

# Toward More Representative Gridded Satellite Products

Adam C. Povey<sup>✉</sup> and Roy G. Grainger

**Abstract**—The most widely used satellite products are averages of data onto a regular spatiotemporal grid, known as Level 3 data. Some atmospheric variables can vary rapidly in response to changing conditions. Over the scales of Level 3 averaging, the combination of observations across different conditions may result in data that is not normally distributed, such that a simple mean is not representative. The problem is illustrated by the distribution of aerosol optical depth from different sensors and algorithms. A simple statistical technique is proposed to better convey the diversity of satellite observations to users whereby a multimodal log-normal distribution is fit to the distribution of data observed within each grid cell. Allowing multiple modes within each cell is shown to improve the agreement between satellite products by highlighting regions of significant variability and isolating systematic differences between instruments.

**Index Terms**—Aerosol, Level 3 satellite data, parametric statistics.

## I. INTRODUCTION

SATELLITE observations are fundamental to our understanding of the composition and dynamics of the atmosphere. Though not without error [1], they provide decades of global coverage necessary to monitor the climate and the processes which influence it [2].

Spaced-based Earth-imaging data typically has a spatial resolution between 1 m and 10 km. An entire data set at such resolution requires terabytes to petabytes of data, which is a significant challenge to catalog and manipulate. To minimize that burden for the majority of users (who do not require full resolution data), most satellite data providers release “Level 3” products [3] which average the data over space and/or time onto a regular grid. Daily and monthly means over a  $1^\circ \times 1^\circ$  latitude-longitude grid are common. Such products are the most popular means of interacting with satellite data as they are easy to store (requiring only megabits), display (being mapped onto a regular, georeferenced grid), and compare to both other data and model outputs.

All manner of statistics are reported to summarize the aggregated data. The most common is an unweighted mean,

Manuscript received February 8, 2018; revised July 21, 2018 and August 17, 2018; accepted November 13, 2018. Date of publication December 12, 2018; date of current version April 22, 2019. This work was supported by the Natural Environment Research Council through the National Center for Earth Observation under Contract PR140015. (Corresponding author: Adam C. Povey.)

The authors are with the National Centre for Earth Observation, University of Oxford, Oxford OX1 3PU, U.K. (e-mail: adam.povey@physics.ox.ac.uk). Color versions of one or more of the figures in this letter are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/LGRS.2018.2881762

the standard deviation, and the number of datum. That presentation implies that the data conformed, at least approximately, to a normal distribution. That would be true for independent observations (by the central limit theorem), but satellite data often are not independent. For example:

- 1) observations are likely correlated because the atmosphere is spatially correlated;
- 2) conversely, a set of measurements may sample different populations such as:
  - a) an area may contain land and sea, which reflect and emit radiation differently;
  - b) clouds form and evolve over minutes, such that successive orbits or scans observe different conditions;
  - c) aerosol plumes can exhibit sharp gradients in particle concentration over tens of meters.
- 3) satellites do not sample the atmosphere randomly. Error sources are not randomly distributed in space (e.g., clear sky conditions may systematically differ from cloudy sky [4]).

Level 3 products can be accurate for data that are spatiotemporally well correlated, such as the concentration of a well-mixed gas, or where the underlying physics are well understood, such as sea surface temperature. However, for many atmospheric variables, the diversity of natural behavior cannot be conveyed by a simple mean over an arbitrary box.

This letter demonstrates the problem of representivity with examples from aerosol remote sensing but should be applicable to many satellite observations of Earth. A simple solution is proposed to fit a curve to the distribution of data within each grid cell, reporting multiple values for each grid cell to represent their multimodal distribution. This is then shown to improve the agreement between independent satellite products.

## II. REPRESENTIVITY

Fig. 1(a) shows the distribution of aerosol optical depth (AOD) observed in a  $5^\circ \times 5^\circ$  cell off the South American coast during January 2008 by the MODerate-resolution Imaging Spectrometers (MODIS, [5]) on the Aqua and Terra platforms. Fig. 1(b) shows, for a different cell, AOD retrieved from the Advanced Along-Track Scanning Radiometer (AATSR, [6]) by two algorithms: the Optimal Retrieval of Aerosol and Cloud (ORAC, [7]) and Swansea University (SU, [8]). A log scale is used as AOD is expected to be log-normally distributed [9]. The mean and standard deviation of each data set is shown below the axis.

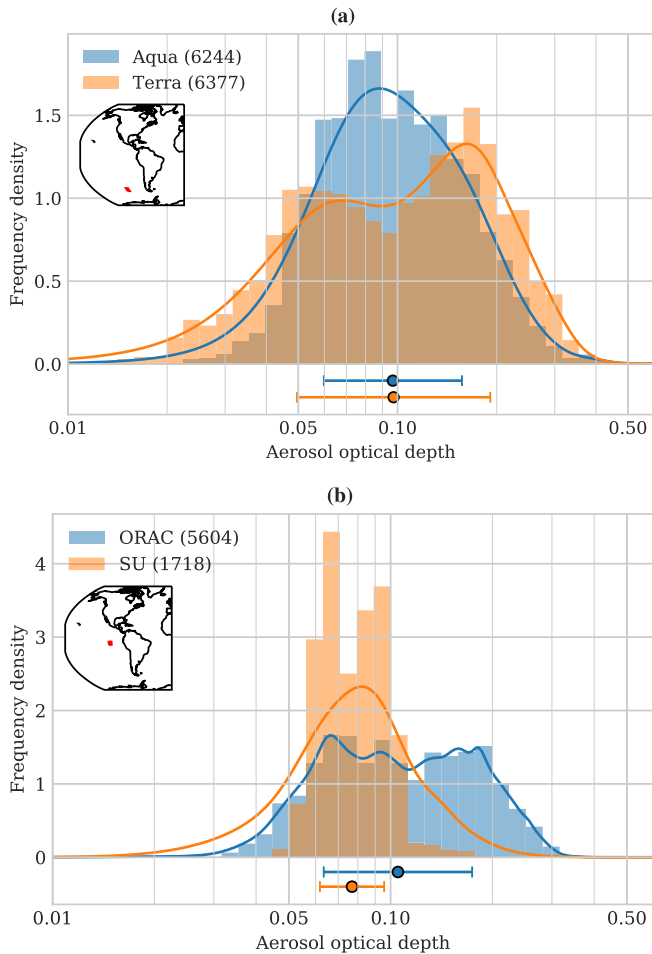


Fig. 1. Histograms of AOD (bars) during January 2008 reported by (a) MODIS Aqua (blue) and Terra (orange) and (b) AATSR from the ORAC (blue) and SU (orange) algorithms. The location of the  $5^\circ \times 5^\circ$  bin is shown in the inset. Lines: equivalent distributions obtained by kernel density estimation. The mean and standard deviation of each data set is presented below the axis with a dot-and-whisker while the total number of observations is given in the legend.

Histograms are not the only means of generating a distribution and can be a poor representation of data that suffers errors of a similar magnitude to the size of the bins (as AOD does). Kernel density estimation [10] represents each datum with a kernel (i.e., a curve), the selection of which is a nontrivial problem. For satellite data with uncertainties provided, it seems reasonable to use a Gaussian with mean equal to the measurement and width equal to the uncertainty. The lines in Fig. 1 are such estimates. They closely follow the histogram while being less prone to sharp peaks (which may have resulted from the chance alignment of bins with the data).

Considering Fig. 1(a), the two sensors appear to have observed different environments in a manner not communicated by their means despite being retrieved with the same algorithm. Aqua saw a single, approximately log-normal, distribution consistent with the Level 3 data meanwhile Terra saw a bimodal distribution with peaks at 0.07 and 0.17. In Fig. 1(b), different algorithms evaluating the same data produce inconsistent means. Considering their histograms,

ORAC observed a roughly bimodal distribution, from which SU only reported the lesser mode.

Such differences are not necessarily a cause for concern: Aqua and Terra have different overpass times while ORAC and SU produced different volumes of data. Differences in filtering, especially of cloud, will result in different distributions as the instruments will take systematically different samples. In particular, monthly averages (as shown here) unevenly weight different days depending on their cloud cover, potentially oversampling a particular aerosol source or region. [This likely caused the difference in Fig. 1(b).]

Regardless, simple Level 3 data may give users a poor understanding of the underlying observations. The distributions of AOD shown are bimodal. Communicating that with a single mean and standard deviation overemphasizes the mean value, which may not be particularly common, and underemphasizes the tails of the distribution. The full distribution can represent multiple regimes of aerosol observed within the grid cell. These could be multiple sources (e.g., in Fig. 1, one mode may represent sea salt aerosols while the other represents particles advected from the land), changing conditions (e.g., precipitation during the month would reduce AOD through wet deposition) or different air masses (e.g., changing wind direction can sample different sources).

These problems are not unknown to the satellite remote sensing community. The MODIS [11] and Multiangle Imaging SpectroRadiometer [12]) Level 3 products contain a variety of histograms but, in the authors' experience, these are rarely used due to the large file size (e.g., one day from [12] is 800 Mb while their  $1^\circ$  means require only 6 Mb). Atmospheric states will be correlated over some distance and time, such that their distributions should be fairly elementary. Thus, the generation of Level 3 data can be thought of as a problem in parametric statistics: to select a functional form that models the distribution of a quantity and then estimate the parameters of that function given a set of observations. Reporting such parameters was proposed in [13], but not implemented.

### III. FITTING METHOD

A simple method is introduced to illustrate the idea. The probability density function of AOD is modeled by either a unimodal, bimodal, or trimodal, log-normal distribution

$$P_1(\tau) = \mathcal{L}(\tau; \bar{\tau}_1, \sigma_1) \quad (1)$$

$$P_2(\tau) = r_2 \mathcal{L}(\tau; \bar{\tau}_2, \sigma_2) + (1 - r_2) P_1(\tau) \quad (2)$$

$$P_3(\tau) = r_3 \mathcal{L}(\tau; \bar{\tau}_3, \sigma_3) + (1 - r_3) P_2(\tau) \quad (3)$$

where  $\tau$  is AOD,  $\mathcal{L}(\tau; \bar{\tau}, \sigma_i)$  is a log-normal distribution with mean  $\bar{\tau}$  and standard deviation  $\sigma$ , and  $r_i$  is the aspect ratio of the  $i$ th mode (i.e., it is magnitude relative to the sum of all previous modes).

This is fit to a histogram using the Python `scipy.interpolate.curve_fit` routine, bounded such that  $0.0032 < \bar{\tau} < 10$ ,  $0.01 < \sigma < 0.5$ , and  $0.01 < r < 0.99$ . The first guess for all bins is  $\bar{\tau}_1 = 0.1$ ,  $\bar{\tau}_2 = 1.0$ ,  $\bar{\tau}_3 = 0.032$ ,  $\sigma_{1,2,3} = 0.1$ , and  $r_{2,3} = 0.5$ .

The fit distribution is then checked for redundant modes to avoid overfitting the data. Using the Holmann test [14], a pair

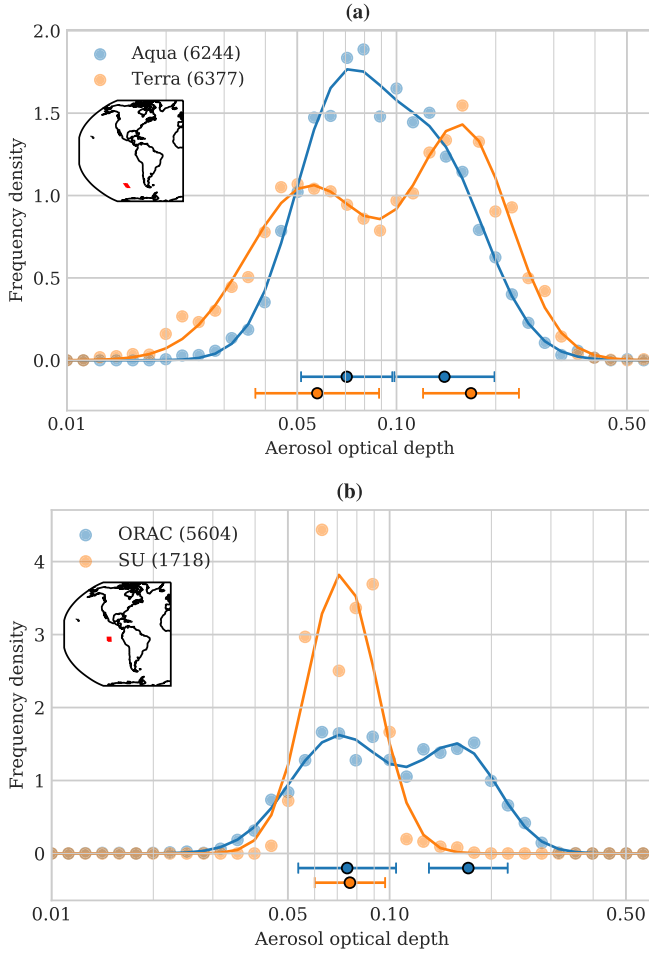


Fig. 2. Fits of unimodal or bimodal log-normal distributions (lines) to the histograms of Fig. 1 (circles). The mean and width of each fitted mode is shown below the axis with dot-and-whiskers.

of modes are independent if  $d > 1$  and

$$\frac{1}{2} |\log(1-r) - \log r| < \log(d - \sqrt{d^2 - 1}) + d\sqrt{d^2 - 1} \quad (4)$$

where  $d = |\bar{\tau}_A - \bar{\tau}_B|/2\sqrt{\sigma_A\sigma_B}$  and  $r$  is the ratio of the mode magnitudes.

If any of the three pairs of modes fail this test, the fit is rejected and a bimodal distribution is fit instead (i.e.,  $r_3 = 0$ ). The bounds are adjusted to  $0.05 < \sigma < 0.5$  and  $0.2 < r < 0.8$  to avoid overfitting narrow or shallow peaks and the first guess is changed to  $\bar{\tau}_2 = 0.01$ . If this pair of modes fails (4), the fit is again rejected and a unimodal distribution is fit instead (i.e.,  $r_2 = r_3 = 0$ ). Fitting is only attempted for histograms with at least five nonzero bins.

It should be emphasized that the distributions shown were selected as illustrations and should not be taken to be representative of general behavior. The parameters of this fitting scheme were selected manually to achieve qualitatively satisfying results and are not intended for general use. Specifically, different bounds were used to allow trimodal distributions in the Saharan dust outflow while discouraging bimodal distributions with a low aspect ratio. The popularity of the latter indicates a log-normal distribution may not be ideal.

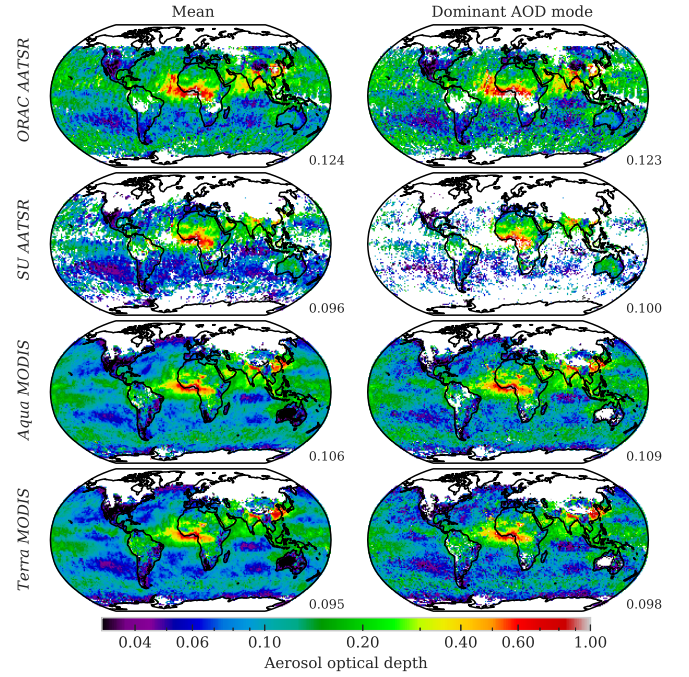


Fig. 3. (Left) Mean AOD and (Right) AOD of the fit mode with the greatest frequency for January 2008 on a  $1^\circ$  latitude-longitude grid. From the top, the rows show AATSR from the ORAC algorithm, AATSR from the SU algorithm, Aqua MODIS, and Terra MODIS. The global mean of each field is given below it.

Other physically sensible statistical distributions could easily be used, especially for satellite products other than AOD, by replacing the  $\mathcal{L}$  terms in (1)-(3). These may require more parameters.

#### IV. RESULTS

Fig. 2 applies this method to the data from Fig. 1. For Fig. 2(a), both sensors required two modes to produce a reasonable agreement with the histograms. The means of these modes are consistent (0.071 and 0.140 for Aqua vs. 0.057 and 0.168 for Terra). The remaining differences may be due to calibration differences (or diurnal variations in AOD). In Fig. 2(b), the AOD of the lower ORAC mode is very similar to the single SU mode (0.075 vs. 0.076).

The fitting method is then applied globally to the four data sets introduced, presented in Fig. 3. The left column shows the simple mean from each cell while the right column shows  $\bar{\tau}$  for the most likely mode (i.e., that with the largest cumulative aspect ratio). The fitted fields are similar to the mean but with an improved agreement between the data sets: the AOD minima west of South America and south of India converge to 0.04 across the data sets. Coverage is lost in the SU field where distributions are exceedingly narrow because only a few days of observations were available; a lower resolution grid may be advisable. Speckle in the fit fields may be reduced by a more carefully designed fitting routine (e.g., making an educated first guess based on surrounding data or applying more stringent convergence criteria).

Fig. 4 shows the AOD of the other two modes. About 45% of pixels fit a second peak. The second mode addresses some

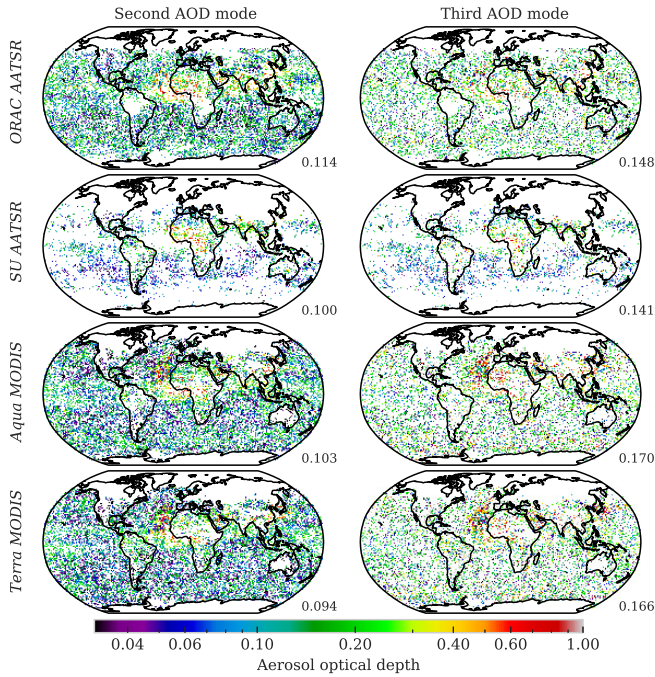


Fig. 4. AOD of the (Left) second and (Right) third most likely fitted modes, following the format of Fig. 3.

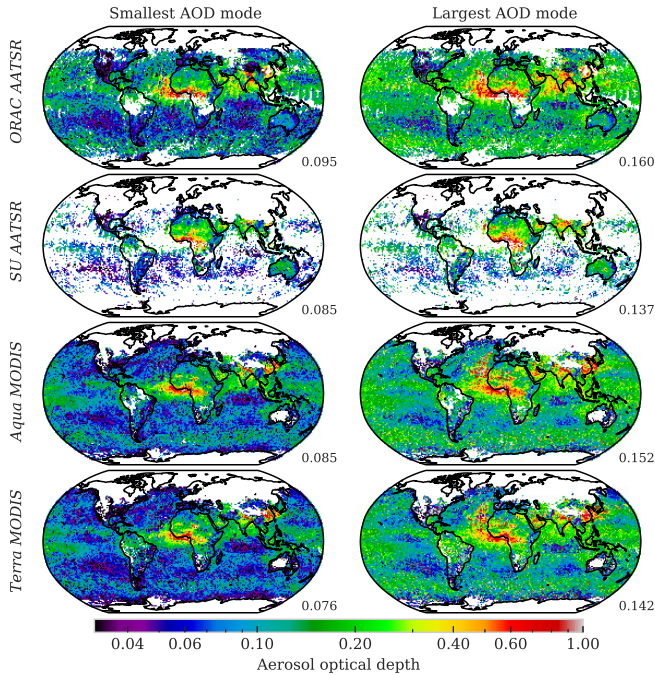


Fig. 5. (Left) Smallest and (Right) largest fitted AOD, following the format of Fig. 3.

of the discrepancies between the ORAC and MODIS fields, resolving the AOD minima west of California and lower AODs around  $60^\circ\text{S}$ . High AODs are now captured over Africa and off the coast of Morocco, where dust is more intermittent. Almost 40% of SU pixels fit three modes while the other data sets required one in only 30% of cases, concentrated in dusty and heavily polluted regions. There is a little spatial

pattern to these fits and their aspect ratio is usually  $< 0.1$ , indicating that a bimodal distribution should be sufficient in many circumstances.

The utility of the fitting is easier to see in Fig. 5, showing the smallest and largest AOD fit to each distribution. The bimodal nature of AOD over much of the ocean is clear, with an ambient AOD of 0.05–0.10 and perturbed AODs of 0.1–0.2. In the Saharan dust outflow, two regimes are evident: an area of consistently high AOD along the Gulf of Guinea and another off the west African coast where high AODs mix with ambient Atlantic aerosol. If, for example, evaluating the representation of dust in a global circulation model, one could consider two runs of the model: one with dust turned off compared to the smaller AOD mode and a second run with dust on compared to the larger AOD mode. This would test how accurately both the dust and ambient aerosols are represented, which would be missed when evaluating against the mean. Alternatively, the fitted modes could be used as the basis of a scene classification scheme, with which to assess differences between collocated measurements.

## V. CONCLUSION

Level 3 products average satellite data onto a regular spatiotemporal grid and are widely used because of their small size and ease of use. This is practical for data that are well correlated in space and time but some atmospheric variables can vary rapidly in response to changing conditions. Over a month or hundreds of kilometers, the combination of observations across different conditions may not be normally distributed, such that a simple mean does not represent the diversity of behavior actually witnessed.

A more meaningful data set can be obtained by using parametric statistics to model the distribution of data. This was demonstrated on AOD observations but is applicable to a range of atmospheric variables. A trimodal, log-normal distribution was fit to monthly observations of AOD, though it was found that a bimodal distribution would have sufficed in many circumstances. This provides the user with mean/standard deviation pairs for each mode (and their aspect ratios). For a simple analysis, a user could simply consider the mean value of the fitted mode with the largest magnitude. Arguably, that is what many users desire from Level 3 data: a value that represents the most common properties observed over some spatiotemporal region. A more detailed analysis could investigate the processes associated with each observed mode of variation, such as different sources or changing meteorological conditions, to understand why data sets differ.

The method presented is intended as a proof of concept rather than a robust technique. Operationally, it would be advisable to apply more constraints to the fitting routine and use a more objective selection of bins for the histogram (or instead fit kernel estimates). Uncertainty in the histogram could also be considered while fitting. If evaluating a variable other than AOD, a physically appropriate probability function would need to be substituted for the log-normal functions used here. In addition, the method shown required only 10% more computational time than generating typical Level 3 data and the additional data is not ubiquitous—only 45% of pixels

fit more than one AOD mode, concentrating the additional information where it is needed to thoroughly understand what the satellite observed.

#### ACKNOWLEDGMENT

The authors would like to thank A. Sayer, for several illuminating conversations at the 2017 AeroSAT meeting, A. Mason, for her numerous useful comments on this letter, and two anonymous reviewers, for their invaluable feedback.

AATSR retrievals were generated for the European Space Agency's Climate Change Initiative and are freely available from <http://esa.cci.int>. Collection 6.1 MODIS aerosol retrievals are available at <https://ladsweb.nascom.nasa.gov>. Figures were generated using matplotlib, seaborn, and cartopy.

#### REFERENCES

- [1] A. C. Povey and R. G. Grainger, "Known and unknown unknowns: Uncertainty estimation in satellite remote sensing," *Atmos. Meas. Techn.*, vol. 8, no. 11, pp. 4699–4718, 2015, doi: [10.5194/amt-8-4699-2015](https://doi.org/10.5194/amt-8-4699-2015).
- [2] C. B. Field *et al.*, Eds., *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. New York, NY, USA: Cambridge Univ. Press, 2014, pp. 1–32.
- [3] R. R. Chase, "Data and information system," Rep. Earth Observing Syst., NASA, Washington, DC, USE, NASA Tech. Memorandum 87777, 1986, vol. 2A. [Online]. Available: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19860021622.pdf>
- [4] A. M. Sayer, G. E. Thomas, P. I. Palmer, and R. G. Grainger, "Some implications of sampling choices on comparisons between satellite and model aerosol optical depth fields," *Atmos. Chem. Phys.*, vol. 10, no. 22, pp. 10705–10716, 2010, doi: [10.5194/acp-10-10705-2010](https://doi.org/10.5194/acp-10-10705-2010).
- [5] R. C. Levy *et al.*, "The collection 6 MODIS aerosol products over land and ocean," *Atmos. Meas. Tech.*, vol. 6, no. 11, pp. 2989–3034, 2013, doi: [10.5194/amt-6-2989-2013](https://doi.org/10.5194/amt-6-2989-2013).
- [6] T. Popp *et al.*, "Development, production and evaluation of aerosol climate data records from European Satellite Observations (Aerosol\_cci)," *Remote Sens.*, vol. 8, no. 5, p. 421, 2016, doi: [10.3390/rs8050421](https://doi.org/10.3390/rs8050421).
- [7] G. E. Thomas *et al.*, "The GRAPE aerosol retrieval algorithm," *Atmos. Meas. Techn.*, vol. 2, no. 2, pp. 679–701, 2009, doi: [10.5194/amt-2-679-2009](https://doi.org/10.5194/amt-2-679-2009).
- [8] S. L. Bevan, P. R. J. North, S. O. Los, and W. M. F. Grey, "A global dataset of atmospheric aerosol optical depth and surface reflectance from AATSR," *Remote Sens. Environ.*, vol. 116, pp. 199–210, Jan. 2012, doi: [10.1016/j.rse.2011.05.024](https://doi.org/10.1016/j.rse.2011.05.024).
- [9] N. T. O'Neill, A. Ignatov, B. N. Holben, and T. F. Eck, "The lognormal distribution as a reference for reporting aerosol optical depth statistics; Empirical tests using multi-year, multi-site AERONET Sunphotometer data," *Geophys. Res. Lett.*, vol. 27, no. 20, pp. 3333–3336, 2000, doi: [10.1029/2000GL011581](https://doi.org/10.1029/2000GL011581).
- [10] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1994.
- [11] MODIS Atmosphere Science Team, *MYD08\_M3—MODIS/Aqua Aerosol Cloud Water Vapor Ozone Monthly L3 Global 1Deg CMG NASA Level 1 and Atmosphere Archive and Distribution System, Collection 6.1*, Goddard Space Flight Centre, Greenbelt, MD, USA, 2017, doi: [10.5067/MODIS/MOD08\\_M3.061](https://doi.org/10.5067/MODIS/MOD08_M3.061).
- [12] D. Diner, "MISR Level 3 Component Global Aerosol product in netCDF format covering a day—Version 4," NASA Langley Atmos. Sci. Data Center DAAC, Hampton, VA, USA, Tech. Rep. MIL3DAEN, 2009, doi: [10.5067/Terra/MISR/MIL3DAEN\\_L3.004](https://doi.org/10.5067/Terra/MISR/MIL3DAEN_L3.004).
- [13] P. Hubanks, S. Platnick, M. King, and B. Ridgway, "MODIS atmosphere L3 gridded product algorithm theoretical basis document (ATBD) & users guide," NASA Goddard Space Flight Center, Greenbelt, MD, USA, Tech. Rep. ATBD-MOD-30, 2016. [Online]. Available: [https://modis-atmosphere.gsfc.nasa.gov/sites/default/files/ModAtmo/L3\\_ATBD\\_C6\\_2018\\_04\\_11.pdf](https://modis-atmosphere.gsfc.nasa.gov/sites/default/files/ModAtmo/L3_ATBD_C6_2018_04_11.pdf)
- [14] H. Holzmann and S. Vollmer, "A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU," *ASTA Adv. Stat. Anal.*, vol. 92, no. 1, pp. 57–69, 2008, doi: [10.1007/s10182-008-0057-2](https://doi.org/10.1007/s10182-008-0057-2).