

Article

An Exploratory Approach Using Regression and Machine Learning in the Analysis of Mass Absorption Cross Section of Black Carbon Aerosols: Model Development and Evaluation

Hanyang Li ^{1,2} and Andrew A. May ^{1,*} 

¹ Department of Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, OH 43210, USA; li.4860@buckeyemail.osu.edu

² Air Quality Research Center, University of California Davis, Davis, CA 95616, USA

* Correspondence: may.561@osu.edu; Tel.: +1-614-688-1206

Received: 7 August 2020; Accepted: 27 October 2020; Published: 2 November 2020



Abstract: Mass absorption cross-section of black carbon (MAC_{BC}) describes the absorptive cross-section per unit mass of black carbon, and is, thus, an essential parameter to estimate the radiative forcing of black carbon. Many studies have sought to estimate MAC_{BC} from a theoretical perspective, but these studies require the knowledge of a set of aerosol properties, which are difficult and/or labor-intensive to measure. We therefore investigate the ability of seven data analytical approaches (including different multivariate regressions, support vector machine, and neural networks) in predicting MAC_{BC} for both ambient and biomass burning measurements. Our model utilizes multi-wavelength light absorption and scattering as well as the aerosol size distributions as input variables to predict MAC_{BC} across different wavelengths. We assessed the applicability of the proposed approaches in estimating MAC_{BC} using different statistical metrics (such as coefficient of determination (R^2), mean square error (MSE), fractional error, and fractional bias). Overall, the approaches used in this study can estimate MAC_{BC} appropriately, but the prediction performance varies across approaches and atmospheric environments. Based on an uncertainty evaluation of our models and the empirical and theoretical approaches to predict MAC_{BC} , we preliminarily put forth support vector machine (SVM) as a recommended data analytical technique for use. We provide an operational tool built with the approaches presented in this paper to facilitate this procedure for future users.

Keywords: black carbon; mass absorption cross section (MAC); aerosol light absorption; machine learning; deep learning

1. Introduction

Black carbon (BC) aerosols are emitted from incomplete combustion processes (e.g., fossil fuel and biomass burning) as fine particles [1,2]. BC has a major role in the climate system due to its ability to absorb solar radiation and interactions with clouds [3–5]. Understanding the properties of BC and quantifying its mass concentration in the atmosphere are essential to estimate its impacts on climate change. One widely used method to determine BC mass concentration is dividing light absorption coefficient (B_{abs}) by mass absorption cross-section (MAC) [6–8]. Although B_{abs} can be measured using various instruments (such as in situ and filter-based optical instruments [7,9]), the value of MAC needs to be estimated or known in advance when performing the numerical calculation of BC mass. Consequently, the derived BC mass is sensitive to the adopted MAC value, so improved estimation of MAC_{BC} will improve estimates of BC mass concentration from ground-based networks providing B_{abs} ,

which can, in turn, result in improvements to chemistry-climate models and our understanding of the radiative impacts of BC.

Historically, MAC of $7.5 \pm 1.2 \text{ m}^2 \text{ g}^{-1}$ (550 nm) for “fresh” and uncoated BC has been recommended [6,10]. In a recent review of 10 published reports of MAC values for freshly emitted BC, Liu et al. [11] conclude that the averaged MAC is $8.0 \pm 0.7 \text{ m}^2 \text{ g}^{-1}$ (550 nm), generally agreeing with the values recommended in Bond and Bergstrom [6]. However, MAC of atmospheric BC particles is more variable and more uncertain than that of fresh BC. For example, MAC of atmospheric BC may be lower (e.g., Nordmann et al. [12] report values of 3.9 to $7.4 \text{ m}^2 \text{ g}^{-1}$ at 550 nm) or larger (e.g., Gyawali et al. [13] and Kondo et al. [8] report values ranging from roughly 10 to $14 \text{ m}^2 \text{ g}^{-1}$ at 550 nm) than the “typical” value of $7.5 \text{ m}^2 \text{ g}^{-1}$. Similarly, the value of MAC used in different climate models spans a broad range (2.3 to $10.5 \text{ m}^2 \text{ g}^{-1}$ (550 nm); e.g., [14] and references therein). Differences in MAC values can exist due to differences in the aerosol mixing state, the underlying size distribution of BC-containing particles, and the effective density of the BC-containing particles. This is further complicated by uncertainty in the complex refractive index of BC; for example, Bond and Bergstrom [6] provide a range of real and imaginary complex refractive indices that may be representative of BC. Moreover, atmospheric aging may result in MAC values that differ from the Bond and Bergstrom [6] value (e.g., [15]); therefore, MAC is likely a function of time, not a constant value (e.g., [16]). Finally, although 550 nm is often the focus for reporting MAC, BC absorbs light across all wavelengths. To extend this value to other wavelengths of light (λ) a power law relationship [3,17,18] is used:

$$\text{MAC}(\lambda) = \text{MAC}(550 \text{ nm}) \times \left(\frac{\lambda}{550} \right)^{-\text{AAE}} \quad (1)$$

where AAE is the absorption Ångström exponent. AAE is commonly assumed to be unity for BC, but it varies in the literature between ~ 0.6 and ~ 1.6 [3,19–21].

The presence of other absorbing materials in the atmosphere further complicates the application of MAC to derive BC mass. Brown carbon (BrC) has been found to absorb light across ultraviolet-visible wavelengths, with MAC lower than that of BC [1,22,23]. Tar balls have attracted considerable attention recently as absorbers of light at visible and near-infrared spectral regions (e.g., [24–26]). One important source of atmospheric BrC and tar balls is biomass burning [6,27,28]. Another source of absorbing aerosols is wind erosion of the Earth’s surface, releasing mineral dusts, which are non-carbonaceous material that weakly absorb light across all wavelengths [9,29–31]. Therefore, due to different spectral dependence of light absorbing materials, simply dividing B_{abs} by MAC may introduce uncertainties to the derived BC mass. Even at longer wavelengths (for example, 870 nm) where light absorption by most BrC is negligible, the computed BC mass may be overestimated if tar balls and mineral dust contribute to the light absorption.

Theoretically, MAC can be calculated from formulations using aerosol complex refractive indices, mixing state, size, and morphology as input variables [11,32–35], all of which are difficult and/or labor-intensive to measure. Necessary assumptions within these calculations can also add uncertainty to the derived MAC. Hence, alternative approaches, such as the use of statistical methods and machine learning approaches may provide a new insight to predict values of MAC with comparable accuracy.

The goal of the present study is to explore and develop a straightforward yet accurate model for predicting temporal variations in the MAC of BC (MAC_{BC}). This model does not require the knowledge of aerosol composition or any assumptions of complex refractive indices and mixing state. In situ measurements are input to the model, including multiple wavelength B_{abs} and light scattering coefficients (B_{scat}) as well as aerosol size distributions. One advantage of incorporating these measurements is that they are likely to be available at the majority of long-term monitoring sites [36–39], while aerosol composition is often not. To build the model, we examine several popular data analytical approaches. These approaches include classical statistical techniques and machine learning techniques (see Section 2.4 for details). We also provide several modules related to data preprocessing (see Section 2.3).

We present here the framework of the model, together with an evaluation of model by testing aerosols from various environments using different statistical metrics. A forthcoming companion work will describe an extension of these models to predict MAC_{BC} across different wavelengths using the model output. This future work will also include a sensitivity analysis of the models to different inputs variables (e.g., if only measurements of single-wavelength B_{abs} and B_{scat} exist) and a discussion of the influence of varying chemical composition and related optical properties (e.g., AAE, single scattering albedo (SSA)).

2. Methodology

2.1. Dataset Description

We applied the data analytical approaches to two publicly available U.S. Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) field campaigns, the Two-Column Aerosol Project (TCAP) and the Cloud, Aerosol and Complex Terrain Interactions (CACTI) project, as well as biomass burning emissions collected during the U.S.'s National Oceanic and Atmospheric Administration (NOAA)-sponsored Fire Influence on Regional to Global Environments Experiment (FIREX) laboratory campaign. A summary of aerosol properties obtained from each dataset is provided in Table 1. Values represent the mean across the entire sampling period along with the standard deviation across the period; hence, there is a temporal variability of ~50% in MAC_{BC} for each dataset.

Table 1. Overview of the studies and aerosol properties. Aerosol properties are expressed as mean (standard deviation) ^a.

Dataset	Number of Data ^b	MAC_{BC} (870 nm) ($m^2 g^{-1}$) ^{c,d,e}		Absorption Ångström Exponent (AAE)	Scattering Ångström Exponent (SAE)	SSA_{528} ^f		D_{50} (nm) ^g	Dv_{50} (nm) ^h
		Li et al. [40]	Bond et al. [41]			Li et al. [40]	Bond et al. [41]		
TCAP (training set)	7321	3.87 (1.92)	7.81 (4.11)	1.50 (0.96)	1.27 (0.46)	0.98 (0.01)	0.95 (0.03)	68.40 (20.12)	1543 (1093)
TCAP (test set)	1830	3.85 (1.90)	7.78 (4.06)	1.48 (0.97)	1.30 (0.46)	0.98 (0.01)	0.96 (0.03)	69.10 (20.48)	1544 (1091)
CACTI	10,709	2.94 (1.68)	7.36 (3.53)	1.32 (0.56)	1.20 (0.64)	0.96 (0.02)	0.92 (0.03)	56.15 (32.83)	2275 (1286)
FIREX ^h	6745	3.35 (1.45)	7.53 (3.25)	2.02 (0.74)	2.26 (0.64)	0.66 (0.25)	0.52 (0.28)	109.05 (46.26)	310 (218)

^a See Figure 1 for the temporal variability of mass absorption cross-section of black carbon (MAC_{BC}); ^b The datasets presented here are reported on a per-minute basis; ^c When calculating the MAC_{BC} in this table, we use the black carbon (BC) mass concentration from the Single Particle Soot Photometers (SP2) and the filter-based light absorption coefficient (B_{abs}) corrected by either Li et al. [40] or Bond et al. [41], which was further modified in Ogren [42]. Hereafter, when we refer to Bond et al. [41], we mean the correction method which has been modified in Ogren [42]; ^d The results are compared to the predicted MAC_{BC} presented in Table S5; ^e Using the “standard” MAC_{BC} from Bond and Bergstrom [6] and an assumed absorption Ångström exponent (AAE) = 1, the “expected” MAC_{BC} at 870 nm is $4.74 \pm 0.75 m^2 g^{-1}$, suggesting that both correction equations may be imperfect; ^f SSA_{528} : single scattering albedo at 528 nm, calculated as $B_{scat}/(B_{abs} + B_{scat})$; ^g D_{50} and Dv_{50} : median diameter of the number and volume size distribution, respectively. The values are reported over the entire size distribution, which may have multiple modes, so D_{50} and Dv_{50} can be different from the geometric mean value of any individual modes; ^h During the Fire Influence on Regional to Global Environments Experiment (FIREX) campaign, only Scanning Mobility Particle Sizers (SMPS) was used to detect aerosol size distribution.

The TCAP (Cape Cod, Massachusetts, USA; roughly 60 miles southeast of Boston) and CACTI (mountain area of north-central Argentina; roughly 50 miles west of the city of Córdoba) field campaigns reported ground measurements of B_{abs} using Particle Soot Absorption Photometer (PSAP), B_{scat} using multi-wavelength nephelometers, aerosol size distributions using Scanning Mobility Particle Sizers (SMPS) and Aerodynamic Particle Sizers (APS) (spanning 11.3–461.4 nm for the SMPS and 444.6–18,747.4 nm for the APS, the latter converted to mobility diameter), and BC mass concentration using Single Particle Soot Photometers (SP2) [43–45]. For both campaigns, we randomly

selected a period when all measurements were available and without bad and/or missing data (TCAP: 16 July 2012 to 14 August 2012, CACTI: 1 December 2018 to 20 December 2018). The major aerosol sources at the two observational sites are likely to be different. Kassianov et al. [45] found that both marine aerosols and aerosols transported from continental North America over the Atlantic Ocean contributed to the observations at TCAP. Different from TCAP, the major sources of aerosols at CACTI likely included local and secondary biomass burning emissions and wind-blown dust [46].

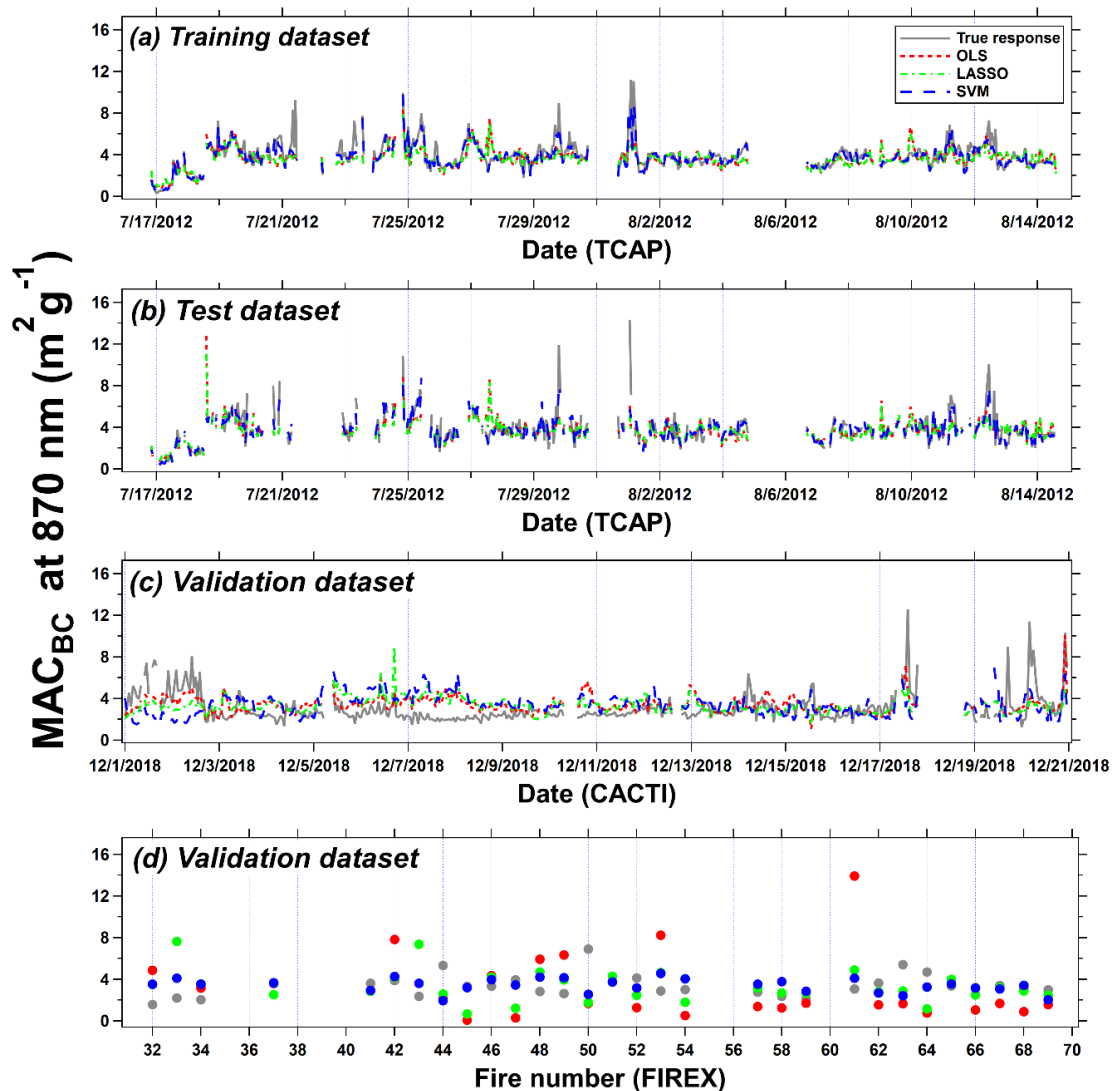


Figure 1. True response and predicted time series of MAC_{BC} using (a) training dataset, (b) test dataset, (c) validation dataset (CACTI), and (d) validation dataset (FIREX). Only the curves for three representative techniques (OLS, LASSO, and SVM) are included in the figure. Note that in panel (d), the OLS technique yields severe overprediction of MAC_{BC} (>10,000 m² g⁻¹) for fire numbers of 33, 37, 41, 43, 51, and 65. To better view the results, the OLS-derived MAC_{BC} for these fires are not shown in the figure (as well as the other figures in this work).

The purpose of the laboratory portion of the FIREX campaign (Fire Sciences Laboratory, Montana, USA) conducted in 2016 was to understand how wildfire smoke influences the atmosphere [47]. There were slight differences in instruments used at FIREX compared to the DOE campaigns. During FIREX, B_{abs} was quantified using a PSAP “successor” (Continuous Light Absorption Photometer, CLAP), B_{scat} was detected by two Photoacoustic Extinctionmeters (PAXs) at 405 and 870 nm, and a

relatively narrower size distribution was obtained by an SMPS (14.6–736.5 nm, mobility diameter); however, BC mass concentrations were still available from an SP2 [48].

In the present work, data preparation was performed in two steps: (1) observations of B_{abs} and B_{scat} were discarded if they fell below the estimated detection limits (PSAP: 0.30 Mm^{-1} [49]; CLAP: 0.60 Mm^{-1} [50]; Nephelometer: 0.29, 0.11, and 0.17 Mm^{-1} at 450, 550, and 700 nm respectively [51]; PAX (B_{scat} only): 0.60 and 0.66 Mm^{-1} at 405 and 870 nm respectively); and (2) all data were converted to a 1-min time interval. For example, B_{abs} and B_{scat} were smoothed from 1-s data into 1-min averages, and the time series of particles size distributions were linearly interpolated from 5-min values to 1-min values.

2.2. Construction of Model

In this section, we first present how we utilize the above datasets to build the model. Then, we summarize the input and output parameters of the model and present the calculation details. Even though we focus on the B_{abs} quantified by PSAP and CLAP in the present work, our model could be also implemented on B_{abs} measured by the other instruments (such as the aethalometer or any photoacoustic technique). Similarly, the particle size distributions derived by other instruments (e.g., the Ultra-High-Sensitivity Aerosol Spectrometer (UHSAS, Droplet Measurement Technologies) or the Laser Aerosol Spectrometer (TSI Inc.)) could be used as input in our model (see Section 2.3.1 for more details). To build the model, we follow a common data partition method used in the machine learning community [52,53]. Specifically, we first randomly split the TCAP data into training set and test set (80% and 20%, respectively). The data analytical techniques are initially applied to the training set to derive parameters and then are evaluated using the test set. The test data are used as new data in order to independently assess the goodness of fit for our data analytical techniques. To further investigate the generalizability of the techniques, we extend the fitted models to another ambient observation site (CACTI) and aerosols obtained from laboratory biomass burning (FIREX). The latter two datasets are referred to as independent validation datasets in this work.

Selecting the appropriate input parameters for model development can affect the quality of MAC_{BC} prediction. Therefore, we use the underlying physics for inspiration in identifying candidate parameters for our statistical models. MAC_{BC} (at 870 nm) can be calculated following:

$$\text{MAC}_{\text{BC}} (870 \text{ nm}) = \frac{B_{\text{abs}} (870 \text{ nm})}{M_{\text{BC}}} \quad (2)$$

where M_{BC} is the mass concentration of BC. Light scattering theories (such as Mie theory and the Rayleigh–Debye–Gans approximation) can be used to calculate B_{abs} using an aerosol population’s complex refractive index and particle number distribution as inputs, while calculations of M_{BC} depend on the aerosol population’s density and particle volume distribution. The complex refractive index and density are both related to aerosol composition. However, because complete aerosol speciation is rare at the majority of long-term monitoring sites, we attempt to use empirical observations of B_{abs} and B_{scat} as an alternative; the basis for this is that previous studies have utilized AAE and scattering Ångström exponent (SAE) (i.e., an “AAE–SAE space”, [30,54–56]) to categorize absorbing aerosols (e.g., BC-dominant, dust-dominant, BC/BrC mixtures). We also include parameterizations of both empirical particle number distributions and particle volume distributions as candidates for our models. Even though these are inherently related, we argue that including representations of both is somewhat analogous to the Two-Moment Aerosol Sectional (TOMAS) model [57], which tracks both number and mass distributions, and has been implemented in chemistry-climate models (e.g., [58,59]) to predict aerosol microphysics. Consequently, there are 20 candidate input variables in our statistical and machine learning models (Table 2). Pre-processing of these variables is described in Section 2.3, and the specific modeling approaches are described in Section 2.4.

Table 2. Input variables in the model.

Absorption Coefficient	Scattering Coefficient	Number Concentration ^a	Fractional Number Concentration ^b	Volume Concentration ^a	Fractional Volume Concentration ^b
B_{abs} (467 nm)	B_{scat} (467 nm)	N_1 (0– d_{N1} nm)	F_{N1} (0– d_{N1} nm)	V_1 (0– d_{V1} nm)	F_{V1} (0– d_{V1} nm)
B_{abs} (528 nm)	B_{scat} (528 nm)	N_2 (d_{N1} – d_{N2} nm)	F_{N2} (d_{N1} – d_{N2} nm)	V_2 (d_{V1} – d_{V2} nm)	F_{V2} (d_{V1} – d_{V2} nm)
B_{abs} (652 nm)	B_{scat} (652 nm)	N_3 ($>d_{N2}$ nm)	F_{N3} ($>d_{N2}$ nm)	V_3 ($>d_{V2}$ nm)	F_{V3} ($>d_{V2}$ nm)
		N_{total}		V_{total}	

^a d_{N1} and d_{N2} represent the user-specified upper or lower bound of each size class. The sum of N_1 , N_2 , and N_3 equals to N_{total} . The same definition and relationship applies for volume concentration. See Section 2.3.1 for more details; ^b Fractional number concentration is derived by dividing the number concentration within each size class by N_{total} , e.g., $F_{N1} = N_1/N_{\text{total}}$. The same definition and relationship applies for fractional volume concentration.

When developing the model, we select the desired output to be MAC_{BC} at 870 nm because this wavelength has traditionally been associated solely with BC (e.g., there is limited influence from other absorbing species), although we acknowledge that tar balls and/or mineral dust does absorb at this wavelength. To yield B_{abs} (870 nm), the B_{abs} values from the filter-based photometers corrected by either Li et al. [40] or Bond et al. [41] are extrapolated to 870 nm, following $B_{\text{abs}} \sim \lambda^{-AAE}$, where AAE is first calculated by the B_{abs} at multiple wavelengths using a power-law fit. To derive the observation-based MAC_{BC} (the true response of MAC_{BC}), M_{BC} quantified by the SP2 at the observational site is utilized.

2.3. Data Preprocessing

In this section, we present a few preprocessing steps (summarized as three modules at <https://doi.org/10.5281/zenodo.3967833> [60]) that are necessary to enable model implementation. Readers should be aware that these modules are optional if they have data formatted as summarized in Table 2 or if they wish to conduct their own data preparation. For example, we provide the implementation of the correction algorithm described in Li et al. [40] and Bond et al. [41] for filter-based B_{abs} , but various other correction techniques exist (e.g., Müller et al. [61] and Virkkula et al. [62] for the PSAP and CLAP; Collaud Coen et al. [63] and Weingartner et al. [64] for aethalometers). As the final step of data preprocessing, we normalize the raw data with different scales of values into scaled data within a specific range [65], which can significantly improve the prediction accuracy of models (see the Supplementary Material for more details).

2.3.1. Preprocessing Size Distribution Data

Both number and volume size distributions are selected as input variables in our model as they can provide different information about the aerosol properties. For example, large particles (such as mineral dust and sea salt) primarily affect the volume concentration, without substantially affecting the number concentration. In this work, aerosol volume distribution ($N_V(D_P)$) is calculated from number size distribution ($N_N(D_P)$) assuming that the aerosols are spherical, following:

$$N_V(D_P) = \frac{\pi}{6} D_P^3 N_N(D_P) \quad (3)$$

where D_P is aerosol mobility diameter (nm). When processing the ARM datasets, we first convert the APS data from aerodynamic to mobility diameters assuming that the particles are spherical with a density of 1.2 g cm^{-3} , same as the procedure used in Ondráček et al. [66] and Shen et al. [67].

Considering that the size bins of the distributions may differ based on pre-defined parameters (such as the selected size range, size resolution, etc.), we use a few size classes to represent the entire size distribution; effectively, we represent both the number and volume size distributions using three bins (see Table 2). The corresponding aerosol concentration and percentage of total aerosol

concentration falling within each class are calculated using Module A, which are later input to our model. Furthermore, the measured particle size range may vary from instrument to instrument. For example, the combination of SMPS and APS can provide a broad size range from roughly 10 nm to 20 μm , while the UHSAS measures particles with sizes ranged from 55 nm to 1 μm . In order to capture the effect of measured size range on the size classes used in our model, the ranges of each size class (d_{N1} and d_{N2} , d_{V1} and d_{V2}) are selected as input parameters in Module A. Our default size boundaries are $d_{N1} = 50$ nm, $d_{N2} = 200$ nm, $d_{V1} = 1000$ nm, and $d_{V2} = 2500$ nm, but users can manually input lower and upper limits of each size class prior to module implementation. For example, when preprocessing the FIREX dataset, we used $d_{N1} = 80$ nm, $d_{N2} = 200$ nm, $d_{V1} = 200$ nm, and $d_{V2} = 800$ nm, since only an SMPS was deployed during the campaign. The sensitivity analysis of the predicted MAC_{BC} due to the selected size bins will be discussed elsewhere.

2.3.2. Preprocessing Absorption and Scattering Measurements

Filter-based absorption photometers have known biases, largely due to the presence of the filter. To minimize measurement biases, we include the correction algorithms developed by Li et al. [40] and Bond et al. [41] as an optional module (Module B). One advantage of the correction algorithm presented in Li et al. [40] is that it updates the correction coefficients based on the optical properties (i.e., SSA and AAE) computed by the input aerosol data. However, considering that the correction method described in Bond et al. [41] is still widely used, we incorporate this method in our module.

Furthermore, because filter-based absorption photometers and nephelometers may operate at different wavelengths, we provide another optional module (Module C) to convert absorption and scattering measurements to “standard” wavelengths (467, 528, and 652 nm in the present work). For example, we use the following equation to convert B_{abs} and B_{scat} from the measured wavelength (λ_{meas}) to 652 nm:

$$B(652 \text{ nm}) = B(\lambda_{\text{meas}}) \times \left(\frac{652}{\lambda_{\text{meas}}} \right)^{-\text{AE}} \quad (4)$$

where B can be either B_{abs} or B_{scat} and AE can be either AAE or SAE. AAE and SAE are calculated by fitting exponential curves on the measured B_{abs} and B_{scat} , respectively [68,69].

2.4. Basic Principles of Data Analytical Techniques

Different data analytical approaches, including multiple linear regression model using ordinary least squares (OLS), stepwise regression, least absolute shrinkage and selection operation (LASSO), support vector machine (SVM), artificial neural network (ANN), and convolutional neural network (CNN), are introduced and applied to our datasets to find the optimal model for MAC_{BC} prediction. The basis, advantages, and limitations of each approach are provided in Table 3. A detailed description of the mechanics for each approach can be found in the Supplementary Material.

2.5. Analytical Tools

The data analysis is performed using the programming language Python (version 3.7.5) for Windows in an “Anaconda” environment. Creating an “Anaconda” environment allows the installation and implementation of libraries and packages needed for processing data analytical techniques mentioned in this work. We provide a “User’s Guide” including a detailed description of procedures to install packages, set the working environment, and apply our modules and models at <https://doi.org/10.5281/zenodo.3967833> [60].

Table 3. Comparison of different data analytical approaches.

Approach	Basis	Advantage	Limitation	References
multiple linear regression using OLS	minimize the sum of squared distances between the “true” observations and predictions (the sum of squared residual errors, SSE)	<ul style="list-style-type: none"> • simple and powerful for linear prediction • backbone for many advanced data analytical methods 	highly depend on the assumptions: <ul style="list-style-type: none"> • The relationship between the output variable and input variables is linear • Residual errors are distributed normally • Input variables have equal variance 	[70]
stepwise regression	choose a group of “best” independent variables based upon some statistical criteria (Akaike Information Criterion (AIC) in this work)	<ul style="list-style-type: none"> • test relationships for different combinations of independent variables • easy to implement 	<ul style="list-style-type: none"> • If two independent variables are highly correlated, the one has the highest correlation with the dependent variable will be left in the model. • The “best stepwise model” may not be the “true model” 	[71–73]
LASSO	automatically select independent variables by adding a penalty term to the SSE	<ul style="list-style-type: none"> • retain the ability of selecting important variables (similar to stepwise regression) • reduce over-fitting through the penalty term 	If several independent variables are inter-correlated, LASSO tends to arbitrarily select only one variable.	[74]
SVM ^a	map the independent variables into a higher dimensional space where linear regression is performed	<ul style="list-style-type: none"> • deal with nonlinear problems with the help of nonlinear kernels • less prone to over-fitting, compared to classical regression techniques 	<ul style="list-style-type: none"> • computational expensive • sensitive to noises and outliers • difficult to interpret the derived model 	[75–77]
ANN	compare the desired output and modelled output (randomly initialized) and then adjust the network weights using back-propagation rule	<ul style="list-style-type: none"> • handle complex nonlinear relationships • not sensitive to noises • require no assumptions about the underlying function 	<ul style="list-style-type: none"> • computational expensive • require “big” data for training • The model is not interpretable (black box) • Model development is empirical 	[78]
CNN ^b	pass a convolutional kernel over the sequence of data and generate a feature map to represent the data	<ul style="list-style-type: none"> • automatically detect important features of the dataset • less complex than a fully connected ANN 	<ul style="list-style-type: none"> • The order of input variable is important • require “big” data for training 	[79,80]

^a A version of support vector machine (SVM) for regression (support vector regression (SVR)) is used in this work to predict MAC_{BC}; ^b We use a modified one-dimensional convolutional neural network (1D-CNN) in this work. The 1D sequence of CNN is the 20 input variables. See the Supplementary Material for more details.

3. Results

As presented in Table 1, B_{abs} corrected by different correction methods can cause variation in the observation-based MAC_{BC} , with mean values differing by roughly a factor of two. Therefore, we considered two versions of model, one using B_{abs} corrected by Li et al. [40] as input variables and the numerator in Equation (2) to calculate MAC_{BC} , and another using B_{abs} corrected by Bond et al. [41] to build the model. In the main text, we focus on the results from the former model as Li et al. [40] observed good agreement between corrected filter-based B_{abs} and reference B_{abs} using photoacoustic instruments. The evaluation of the latter model can be found in the Supplementary Material (Tables S9 and S10, Figures S10–S12). We also input B_{abs} corrected by Bond et al. [41] into the first model (without rebuilding the model from scratch) and compare the predicted MAC_{BC} against our original results (Tables S11–S13).

3.1. Performance of Models on the Training and Test Datasets

The parameters of the seven data analytical techniques are trained using 7321 TCAP records and tested using 1830 remaining TCAP records. The results of coefficient values output from the models can be found in the Supplementary Material (Figure S6). The performance of the techniques is assessed according to statistical criteria, including coefficient of determination (R^2) and mean square error (MSE). As seen in Table 4, all techniques are able to predict the changes of MAC_{BC} to some extent, although some models are better than others based on these statistical criteria. We have also included results considering the use of the “standard” assumptions ($\text{MAC}_{\text{BC}} = 7.5 \text{ m}^2 \text{ g}^{-1}$ and $\text{AAE}_{\text{BC}} = 1$) to determine MAC_{BC} at 870 nm; based on the statistical criteria, this is unequivocally the worst-performing approach to predict the hourly MAC_{BC} for any dataset.

Table 4. The performance of the seven data analytical techniques along with the “standard” assumption of MAC_{BC} ^a.

Approach	R^2				MSE			
	Training	Test	Validation (CACTI)	Validation (FIREX)	Training	Test	Validation (CACTI)	Validation (FIREX)
OLS	0.467	0.465	0.272	0.50 ^b	0.932	0.961	1.738	9×10^{21} ^b
Forward stepwise	0.459	0.461	0.218	0.354	0.965	1.015	2.148	6.442
Backward stepwise	0.468	0.468	0.283	0.343	0.932	0.944	2.538	5.914
LASSO	0.420	0.412	0.201	0.321	0.964	0.970	2.175	4.977
SVM	0.781	0.766	0.247	0.413	0.284	0.375	2.991	2.299
ANN	0.798	0.778	0.251	0.367	0.227	0.370	2.715	6.401
CNN	0.735	0.728	0.222	0.390	0.488	0.579	3.440	3.400
“Standard”	0.246	0.279	<0 ^c	0.258	2.083	2.075	4.242	7.018

^a The results are generated with lower time resolutions (TCAP and CACTI: hourly-averaged data (687 and 480 timestamps, respectively); FIREX: fire-averaged data (30 fires)). The results generated using minute-basis can be found in Table S6; ^b It is important not to misinterpret the R^2 calculated by the OLS model implemented on the FIREX dataset, because an extreme value of mean square error (MSE) (9×10^{21}) is obtained from this analysis; ^c A negative coefficient of determination (R^2) is because the “standard” MAC_{BC} ($4.74 \text{ m}^2 \text{ g}^{-1}$ at 870 nm) is greater than the “true” response of MAC_{BC} at the site ($2.94 \pm 1.68 \text{ m}^2 \text{ g}^{-1}$).

During the training phase, ANN features the highest statistical accuracy ($R^2 = 0.798$ and $\text{MSE} = 0.227$), although the second and third highest accuracy techniques (SVM and CNN) are not very different from ANN. However, traditional statistical approaches (OLS, forward and backward regressions) and LASSO have lower accuracies ($R^2 \approx 0.450$ and $\text{MSE} \approx 0.950$). The complicated model structures of ANN, SVM and CNN allow them to better capture nonlinear relationships between MAC_{BC} and the aerosol properties than traditional statistical models. Moreover, the slightly better performance of ANN than CNN is consistent with our expectation, since the ANN algorithm has a deeper network structure (more parameters in the network) than the CNN algorithm.

During the testing phase, there are no major differences in the techniques’ performance compared to the results from the training phase (i.e., R^2 and MSE are similar for both phases). Overall, the similar

performance of the models on training and test datasets suggests that the models have achieved a satisfactory generalization capability for new and “unseen” MAC_{BC} data at the TCAP observational site.

Figure 1a,b shows the time series of predicted MAC_{BC} by different techniques from the training and test sets, respectively, as well as the “true” MAC_{BC} . We only present the results of OLS, LASSO, and SVM in this figure; results for the other models are provided in Figures S7 and S8. We apply the same time resolution to generate the figures as we calculated the results in Table 4. For both the training and test sets, the predicted MAC_{BC} follows closely to “true” MAC_{BC} with no signs of systematic errors.

3.2. Performance of Models on the Independent Validation Datasets

As the ultimate goal of comparing different techniques is to evaluate their generalizability on new datasets other than TCAP data, we apply all techniques to two independent validation sets: field data from the CACTI campaign and laboratory data from the FIREX campaign.

The CACTI campaign deployed the same instrumentation as the TCAP campaign but was conducted at a different observational site, resulting in different aerosol properties (Table 1). Hence, the CACTI data provide us with an opportunity to explore how the models may be applicable to different sampling locations and ambient aerosol sources. As seen in Figure 1c, Figure 2a and Figure S9, the seven approaches are capable of predicting MAC_{BC} within a factor of two of the “true” MAC_{BC} (i.e., predicted errors $< \pm 50\%$) for most of the sampling days, which suggests that our models may be valid for new datasets. However, there are some times during which our models are consistently biased high (e.g., 6–9 December) or low (e.g., a ~6 h window on 20 December). Yuan et al. [16] demonstrate that MAC_{BC} can vary with BC mixing state, so these discrepancies could be the result of an aerosol mixing state that differs from the training data. We will explore the role of mixing state and other potential explanatory variables for these systematic biases in a forthcoming companion paper.

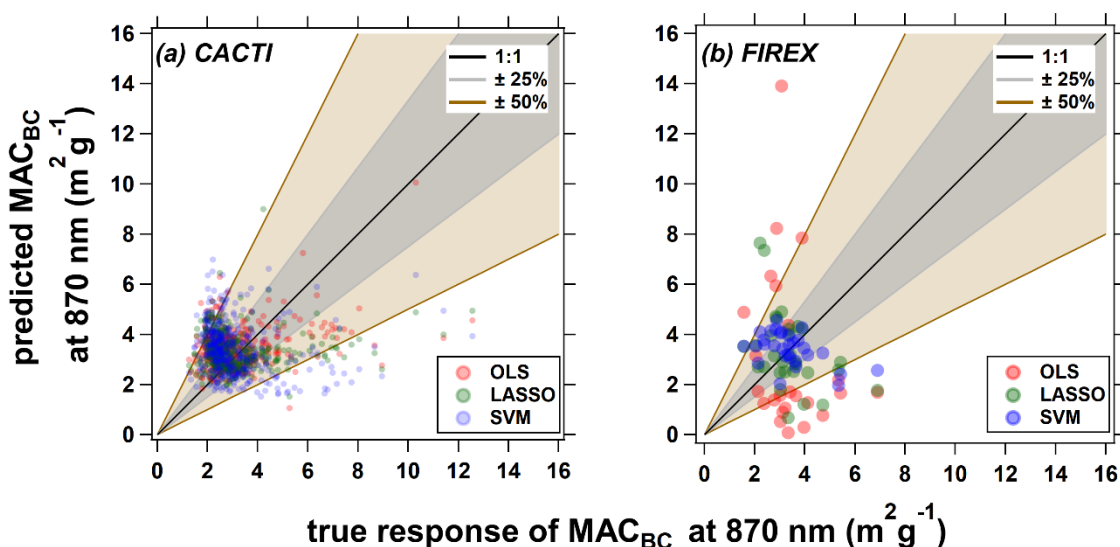


Figure 2. Comparison between the predicted MAC_{BC} and “true” MAC_{BC} using independent validation CACTI (a) and FIREX (b) datasets. Only three representative techniques (OLS, LASSO, and SVM) are included in the figure. The results of the other techniques can be found in Figure S9. The grey and brown shaded regions represent predicted errors $< \pm 25\%$ and $< \pm 50\%$ referenced to the line of perfect agreement (1:1).

We quantitatively assess the model scatter by calculating the percentage of points that lay within a factor of two of the true response (Table S7). Overall, over 80% of the MAC_{BC} from CACTI can be estimated by our models within a factor of two, and the OLS yields the greatest percentage (92.4%). When comparing the R^2 and MSE across the approaches for the CACTI dataset (Table 4), we find that the traditional statistical approaches tend to yield similar or even better results than the deep learning

approaches, which is different from the TCAP results. However, in the case of these CACTI data, the performance of all models is arguably similar (see also Figure 3 and relevant discussion).

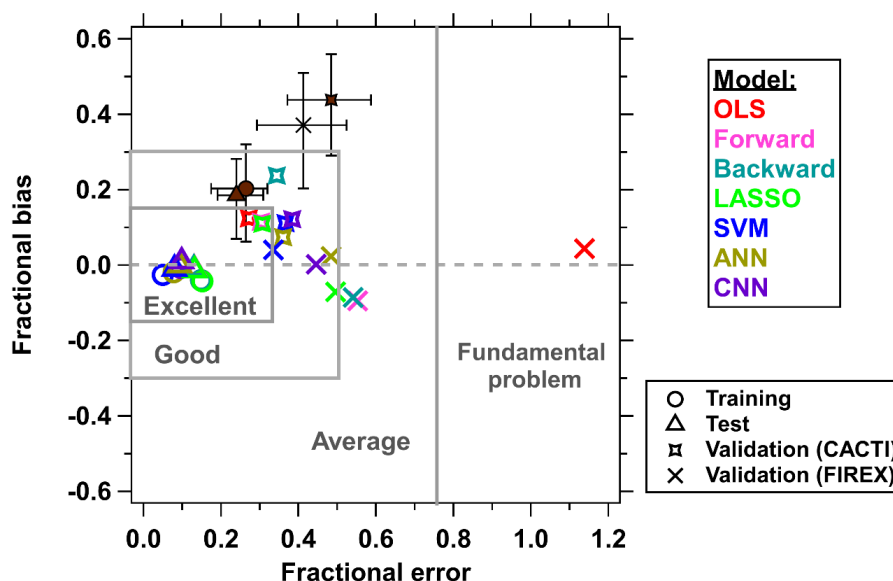


Figure 3. Fractional bias against fractional error for different combinations of datasets and data analytical approaches. When deriving the points with error bars, the “standard” MAC_{BC} from Bond and Bergstrom [6] is used as predicted MAC_{BC}, and “true” MAC_{BC} is used as desired output of MAC_{BC}.

We also applied these models to our FIREX dataset (laboratory study of biomass burning) to determine how our models work for aerosols that are vastly different than the ambient datasets, e.g., the mean SSA₅₃₂ is 0.66 ± 0.32 for the FIREX dataset, whereas SSA₅₃₂ for the ambient datasets tends to be >0.95 (Table 1). Consequently, the FIREX data provide a challenge for some, but not all, of our models. This could, at least in part, be due to an inadequate representation of chemical composition by B_{abs} and B_{scat} for these biomass burning aerosols. However, it could also be related to differences between the mixing state of the aerosols, the presence of tar balls, or the fact that we are extrapolating the model away from the (presumably) aged BC in the training data to fresh BC in the emissions data. We will further investigate some of these factors in our forthcoming companion paper. Nevertheless, our SVM model does not appear to be very sensitive to the observed differences in aerosol optical properties or the likely differences in aerosol chemical composition (Figure 2).

3.3. Comparison across Datasets

We next adopt the model performance criteria proposed by Morris et al. [81] to quantitatively inter-compare our models across different datasets. According to the criteria, the models can be classified as “excellent”, “good”, “average”, and “[having a] fundamental problem”, based on the fractional error and fractional bias. The fractional error is defined as:

$$\frac{2}{N} \sum \frac{|P_i - T_i|}{P_i + T_i} \tag{5}$$

while fractional bias is defined as:

$$\frac{2}{N} \sum \frac{P_i - T_i}{P_i + T_i} \tag{6}$$

where P_i is the predicted MAC_{BC} and T_i is the “true” MAC_{BC}.

The results for our models are provided in Figure 3 (see Table S8 for the detailed results). All models applied on the training and test datasets meet the criteria for “excellent”, with fractional bias close to zero and fractional error around 0.1. The models also have similar performance for the CACTI dataset,

falling along the “excellent”/“good” boundary with fractional error of roughly 0.35 and fractional bias of roughly 0.15. For the domains around the boundary, the traditional statistical approaches applied on the CACTI data lay inside the “excellent” region, while the SVM and neural networks fall outside (but within the “good” region), which is consistent with the results of R^2 and MSE in Table 4. Performance begins to diverge on the FIREX data set; while the fractional biases are within ± 0.10 , the fractional errors range from roughly 0.35 (SVM) to roughly 1.15 (OLS), with the majority falling around the “good”/“average” boundary. The greater value of fractional error in the case of the OLS approach is because it tends to predict a MAC_{BC} value that is the most different compared to the true value (e.g., Figures 1d and 2b). While the performance of the other approaches remains statistically worse than those applied to the training and test datasets (Table 4), their ability to predict MAC_{BC} still seems reasonable for FIREX (either “good” or “average”).

For the datasets included in Figure 3, we also compute the fractional error and fractional bias using the “standard” MAC_{BC} values ($4.74 \pm 0.76 \text{ m}^2 \text{ g}^{-1}$, 870 nm) as the “predicted” MAC_{BC} . The “standard” MAC_{BC} is derived using MAC_{BC} (550 nm) of $7.5 \pm 1.2 \text{ m}^2 \text{ g}^{-1}$ (recommended by Bond and Bergstrom [6]) and AAE of 1, following Equation (1). The “standard” MAC_{BC} results in greater fractional bias and fractional error than most of the combinations of our dataset and approaches. The better performance of our approaches could be explained by their ability to incorporate temporally varying input parameters in the calculation of MAC_{BC} . Such variations in MAC_{BC} cannot be captured by the “standard” MAC_{BC} since it is reported as a single value with a standard deviation and was recommended for fresh and uncoated BC particles.

We explore bias as a function of time in the Supplementary Materials. Specifically, we compare the time series residuals (residuals = predicted MAC_{BC} – $MAC_{BC,true}$) computed by our models and the one using the “standard” MAC_{BC} (see Table S15 and Figure S13). We find that our models have residuals closer to zero than the standard approach for all datasets. However, our models have a greater standard deviation than the standard approach for the validation datasets, which may be related to the fact that the standard approach is generally biased high. Conversely, the SVM model for the CACTI dataset, for example, has periods where it is biased high (e.g., that 6–9 December window), biased low (e.g., 1–3 December), and has practically no bias (e.g., 15–17 December). Nevertheless, these results suggest that our models, while uncertain, may have the potential to capture temporal variability in MAC_{BC} .

Consequently, our results indicate that the proposed data analytical approaches work reasonably well on datasets that are not very different from our training dataset (e.g., ambient measurements from different environmental conditions) and the MAC_{BC} estimated by these approaches are very likely to provide better predictions of BC mass than using a constant value of MAC_{BC} . Furthermore, under this condition, the performance of different models is not very different. For datasets that are vastly different than our training dataset (e.g., aerosols with high BrC concentrations as in biomass burning), the machine learning approaches (e.g., SVM and ANN) yield lower fractional errors and fractional biases closer to zero compared to the traditional statistical approaches and the “standard” MAC_{BC} . Moreover, our FIREX dataset (laboratory emissions) resulted in a mean SSA of 0.66 while field measurements of the aerosol optical properties in biomass burning plumes tend to report SSA > 0.9 (e.g., [82–85] and references within [86]); thus, it is possible that we have used an “extreme” dataset that may not be realistic for real world biomass burning plumes. In other words, our models may have limited utility for measurements of fresh emissions from combustion sources.

3.4. Comparison to a Different Correction for B_{abs}

In this section, we provide a brief discussion of the results using the Bond et al. [41] correction on the model outputs. We rebuilt the models using this correction, providing values of R^2 and MSE in Table S9 and fractional error and fraction bias in Table S10. Compared to the results using the models built using the Li et al. [40] correction, the values of R^2 are slightly better for the training, test, CACTI, and some of the FIREX data in these rebuilt models, but MSE is, on average, about a factor of 5 larger.

Likewise, the fractional error is roughly 0.03 greater (on average) for the rebuilt models in the hourly training and test data, while in the validation data, the fractional error is sometimes lower (e.g., 0.484 in the original ANN model and 0.418 in the rebuilt ANN model). There are similar results for fractional bias, with the original models having slightly lower fractional bias for the training and test hourly data, while the comparison for the validation data sets are more varied (considering both magnitude and direction). Interestingly, the rebuilt machine learning models all have slightly lower fractional bias in the CACTI data but slightly larger fractional bias in the FIREX data.

To understand how the choice of B_{abs} correction affects model output when simply running the models, we corrected B_{abs} values using Bond et al. [41] and input these into our original models built using the Li et al. [40] correction; statistical metrics (R^2 , MSE, fractional error, and fractional bias) in Tables S11 and S12. When comparing these to the original model results presented in earlier sections, R^2 is typically lower using the Bond et al. [41] correction. Interestingly, the Li et al. [40] correction nearly always results in better performance (lower MSE, lower fractional error, fractional bias closer to zero) for the CACTI data, while the Bond et al. [41] correction nearly always has less error for the FIREX data (fractional bias varies by model).

As a further comparison between these two correction equations, we conducted t-tests using the 1-min data to determine if their mean MAC_{BC} values were significantly different from each other in the original models (using the Li et al. [40] correction to build the model); results are provided in Table S13. All models were significantly different ($p < 0.05$) for the TCAP dataset, for two models in the CACTI dataset (OLS, backward stepwise), and for all but the SVM and ANN models in the FIREX dataset. Some of these statistical differences may be related to the relatively large sample sizes (>6700), because the relative percent different between the means ranged from 0.0 to 5.2%. Moreover, these differences are small compared to the estimated uncertainties in the model outputs (see Section 4).

Consequently, a holistic consideration of these metrics suggests that the Li et al. [40] correction appears to result in better performance when building the models. However, there is no strong evidence that suggests either correction model is better when running the models to predict MAC_{BC} .

4. Discussion

In this section, we present an uncertainty assessment of MAC_{BC} quantified by our data analytical models. For comparison, we also compute the uncertainties of MAC_{BC} derived by various empirical and theoretical approaches. At the end of this section, we provide some insights for the application of these approaches.

Empirical MAC_{BC} determined from direct measurements of BC mass and absorption is probably the least controversial, as we can estimate a 24% uncertainty in MAC_{BC} via quadrature, based on 10% in B_{abs} [40] and 22% in SP2-measured BC mass concentrations [87]. However, for the data analytical approaches developed in this paper, the “standard” MAC_{BC} values reviewed by Bond and Bergstrom [6], and the theoretical approaches, each of the contributing terms may propagate uncertainties to the derived MAC_{BC} . Therefore, we provide the following uncertainty estimations using both quadrature (because all models will be affected by random error in the input variables) and Monte Carlo approaches [88–90].

4.1. Uncertainty in Data Analytical Approaches

All models will have some uncertainty due to random error in the model inputs. To compare to the uncertainty of empirical MAC_{BC} , we consider B_{abs} to have 10% uncertainty [40] and B_{scat} to have 10% uncertainty [91], and we assume that all of the parameters related to aerosol size have 10% uncertainty, based on [92] for SMPS and [93] for APS. Using quadrature, the models including all 20 input parameters (OLS, SVM, ANN, and CNN) have an estimated uncertainty of roughly 45%, while LASSO, with only 9 input parameters, has an estimated uncertainty of roughly 30%. This is likely an imperfect approach because the machine learning models, in particular, are very complex and because some of the model inputs are related to each other. For example, the total number (volume)

concentration is equal to the sum of the number (volume) concentrations within the coarse bins, while the fractional number concentration is equal to the number (volume) concentration in a given bin divided by the total number (volume) concentration. Moreover, there are analogous equations to Equation (1) that describe the relationship between B_{scat} and B_{abs} across the different wavelengths.

We also consider an uncertainty analysis using the bootstrap approach (a nonparametric Monte Carlo analysis [89,90]). In our analysis, bootstrap mimics the process of obtaining new data from the TCAP site, and these data have the same underlying distribution as our original TCAP dataset. We then rebuild different data analytical models using the new datasets. The relative standard deviation (i.e., standard deviation/mean) of R^2 and MSE derived by the new models are presented in Table 5. Using the resampled datasets, the rebuilt models based for SVM and CNN tend to be more stable (i.e., lower relative standard deviations) than the traditional regression approaches and the other machine learning approaches when considering the validation datasets, so there appears to be lower model uncertainty when applying those models to new datasets.

Table 5. The relative standard deviation of R^2 and MSE derived by the bootstrap resampling approach^a.

Approach	R^2				MSE			
	Training	Test	Validation (CACTI)	Validation (FIREX)	Training	Test	Validation (CACTI)	Validation (FIREX)
OLS	3.39%	9.28%	/ ^b	/ ^b	4.03%	12.60%	/ ^b	/ ^b
Forward stepwise	2.86%	9.20%	5.56%	11.73%	3.54%	12.10%	35.10%	21.35%
Backward stepwise	3.11%	9.26%	33.06%	/ ^b	3.91%	12.49%	52.51%	/ ^b
LASSO	3.29%	9.34%	/ ^b	20.98%	5.45%	13.11%	/ ^b	28.74%
SVM	5.16%	5.84%	8.06%	3.98%	16.67%	17.87%	19.03%	6.49%
ANN	3.80%	2.48%	5.12%	4.81%	16.22%	15.15%	10.82%	35.77%
CNN	3.57%	5.51%	7.00%	14.72%	9.57%	18.95%	6.39%	18.85%

^a The Bootstrap resampling process is repeated five times. The mean and standard deviation of R^2 and MSE can be found in Table S14; ^b the result is not interpretable, because extreme values of MSE are computed when using the new datasets to rebuild the model. We have also excluded the R^2 value for these cases.

In Figure 2 and Figure S9 and in Table S7, we demonstrate that all of our data analytical approaches can predict the majority (>63%) of the empirical MAC_{BC} within $\pm 50\%$, with one exception (the OLS model for the FIREX data). Considering that none of these models include aerosol composition or explicitly consider aerosol optical properties such as SSA and AAE, we argue that this performance is robust. Moreover, all of these models may capture temporal variability due to their use of time-resolved input data. Consequently, our models provide considerable improvements over the assumption of a constant value of MAC_{BC} .

However, given that our models have not been evaluated using, e.g., urban aerosols, the model performance for some atmospheric environments is unknown. Considering that the inherent limitation of machine learning approaches is that they tend to perform worse for extrapolation than interpolation problems [94], our models may only have utility for environments that are similar to the training data (e.g., highly aged/mixed BC).

4.2. Uncertainty in the “Standard Approach”

As discussed previously, the standard approach to calculate MAC_{BC} is based on MAC_{BC} at $550 \text{ nm} = 7.5 \text{ m}^2 \text{ g}^{-1}$ and $AAE = 1$, although discrepancies with both of these values exist in the literature. Therefore, uncertainties may exist when extending MAC_{BC} to 870 nm . We conduct a Monte Carlo simulation to quantify this uncertainty, assuming that MAC_{BC} values are normally distributed about $7.5 \text{ m}^2 \text{ g}^{-1}$ with a standard deviation of 1.2 and that AAE is uniformly distributed between 0.6 and 1.6. Using 10,000 iterations, we estimate an uncertainty of roughly 21% in MAC_{BC} at 870 nm based on the standard assumption. However, it is also worth noting that this represents the uncertainty for a

constant value, and therefore, cannot capture any spatial or temporal variability in the “true” MAC_{BC} , which may be around 50% following Table 1.

4.3. Uncertainty in Theoretical Approaches

It is difficult to quantify uncertainties in theoretical approaches due to the number of requisite assumptions. For any of the number of optical models that could be used to predict MAC_{BC} (e.g., [9] and references therein), one requires the BC mixing state, BC complex refractive index, and BC density, along with the aerosol size distribution and a representation of BC-containing particles across that distribution. Moreover, the choice of a given optical model itself depends on the BC morphology. In the literature, Curci et al. [95] estimate an uncertainty of ~30–35% in calculations of SSA and aerosol optical depth due to different model input assumptions, and mixing state plays a large role in these uncertainties. Similarly, Tuccella et al. [96] find the BC direct radiative effect varies by 50%, depending on the assumed mixing state.

We attempt to estimate uncertainty in MAC_{BC} from a theoretical approach as follows. We assume that the calculation incorporates an empirical aerosol size distribution, with 10% uncertainty. We conservatively estimate that mixing state contributes ~25% uncertainty based on Curci et al. [95] and Tuccella et al. [96], and we consider the range in recommended complex refractive index and density recommended by Bond and Bergstrom [6] to each contribute ~10% uncertainty. We have no *a priori* estimates for uncertainty related to the apportionment of BC across the aerosol size distribution or BC morphology, so we conservatively estimate 10% uncertainty for each. Therefore, we estimate roughly 34% uncertainty (via quadrature) in predicted MAC_{BC} from a theoretical approach.

4.4. Recommendations

There are benefits and limitations to each of the approaches discussed above. Experimental MAC_{BC} requires the concurrent measurements of B_{abs} and M_{BC} , which is not often available at long-term monitoring sites. Our data analytical approaches may capture temporal variations in MAC_{BC} using empirical observations of B_{abs} , B_{scat} , and aerosol size distributions, but they have the largest estimated uncertainties propagated from the input variables and their performance for different aerosol types than what we consider here is unknown. Conversely, the standard approach has the lowest estimated uncertainty, but because this uses a constant value, it will never capture any temporal variability in MAC_{BC} at a given location. Moreover, while the theoretical approaches are physics-based models, each require major assumptions that may not be constrained by empirical observations.

Considering all of these factors, we recommend the use of our SVM model for further exploration in the calculation of MAC_{BC} . Any of our models can predict MAC_{BC} using data common to ground-based monitoring networks, but the SVM model appears to be the least sensitive to different aerosol types, based on a holistic consideration of R^2 , MSE, fractional error, and fractional bias for both evaluation datasets, and its performance is relatively stable when rebuilding the model using the bootstrap approach. Moreover, SVM is the least sensitive to variations in empirical limitations (e.g., if a single wavelength of B_{abs} and B_{scat} are available, if only an APS is available), which will be discussed in a forthcoming companion study. Finally, while an argument could be made for the ANN or CNN models, the SVM model is more computationally efficient than those other two machine learning models, which suggests utility in processing large datasets from long-term field sites.

However, we add the caveat that when applying our model to aerosols whose properties are very different from our training datasets (see Table 1), the model performance is unknown and could result in substantial error. For example, the analysis of fresh aerosol emissions from fossil fuel combustion (such as diesel engines) may result in similar issues to those in some of our FIREX data; therefore, when the aerosols are likely dominated by BC ($SSA < 0.5$), using the standard assumptions ($MAC_{BC} = 7.5 \text{ m}^2 \text{ g}^{-1}$; $AAE_{BC} = 1$) may be more appropriate than any of our models.

5. Conclusions

In this exploratory study we apply OLS regression, forward stepwise regression, backward stepwise regression, LASSO, SVM, ANN and CNN to three datasets (TCAP and CACTI ambient data and FIREX laboratory biomass burning data), in order to study their applicability to estimate MAC_{BC} of black carbon aerosols. The inputs of the model simply include B_{abs} , B_{scat} , and aerosol size distribution. The nonlinear approaches (SVM, ANN and CNN) perform much better than the traditional regressions and the LASSO approach for the training and test set (TCAP data). When applying the approaches on independent validation datasets, the models can capture the overall trends in MAC_{BC} , although there is some degree of decreased performance for some of the data. Arguably, no model is unequivocally the best for the CACTI data when holistically considering R^2 , MSE, fractional error, and fractional bias. However, SVM and CNN appear to perform the best for the FIREX data, and therefore, may be less sensitive to extreme variations in aerosol optical properties. Notably, there appears to be no obvious difference in model outputs when the input B_{abs} data are corrected using a different approach for any of the models.

Our uncertainty analysis (via quadrature) suggests ~24% uncertainty in empirically-derived MAC_{BC} , compared to ~21% in the standard approach, ~34% in a theoretical formulation, and up to ~45% in our statistical models due to the random error of our 20 input variables. However, the standard approach cannot account for temporal variations in MAC_{BC} , which are on the order of $\pm 50\%$ (Table 1), while numerous assumptions are required for any theoretical formulations. Therefore, even though our analysis suggests the highest propagated uncertainty, we recommend the use of our SVM model for the calculation of MAC_{BC} for long-term field monitoring sites because it requires no assumptions, and it incorporates temporally varying aerosol properties and therefore, has the potential to capture temporal variations in MAC_{BC} . Ultimately, we anticipate that the exploratory approaches in this work can pave the way for design of data-driven models to estimate complex aerosol properties in future studies.

In a forthcoming paper, we will discuss the extension of the predicted MAC_{BC} to different wavelengths. The extension requires underlying assumptions to establish the relationship among MAC_{BC} , wavelength, and AAE_{BC} (Equation (1)), where AAE_{BC} can be assumed either to be constant or to vary with local aerosol properties. We will also explore how the predicted MAC_{BC} are impacted by varying atmospheric parameters and changing input variables (such as using single-wavelength B_{abs} , rather than three-wavelength B_{abs}), in order to quantify to what extent our model results are sensitive and robust.

Supplementary Materials: The supplementary materials are available online at <http://www.mdpi.com/2073-4433/11/11/1185/s1>.

Author Contributions: Conceptualization, H.L. and A.A.M.; Formal analysis, H.L.; Funding acquisition, A.A.M.; Methodology, H.L. and A.A.M.; Project administration, A.A.M.; Software, H.L.; Writing—first draft, H.L.; Writing—review & editing, H.L. and A.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NOAA Climate Program Office Grant NA16OAR4310109.

Acknowledgments: The ambient data at TCAP and CACTI sites were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Office of Biological and Environmental Research. The authors would like to thank Rongjun Qin (The Ohio State University) and Adam Varble (Pacific Northwest National Laboratory) for useful discussions on machine learning and the CACTI campaign, respectively. The authors also thank the anonymous reviewers for constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Code and Data Availability: The model and modules developed in this work are available at <https://doi.org/10.5281/zenodo.3967833> [60]. The TCAP and CACTI aerosol products are available at <https://www.archive.arm.gov/discovery/>. Datasets were last accessed on 6 September 2019 and 1 November 2019, respectively. The FIREX data are available at <https://www.esrl.noaa.gov/csl/projects/firex/firelab/>.

References

1. Andreae, M.O.; Gelencsér, A. Black carbon or brown carbon? The nature of light-absorbing carbonaceous aerosols. *Atmos. Chem. Phys.* **2006**, *6*, 3131–3148. [[CrossRef](#)]
2. Cooke, W.F.; Wilson, J.J.N. A global black carbon aerosol model. *J. Geophys. Res. Atmos.* **1996**, *101*, 19395–19409. [[CrossRef](#)]
3. Bond, T.C.; Doherty, S.J.; Fahey, D.W.; Forster, P.M.; Berntsen, T.; DeAngelo, B.J.; Flanner, M.G.; Ghan, S.; Kärcher, B.; Koch, D.; et al. Bounding the role of black carbon in the climate system: A scientific assessment. *J. Geophys. Res. Atmos.* **2013**, *118*, 5380–5552. [[CrossRef](#)]
4. Hansen, A.D.A.; Rosen, H.; Novakov, T. The aethalometer—An instrument for the real-time measurement of optical absorption by aerosol particles. *Sci. Total Environ.* **1984**, *36*, 191–196. [[CrossRef](#)]
5. Ramanathan, V.; Carmichael, G. Global and regional climate changes due to black carbon. *Nat. Geosci.* **2008**, *1*, 221–227. [[CrossRef](#)]
6. Bond, T.C.; Bergstrom, R.W. Light Absorption by Carbonaceous Particles: An Investigative Review. *Aerosol Sci. Technol.* **2006**, *40*, 27–67. [[CrossRef](#)]
7. Lack, D.A.; Moosmüller, H.; McMeeking, G.R.; Chakrabarty, R.K.; Baumgardner, D. Characterizing elemental, equivalent black, and refractory black carbon aerosol particles: A review of techniques, their limitations and uncertainties. *Anal. Bioanal. Chem.* **2014**, *406*, 99–122. [[CrossRef](#)]
8. Kondo, Y.; Sahu, L.; Kuwata, M.; Miyazaki, Y.; Takegawa, N.; Moteki, N.; Imaru, J.; Han, S.; Nakayama, T.; Oanh, N.T.K.; et al. Stabilization of the Mass Absorption Cross Section of Black Carbon for Filter-Based Absorption Photometry by the use of a Heated Inlet. *Aerosol Sci. Technol.* **2009**, *43*, 741–756. [[CrossRef](#)]
9. Moosmüller, H.; Chakrabarty, R.K.; Arnott, W.P. Aerosol light absorption and its measurement: A review. *J. Quant. Spectrosc. Radiat. Transf.* **2009**, *110*, 844–878. [[CrossRef](#)]
10. Sheridan, P.J.; Arnott, W.P.; Ogren, J.A.; Andrews, E.; Atkinson, D.B.; Covert, D.S.; Moosmüller, H.; Petzold, A.; Schmid, B.; Strawa, A.W.; et al. The Reno Aerosol Optics Study: An Evaluation of Aerosol Absorption Measurement Methods. *Aerosol Sci. Technol.* **2005**, *39*, 1–16. [[CrossRef](#)]
11. Liu, F.; Yon, J.; Fuentes, A.; Lobo, P.; Smallwood, G.J.; Corbin, J.C. Review of recent literature on the light absorption properties of black carbon: Refractive index, mass absorption cross section, and absorption function. *Aerosol Sci. Technol.* **2020**, *54*, 33–51. [[CrossRef](#)]
12. Nordmann, S.; Birmili, W.; Weinhold, K.; Müller, K.; Spindler, G.; Wiedensohler, A. Measurements of the mass absorption cross section of atmospheric soot particles using Raman spectroscopy. *J. Geophys. Res. Atmos.* **2013**, *118*, 12–075. [[CrossRef](#)]
13. Gyawali, M.; Arnott, W.; Zaveri, R.; Song, C.; Flowers, B.; Dubey, M.; Setyan, A.; Zhang, Q.; China, S.; Mazzoleni, C.; et al. Evolution of Multispectral Aerosol Absorption Properties in a Biogenically-Influenced Urban Environment during the CARES Campaign. *Atmosphere* **2017**, *8*, 217. [[CrossRef](#)]
14. Koch, D.; Schulz, M.; Kinne, S.; McNaughton, C.; Spackman, J.R.; Balkanski, Y.; Bauer, S.; Berntsen, T.; Bond, T.C.; Boucher, O.; et al. Evaluation of black carbon estimations in global aerosol models. *Atmos. Chem. Phys.* **2009**, *9*, 9001–9026. [[CrossRef](#)]
15. Subramanian, R.; Kok, G.L.; Baumgardner, D.; Clarke, A.; Shinozuka, Y.; Campos, T.L.; Heizer, C.G.; Stephens, B.B.; de Foy, B.; Voss, P.B.; et al. Black carbon over Mexico: The effect of atmospheric transport on mixing state, mass absorption cross-section, and BC/CO ratios. *Atmos. Chem. Phys.* **2010**, *10*, 219–237. [[CrossRef](#)]
16. Yuan, J.; Modini, R.L.; Zanatta, M.; Herber, A.; Müller, T.; Wehner, B.; Poulain, L.; Tuch, T.; Baltensperger, U.; Gysel-Ber, M. Variability in the mass absorption cross-section of black carbon (BC) aerosols is driven by BC internal mixing state at a central European background site (Melpitz, Germany) in winter. *Atmos. Chem. Phys. Discuss.* **2020**, 1–36. [[CrossRef](#)]
17. Kahnert, M. Modelling the optical and radiative properties of freshly emitted light absorbing carbon within an atmospheric chemical transport model. *Atmos. Chem. Phys.* **2010**, *10*, 1403–1416. [[CrossRef](#)]
18. Zangmeister, C.D.; You, R.; Lunny, E.M.; Jacobson, A.E.; Okumura, M.; Zachariah, M.R.; Radney, J.G. Measured in-situ mass absorption spectra for nine forms of highly-absorbing carbonaceous aerosol. *Carbon* **2018**, *136*, 85–93. [[CrossRef](#)]
19. Liu, C.; Chung, C.E.; Yin, Y.; Schnaiter, M. The absorption Ångström exponent of black carbon: From numerical aspects. *Atmos. Chem. Phys.* **2018**, *18*, 6259–6273. [[CrossRef](#)]

20. Kirchstetter, T.W.; Novakov, T.; Hobbs, P.V. Evidence that the spectral dependence of light absorption by aerosols is affected by organic carbon. *J. Geophys. Res.* **2004**, *109*, D21208. [[CrossRef](#)]
21. Gyawali, M.; Arnott, W.P.; Lewis, K.; Moosmüller, H. In situ aerosol optics in Reno, NV, USA during and after the summer 2008 California wildfires and the influence of absorbing and non-absorbing organic coatings on spectral light absorption. *Atmos. Chem. Phys.* **2009**, *9*, 8007–8015. [[CrossRef](#)]
22. Saleh, R.; Cheng, Z.; Atwi, K. The Brown–Black Continuum of Light-Absorbing Combustion Aerosols. *Environ. Sci. Technol. Lett.* **2018**, *5*, 508–513. [[CrossRef](#)]
23. Cho, C.; Kim, S.-W.; Lee, M.; Lim, S.; Fang, W.; Gustafsson, Ö.; Andersson, A.; Park, R.J.; Sheridan, P.J. Observation-based estimates of the mass absorption cross-section of black and brown carbon and their contribution to aerosol light absorption in East Asia. *Atmos. Environ.* **2019**, *212*, 65–74. [[CrossRef](#)]
24. Hoffer, A.; Tóth, Á.; Pósfai, M.; Chung, C.E.; Gelencsér, A. Brown carbon absorption in the red and near-infrared spectral region. *Atmos. Meas. Tech.* **2017**, *10*, 2353–2359. [[CrossRef](#)]
25. Pósfai, M.; Gelencsér, A.; Simonics, R.; Arató, K.; Li, J.; Hobbs, P.V.; Buseck, P.R. Atmospheric tar balls: Particles from biomass and biofuel burning. *J. Geophys. Res. Atmos.* **2004**, *109*. [[CrossRef](#)]
26. Corbin, J.C.; Czech, H.; Massabò, D.; de Mongeot, F.B.; Jakobi, G.; Liu, F.; Lobo, P.; Mennucci, C.; Mensah, A.A.; Orasche, J.; et al. Infrared-absorbing carbonaceous tar can dominate light absorption by marine-engine exhaust. *NPJ Clim. Atmos. Sci.* **2019**, *2*, 12. [[CrossRef](#)]
27. Chakrabarty, R.K.; Moosmüller, H.; Chen, L.-W.A.; Lewis, K.; Arnott, W.P.; Mazzoleni, C.; Dubey, M.K.; Wold, C.E.; Hao, W.M.; Kreidenweis, S.M. Brown carbon in tar balls from smoldering biomass combustion. *Atmos. Chem. Phys.* **2010**, *10*, 6363–6370. [[CrossRef](#)]
28. Adachi, K.; Buseck, P.R. Atmospheric tar balls from biomass burning in Mexico. *J. Geophys. Res.* **2011**, *116*, D05204. [[CrossRef](#)]
29. Barnard, J.C.; Volkamer, R.; Kassianov, E.I. Estimation of the mass absorption cross section of the organic carbon component of aerosols in the Mexico City Metropolitan Area. *Atmos. Chem. Phys.* **2008**, *8*, 6665–6679. [[CrossRef](#)]
30. Cazorla, A.; Bahadur, R.; Suski, K.J.; Cahill, J.F.; Chand, D.; Schmid, B.; Ramanathan, V.; Prather, K.A. Relating aerosol absorption due to soot, organic carbon, and dust to emission sources determined from in-situ chemical measurements. *Atmos. Chem. Phys.* **2013**, *13*, 9337–9350. [[CrossRef](#)]
31. Caponi, L.; Formenti, P.; Massabó, D.; Di Biagio, C.; Cazaunau, M.; Panguì, E.; Chevaillier, S.; Landrot, G.; Andreae, M.O.; Kandler, K.; et al. Spectral- and size-resolved mass absorption efficiency of mineral dust aerosols in the shortwave spectrum: A simulation chamber study. *Atmos. Chem. Phys.* **2017**, *17*, 7175–7191. [[CrossRef](#)]
32. García Fernández, C.; Picaud, S.; Devel, M. Calculations of the mass absorption cross sections for carbonaceous nanoparticles modeling soot. *J. Quant. Spectrosc. Radiat. Transf.* **2015**, *164*, 69–81. [[CrossRef](#)]
33. Zanutta, M.; Laj, P.; Gysel, M.; Baltensperger, U.; Vratolis, S.; Eleftheriadis, K.; Kondo, Y.; Dubuisson, P.; Winiarek, V.; Kazadzis, S.; et al. Effects of mixing state on optical and radiative properties of black carbon in the European Arctic. *Atmos. Chem. Phys.* **2018**, *18*, 14037–14057. [[CrossRef](#)]
34. Lack, D.A.; Cappa, C.D. Impact of brown and clear carbon on light absorption enhancement, single scatter albedo and absorption wavelength dependence of black carbon. *Atmos. Chem. Phys.* **2010**, *10*, 4207–4220. [[CrossRef](#)]
35. Forestieri, S.D.; Helgestad, T.M.; Lambe, A.T.; Renbaum-Wolff, L.; Lack, D.A.; Massoli, P.; Cross, E.S.; Dubey, M.K.; Mazzoleni, C.; Olfert, J.S.; et al. Measurement and modeling of the multiwavelength optical properties of uncoated flame-generated soot. *Atmos. Chem. Phys.* **2018**, *18*, 12141–12159. [[CrossRef](#)]
36. McComiskey, A.; Sisterton, D. *ARM Aerosol Measurement Science Group Strategic Planning Workshop 2017*. No. DOE/SC-ARM-TR-207; DOE Office of Science Atmospheric Radiation Measurement (ARM) Program (United States): Oliktok Point, AK, USA, 2018. [[CrossRef](#)]
37. Uin, J.; Aiken, A.C.; Dubey, M.K.; Kuang, C.; Pekour, M.; Salwen, C.; Sedlacek, A.J.; Senum, G.; Smith, S.; Wang, J.; et al. Atmospheric Radiation Measurement (ARM) Aerosol Observing Systems (AOS) for Surface-Based In Situ Atmospheric Aerosol and Trace Gas Measurements. *J. Atmos. Ocean. Technol.* **2019**, *36*, 2429–2447. [[CrossRef](#)]
38. Schmeisser, L.; Andrews, E.; Ogren, J.A.; Sheridan, P.; Jefferson, A.; Sharma, S.; Kim, J.E.; Sherman, J.P.; Sorribas, M.; Kalapov, I.; et al. Classifying aerosol type using in situ surface spectral aerosol optical properties. *Atmos. Chem. Phys.* **2017**, *17*, 12097–12120. [[CrossRef](#)]

39. Norwegian Institute for Air Research EBAS Database. Available online: <http://ebas.nilu.no/> (accessed on 29 September 2020).
40. Li, H.; McMeeking, G.R.; May, A.A. Development of a new correction algorithm applicable to any filter-based absorption photometer. *Atmos. Meas. Tech.* **2020**, *13*, 2865–2886. [[CrossRef](#)]
41. Bond, T.C.; Anderson, T.L.; Campbell, D. Calibration and Intercomparison of Filter-Based Measurements of Visible Light Absorption by Aerosols. *Aerosol Sci. Technol.* **1999**, *30*, 582–600. [[CrossRef](#)]
42. Ogren, J.A. Comment on “ Calibration and Intercomparison of Filter-Based Measurements of Visible Light Absorption by Aerosols”. *Aerosol Sci. Technol.* **2010**, *44*, 589–591. [[CrossRef](#)]
43. Varble, A.; Nesbitt, S.; Salio, P.; Avila, E.; Borque, P.; DeMott, P.; McFarquhar, G.; van den Heever, S.; Zipser, E.; Gochis, D. *Cloud, Aerosol, and Complex Terrain Interactions (CACTI) Field Campaign Report (No. DOE/SC-ARM-19-028)*; ARM Data Center, Oak Ridge National Laboratory (ORNL): Oak Ridge, TN, USA, 2019. [[CrossRef](#)]
44. Titos, G.; Jefferson, A.; Sheridan, P.J.; Andrews, E.; Lyamani, H.; Alados-Arboledas, L.; Ogren, J.A. Aerosol light-scattering enhancement due to water uptake during the TCAP campaign. *Atmos. Chem. Phys.* **2014**, *14*, 7031–7043. [[CrossRef](#)]
45. Kassianov, E.; Barnard, J.; Pekour, M.; Berg, L.K.; Shilling, J.; Flynn, C.; Mei, F.; Jefferson, A. Simultaneous retrieval of effective refractive index and density from size distribution and light-scattering data: Weakly absorbing aerosol. *Atmos. Meas. Tech.* **2014**, *7*, 3247–3261. [[CrossRef](#)]
46. Camponogara, G.; Silva Dias, M.A.F.; Carrió, G.G. Relationship between Amazon biomass burning aerosols and rainfall over the La Plata Basin. *Atmos. Chem. Phys.* **2014**, *14*, 4397–4407. [[CrossRef](#)]
47. Selimovic, V.; Yokelson, R.J.; Warneke, C.; Roberts, J.M.; de Gouw, J.; Reardon, J.; Griffith, D.W.T. Aerosol optical properties and trace gas emissions by PAX and OP-FTIR for laboratory-simulated western US wildfires during FIREX. *Atmos. Chem. Phys.* **2018**, *18*, 2929–2948. [[CrossRef](#)]
48. Li, H.; Lamb, K.D.; Schwarz, J.P.; Selimovic, V.; Yokelson, R.J.; McMeeking, G.R.; May, A.A. Inter-comparison of black carbon measurement methods for simulated open biomass burning emissions. *Atmos. Environ.* **2019**. [[CrossRef](#)]
49. Springston, S.R. *Radiance Research Particle Soot/Absorption Photometer Instrument Handbook*. No. DOE/SC-ARM-TR-176; DOE ARM Climate Research Facility: Washington, DC, USA, 2016. [[CrossRef](#)]
50. Ogren, J.A.; Wendell, J.; Andrews, E.; Sheridan, P.J. Continuous light absorption photometer for long-term studies. *Atmos. Meas. Tech.* **2017**, *10*, 4805–4818. [[CrossRef](#)]
51. Müller, T.; Laborde, M.; Kassell, G.; Wiedensohler, A. Design and performance of a three-wavelength LED-based total scatter and backscatter integrating nephelometer. *Atmos. Meas. Tech.* **2011**, *4*, 1291–1303. [[CrossRef](#)]
52. Raschka, S.; Patterson, J.; Nolet, C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information* **2020**, *11*, 193. [[CrossRef](#)]
53. Westphal, M.; Brannath, W. Evaluation of multiple prediction models: A novel view on model selection and performance assessment. *Stat. Methods Med. Res.* **2020**, *29*, 1728–1745. [[CrossRef](#)]
54. Cappa, C.D.; Kolesar, K.R.; Zhang, X.; Atkinson, D.B.; Pekour, M.S.; Zaveri, R.A.; Zelenyuk, A.; Zhang, Q. Understanding the optical properties of ambient sub- and supermicron particulate matter: Results from the CARES 2010 field study in northern California. *Atmos. Chem. Phys.* **2016**, *16*, 6511–6535. [[CrossRef](#)]
55. Sorribas, M.; Andrews, E.; Ogren, J.A.; del Águila, A.; Fraile, R.; Sheridan, P.; Yela, M. Climatological study for understanding the aerosol radiative effects at southwest Atlantic coast of Europe. *Atmos. Environ.* **2019**, *205*, 52–66. [[CrossRef](#)]
56. Romano, S.; Perrone, M.R.; Pavese, G.; Esposito, F.; Calvello, M. Optical properties of PM_{2.5} particles: Results from a monitoring campaign in southeastern Italy. *Atmos. Environ.* **2019**, *203*, 35–47. [[CrossRef](#)]
57. Adams, P.J.; Seinfeld, J.H. Predicting global aerosol size distributions in general circulation models. *J. Geophys. Res.* **2002**, *107*, 4370. [[CrossRef](#)]
58. Lee, Y.H.; Chen, K.; Adams, P.J. Development of a global model of mineral dust aerosol microphysics. *Atmos. Chem. Phys.* **2009**, *9*, 2441–2458. [[CrossRef](#)]
59. D’Andrea, S.D.; Häkkinen, S.A.K.; Westervelt, D.M.; Kuang, C.; Levin, E.J.T.; Kanawade, V.P.; Leaitch, W.R.; Spracklen, D.V.; Riipinen, I.; Pierce, J.R. Understanding global secondary organic aerosol amount and size-resolved condensational behavior. *Atmos. Chem. Phys.* **2013**, *13*, 11519–11534. [[CrossRef](#)]

60. Li, H.; May, A.A. Application of regression and machine learning approaches in the analysis of mass absorption cross section of black carbon aerosols. *Zenodo* **2020**. [[CrossRef](#)]
61. Müller, T.; Virkkula, A.; Ogren, J.A. Constrained two-stream algorithm for calculating aerosol light absorption coefficient from the Particle Soot Absorption Photometer. *Atmos. Meas. Tech.* **2014**, *7*, 4049–4070. [[CrossRef](#)]
62. Virkkula, A.; Ahlquist, N.C.; Covert, D.S.; Arnott, W.P.; Sheridan, P.J.; Quinn, P.K.; Coffman, D.J. Modification, Calibration and a Field Test of an Instrument for Measuring Light Absorption by Particles. *Aerosol Sci. Technol.* **2005**, *39*, 68–83. [[CrossRef](#)]
63. Collaud Coen, M.; Weingartner, E.; Apituley, A.; Ceburnis, D.; Fierz-Schmidhauser, R.; Flentje, H.; Henzing, J.S.; Jennings, S.G.; Moerman, M.; Petzold, A.; et al. Minimizing light absorption measurement artifacts of the Aethalometer: Evaluation of five correction algorithms. *Atmos. Meas. Tech.* **2010**, *3*, 457–474. [[CrossRef](#)]
64. Weingartner, E.; Saathoff, H.; Schnaiter, M.; Streit, N.; Bitnar, B.; Baltensperger, U. Absorption of light by soot particles: Determination of the absorption coefficient by means of aethalometers. *J. Aerosol Sci.* **2003**, *34*, 1445–1463. [[CrossRef](#)]
65. Grisel, O.; Mueller, A.; Buitinck, L.; Gramfort, A.; Louppe, G.; Prettenhofer, P.; Blondel, M.; Niculae, V.; Nothman, J.; Joly, A.; et al. scikit-learn/scikit-learn: Scikit-learn 0.23.1 (Version 0.23.1). *Zenodo* **2020**. [[CrossRef](#)]
66. Ondráček, J.; Ždímal, V.; Smolík, J.; Lazaridis, M. A Merging Algorithm for Aerosol Size Distribution from Multiple Instruments. *Water Air Soil Pollut.* **2009**, *199*, 219–233. [[CrossRef](#)]
67. Shen, X.J.; Sun, J.Y.; Zhang, Y.M.; Wehner, B.; Nowak, A.; Tuch, T.; Zhang, X.C.; Wang, T.T.; Zhou, H.G.; Zhang, X.L.; et al. First long-term study of particle number size distributions and new particle formation events of regional aerosol in the North China Plain. *Atmos. Chem. Phys.* **2011**, *11*, 1565–1580. [[CrossRef](#)]
68. Bergstrom, R.W.; Pilewskie, P.; Russell, P.B.; Redemann, J.; Bond, T.C.; Quinn, P.K.; Sierau, B. Spectral Absorption Properties of Atmospheric Aerosols. *Atmos. Chem. Phys.* **2007**, *7*, 5937–5943. [[CrossRef](#)]
69. Byčenkienė, S.; Ulevicius, V.; Dudoitis, V.; Pauraitė, J. Identification and Characterization of Black Carbon Aerosol Sources in the East Baltic Region. *Adv. Meteorol.* **2013**, *2013*, 380614. [[CrossRef](#)]
70. Kleinbaum, D.G.; Kupper, L.L.; Muller, K.E.; Nizam, A. *Applied Regression Analysis and other Multivariable Methods*; Duxbury Press: Belmont, CA, USA, 1988.
71. Johnsson, T. A procedure for stepwise regression analysis. *Stat. Pap.* **1992**, *33*, 21–29. [[CrossRef](#)]
72. Thompson, M.L. Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *Int. Stat. Rev./Rev. Int. Stat.* **1978**, *46*, 1. [[CrossRef](#)]
73. Akaike, H. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Stat. Math.* **1978**, *30*, 9–14. [[CrossRef](#)]
74. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [[CrossRef](#)]
75. Drucker, H.; Burges, C.J.C.; Kaufman, L.; Smola, A.J.; Vapnik, V. Support vector regression machines. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; pp. 155–161.
76. Smola, A.J.; Schölkopf, B. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica* **1998**, *22*, 211–231. [[CrossRef](#)]
77. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
78. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [[CrossRef](#)]
79. Fukushima, K.; Miyake, S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.* **1982**, *15*, 455–469. [[CrossRef](#)]
80. Kiranyaz, S.; Ince, T.; Abdeljaber, O.; Avci, O.; Gabbouj, M. 1-D Convolutional Neural Networks for Signal Processing Applications. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8360–8364.
81. Morris, R.E.; McNally, D.E.; Tesche, T.W.; Tonnesen, G.; Boylan, J.W.; Brewer, P. Preliminary Evaluation of the Community Multiscale Air Quality Model for 2002 over the Southeastern United States. *J. Air Waste Manag. Assoc.* **2005**, *55*, 1694–1708. [[CrossRef](#)]
82. Kleinman, L.; Sedlacek, A.J., III; Adachi, K.; Buseck, P.R.; Collier, S.; Dubey, M.K.; Hodshire, A.L.; Lewis, E.; Onasch, T.B.; Pierce, J.R.; et al. Rapid Evolution of Aerosol Particles and their Optical Properties Downwind of Wildfires in the Western U.S. *Atmos. Meas. Tech. Discuss.* **2020**. [[CrossRef](#)]

83. Romonosky, D.E.; Gomez, S.L.; Lam, J.; Carrico, C.M.; Aiken, A.C.; Chylek, P.; Dubey, M.K. Optical Properties of Laboratory and Ambient Biomass Burning Aerosols: Elucidating Black, Brown, and Organic Carbon Components and Mixing Regimes. *J. Geophys. Res. Atmos.* **2019**, *124*, 5088–5105. [\[CrossRef\]](#)
84. Liu, S.; Aiken, A.C.; Arata, C.; Dubey, M.K.; Stockwell, C.E.; Yokelson, R.J.; Stone, E.A.; Jayarathne, T.; Robinson, A.L.; Demott, P.J.; et al. Aerosol single scattering albedo dependence on biomass combustion efficiency: Laboratory and field studies. *Geophys. Res. Lett.* **2014**, *41*, 742–748. [\[CrossRef\]](#)
85. Holder, A.L.; Hagler, G.S.W.; Aurell, J.; Hays, M.D.; Gullett, B.K. Particulate matter and black carbon optical properties and emission factors from prescribed fires in the southeastern United States. *J. Geophys. Res. Atmos.* **2016**, *121*, 3465–3483. [\[CrossRef\]](#)
86. Reid, J.S.; Eck, T.F.; Christopher, S.A.; Koppmann, R.; Dubovik, O.; Eleuterio, D.P.; Holben, B.N.; Reid, E.A.; Zhang, J. A review of biomass burning emissions part III: Intensive optical properties of biomass burning particles. *Atmos. Chem. Phys.* **2005**, *5*, 827–849. [\[CrossRef\]](#)
87. May, A.A.; McMeeking, G.R.; Lee, T.; Taylor, J.W.; Craven, J.S.; Burling, I.; Sullivan, A.P.; Akagi, S.; Collett, J.L.; Flynn, M.; et al. Aerosol emissions from prescribed fires in the United States: A synthesis of laboratory and aircraft measurements. *J. Geophys. Res. Atmos.* **2014**, *119*, 11826–11849. [\[CrossRef\]](#)
88. Bell, S.A. *A Beginner's Guide to Uncertainty of Measurement*; National Physical Laboratory: Middlesex, UK, 2001.
89. Manly, B.F.J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*; Chapman and Hall/CRC: New York, NY, USA, 2018; ISBN 9781315273075.
90. Gine, E.; Zinn, J. Bootstrapping General Empirical Measures. *Ann. Probab.* **1990**, *18*, 851–869. [\[CrossRef\]](#)
91. Anderson, T.L.; Covert, D.S.; Marshall, S.F.; Laucks, M.L.; Charlson, R.J.; Waggoner, A.P.; Ogren, J.A.; Caldow, R.; Holm, R.L.; Quant, F.R.; et al. Performance Characteristics of a High-Sensitivity, Three-Wavelength, Total Scatter/Backscatter Nephelometer. *J. Atmos. Ocean. Technol.* **1996**, *13*, 967–986. [\[CrossRef\]](#)
92. Wiedensohler, A.; Wiesner, A.; Weinhold, K.; Birmili, W.; Hermann, M.; Merkel, M.; Müller, T.; Pfeifer, S.; Schmidt, A.; Tuch, T.; et al. Mobility particle size spectrometers: Calibration procedures and measurement uncertainties. *Aerosol Sci. Technol.* **2018**, *52*, 146–164. [\[CrossRef\]](#)
93. Pfeifer, S.; Müller, T.; Weinhold, K.; Zikova, N.; Martins dos Santos, S.; Marinoni, A.; Bischof, O.F.; Kykal, C.; Ries, L.; Meinhardt, F.; et al. Intercomparison of 15 aerodynamic particle size spectrometers (APS 3321): Uncertainties in particle sizing and number size distribution. *Atmos. Meas. Tech.* **2016**, *9*, 1545–1551. [\[CrossRef\]](#)
94. Barnard, E.; Wessels, L.F. Extrapolation and interpolation in neural network classifiers. *IEEE Control Syst.* **1992**, *12*, 50–53. [\[CrossRef\]](#)
95. Curci, G.; Hogrefe, C.; Bianconi, R.; Im, U.; Balzarini, A.; Baró, R.; Brunner, D.; Forkel, R.; Giordano, L.; Hirtl, M.; et al. Uncertainties of simulated aerosol optical properties induced by assumptions on aerosol physical and chemical properties: An AQMEII-2 perspective. *Atmos. Environ.* **2015**, *115*, 541–552. [\[CrossRef\]](#)
96. Tuccella, P.; Curci, G.; Pitari, G.; Lee, S.; Jo, D.S. Direct Radiative Effect of Absorbing Aerosols: Sensitivity to Mixing State, Brown Carbon, and Soil Dust Refractive Index and Shape. *J. Geophys. Res. Atmos.* **2020**, *125*. [\[CrossRef\]](#)

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).