A neural-network-based forward model to improve air quality estimation from spaceborne polarimeters

Abhinay Dommalapati
School of Data Science
University of Virginia
Charlottesville, Virginia
ad4bu@virginia.edu

Anura Ranasinghe
School of Data Science
University of Virginia
Charlottesville, Virginia
sar2jf@virginia.edu

Jack Peele
School of Data Science
University of Virginia
Charlottesville, Virginia
jcp2jf@virginia.edu

Stephen Whetzel
School of Data Science
University of Virginia
Charlottesville, Virginia
sjw5ke@virginia.edu

Michael Jones
NASA Langley Research Center
Hampton, Virginia
michael.jones-4@nasa.gov

Adam Bell
NASA Langley Research Center
Hampton, Virginia
adam.d.bell@nasa.gov

Eduard Chemyakin
NASA Langley Research Center
Hampton, Virginia
eduard.v.chemyakin@nasa.gov

Snorre Stamnes
NASA Langley Research Center
Hampton, Virginia
snorre.a.stamnes@nasa.gov

Heman Shakeri
School of Data Science
University of Virginia
Charlottesville, Virginia
hs9hd@virginia.edu

***Abstract*—A growing awareness of the adverse effects of high concentrations of aerosol pollutants on human health [1] motivates the need to accurately measure and forecast the amount of PM2.5 in the air; that is the particulate matter of aerosol particles with size 2.5 microns or less in diameter [2]. Quantifying concentrations of aerosols, particularly near the surface, is foundational to the understanding of the sources, evolution, and transport of PM2.5 and will help to support environmental justice for communities across America and the world. Moreover, developing improved algorithms to accurately invert or retrieve surface-level PM2.5 from satellite remote sensing is critical to improve neighborhood-scale estimates of air quality [3]. In particular, past and future satellite polarimeter and lidar measurements will be key to understanding surface-level PM2.5 conditions in real-time across the globe.**

A current solution to the retrieval of accurate aerosol properties from satellite polarimeter measurements has been developed by NASA for the Plankton, Aerosols, Clouds and Ecosystems mission (PACE) mission in the form of the Microphysical Aerosol Properties from Polarimetry (PACE-MAPP) algorithm [4]. However, because solving the vector radiative transfer is numerically intensive, and solving the non-linear inverse problem requires an iterative approach that for multiple channels involves hundreds of vector radiative transfer calls, this approach delivers products at a rate that has latencies too large for and is prohibitively inefficient for the large-scale datasets that will be needed to resolve PM2.5 at neighborhood-scale resolutions of less than 1 km by 1 km. PACE-MAPP solves this problem by developing a neural network framework to replace the complex and time-consuming vector radiative transfer calculations at each iteration.

In this study, we apply the PACE-MAPP framework to polarimetry data gathered from the POLDER instrument (POLarization and Directionality of the Earth's Reflectances) [5] onboard PARASOL, a satellite that flew from 2006 to 2013 as a part of efforts to understand the effects of clouds and aerosols on the Earth's climate [6] [7], and demonstrate for the first time ever that a neural-network-based approach using coupled atmosphere-ocean vector radiative transfer can be applied to retrieve aerosol properties from satellite polarimeter data, and to take the first step toward evaluating the algorithm's performance at producing air quality products such as PM2.5. We further demonstrate the feasibility of deploying neural networks to solve the numerical inefficiencies that plague satellite polarimeter retrievals while maintaining high accuracy, and expect to cut the speed of acquisition by a factor of 1000.

I. Introduction

The ultimate goal of this project is to more accurately and efficiently retrieve surface-level PM2.5 from satellite remote sensing polarimeter and lidar measurements. PM2.5 refers to the mass concentration of any particles in the air that are two and a half microns or less in aerodynamic diameter [2]. These particulates can be produced by natural causes such as wildfires, volcanic emissions and wind-driven dust. They are also produced by the burning of fuels, such as exhaust from automobiles, factories, power plants, etc. Other notable sources include rubber particles emitted by vehicle tires, wind-driven coal dust from open coal storage facilities nearby residential areas, and emissions from factories. Regardless of the source, PM2.5 can be carried over long distances by wind and can be found hundreds of miles away from the source [8]. Because PM2.5 particles are so small in size, they can easily travel deep into the human respiratory tract. Such exposure affects lung function and worsens medical conditions like asthma and heart disease, and as such, poses significant health risks to humans.

In order to more efficiently estimate surface-level PM2.5 from polarimeter satellite measurements, we must first improve the speed of the forward model, which calculates the multi-wavelength and multi-angular polarimetry measurements from an input state vector. Then, we can solve the inverse problem more efficiently through iterative Bayesian approaches such as optimal estimation. It is well established that there exist highly accurate vector radiative transfer codes that can map input atmospheric and oceanic state vector parameters to polarimeter measurements [4]. However, vector radiative transfer models are computationally inefficient for large datasets and are thus unable to be used to solve the inverse problem at scale without significant HPC resources.

Therefore, we aim to apply a neural network to the forward problem that will significantly improve the speed at which polarimeter satellite measurements can be modeled. We will then use optimal estimation to iteratively update the state vector until the modeled polarimeter measurements converge to the real polarimeter measurements. Once we reach convergence, we can then use the aerosol parameters of the final state vector, namely the AOD and the aerosol size distribution parameters, ı.e. effective radius and effective variance, to calculate PM2.5.

The estimation of PM2.5 from PARASOL satellite data will be accomplished by adapting the PACE-MAPP algorithm created for the two polarimeters onboard NASA's prospective Plankton, Aerosol, Cloud, and ocean Ecosystem mission (PACE) [9]. Like PARASOL, these two polarimeter instruments onboard PACE, the Hyper-Angular Rainbow Polarimeter instrument (HARP2) and the Spectropolarimeter for Planetary EXploration one (SPEXone), measure the total and polarized reflectance from sunlight reflected off of the Earth's surface as it intercepts the instruments from multiple angles. The polarimeter measurements at multiple wavelengths can be used to retrieve important ocean and atmospheric properties such as AOD, aerosol complex refractive index, aerosol top height, ocean surface roughness, and chlorophyll-a concentration [4]. PACE-MAPP simultaneously solves for the ocean and atmospheric state parameters using multi-angular, multi-channel polarimeter measurements taken from these two polarimeters. The PACE-MAPP algorithm uses a coupled atmosphere-ocean radiative transfer model to simultaneously estimate atmosphere and ocean state parameters by iteratively fitting an optimal solution. Accounting for the radiative coupling of the atmosphere and ocean in the radiative transfer model allows for simultaneous parameter retrieval as well as greater accuracy and plausibility in the predicted state vectors.

However, the PACE-MAPP algorithm was not designed to perform retrievals on PARASOL polarimeter satellite data, which uses different wavelengths and a different data format. Furthermore, PACE-MAPP was not funded to produce or validate the resulting air quality products that can be generated from its retrieved aerosol properties, and was designed to retrieve aerosol products over ocean, and not over land, as needed for air quality applications. Lastly, PACE-MAPP was not designed to use lidar measurements to improve the estimation of surface-level PM2.5. In this paper, we complete the initial steps required to perform PARASOL-MAPP aerosol optical and microphysical retrievals using data from the PARASOL polarimeter and produce the PM2.5 air quality product for validation against ground-based sites. These first steps include preprocessing the PARSOL data to the PACE L1C format, and training an accurate and fast neural network that can be used to simulate the PARASOL polarimeter measurements, which is required by the MAPP optimal estimation framework.

Our approach to developing a neural network model for the PARASOL satellite polarimeter to solve this inefficiency involved first producing an appropriately scaled set of PARASOL training data. These data are simulated by randomly cycling through the distribution of input state vectors, including aerosol optical depth (AOD) and PM2.5 along with the other ocean and atmospheric parameters. NASA's vector radiative transfer model is used to generate the corresponding simulated polarimetric measurements as observed by PARASOL. These simulated data are used to train a neural network that solves the forward problem of generating polarimetry data from an input state vector, which then replaces the vector radiative transfer model that is used to model the PARASOL measurements and the Jacobian matrix of the measurements, which are used by the optimal estimation algorithm [4].

II. Data Description

One of the crucial parts of this project is to understand how polarimetry data are stored and accessed. The University of Lille stores large volumes of satellite polarimeter data through ICARE, one of the four data centers under AERIS, a French-based data infrastructure dedicated towards atmospheric dynamics, physics, and chemistry [10]. The data is stored in heirarchical data format (HDF) files. Each satellite orbit contains polarimetric measurements at multiple wavelengths imaged over a swath of longitudinal and latitudinal coordinates, referred to as pixels. Moreover, each pixel is observed from multiple viewing angles as the satellite moves over the surface of the Earth. Figures 1 and 2 illustrate this multi-angular aspect of the PARASOL polarimeter data, and show the total normalized radiance (intensity) and degree of linear polarization (DOLP), respectively, at multiple wavelengths as a function of viewing angle mapped onto a specific pixel.

In order to develop a neural network capable of fulfilling the tasks of the existing vector radiative transfer model, a robust set of training data must be developed from which the model can learn. Note that while the vector radiative transfer model is relatively slow, and will ultimately be replaced by the trained neural network for retrieval purposes, the vector radiative transfer model is instrumental in generating labeled data to train the neural network in a process detailed below.

Input data consisting of randomly generated input parameters or state vectors are passed to the vector radiative transfer

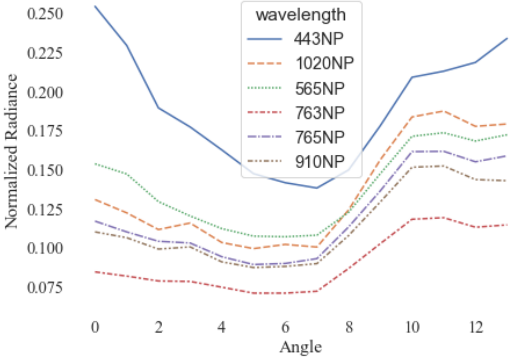

Fig. 1. Parasol normalized radiances (or intensities, I) for non-polarized wavelengths as a function of viewing angle

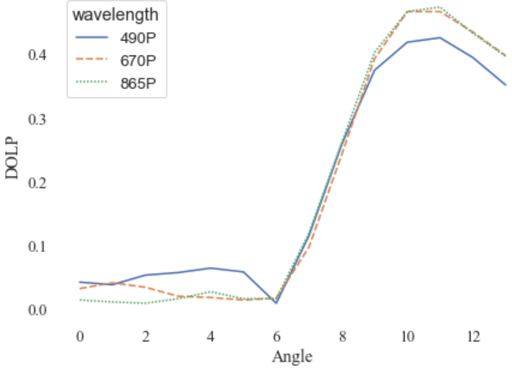

Fig. 2. Parasol degree of linear polarization (DOLP) as a function of viewing angle

model. The input parameters result in synthetic measurements that simulate real polarimeter measurements and vary across a wide range of realistic values to ensure a large, and widely applicable training set. The input parameters are then passed into the vector radiative transfer model, which outputs pixel-by-pixel polarimeter measurements of I and DOLP corresponding to the aerosol/ocean properties. The output of the vector radiative transfer model is stored in HDF5 format. The output of the vector radiative transfer model is then passed through a pre-processing script that prepares the data for training of the neural network. Pre-processing results in a series of 179 HDF files, each with 200,000 observations for nearly 36 million training observations.

III. Methodology

A. RSP-MAPP Algorithm

NASA originally developed the RSP-MAPP algorithm to simultaneously retrieve ocean and aerosol microphysical properties from the airborne NASA Research Scanning Polarimeter (RSP) that collects multi-wavelength, hyperangular total and polarized measurements radiance measurements [2]. The RSP-MAPP algorithm, and the PACE-MAPP algorithm which was adapted from it, use optimal estimation to iteratively fit a state vector of the ocean and aerosol optical and microphysical properties by comparing the output of the vector radiative transfer model, which maps a state vector to the intensity and DOLP polarimeter measurements. The state vector is iteratively updated through an optimal estimation procedure using the Jacobian matrix until convergence between the actual measurements and the modeled measurements is achieved by minimizing the residual with measurement uncertainty taken into account (Figure 3).

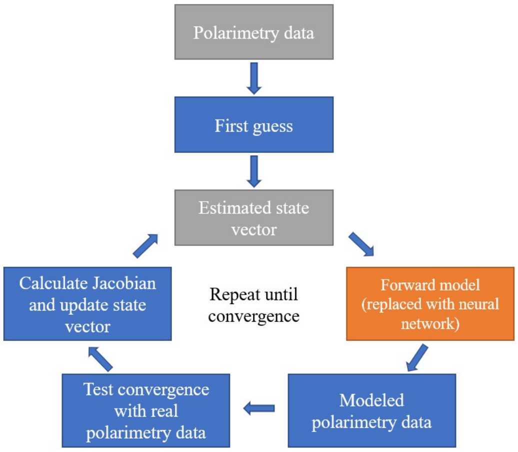
Polarimetry data
First guess
Estimated state vector
Calculate Jacobian and update state vector
Repeat until convergence
Forward model (replaced with neural network)
Test convergence with real polarimetry data
Modeled polarimetry data

Fig. 3. MAPP optimal estimation flowchart

The numerically-intensive vector radiative transfer calculations together with the iterative nature of this process means that the retrieval of the aerosol properties is very slow using the RSP-MAPP algorithm, on the order of 1 hour per retrieval per core. This acquisition latency rules out real-time or low-latency retrieval of the aerosol parameters required by air quality and weather forecasting algorithms. To achieve low-latency, the approach taken by PACE-MAPP is to generate simulated and labeled polarimetry data to train a neural network that replaces the radiative transfer model in the iterative optimal estimation process, leading to near real-time acquisitions. This approach is being pursued for the PACE mission, but PACE will not launch until 2023. Our research applies the PACE-MAPP approach to the PARASOL satellite to analyze and validate the air quality products that can be produced from the retrieved aerosol products. We expect this effort to demonstrate and eventually allow for the rapid acquisition of air quality products such as PM2.5 from NASA satellites and airborne polarimeter and lidar sensors.

B. Generating Simulated Data

One of the main issues in generating a neural network model to solve the forward problem was the lack of a sufficiently large, labeled dataset for PARASOL. While existing PARASOL data products could have been used, if the dataset is not adequately diverse the resulting model will not be capable of handling combinations of state parameters that have not been seen before. The solution is to simulate labeled polarimetry data using NASA's existing vector radiative

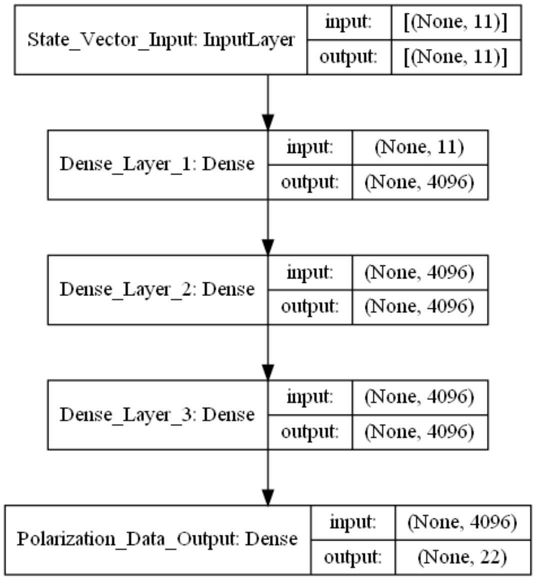

Fig. 4. PARASOL-MAPP forward model neural network architecture

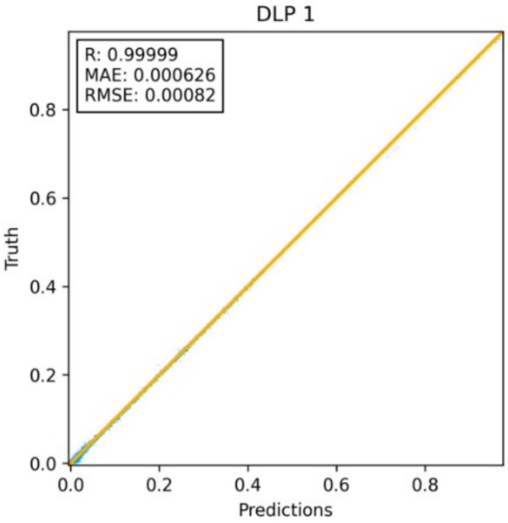

Fig. 5. PARASOL-MAPP neural network forward model prediction of DOLP

transfer model. To create the simulated data, draws are taken from a uniform distribution of realistic state parameters such as AOD, PM2.5, ocean surface roughness, etc. to create a set of diverse yet plausible state vectors. These state vectors are then used as input in the vector radiative transfer model which models the multi-angle, multi-wavelength PARASOL polarimeter measurements (nine intensity wavelengths, and three DOLP wavelengths). Finally, Gaussian noise is applied to the simulated polarimetry measurements when performing synthetic retrievals to produce realistic performance estimates.

C. Neural Network Development

We use a fully connected neural network to achieve the goal of speeding up the acquisition of intensity and DOLP measurements from our input state vectors. A neural network architecture of 3 hidden layers with 4,096 nodes on each hidden layer is used, a design found to be sufficiently complex (33,701,910 trainable parameters) to appropriately model the intricate mathematical relationships between our inputs and outputs (minimizing model bias) while simple enough to make training on an extremely large dataset computationally efficient (Figure 4).

Regularization is assured using a validation data set, with 10 percent of all simulated training data being withheld from training to test our model loss after each training epoch. A constant learning rate is found to work well to efficiently decrease the mean square error using the “Adam” optimizer. Using this trained network, we can use the validation dataset to examine the validation accuracy and loss of each of the elements of our output vector to assess acquisition accuracy across each parameter and the appropriateness of the model.

IV. Results & Discussion

The PARASOL neural network model maps 11 state parameters and 3 solar geometries to 9 DOLP and 9 intensity (I) values. Figures 5 and 6 show the model’s performance for predicting DOLP and I for a given wavelength. These figures show plots of the network’s estimates against the values in the simulated data. Shown in each figure are the following accuracy measures: correlation coefficient (R), mean average error (MAE), and root mean squared error (RMSE).

Across all of the 9 different wavelength outputs, the average R-value is 0.999989 for DOLP and 0.999926 for intensity. The average MAE value is 0.000644 for DOLP and 0.000485 for Intensity. The average RMSE value is 0.000852 for DOLP and 0.000663 for intensity. The PARASOL neural network architecture is an efficient solution to the forward problem, generating outputs 1000x faster than the vector radiative transfer model.

Using the PARASOL neural network as our forward model, we can now rapidly perform retrievals using simulated data. Figure 7 shows the retrieved values of AOD from simulated PARASOL data plotted against the true AOD values. These preliminary results for converged PARASOL retrievals are a proof of concept demonstration, do not yet include realistic PARASOL measurement uncertainties, and will require more work to assure optimal performance. It may be necessary to further train the neural network and to tune the step size in the computation of the finite difference derivatives needed to construct that Jacobian matrix, and that is used to update the state vector.

A key goal of this project is to improve the speed at which we solve the forward problem, as it is the bottleneck in retrieving aerosol properties from polarimeter data. The forward model is repeatedly called until convergence is achieved in the multi-parameter retrieval. The neural net-

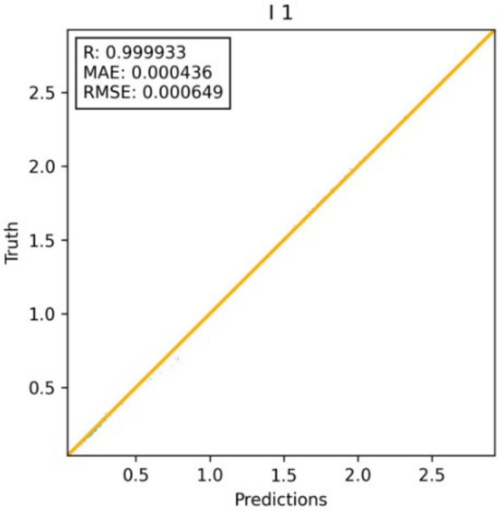

Fig. 6. PARASOL-MAPP neural network forward model prediction of intensity, I

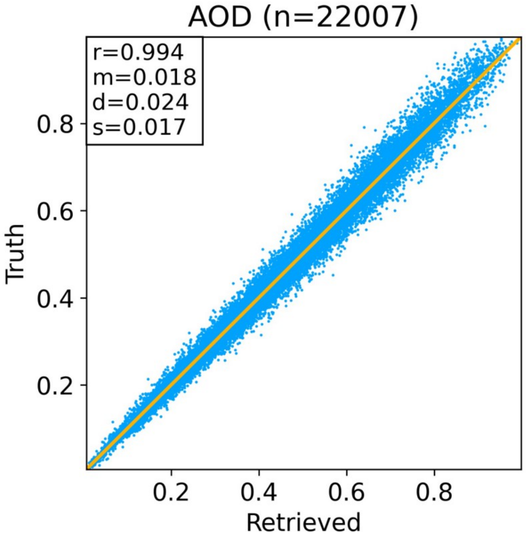

Fig. 7. PARASOL-MAPP retrieval of AOD using optimal estimation

work's enhanced speed compared to the vector radiative transfer model means that the iterative process completes for each pixel of PARASOL polarimeter data in seconds rather than hours. Further, the neural network framework does not sacrifice appreciable accuracy, thereby increasing the scale at which we can reliably study surface-level air quality.

V. Conclusion & Future Work

The ultimate goal of this research is to apply a neural network forward model to real PARASOL polarimeter data in order to produce more accurate but low-latency air quality products including PM2.5. This retrieval capability will allow us to extract aerosol properties and PM2.5 for locations of interest. However, the PARASOL-MAPP retrieval framework requires measurements to be in a specific format, which differs from the native PARASOL satellite polarimeter data format. As a result, future work involves completing the algorithm that converts PARASOL data to NASA's PACE L1C format, another hierarchical data format that uses a different organizational structure than the PARASOL data. Once the infrastructure is developed to convert the PARASOL data to the NASA L1C format reliably, we can then use the PARASOL-MAPP retrieval framework to perform retrievals on real PARASOL data.

Future work will also include using collocated PARASOL polarimeter and lidar data taken from CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization), a two-wavelength polarization lidar instrument aboard the CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations) satellite [11]. The PARASOL polarimeter data delivers the fine and coarse-mode aerosol optical depth and size distribution for the column-averaged properties of the aerosols in the total atmospheric column, but has relatively weak sensitivity to the altitude of the aerosol compared to CALIOP lidar measurements. It is expected that CALIOP lidar data can significantly improve estimates of aerosol number concentration and PM 2.5 at the surface level, which will be critical for validating and improving our forecasts of PM2.5 from PARASOL. The ICARE data archive includes collocated PARASOL-CALIOP datasets and includes relatively new GRASP retrievals of the PARASOL data, including PM2.5. The next step will be to incorporate the CALIOP data of regional aerosol location and compare the updated surface-level PM2.5 results against the existing aerosol location retrieval and PM2.5 estimates. Additional future work will involve training a neural network forward model to include the reflectance from land surface types of interest, such as urban and desert surfaces, further expanding the neural network model's usefulness across the globe.

References

[1] S. Feng, D. Gao, F. Liao, F. Zhou, and X. Wang, "The health effects of ambient PM2.5 and potential mechanisms," Ecotoxicology and Environmental Safety, vol. 128, pp. 67–74, 2016.

[2] V. A. Southerland, M. Brauer, A. Mohegh, M. S. Hammer, A. van Donkelaar, R. V. Martin, J. S. Apte, and S. C. Anenberg, "Global Urban Temporal Trends in fine particulate matter (PM2·5) and attributable health burdens: Estimates from Global Datasets," The Lancet Planetary Health, vol. 6, no. 2, Feb. 2022.

[3] R. V. Martin, "Satellite remote sensing of Surface Air Quality," Atmospheric Environment, vol. 42, no. 34, pp. 7823–7843, 2008.

[4] S. Stamnes, C. Hostetler, R. Ferrare, S. Burton, X. Liu, J. Hair, Y. Hu, A. Wasilewski, W. Martin, B. van Diedenhoven, J. Chowdhary, I. Cetinić, L. K. Berg, K. Stamnes, and B. Cairns, "Simultaneous polarimeter retrievals of Microphysical Aerosol and ocean color parameters from the 'MAPP' algorithm with comparison to high-spectral-resolution lidar aerosol and Ocean Products," Applied Optics, vol. 57, no. 10, p. 2394, 2018.

[5] S. Shi, T. Cheng, X. Gu, H. Guo, Y. Wu, Y. Wang, F. Bao, and X. Zuo, "Probing the dynamic characteristics of aerosol originated from South Asia biomass burning using polder/grasp satellite data with relevant accessory technique design," Environment International, vol. 145, p. 106097, 2020.

[6] P. Lier and M. Bach, “Parasol a microsatellite in the A-train for earth atmospheric observations,” Acta Astronautica, vol. 62, no. 2-3, pp. 257–263, 2008.

[7] Parasol, 24-Apr-2015. [Online]. Available: https://parasol.cnes.fr/en/PARASOL/index.htm. [Accessed: 08-Apr-2022].

[8] G. Shaddick, M. L. Thomas, A. Jobling, M. Brauer, A. van Donkelaar, R. Burnett, H. Chang, A. Cohen, R. Van Dingenen, C. Dora, S. Gumy, Y. Liu, R. Martin, L. A. Waller, J. West, J. V. Zidek, and A. Prüss-Ustün, “Data Integration Model for air quality: A hierarchical approach to the global estimation of exposures to ambient air pollution,” Applied Statistics, 2017.

[9] E. Gorman, D. A. Kubalak, P. Deepak, A. Dress, D. B. Mott, G. Meister, and J. Werdell, “The NASA plankton, aerosol, cloud, ocean ecosystem (PACE) mission: An emerging era of global, Hyperspectral Earth System Remote Sensing,” Sensors, Systems, and Next-Generation Satellites XXIII, 2019.

[10] “Parasol”, distributed by ICARE On-line Data Archive, https://www.icare.univ-lille.fr/asd-content/archive/?dir=PARASOL/

[11] D. M. Winker, M. A. Vaughan, A. Omar, Y. Hu, K. A. Powell, Z. Liu, W. H. Hunt, and S. A. Young, “Overview of the CALIPSO mission and CALIOP data processing algorithms,” Journal of Atmospheric and Oceanic Technology, vol. 26, no. 11, pp. 2310–2323, 2009.