; (g, h) Case#04 on March 17, 2022. On the left side, MODIS Terra and Aqua/NASA True
 921 color satellite images (<https://wvs.earthdata.nasa.gov>); and on the right the cluster spatial
 922 distribution prescribed by the model.

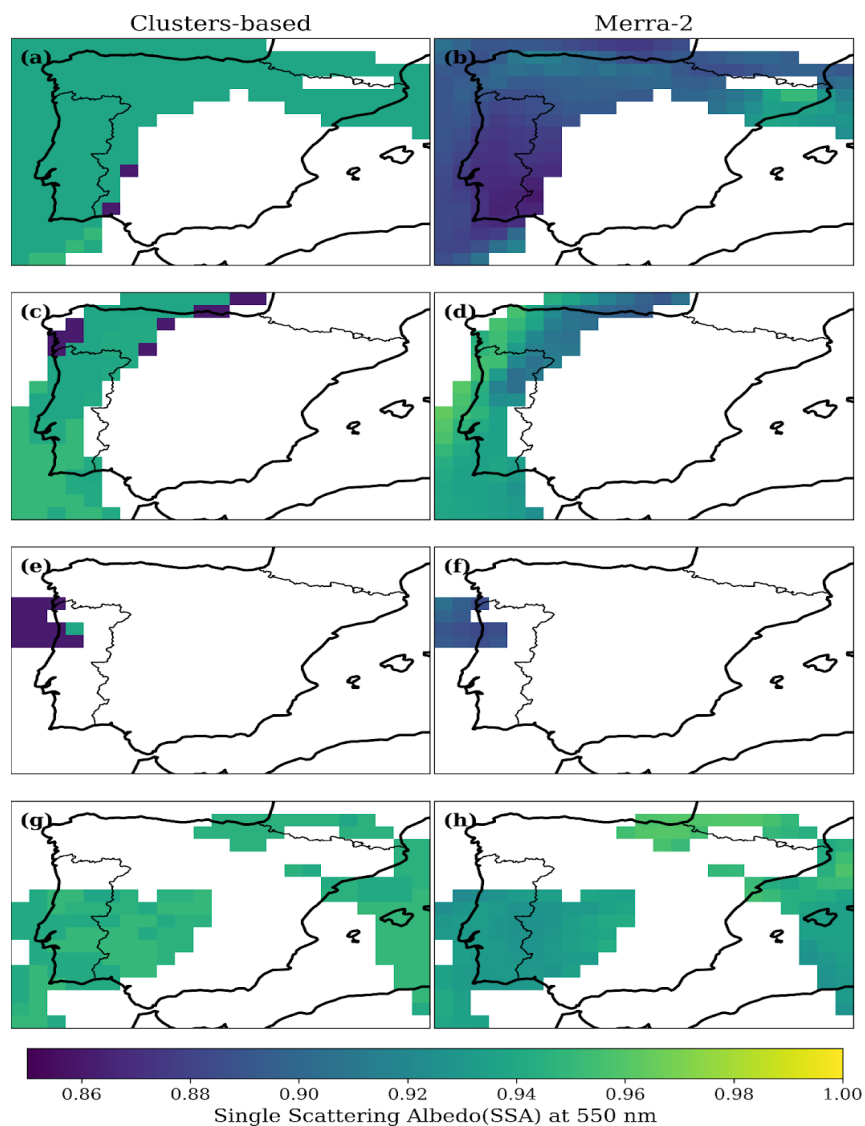
923

924 **Figure 12** shows the single scattering albedo at 550 nm, comparing the current approach
925 and MERRA-2 reanalysis results. The MERRA-2 columnar total SSA was calculated based on
926 the ratio of total scattering aerosol optical depth to total extinction aerosol optical depth,
927 both provided in MERRA-2 aerosol products. For smoke scenarios on June 27, 2023,
928 MERRA-2 indicated a more absorbing optical regime (SSA at 550 nm $\sim 0.86 - 0.90$)
929 compared to the current approach (SSA at 550 nm ~ 0.95). On this day, the average SSA at
930 550 nm over the AERONET site in Coruna City, which was directly affected by Canadian
931 smoke, exceeded 0.95. The opposite was observed for the strong smoke emission event that
932 occurred over northern Portugal on August 11, 2016. Current strategy prescribed a much
933 lower SSA, therefore the strongest absorbing regime, when compared with the MERRA-2
934 calculation. Due to the absence of a site in the northern part of Portugal, we were not able to
935 compare the prescribed and simulated values with AERONET data. For the dust scenario,
936 on March 17, 2022, the current approach prescribed a less absorbing optical regime (SSA at
937 550 nm $\sim 0.94 - 0.95$) compared to MERRA-2, which reported a SSA at 550 nm of roughly
938 $0.92 - 0.94$. The analysis of SSA at 550 nm over AERONET sites affected by the dust event
939 surpassed 0.94. While these cases highlight differences between the prescriptions based on
940 the clusters and MERRA-2 results, they are only sufficient to warrant further investigation.
941 To gain a statistical perspective on whether the findings from these case studies are isolated
942 incidents or indicative of a trend, we compare a much larger sample of MERRA-2 SSA at 550
943 nm across various AERONET sites in the Iberian Peninsula using the clusters approach. We
944 focused only on MERRA-2 aerosol scenarios for AOD at 550 nm larger than 0.3, which
945 correspond to AOD higher than 0.4 at 440 nm previously mentioned, and conducted the
946 comparison segmented by the optical regimes defined by the clusters.

947 **Figure 13** shows the count distribution of MERRA-2 SSA at 550 nm for the aerosol regimes
948 represented by the clusters C0, C1, C3, and C4, as classified by the random forest classifier
949 we developed. Histograms of clusters of SSA at 550 nm presented in Figure 13 were
950 generated following a Gaussian distribution, considering the cluster average as the central
951 value of each optical regime cluster and standard deviation as the typical spread. A similar
952 analysis was conducted for the Angstrom Exponent (**Figure 14**) to evaluate aspects related
953 to particle size distribution. Based on **Figure 13**, we found that, on average, our aerosol
954 optical regime prescription based on the clusters (AERONET) is less absorbing than
955 MERRA-2 for aerosol regimes C0, C1, C3, and C4. More significant differences are observed
956 for C1, C3, and C4. Cluster C1 corresponds to a dust scenario closer to pure dust, while C4 is
957 dominated by smoke. Regarding the particle size indicator (AE), it was observed that
958 MERRA-2 has a lower contribution of coarse particles in the dust regimes compared to the
959 cluster-based prescriptions (**Figure 14a, b**). This finding aligns with Adebisi et al. (2023),
960 who noted that climate models tend to underestimate large dust particles, mainly when
961 representing North African dust plumes. Conversely, for the non-dust regimes (C3, C4),
962 MERRA-2 shows a larger relative contribution of coarse particles than the clusters-based
963 prescription (**Figure 14c, d**). **Figure 15** shows the results for C2. For this specific regime,
964 on average, prescriptions based on the cluster (AERONET) are more absorbing than
965 MERRA-2, opposite to the findings of the other clusters. Regarding AE, under the C2 regime,
966 MERRA-2's mean AE is lower than that prescribed from the cluster, suggesting a lower
967 relative contribution of fine mode in the reanalysis simulations. This is similar to the
968 findings related to the two other fine-mode dominant regimes (C3 and C4).

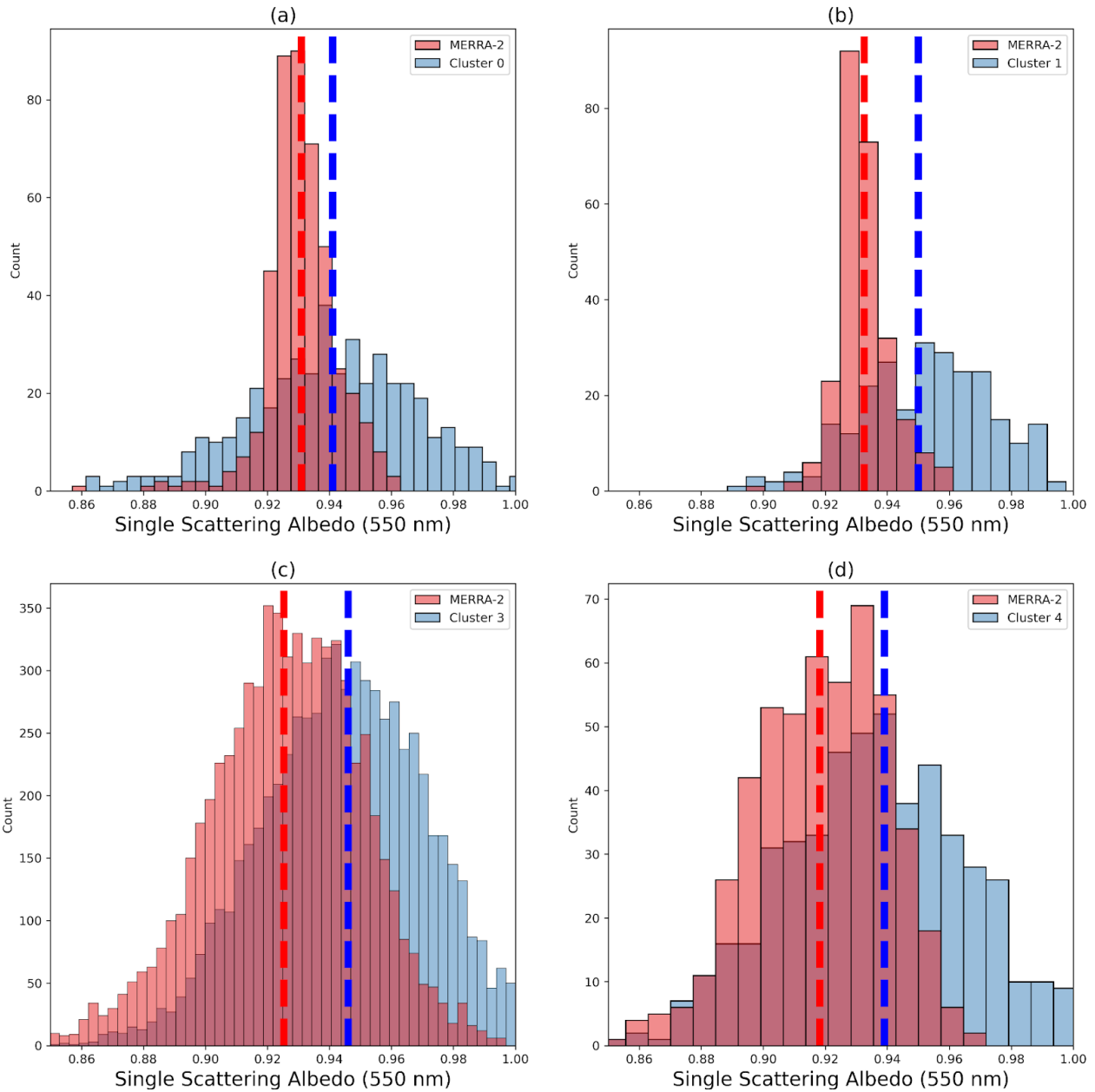
969 As demonstrated by the SSA and AE distributions (Figures 13, 14, 15) and evaluated from
 970 Tables 1 and 4, the model can also predict the occurrence of the minority cluster C2 (3–4
 971 percent of samples). The model preserves the distribution of optical properties of less
 972 frequent aerosol regimes while capturing MERRA-2 features without the need for explicit
 973 class imbalance treatment, with C2's highly absorbing and dominant fine mode conditions
 974 reflected in both SSA and AE predictions, with the distributions of values across clusters
 975 showing coherence with MERRA-2 values. With C2's highly absorbing and dominant fine
 976 mode conditions reflected in both SSA and AE predictions, the distributions across clusters
 977 demonstrate agreement between expected and observed distribution values.

978



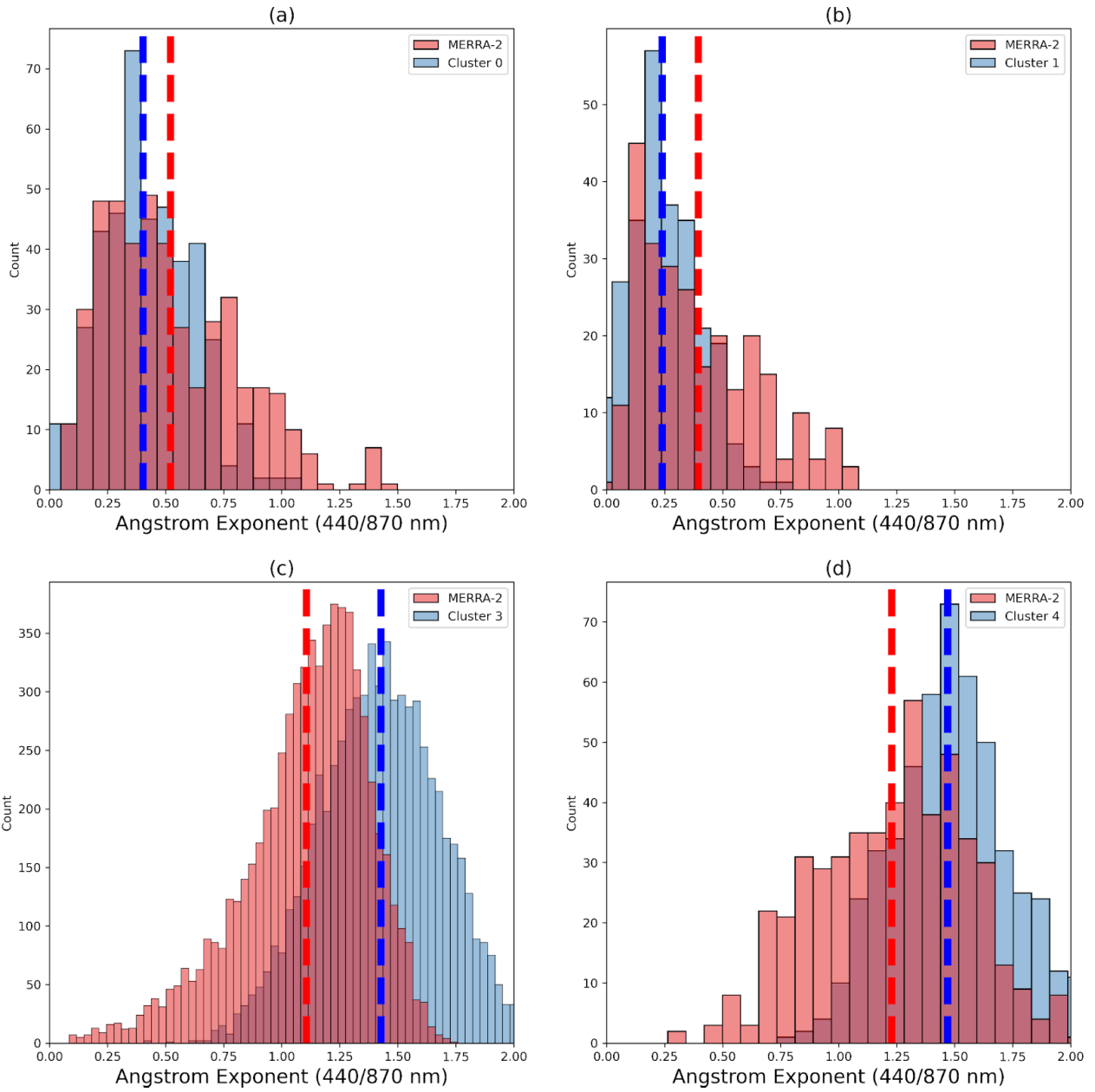
979

980 **Figure 12:** Single Scattering Albedo (SSA) prescription based on the current study approach
 981 (left) and that simulated by MERRA-2 (right) for the selected case studies of Table 2: (a, b)
 982 Case#01 on June 27, 2023; (c, d) Case#02 on October 16, 2017; (e, f) Case#03 on August 11,
 983 2016; (g, h) Case#04 on March 17, 2022.



985

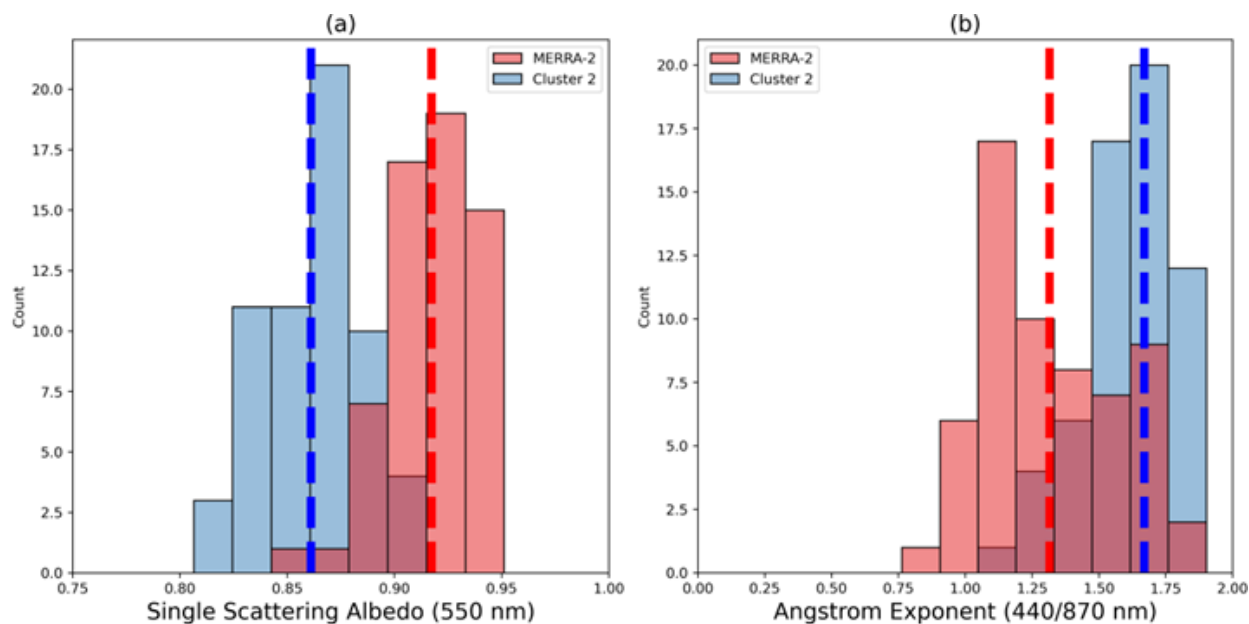
986 **Figure 13:** Current study prescription and MERRA-2 simulation of Single Scattering Albedo
 987 (SSA) frequency distribution as function of the optical regime (clusters): a) Cluster 0; b)
 988 Cluster 1; c) Cluster 3; e) Cluster 4. The dashed lines represent the mean values.



989

990 **Figure 14:** Current study prescription and MERRA-2 simulation of Angstrom Exponent (AE)
 991 frequency distribution as function of the optical regime (clusters): a) Cluster 0; b) Cluster 1; c)
 992 Cluster 3; e) Cluster 4. The dashed lines represent the mean values.

993



994

995 **Figure 15:** Current study prescription and MERRA-2 simulation of (a) Single Scattering
 996 Albedo and (b) Angstrom Exponent (AE) frequency distribution for the Cluster 2 scenario. The
 997 dashed lines represent the mean values.

998

999 4. Conclusions

1000 This study emphasizes the importance of observational-based research to constrain the
 1001 prescription of aerosol-intensive properties in atmospheric models. We aimed to
 1002 characterize the typical aerosol intensive optical properties affecting the Iberian Peninsula
 1003 (IP) using data from the atmospheric column AERONET sky inversion products. We
 1004 employed K-means clustering to analyze historical aerosol intensive properties across all
 1005 AERONET sites that operated for at least two years and had the highest quality dataset level
 1006 (2.0) available. We identified five distinct clusters (C0, C1, C2, C3, and C4) representing
 1007 different optical regimes, illustrating the predominant aerosol scenarios in the IP. The key
 1008 difference among these clusters lies in the contribution of coarse-mode particles and their
 1009 absorption efficiency. Clusters C0 and C1 are dominated by coarse-mode particles and
 1010 classified as dust regimes due to their association with Saharan dust transport. In
 1011 particular, the optical properties of C1 closely resemble a pure dust scenario, while C0
 1012 indicates a more mixed situation, which we refer to as dusty. On the other hand, clusters C2
 1013 and C4 are identified as non-dust regimes, linked to strong and moderate absorption related
 1014 to smoke plumes. Cluster C3, also a non-dust regime, is more frequently observed in the
 1015 eastern part of the IP and differs from C4 mainly by having a much lower real part of the
 1016 refractive index. After identifying the typical aerosol regimes affecting the IP, we utilized
 1017 aerosol-type columnar mass density data (dust, organic carbon, black carbon, sea salt, and
 1018 sulfates) from MERRA-2 to predict the aerosol optical regime at each grid point using the
 1019 supervised learning methodology Random Forest. We tested the performance of the trained

1020 model under various aerosol scenarios. The accuracy of the predictions for the aerosol
1021 optical regimes ranged from 60% to 75%, depending on the regime, with an average
1022 accuracy of 70%. Notably, the accuracy exceeded 90% when predicting solely dust or
1023 non-dust optical regimes.

1024 An analysis of MERRA-2 simulations alongside this study's AERONET cluster-based
1025 prescriptions of optical regime indicators, such as absorption (SSA) and size (AE), reveals
1026 that MERRA-2 is generally more absorbing for the aerosol optical regimes (C0, C1, C3, and
1027 C4) impacting the atmosphere of the Iberian Peninsula, except for the most absorbing
1028 regime (C2). Specifically, the reanalysis simulations indicate higher absorption under the
1029 non-dust regimes C3 and C4. When examining the relative contributions of fine and coarse
1030 modes, the cluster-based prescription indicates a larger average contribution of coarse
1031 particles than the MERRA-2 under dust regimes (C0, C1). Conversely, for the non-dust
1032 regimes (C2, C3, C4), MERRA-2 shows a lower relative contribution from the fine mode
1033 compared to the clusters-based prescription.

1034 Our findings contribute to enhancing the understanding of the dynamic aerosol optical
1035 properties over the Iberian Peninsula and highlight the potential of machine-learning
1036 approaches to improve the representation of aerosol radiative forcing in atmospheric
1037 models. Many atmospheric modelling systems are not designed to simulate
1038 aerosol-intensive microphysical and optical properties in real time. Additionally,
1039 computational cost remains a common limitation worldwide. Our approach integrates
1040 AERONET-derived intensive properties based on climatological optical regimes to refine the
1041 model, coupled with predicted aerosol-type columnar mass density. This integration can
1042 help reduce regional uncertainty in the simulation of aerosol radiative forcing.

1043 Nevertheless, we acknowledge that additional research and analysis are necessary to build
1044 on the developments and findings presented here. Among the potential limitations and
1045 directions for future work, we emphasize the importance of better understanding the
1046 impact of AERONET parameter uncertainties on the clustering process, as well as
1047 conducting an intercomparison between basic and more advanced clustering approaches. A
1048 natural extension of this study would be the development of a comprehensive investigation
1049 focused on radiative transfer calculations, within which the proposed method could be
1050 thoroughly evaluated.

1051

1052 **Competing interests**

1053 The authors declare that they have no conflict of interest.

1054 **Acknowledgements and financial support**

1055 The authors acknowledge the financial support of FCT—Science and Technology Portuguese
1056 Foundation, which funded the project FIRESMOKE
1057 (<http://doi.org/10.54499/PTDC/CTA-MET/3392/2020>) through national funds. Thanks
1058 are also owed to the financial support given to CESAM by FCT (UID Centro de Estudos do

1059 Ambiente e Mar (CESAM) + LA/P/0094/2020) through national funds. We also
1060 acknowledge the financial support of CNPq - National Council for Scientific and
1061 Technological Development (CNPq) through the funding processes CNPq N° 441851/2023-1
1062 and CNPq N° 172486/2023-8. Author HFCV also thanks to the CNPq grant No
1063 315349/2023-9. We thank AERONET and MERRA-2 PIs and teams for their effort in
1064 establishing and maintaining the sites and the reanalysis development used in this study.
1065 We acknowledge the use of imagery from the Worldview Snapshots application
1066 (<https://wvs.earthdata.nasa.gov>), part of the Earth Observing System Data and Information
1067 System (EOSDIS).

1068

1069 **Author contributions**

1070 NR, KL and PT designed and performed the research, analyzed the data, and wrote the first
1071 version of the paper. MY, SF, LF, OM, HFCV contributed to writing, discussion, review and
1072 editing. ICM and AIM conceptualization and coordination of the Project FIRESMOKE,
1073 discussion, review and editing.

1074 **Code and data availability.**

1075 All the datasets (AERONET and MERRA-2) used in this study are publicly available and were
1076 downloaded from their respective websites (<https://aeronet.gsfc.nasa.gov/>; and
1077 <https://disc.gsfc.nasa.gov/datasets?project=MERRA-2>). Code and dataset required to
1078 conduct the analyses herein is available at <https://doi.org/10.5281/zenodo.15178347>
1079 (Rosario, 2025).

1080 **References**

1081 Abraham, A, F Pedregosa, M Eickenberg, P Gervais, A Mueller, J Kossaifi, A Gramfort, B
1082 Thirion, and G Varoquaux. 2014. "Machine Learning for Neuroimaging with Scikit-Learn."
1083 *Front Neuroinform* 8: 14.

1084 Adebiyi, A.A., Huang, Y., Samset, B.H. et al. Observations suggest that North African dust
1085 absorbs less solar radiation than models estimate. *Commun Earth Environ* 4, 168 (2023).
1086 <https://doi.org/10.1038/s43247-023-00825-2>.

1087 Alvarez, Albert, Judit Lecina-Diaz, Enric Batllori, Andrea Duane, Lluís Brotons, Javier
1088 Retana, Spatiotemporal patterns and drivers of extreme fire severity in Spain for the period
1089 1985–2018, *Agricultural and Forest Meteorology*, Volume 358, 2024, 110185, ISSN
1090 0168-1923, <https://doi.org/10.1016/j.agrformet.2024.110185>

1091 Asfaw, H. W., McGee, T. K., & Correia, F. J. (2022). Wildfire preparedness and response during
1092 the 2016 Arouca wildfires in rural Portugal. *International Journal of Disaster Risk
1093 Reduction*, 73, 102-895. <https://doi.org/10.1016/j.ijdrr.2022.102895>

1094 Breiman, Leo. 2001. "Random Forests". *Machine Learning* 45 (1): 5–32.
1095 <https://doi.org/10.1023/a:1010933404324>.

1096 Brown H, Liu X, Pokhrel R, Murphy S, Lu Z, Saleh R, Mielonen T, Kokkola H, Bergman T,
1097 Myhre G, Skeie RB, Watson-Paris D, Stier P, Johnson B, Bellouin N, Schulz M, Vakkari V,
1098 Beukes JP, van Zyl PG, Liu S, Chand D. Biomass burning aerosols in most climate models are
1099 too absorbing. *Nat Commun.* 2021 Jan 12;12(1):277. doi: 10.1038/s41467-020-20482-9.
1100 PMID: 33436592; PMCID: PMC7804930.

1101 Buchard-Marchant, V.J., C.A. Randles, A.M. da Silva, A. Darmenov, P.R. Colarco, R. Govindaraju,
1102 R.A. Ferrare, J. Hair, A. Beyersdorf, L.D. Ziemba, and H. Yu (2017), The MERRA-2 Aerosol
1103 Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies, *J. Climate*, 30, 6851-6872,
1104 doi:10.1175/JCLI-D-16-0613.1.

1105 Cachorro, V. E., Burgos, M. A., Mateos, D., Toledano, C., Bennouna, Y., Torres, B., de Frutos, Á.
1106 M., and Herguedas, Á.: Inventory of African desert dust events in the north-central Iberian
1107 Peninsula in 2003–2014 based on sun-photometer–AERONET and particulate-mass–EMEP
1108 data, *Atmos. Chem. Phys.*, 16, 8227–8248, <https://doi.org/10.5194/acp-16-8227-2016>,
1109 2016.

1110 Chen, G., Wang, J., Wang, Y., Wang, J., Jin, Y., Cheng, Y., et al. (2023). An aerosol optical module
1111 with observation-constrained black carbon properties for global climate models. *Journal of*
1112 *Advances in Modeling Earth Systems*, 15, e2022MS003501.
1113 <https://doi.org/10.1029/2022MS003501>

1114 Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., Martin, R. V., Logan, J. A.,
1115 Higurashi, A., and Nakajima, T.: Tropospheric aerosol optical thickness from the GOCART
1116 model and comparisons with satellite and sun photometer measurements, *J. Atmos. Sci.*, 59,
1117 461–483, [https://doi.org/10.1175/1520-0469\(2002\)059<0461:taotft>2.0.co;2](https://doi.org/10.1175/1520-0469(2002)059<0461:taotft>2.0.co;2), 2002.

1118 Colarco, P., Da Silva, A., Chin, M., and Diehl, T.: Online simulations of global aerosol
1119 distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based
1120 aerosol optical depth, *J. Geophys. Res.-Atmos.*, 115, D14207,620
1121 <https://doi.org/10.1029/2009JD012820>, 2010.

1122 Colarco, P. R., Nowottnick, E. P., Randles, C. A., Yi, B., Yang, P., Kim, K.-M., Smith, J. A., and
1123 Bardeen, C. G.: Impact of radiatively interactive dust aerosols in the NASA GEOS-5 climate
1124 model: Sensitivity to dust particle shape and refractive index, *J. Geophys. Res.-Atmos.*, 119,
1125 753–786, <https://doi.org/10.1002/2013JD020046>, 2014

1126 Dubovik, O., B. Holben, T. F. Eck, A. Smirnov, Y. J. Kaufman, M. D. King, D. Tanré, and I.
1127 Slutsker, 2002: Variability of Absorption and Optical Properties of Key Aerosol Types
1128 Observed in Worldwide Locations. *J. Atmos. Sci.*, 59, 590–608,
1129 [https://doi.org/10.1175/1520-0469\(2002\)059<0590:VOAAOP>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<0590:VOAAOP>2.0.CO;2).

1130 Eck, T. F., Holben, B. N., Reid, J. S., Dubovik, O., Smirnov, A., O'Neill, N. T., Slutsker, I., and
1131 Kinne, S.: Wavelength dependence of the optical depth of biomass burning, urban, and

1132 desert dust aerosols, *J. Geophys. Res.*, 104, 31333–31349, doi:10.1029/1999jd900923,
1133 1999.

1134 Elias, Thierry Ghislain, Ana Maria Silva, Maria João Figueira, Nuno Belo, Sergio Pereira, Paola
1135 Formenti, Gunter Helas, "Aerosol extinction and absorption in Évora, Portugal, during the
1136 European 2003 summer heat wave," *Proc. SPIE 5571, Remote Sensing of Clouds and the*
1137 *Atmosphere IX*, (30 November 2004); <https://doi.org/10.1117/12.566579>

1138 Ermitão, T.; Páscoa, P.; Trigo, I.; Alonso, C.; Gouveia, C. Mapping the Most Susceptible Regions
1139 to Fire in Portugal. *Fire* 2023, 6, 254. <https://doi.org/10.3390/fire6070254>

1140 Fan, Y. , X. Sun, H. Huang, R. Ti, X. Liu The primary aerosol models and distribution
1141 characteristics over China based on the AERONET data *J. Quant. Spectrosc. Ra.*, 275 (2021),
1142 10.1016/j.jqsrt.2021.107888

1143 Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and
1144 Applications, Version 2 (MERRA-2). *J. Climate*, 30, 5419–5454,
1145 <https://doi.org/10.1175/JCLI-D-16-0758.1>.

1146 Gómez-Amo, J. L., Estellés, V., Marcos, C., Segura, S., Esteve, A. R., Pedrós, R., Utrillas, M. P., and
1147 Martínez-Lozano, J. A.: Impact of dust and smoke mixing on column-integrated aerosol
1148 properties from observations during a severe wildfire episode over Valencia (Spain),
1149 *Science Total Environ.*, 599–600, 2121–2134,
1150 <https://doi.org/10.1016/j.scitotenv.2017.05.041>, 2017.

1151 Groß, S., Tesche, M., Freudenthaler, V., Toledano, C., Wiegner, M., Ansmann, A., Althausen, D.
1152 and Seefeldner, M. (2011) 'Characterization of Saharan dust, marine aerosols and mixtures
1153 of biomass-burning aerosols and dust by means of multi-wavelength depolarization and
1154 Raman lidar measurements during SAMUM 2', *Tellus B: Chemical and Physical Meteorology*,
1155 63(4), p. 706-724. Available at: <https://doi.org/10.1111/j.1600-0889.2011.00556.x>.

1156 Hammed, R.A.; Alawode, G.L.; Montoya, L.E.; Krasovskiy, A.; Kraxner, F. Exploring Drivers of
1157 Wildfires in Spain. *Land* 2024, 13, 762. <https://doi.org/10.3390/land13060762>

1158 Henok Workeye Asfaw, Tara K. McGee, Fernando Jorge Correia, Wildfire preparedness and
1159 response during the 2016 Arouca wildfires in rural Portugal, *International Journal of*
1160 *Disaster Risk Reduction*, Volume 73, 2022, 102895, ISSN 2212-4209,
1161 <https://doi.org/10.1016/j.ijdrr.2022.102895>.

1162 Hess, M., P. Koepke, and I. Schult, 1998: Optical properties of aerosols and clouds: The
1163 software package OPAC. *Bull. Amer. Meteor. Soc.*, 79, 831–844.

1164 Hoelzemann, J. J., Longo, K. M., Fonseca, R. M., do Rosario, N. M. E., Elbern, H., Freitas, S. R.,
1165 and Pires, C.: Regional representativity of AERONET observation sites during the biomass
1166 burning season in South America determined by correlation studies with MODIS Aerosol
1167 Optical Depth, *J. Geophys. Res.*, 114, D13301, doi:10.1029/2008jd010369, 2009

1168 Holben, B. N., Eck, T. F., Slutsker, I., Tanre, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A.,
1169 Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET- A
1170 Federated Instrument Network and Data Archive for Aerosol Characterization, *Remote Sens.*
1171 *Environ.*, 66, 1–16, doi:10.1016/s0034-4257(98)00031-5, 1998.

1172 Illingworth, A. J., and Coauthors, 2015: The EarthCARE Satellite: The Next Step Forward in
1173 Global Measurements of Clouds, Aerosols, Precipitation, and Radiation. *Bull. Amer. Meteor.*
1174 *Soc.*, 96, 1311–1332, <https://doi.org/10.1175/BAMS-D-12-00227.1>.

1175 IPCC, 2021: Climate Change 2021 - the Physical Science Basis, Contribution of Working
1176 Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change
1177 [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L.
1178 Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T.
1179 Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, In Press,
1180 Published: 9 August 2021.

1181 Kanitz, T., A. Ansmann, R. Engelmann, and D. Althausen, 2013: North-south cross sections of
1182 the vertical aerosol distribution over the Atlantic Ocean from multiwavelength
1183 Raman/polarization lidar during Polarstern cruises. *J. Geophys. Res. Atmos.*, 118,
1184 2643–2655, doi:10.1002/jgrd.50273.

1185 Kim, D. and Ramanathan, V. (2008) Solar Radiation Budget and Radiative Forcing Due to
1186 Aerosols and Clouds. *Journal of Geophysical Research: Atmospheres*, 113, D02203.
1187 <https://doi.org/10.1029/2007JD008434>

1188 Koepke, P., M. Hess, I. Schult, and E. P. Shettle (1997), Global aerosol data set, *Rep. 243*,
1189 Max-Planck-Inst. für Meteorol., Hamburg, Germany.

1190 Levy, R. C., Remer, L. A., Kleidman, R. G., Mattoo, S., Ichoku, C., Kahn, R., and Eck, T. F.: Global
1191 evaluation of the Collection 5 MODIS dark-target aerosol products over land, *Atmos. Chem.*
1192 *Phys.*, 10, 10399–10420, doi:10.5194/acp-10-10399-2010, 2010

1193 Levy, R. C., Remer, L. A., and Dubovik, O.: Global aerosol optical properties and application to
1194 Moderate Resolution Imaging Spectroradiometer aerosol retrieval over land, *J. Geophys.*
1195 *Res.-Atmos.*, 112, D13210, <https://doi.org/10.1029/2006JD007815>, 2007.

1196 Li, J., Carlson, B.E., Yung, Y.L. et al. Scattering and absorbing aerosols in the climate system.
1197 *Nat Rev Earth Environ* 3, 363–379 (2022). <https://doi.org/10.1038/s43017-022-00296-7>

1198 Li, J., L. Liu, A. A. Lacis, and B. E. Carlson (2010), An optimal fitting approach to improve the
1199 GISS ModelE aerosol optical property parameterization using AERONET data, *J. Geophys.*
1200 *Res.*, 115, D16211, doi:10.1029/2010JD013909.

1201 Li, Z.; Zhang, Y.; Xu, H.; Li, K.; Dubovik, O.; Goloub, P. The Fundamental Aerosol Models Over
1202 China Region: A Cluster Analysis of the Ground-Based Remote Sensing Measurements of
1203 Total Columnar Atmosphere. *Geophys. Res. Lett.* 2019, 46, 4924–4932

1204 Martins, J. V., Artaxo, P., Kaufman, Y. J., Castanho, A. D., and Remer, L. A.: Spectral absorption
1205 properties of aerosol particles from 350–2500 nm, *Geophys. Res. Lett.*, 36, L13810,
1206 <https://doi.org/10.1029/2009GL037435>, 2009.

1207 Moise, T., Flores, J. M., and Rudich, Y.: Optical properties of secondary organic aerosols and
1208 their changes by chemical processes, *Chem. Rev.*, 115, 4400–4439, 2015.

1209 Osborne, M., Malavelle, F. F., Adam, M., Buxmann, J., Sugier, J., Marengo, F., and Haywood, J.:
1210 Saharan dust and biomass burning aerosols during ex-hurricane Ophelia: observations from
1211 the new UK lidar and sun-photometer network, *Atmos. Chem. Phys.*, 19, 3557–3578,
1212 <https://doi.org/10.5194/acp-19-3557-2019>, 2019.

1213 Proske, U., Ferrachat, S., and Lohmann, U.: Developing a climatological simplification of
1214 aerosols to enter the cloud microphysics of a global climate model, *Atmos. Chem. Phys.*, 24,
1215 5907–5933, <https://doi.org/10.5194/acp-24-5907-2024>, 2024.

1216 Ramanathan, V., P. J. Crutzen, J. T. Kiehl, and D. Rosenfeld. 2001. “Aerosols, Climate, and the
1217 Hydrological Cycle”. *Science* 294 (5549). <https://doi.org/10.1126/science.1064034>.

1218 Reid, J. S. and Hobbs, P. V.: Physical and optical properties of smoke from individual biomass
1219 fires in Brazil, *J. Geophys. Res.*, 103, 32 013–32 031, 1998

1220 Reid, J. S., Eck, T. F., Christopher, S. A., Koppmann, R., Dubovik, O., Eleuterio, D. P., Holben, B.
1221 N., Reid, E. A., and Zhang, J.: A review of biomass burning emissions part III: intensive optical
1222 properties of biomass burning particles, *Atmos. Chem. Phys.*, 5, 827–849,
1223 <https://doi.org/10.5194/acp-5-827-2005>, 2005.

1224 Rodríguez, S. and López-Darias, J.: Extreme Saharan dust events expand northward over the
1225 Atlantic and Europe, prompting record-breaking PM₁₀ and PM_{2.5} episodes, *Atmos. Chem.*
1226 *Phys.*, 24, 12031–12053, <https://doi.org/10.5194/acp-24-12031-2024>, 2024.

1227 Rosario, N. E.: Machine learning-driven characterization and prescription of aerosol optical
1228 properties for atmospheric models, Zenodo [code],
1229 <https://doi.org/10.5281/zenodo.14825197>, 2025.

1230 Rosário, N. E., Longo, K. M., Freitas, S. R., Yamasoe, M. A., and Fonseca, R. M.: Modeling the
1231 South American regional smoke plume: aerosol optical depth variability and surface
1232 shortwave flux perturbation, *Atmos. Chem. Phys.*, 13, 2923–2938,
1233 <https://doi.org/10.5194/acp-13-2923-2013>, 2013.

1234 Russell, P. B., Kacenelenbogen, M., Livingston, J. M., Hasekamp, O. P., Burton, S. P., Schuster, G.
1235 L., Johnson, M. S., Knobelspiesse, K. D., Redemann, J., Ramachandran, S., and Holben, B.: A
1236 multiparameter aerosol classification method and its application to retrievals from
1237 spaceborne polarimetry, *J. Geophys. Res.-Atmos.*, 119, 9838–9863,
1238 <https://doi.org/10.1002/2013JD021411>, 2014

1239 Saleh, R., Robinson, E. S., Tkacik, D. S., Ahern, A. T., Liu, S., Aiken, A. C., Sullivan, R. C., Presto,
1240 A. A., Dubey, M. K., Yokelson, R. J., Donahue, N. M., & Robinson, A. L. (2014). Brownness of
1241 organics in aerosols from biomass burning linked to their black carbon content. *Nature*
1242 *Geoscience*, 7(9), 647-650. <https://doi.org/10.1038/ngeo2220>

1243 Samset, B.H., Stjern, C.W., Andrews, E. et al. Aerosol Absorption: Progress Towards Global
1244 and Regional Constraints. *Curr Clim Change Rep* 4, 65–83 (2018).
1245 <https://doi.org/10.1007/s40641-018-0091-4>

1246 Sand, M., Samset, B. H., Myhre, G., Gliß, J., Bauer, S. E., Bian, H., Chin, M., Checa-Garcia, R.,
1247 Ginoux, P., Kipling, Z., Kirkevåg, A., Kokkola, H., Le Sager, P., Lund, M. T., Matsui, H., van Noije,
1248 T., Olivie, D. J. L., Remy, S., Schulz, M., Stier, P., Stjern, C. W., Takemura, T., Tsigaridis, K., Tsyro,
1249 S. G., and Watson-Parris, D.: Aerosol absorption in global models from AeroCom phase III,
1250 *Atmos. Chem. Phys.*, 21, 15929–15947, <https://doi.org/10.5194/acp-21-15929-2021>, 2021.

1251 Schwink SK, Mael LE, Dunnington TH, Schmid MJ, Silberstein JM, Heck A, Gotlib N, Hannigan
1252 MP, Vance ME. Impacts of Aging and Relative Humidity on Properties of Biomass Burning
1253 Smoke Particles. *ACS EST Air*. 2024 Dec 6;2(1):109-118. doi: 10.1021/acsestair.4c00224.
1254 PMID: 39817254; PMCID: PMC11730893.

1255 Shettle, E. P. and Fenn, R. W.: Models for the Aerosols of the Lower Atmosphere and the
1256 Effects of Humidity Variations on Their Optical Properties, AFGL-TR-79-0214, 94, 1979

1257 Shi, C., Wei, B., Wei, S. et al. A quantitative discriminant method of elbow point for the
1258 optimal number of clusters in clustering algorithm. *J Wireless Com Network* 2021, 31
1259 (2021). <https://doi.org/10.1186/s13638-021-01910-w>

1260 Shin, S.-K., Tesche, M., Kim, K., Kezoudi, M., Tatarov, B., Müller, D., and Noh, Y.: On the spectral
1261 depolarisation and lidar ratio of mineral dust provided in the AERONET version 3 inversion
1262 product, *Atmos. Chem. Phys.*, 18, 12735–12746,
1263 <https://doi.org/10.5194/acp-18-12735-2018>, 2018.

1264 Silva, P.; Carmo, M.; Rio, J.; Novo, I. Changes in the Seasonality of Fire Activity and Fire
1265 Weather in Portugal: Is the Wildfire Season Really Longer? *Meteorology* 2023, 2, 74-86.
1266 <https://doi.org/10.3390/meteorology2010006>

1267 Sinyuk, A., Holben, B. N., Eck, T. F., Giles, D. M., Slutsker, I., Korkin, S., Schafer, J. S., Smirnov, A.,
1268 Sorokin, M., and Lyapustin, A.: The AERONET Version 3 aerosol retrieval algorithm,
1269 associated uncertainties and comparisons to Version 2, *Atmos. Meas. Tech.*, 13, 3375–3411,
1270 <https://doi.org/10.5194/amt-13-3375-2020>, 2020.

1271 Smirnov, A., B. N. Holben, Y. J. Kaufman, O. Dubovik, T. F. Eck, I. Slutsker, C. Pietras, and R. N.
1272 Halthore, 2002: Optical Properties of Atmospheric Aerosol in Maritime Environments. *J.*
1273 *Atmos. Sci.*, 59, 501–523,
1274 [https://doi.org/10.1175/1520-0469\(2002\)059<0501:OPOAAI>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<0501:OPOAAI>2.0.CO;2).

- 1275 Spencer, RS, RC Levy, LA Remer, S Mattoo, GT Arnold, DL Hlavka, KG Meyer, A Marshak, EM
1276 Wilcox, and SE Platnick. 2019. "Exploring Aerosols near Clouds with
1277 High-Spatial-Resolution Aircraft Remote Sensing during SEAC(4)RS". *J Geophys Res Atmos*
1278 124: 2148–73.
- 1279 Toledano, C., Cachorro, V. E., de Frutos, A. M., Sorribas, M., and Prats, N.: Inventory of African
1280 Desert Dust Events Over the Southwestern Iberian Peninsula in 2000–2005 with an
1281 AERONET Cimel Sun Photometer, *J. Geophys. Res.*, 112, D21201,
1282 doi:10.1029/2006JD008307, 2007
- 1283 Zhao, G., Tan, T., Zhao, W., Guo, S., Tian, P., and Zhao, C.: A new parameterization scheme for
1284 the real part of the ambient urban aerosol refractive index, *Atmos. Chem. Phys.*, 19,
1285 12875–12885, <https://doi.org/10.5194/acp-19-12875-2019>, 2019.
1286 <https://acp.copernicus.org/articles/19/12875/2019/>
- 1287 Zhong Q, Schutgens N, van der Werf GR, van Noije T, Bauer SE, Tsigaridis K, Mielonen T,
1288 Checa-Garcia R, Neubauer D, Kipling Z, Kirkevåg A, Olivie DJL, Kokkola H, Matsui H, Ginoux P,
1289 Takemura T, Le Sager P, Rémy S, Bian H, Chin M. Using modelled relationships and satellite
1290 observations to attribute modelled aerosol biases over biomass burning regions. *Nat*
1291 *Commun.* 2022 Oct 7;13(1):5914. doi: 10.1038/s41467-022-33680-4. PMID: 36207322;
1292 PMCID: PMC9547058.
- 1293 Zhou, P.; Wang, Y.; Liu, J.; Xu, L.; Chen, X.; Zhang, L. Difference between global and regional
1294 aerosol model classifications and associated implications for spaceborne aerosol optical
1295 depth retrieval. *Atmos. Environ.* **2023**, *300*, 119674.