Supplementary material

A Machine Learning Approach to Retrieving Aerosol Optical Depth Using Solar Radiation Measurements

Stavros-Andreas Logothetis[1], Vasileios Salamalikis[2] and Andreas Kazantzidis[1]*

[1]Laboratory of Atmospheric Physics, Physics Department, University of Patras, 26500, Patras, Greece

[2]NILU, P. O. Box 100, Kjeller 2027, Norway

*corresponding author: akaza@upatras.gr

Table S1: Information about the BSRN and AERONET stations' names, BSRN stations' acronym, location, train and test period, data availability (at 1-min resolution) and machine learning algorithm with the highest performance.

BSRN Station	AERONET Station	BSRN ACRO	Location	Train period	Test period	N	MLA
Equatorial (A)							
Darwin	Darwin ARM Darwin	DAR	Australia	04/2004−04/2013	05/2013−12/2014	28461	RF
Arid (B)							
Boulder	Table Mountain	BOS	Colorado, USA	02/2009−07/2015	08/2015−11/2019	59892	MARS
Boulder	BSRN BAO Boulder	BOU	Colorado, USA	09/2004−05/2014	06/2014−06/2016	27473	MARS
Desert Rock	Frenchman Flat	DRA	Nevada, USA	12/2006−04/2013	05/2013−01/2015	22424	ANN
Gobabeb	Gobabeb	GOB	Namib Desert, Namibia	11/2014−08/2018	09/2018−09/2019	147188	ANN
Petrolina	Petrolina SONDA	PTR	Brazil	05/2007−04/2015	05/2015−08/2016	12387	MARS
Sede Boqer	SEDE_BOKER	SBO	Israel	01/2004−09/2010	10/2010−12/2012	53473	RF
Tamanrasset	Tamanrasset INM Tamanrasset TMP	TAM	Algeria	09/2006−09/2014	10/2014−02/2018	52901	RF
Xianghe	Xianghe	XIA	China	01/2005−07/2014	08/2014−10/2015	24826	RF
Warm Temperate (C)							
Billings	Cart Site	BIL	Oklahoma, USA	09/2004−07/2013	08/2013−04/2018	50509	MARS
Bondville	BONDVILLE	BON	Illinois, USA	01/2004−10/2016	11/2016−12/2019	21723	ANN
Carpentras	Carpentras	CAR	France	06/2004−03/2016	04/2016−11/2018	82735	ANN
Chesapeake Light	COVE SEAPRISM COVE	CLH	North Atlantic Ocean	01/2004−10/2006	11/2006−01/2008	9303	ANN
Cener	CENER	CNR	Spain, Sarriguren, Navarra	06/2013−10/2016	11/2016−10/2018	16218	ANN
Southern Great Plains	Cart Site ARM_SGP	E13	Oklahoma, USA	01/2004−01/2015	02/2015−05/2018	55098	ANN
Fukuoka	Fukuoka	FUA	Japan	06/2012−08/2015	09/2015−01/2019	29379	ANN
Lindenberg	MetObs Lindenberg	LIN	Germany	09/2013−07/2017	08/2017−06/2018	4529	ANN

Langley Research Center	NASA LaRC	LRC	Hampton, Virginia, USA	12/2014−11/2018	12/2018−12/2019	30426	MARS
Observatory of Huancayo	Huancayo-IGP	OHY	Peru	08/2017−07/2018	08/2018−03/2019	10897	MARS
Palaiseau, SIRTA Observatory	Palaiseau	PAL	France	10/2015−09/2018	10/2018−10/2019	26935	ANN
São Martinho da Serra	Sao Martinho SONDA	SMS	Brazil	11/2008−07/2013	08/2013−11/2014	6415	ANN
Snow (D)							
Regina	Bratts_Lake	REG	Canada	01/2004−06/2010	07/2010−12/2011	61898	RF
Sapporo	Hokkaido University	SAP	Japan	10/2015−08/2018	09/2018−09/2019	23845	ANN
Sioux Falls	Sioux Falls Sioux Falls X	SXF	South Dakota, USA	06/2004−03/2016	04/2016−08/2018	20593	MARS
Toravere	Toravere	TOR	Estonia	02/2004−03/2016	04/2016−07/2019	22589	ANN
Polar (E)							
Barrow	Barrow NEON BARR	BAR	Alaska, USA	05/2004−07/2012	08/2012−07/2017	7839	ANN

Table S2: Hyperparameters that are tuned during the training of machine learning algorithms**.**

MLA	Hyperparameters
LGBM	**n_estimators** = [50, 100, 500, 1000, 1500, 2000] **learning_rate** = [0.3, 0.2, 0.1, 0.05, 0.01, 0.001, 0.0001] **max_depth** = [2, 5, 10, 12, 15, 20] **num_leaves** = [2, 5, 10, 12, 15, 20]
RF	**n_estimator** = [100, 300, 500, 700, 900, 1000, 1500, 2000] **max_features** = [1, 2, 3] **max_depth** = [None, 10, 15, 20, 25] **min_samples_split** = [2, 3, 4, 5, 6] **min_samples_leaf** = [1, 2, 3, 4, 5, 6]
MARS	**max_terms** = [1, 2, 3] **max_degree** = [1, 2, 3]
KNN	**n_neighbors** = [neighbors from 1 to 30 per 1]
ANN	**Hidden layer 1** = [nodes from 8 to 64 per 4] **Hidden layer 2** = [nodes from 8 to 32 per 4]

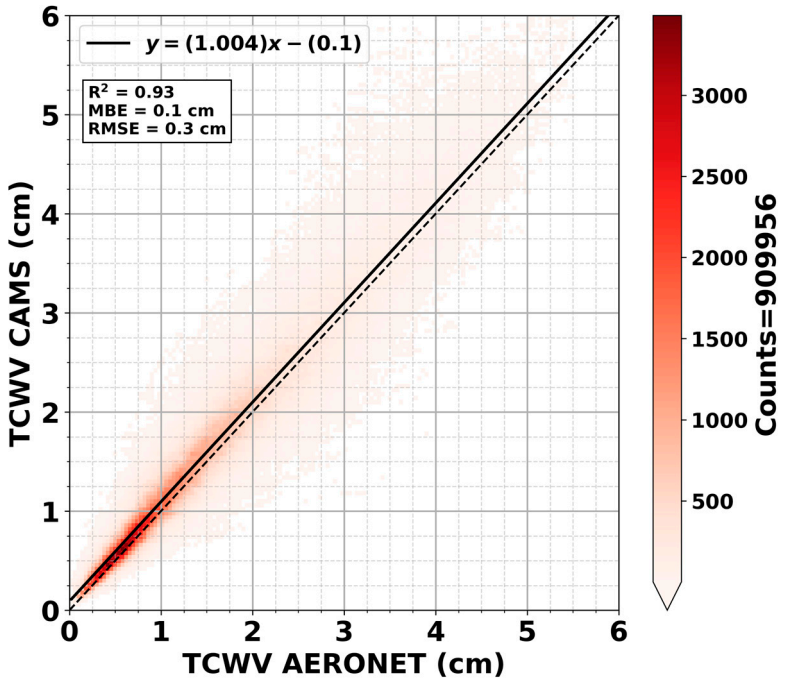

Figure S1: Linear correlation between CAMS and AERONET total column water vapor retrievals.

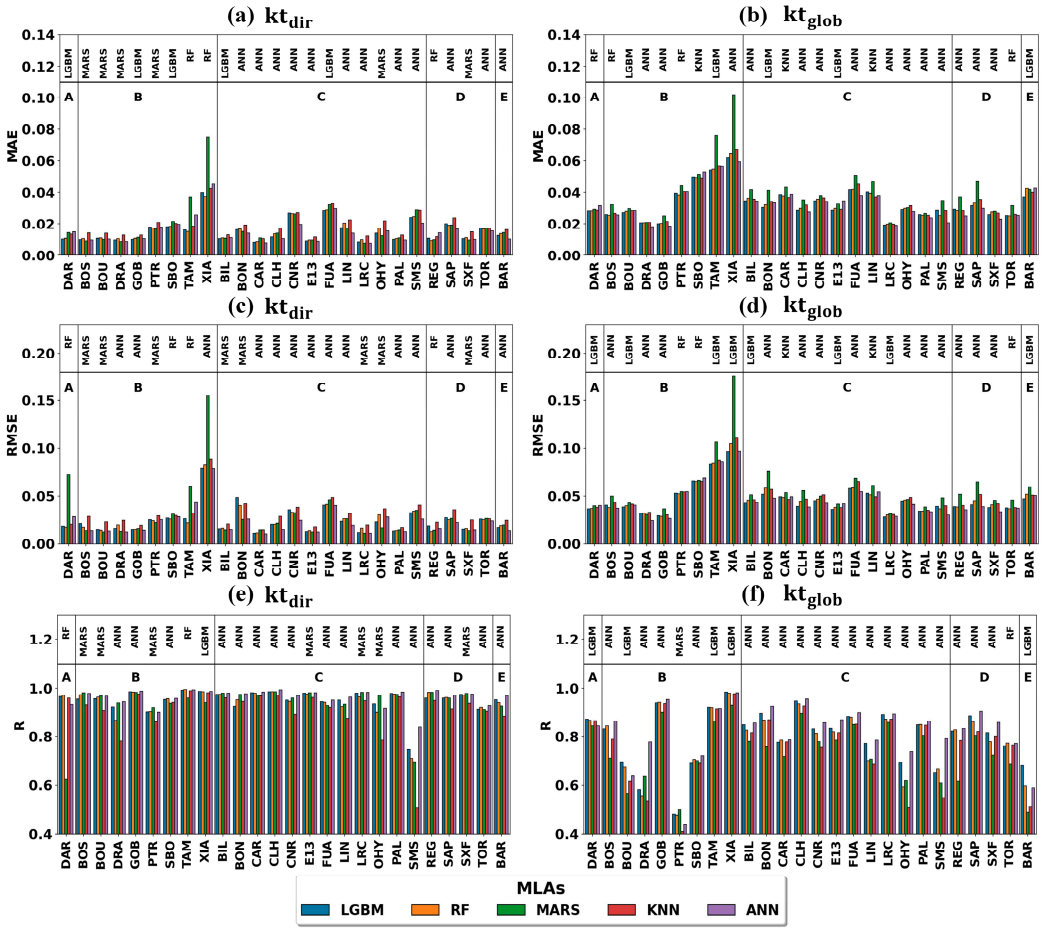

Figure S2. Statistical indices (MAE, RMSE and R) for Scenario 1 (a, c, e) and Scenario 2 (b, d, f) for each machine learning algorithm (MLA). The horizontal lines separate the stations based on the KG classes. The MLA with the best performance is added above each station.

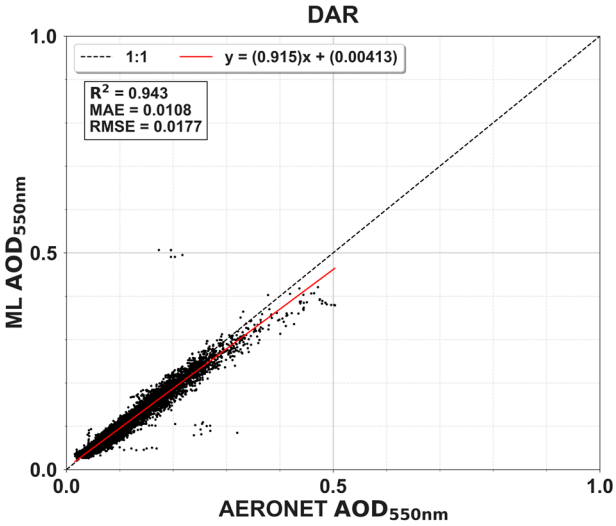

Figure S3. Linear correlation between ML-AOD and AERONET AOD for the stations in Equatorial (A) KG climate class.

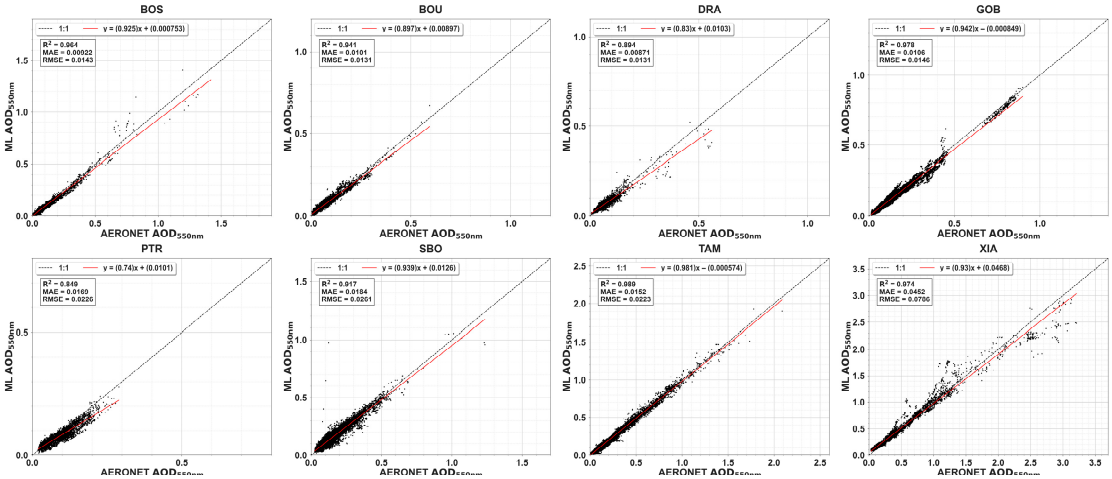

Figure S4. Linear correlation between ML-AOD and AERONET AOD for the stations in Arid (B) KG climate class.

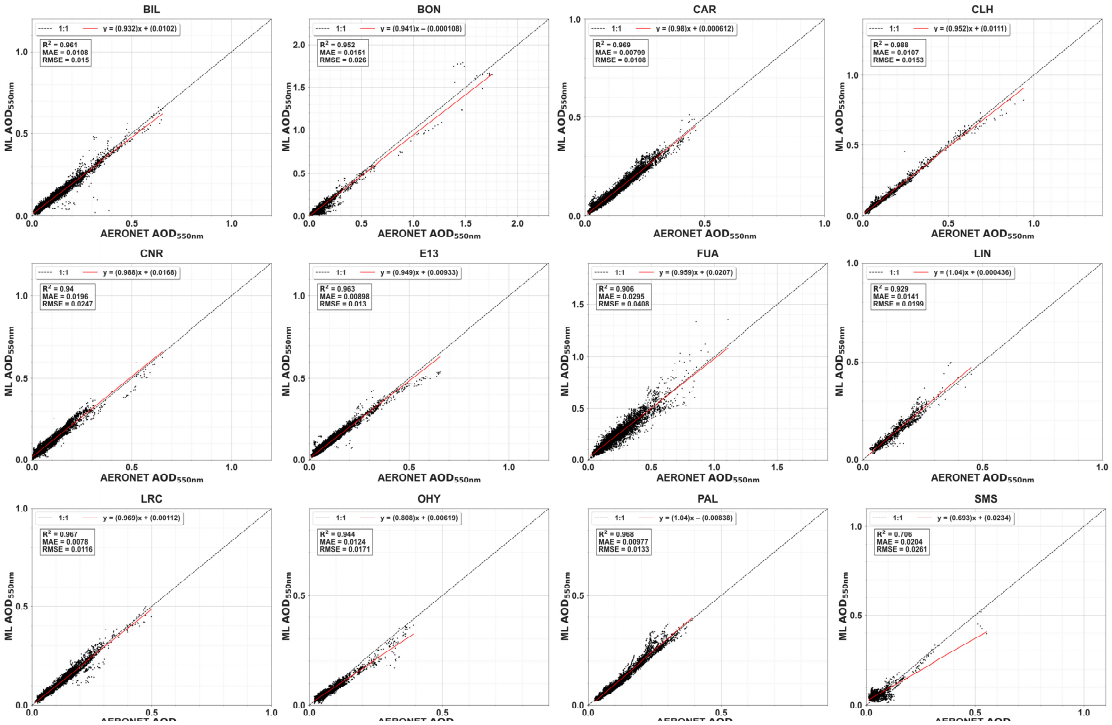

Figure S5. Linear correlation between ML-AOD and AERONET AOD for the stations in Warm temperate (C) KG climate class.

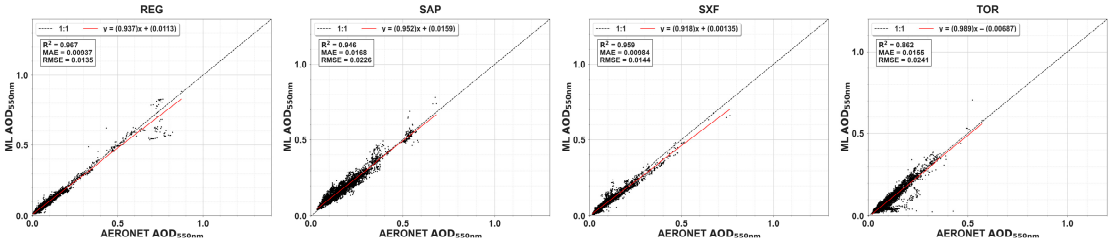

Figure S6. Linear correlation between ML-AOD and AERONET AOD for the stations in Snow (D) KG climate class.

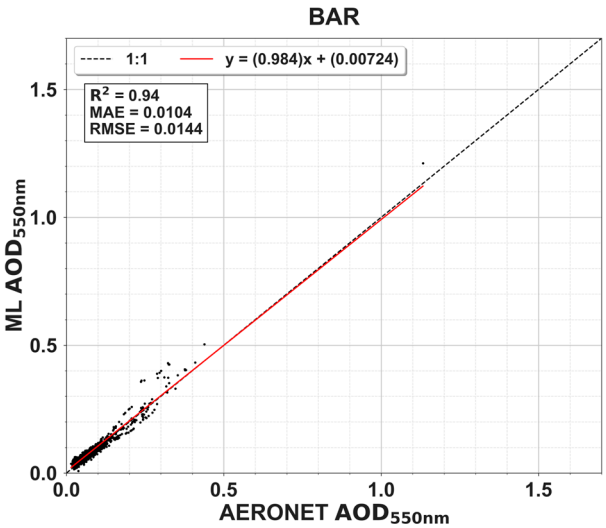

Figure S7. Linear correlation between ML-AOD and AERONET AOD for the stations in Polar (E) KG climate class.

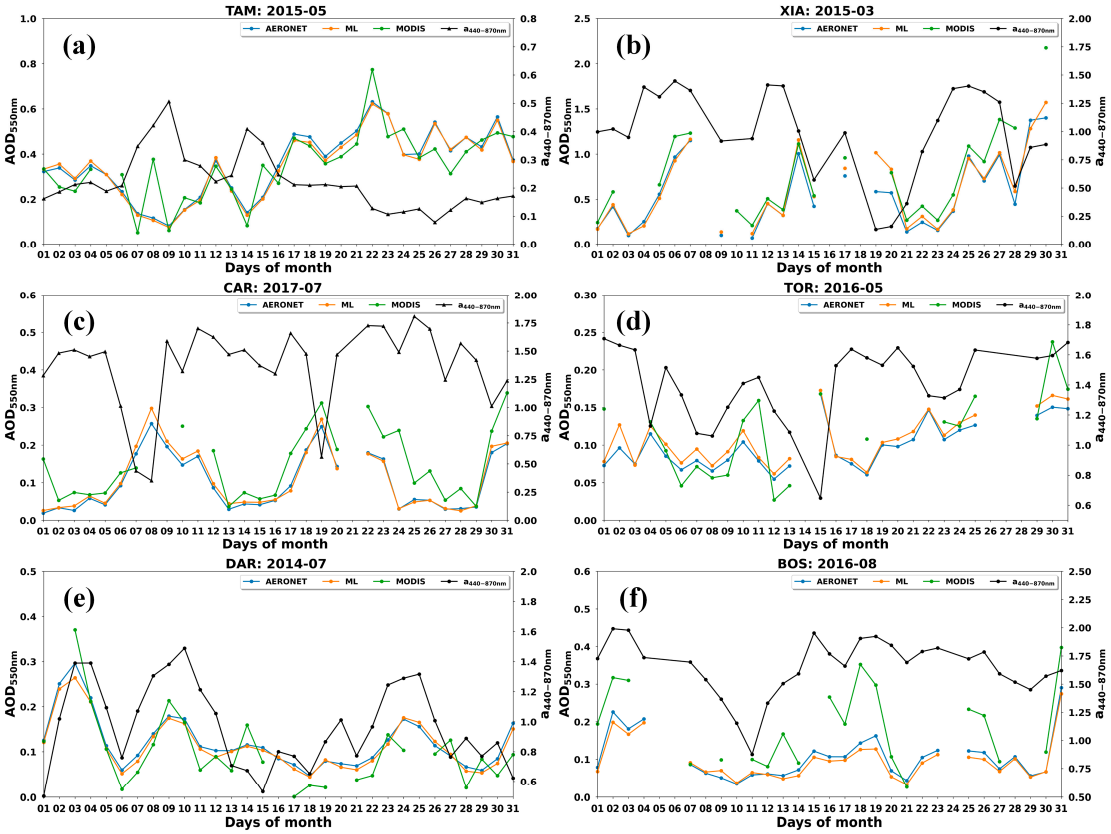

Figure S8. AERONET, ML-AOD, and MODIS AOD_{550nm} daily variability for specific months at (a) TAM, (b) XIA, (c) CAR, (d) TOR, (e) DAR, and (f) BOS sites. The second y-axis shows the daily variability of $a_{440\text{-}870nm}$.