



# Classification of iron oxide aerosols by a single particle soot photometer using supervised machine learning

Kara D. Lamb<sup>1,2</sup>

<sup>1</sup>Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA

<sup>2</sup>NOAA Earth System Research Laboratory Chemical Sciences Division, Boulder, CO, USA

**Correspondence:** Kara D. Lamb (kara.lamb@noaa.gov)

Received: 15 March 2019 – Discussion started: 22 March 2019

Revised: 20 June 2019 – Accepted: 21 June 2019 – Published: 15 July 2019

**Abstract.** Single particle soot photometers (SP2) use laser-induced incandescence to detect aerosols on a single particle basis. SP2s that have been modified to provide greater spectral contrast between their narrow and broad-band incandescent detectors have previously been used to characterize both refractory black carbon (rBC) and light-absorbing metallic aerosols, including iron oxides ( $\text{FeO}_x$ ). However, single particles cannot be unambiguously identified from their incandescent peak height (a function of particle mass) and color ratio (a measure of blackbody temperature) alone. Machine learning offers a promising approach for improving the classification of these aerosols. Here we explore the advantages and limitations of classifying single particle signals obtained with a modified SP2 using a supervised machine learning algorithm. Laboratory samples of different aerosols that incandesce in the SP2 (fullerene soot, mineral dust, volcanic ash, coal fly ash,  $\text{Fe}_2\text{O}_3$ , and  $\text{Fe}_3\text{O}_4$ ) were used to train a random forest algorithm. The trained algorithm was then applied to test data sets of laboratory samples and atmospheric aerosols. This method provides a systematic approach for classifying incandescent aerosols by providing a score, or conditional probability, that a particle is likely to belong to a particular aerosol class (rBC,  $\text{FeO}_x$ , etc.) given its observed single particle features. We consider two alternative approaches for identifying aerosols in mixed populations based on their single particle SP2 response: one with specific class labels for each species sampled, and one with three broader classes (rBC, anthropogenic  $\text{FeO}_x$ , and dust-like) for particles with similar SP2 responses. Predictions of the most likely particle class (the one with the highest mean probability) based on applying the trained random forest algorithm to the single particle features for test data sets comprising examples of

each class are compared with the true class for those particles to estimate generalization performance. While the specific class approach performed well for rBC and  $\text{Fe}_3\text{O}_4$  ( $\geq 99\%$  of these aerosols are correctly identified), its classification of other aerosol types is significantly worse (only 47%–66% of other particles are correctly identified). Using the broader class approach, we find a classification accuracy of 99% for  $\text{FeO}_x$  samples measured in the laboratory. The method allows for classification of  $\text{FeO}_x$  as anthropogenic or dust-like for aerosols with effective spherical diameters from 170 to  $> 1200$  nm. The misidentification of both dust-like aerosols and rBC as anthropogenic  $\text{FeO}_x$  is small, with  $< 3\%$  of the dust-like aerosols and  $< 0.1\%$  of rBC misidentified as  $\text{FeO}_x$  for the broader class case. When applying this method to atmospheric observations taken in Boulder, CO, a clear mode consistent with  $\text{FeO}_x$  was observed, distinct from dust-like aerosols.

## 1 Introduction

The single particle soot photometer (SP2) has been used over the past decade to quantify refractory black carbon (rBC) mass and internal mixing on a single particle basis (Stephens et al., 2003; Schwarz et al., 2006). Recently, the SP2 has been increasingly used to quantify other light-absorbing refractory aerosols (e.g., Moteki et al., 2017; Liu et al., 2018). In particular, observations in source regions have shown that iron-oxide-containing aerosols from anthropogenic origins are present in the atmosphere (Liati et al., 2015; Dall’Osto et al., 2016; Adachi et al., 2016; Li et al., 2017), and these aerosols can be detected via laser-induced incandescence

with an SP2 (Yoshida et al., 2016; Moteki et al., 2017). These aerosols have been found to be mostly pure iron oxides that are fractal aggregates of  $\sim 100$  nm spheroids, internally mixed (heterogeneously) with nitrate or sulfate (Dall'Osto et al., 2016; Adachi et al., 2016; Li et al., 2017). They have been linked to transportation sources (engine exhaust, traffic brake wear) and industrial sources such as steel processing (Ohata et al., 2018). Iron oxide aerosols quantified by the SP2 were referred to as  $\text{FeO}_x$  in past literature (e.g., Moteki et al., 2017), and we continue this convention here. In general,  $\text{FeO}_x$  as quantified by the SP2 in atmospheric observations in past studies potentially included aerosols from both anthropogenic and nonanthropogenic sources. The mass mixing ratio and size distribution of  $\text{FeO}_x$  has been quantified in East Asia, where observations suggested these aerosols were mainly from anthropogenic sources, and were also observed to be significantly more prevalent than previously believed (Yoshida et al., 2016, 2018; Moteki et al., 2017; Ohata et al., 2018). These measurements have important implications for the climatic effects associated with these aerosols: the direct radiative climate effects of anthropogenic  $\text{FeO}_x$  may be as important as brown carbon in some regions (Moteki et al., 2017; Matsui et al., 2018), and modeling studies based on these measurements indicate these aerosols could also be an important source of particulate iron for the oceanic biogeochemical cycle (Matsui et al., 2018; Ito et al., 2018).

Improving the detection of iron oxide aerosols linked to anthropogenic sources is key to understanding their potential impact on the climate. The SP2 offers a promising method for real-time quantification of these aerosols, as previous detection techniques are limited to offline methods such as X-ray spectrometry (Adachi et al., 2016). However, while laser-induced incandescence can be used to quantify the mass of pure magnetite ( $\text{Fe}_3\text{O}_4$ ) and, to a lesser extent, hematite ( $\text{Fe}_2\text{O}_3$ ) and wüstite ( $\text{FeO}$ ) (Yoshida et al., 2016), the interpretation of ambient SP2 observations has been limited by the misclassification of other aerosols as  $\text{FeO}_x$ , including both rBC and aerosols containing metallic components from nonanthropogenic sources.

To first order,  $\text{FeO}_x$  can be differentiated from refractory black carbon (rBC) because of differences between the blackbody temperature and peak incandescent signal (relative to the particle mass) associated with single particles incandescing in the laser of the SP2. However, the temperature and incandescent peak height alone are not sufficient to unambiguously identify  $\text{FeO}_x$ . Because  $\text{FeO}_x$  has a higher mass-to-incandescent-signal relationship than rBC (Yoshida et al., 2016), and is generally significantly rarer than rBC in the atmosphere in a source region by a factor of  $\sim 250\times$  (Moteki et al., 2017), the misclassification of even a small fraction of rBC as  $\text{FeO}_x$  can bias the retrieved mass mixing ratio.

In addition, other types of metallic aerosols (e.g., tungsten, silicon, chromium, niobium, gold, and aluminum) can be detected via laser-induced incandescence (Stephens et al.,

2003; Schwarz et al., 2006). Although these aerosols are unlikely to be significantly present in the atmosphere, more common aerosol classes, such as coal fly ash, mineral dust, and volcanic ash, also can contain metallic inclusions that are detected with low efficiency by the SP2 and in some cases have similar blackbody temperatures/incandescent peak heights to  $\text{FeO}_x$ . Laboratory tests performed for this study on different samples of fly ash (Miami F, Welsh C, and Clifty-F), a coal combustion product that is a significant source of metallic particles to the atmosphere, indicated these aerosols incandesce with low efficiency (i.e., only a small fraction of the particles have a nonzero incandescent signal in the SP2). This low detection efficiency likely indicates that only a small fraction of these aerosols contain sufficient quantities of materials that can be heated to detectable incandescence in the SP2 laser. Approximately 5%–10% of volcanic ash particles from the Eyjafjallajökull volcano incandesced in the SP2 during the SOOT 11 campaign, with greater incidence of incandescence for larger particles (Heimerl et al., 2012). Incandescent particles in Icelandic mineral dust and Taklamakan Desert dust also have been detected with low efficiency using the SP2 (Yoshida et al., 2016). An SP2 was used in one study to estimate the hematite content in Saharan dust measured off the west coast of Africa and demonstrated good closure between the hematite concentration associated with single dust particles and the optical properties observed in the dust plume (Liu et al., 2018). While previous work focusing on anthropogenic  $\text{FeO}_x$  has relied on using the optical size of these aerosols after any volatile coatings have evaporated as additional criteria to differentiate anthropogenic and nonanthropogenic aerosols with metallic components (Yoshida et al., 2016, 2018; Moteki et al., 2017; Ohata et al., 2018), this method is limited to the range over which the SP2 optically sizes these aerosols (generally limited to  $\sim 170$ – $350$  nm volume equivalent diameter for  $\text{FeO}_x$ ). Previous studies using the SP2 to quantify  $\text{FeO}_x$  associated with anthropogenic sources have also not provided quantitative measures of classification performance for these aerosols.

Here we demonstrate that supervised machine learning can be used to differentiate laboratory samples of pure  $\text{FeO}_x$  from other types of incandescent aerosols expected in the ambient. Machine learning refers to a number of related algorithms using optimization techniques based in probability theory to directly extract information from observations, without relying on a priori knowledge of underlying physical models. Supervised machine learning methods, which use labeled data sets to initially train algorithms, are particularly suited to classification problems. These methods are used, for example, to classify images, for text-to-speech applications, and for identifying handwritten digits and are also increasingly being applied to scientific applications, including atmospheric aerosol measurements. While machine learning approaches have been used to classify single particle aerosol mass spectra (Zawadowicz et al., 2017; Christopoulos et al.,

2018) and biological aerosols detected via ultraviolet-light-induced fluorescence (Robinson et al., 2013; Ruske et al., 2017, 2018; Savage and Huffman, 2018), they have not yet been applied to the problem of classifying aerosols detected via laser-induced incandescence. We review detection of incandescing aerosols with the SP2 and describe measurements on laboratory samples in Sect. 2. We discuss how features derived from single particle signals can be used as input to a supervised learning algorithm and describe a method for training and optimizing this algorithm in Sect. 3. In Sect. 4, we discuss the performance of the trained random forest algorithm on laboratory samples and atmospheric observations. This method extends the classification of  $\text{FeO}_x$  associated with anthropogenic sources beyond the range over which the SP2 can optically size these aerosols and also reduces the misclassification of other aerosols as anthropogenic  $\text{FeO}_x$ .

## 2 SP2 detection of incandescing aerosols

To optimize classification of aerosols measured by the SP2, we first describe the detector configuration, calibrations, and laboratory measurements used in this study and discuss how different aerosols are detected by the SP2. SP2s operate by using laser-induced incandescence (LII) to detect submicron incandescing aerosols on a single particle basis (Stephens et al., 2003). Their operation has been discussed in detail elsewhere (Schwarz et al., 2006, 2010; Moteki and Kondo, 2010). The SP2 determines the mass of the incandescent portion of single aerosol particles by using an Nd:YAG laser (1064 nm) to heat refractory particles with a sufficient absorption cross section to vaporization. Aerosol particles are observed by four detectors as they traverse the laser beam, with two detectors measuring the incandescent signal in the visible, and two measuring scattered light at 1064 nm. For the study, we define “incandescent” aerosols as those that have a nonzero signal in the two incandescent channels.

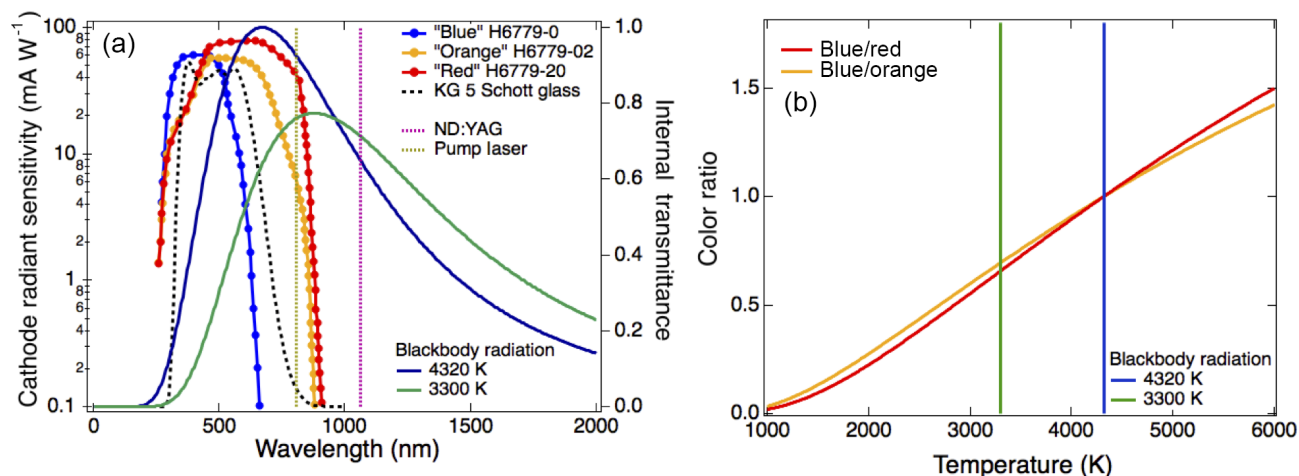
### 2.1 NOAA SP2 detector configuration

The two incandescent detectors measure light emitted from the particles in distinct wavelength bands, providing a measure of the spectral dependence of incandescence, which can be converted to a temperature (Moteki and Kondo, 2010). For this study we use a customized SP2 (the NOAA SP2) whose detector configuration differs slightly from the commercial versions (Droplet Measurement Technology, Longmont, CO), and which was previously described in Schwarz et al. (2006, 2010). This SP2 is operated with four detector channels and a 5 MHz acquisition rate. In the typical configuration of the NOAA SP2, a “red” incandescent detector is a photomultiplier tube (PMT) with a peak sensitivity at 630 nm (450–650 nm, Hamamatsu H6779-20) and a “blue” detector is a PMT with peak sensitivity at 420 nm (350–450 nm, Hamamatsu H6779). Alternatively a PMT (Hama-

matsu H6779-02) with a peak sensitivity of 500 nm and a smaller range (450–630 nm, “orange” detector) is sometimes used in place of the red detector in the NOAA SP2. An additional Schott glass band-pass filter (KG5, 330–665 nm) in front of the red (orange) detector removes light at wavelengths longer than 750 nm to avoid detection of scattered pump laser light (at 807 nm; see Fig. 1 for detector sensitivity ranges). The NOAA SP2 also uses a  $2 \times 1$  mm aperture and a shortwave pass filter (SWP-730, Spectrogon) in front of the red detector to further reduce sensitivity to scattered light from the pump laser. Color temperature ratio is calculated from the ratio (blue:red or blue:orange) of the measured signals at the peak of incandescence (in practice, an average over a small range around each peak is used to reduce sensitivity to high-frequency noise). In this work, the gains on the SP2 blue and red channels were chosen so that the distribution of the color ratios for ambient black carbon is centered near 1. Due to a shift towards the red of blackbody radiation for cooler objects (see Fig. 1a), the characteristic boiling temperature for iron oxide aerosols ( $\sim 3300$  K vs.  $\sim 4320$  K for rBC) corresponds to a color temperature ratio of  $\sim 0.7$ , relative to rBC at 1.0 (Fig. 1b).

The mass of the portion of the particle that incandesces can be determined from the peak height of either incandescent signal, which in the case of rBC is linearly proportional to its mass over most of the accumulation mode (Schwarz et al., 2006; Moteki and Kondo, 2010; Gysel et al., 2012). In this work, we use the blue incandescent peak amplitude to derive single particle incandescent mass and show incandescent peak height (linearly) scaled based on the rBC mass calibration (as this provides a physical metric that is not dependent on the detector gain settings). The detection efficiency of  $\text{FeO}_x$  is dependent on the SP2’s laser power. Although magnetite can be detected with nearly 100 % efficiency under typical conditions (Yoshida et al., 2016), the detection efficiency of hematite is lower and dependent on the particle’s total hematite mass. Up to a point, higher laser power increases the efficiency with which the smaller  $\text{FeO}_x$  aerosols can be detected, as the higher laser power compensates for their smaller absorption cross sections. The SP2 is insensitive to goethite and ferrihydrite, as their absorption cross sections at the wavelength of the Nd:YAG laser are not sufficient for these aerosols to be heated to incandescence (Yoshida et al., 2016).

Incandescing aerosols can be simultaneously optically sized using the scattering channels in the SP2. The optical size is determined by an avalanche photodiode (APD) with sensitivity at 1064 nm (model C30916E, Perkin-Elmer Optoelectronics, Quebec, Canada). The SP2 additionally uses a position-sensitive detector (a four quadrant silicon APD, Perkin-Elmer C30927E-01) to determine the position of the particles in the beam with respect to the center of the laser as has been described in detail in Gao et al. (2007). The SP2 used in this study can be run with either a high gain scattering channel setting or a low gain scattering channel setting. The



**Figure 1.** Determination of single particle blackbody temperature from SP2 incandescent detectors. (a) Normalized blackbody curves for  $T = 4320$  K (typical of rBC) and  $T = 3300$  K (typical of  $\text{FeO}_x$ ) are shown as solid blue and green curves, along with the cathode radiant sensitivity of the PMTs typically used in the SP2, as blue, orange, and red lines and markers. The dashed black line gives the transmissivity of the glass filter used with the red and orange detectors. The yellow dashed line is the wavelength of the pump light, and the maroon dashed line is the wavelength of the Nd:YAG laser. (b) The color ratio as a function of the particle's characteristic blackbody temperature derived for two different detector configurations (blue : red, blue : orange) is shown, assuming the color ratio is scaled to 1 at 4320 K.

high gain setting ( $5\times$  higher than the low gain setting) is optimized for detection of rBC in the accumulation mode (typically  $\sim 90\text{--}550$  nm). The low gain setting allows the SP2 to optically size larger aerosols, although a significant fraction of the non-rBC materials cannot be optically sized even with this lower gain setting. The measurements used in this study use the high gain setting as these are the typical settings used during past aircraft campaigns.

## 2.2 Calibrations

The laser power of the SP2 used in the analysis was calibrated with 220 nm polystyrene latex spheres before and after each data set was taken (Schwarz et al., 2010). The incandescent-signal-to-mass relationship for rBC was calibrated by using fullerene soot (Lot #F12SO11) size selected at different mobility diameters (between 150 and 350 nm) through a differential mobility analyzer (DMA), along with the mass-to-mobility-diameter relationship for rBC from Moteki and Kondo (2010). The incandescent-to-mass relationships for laboratory samples of both magnetite and hematite have previously been characterized (Yoshida et al., 2016), and we determine the  $\text{FeO}_x$  mass relative to the rBC mass calibration using those relationships (see Fig. 2b).

## 2.3 Preparation of laboratory data sets

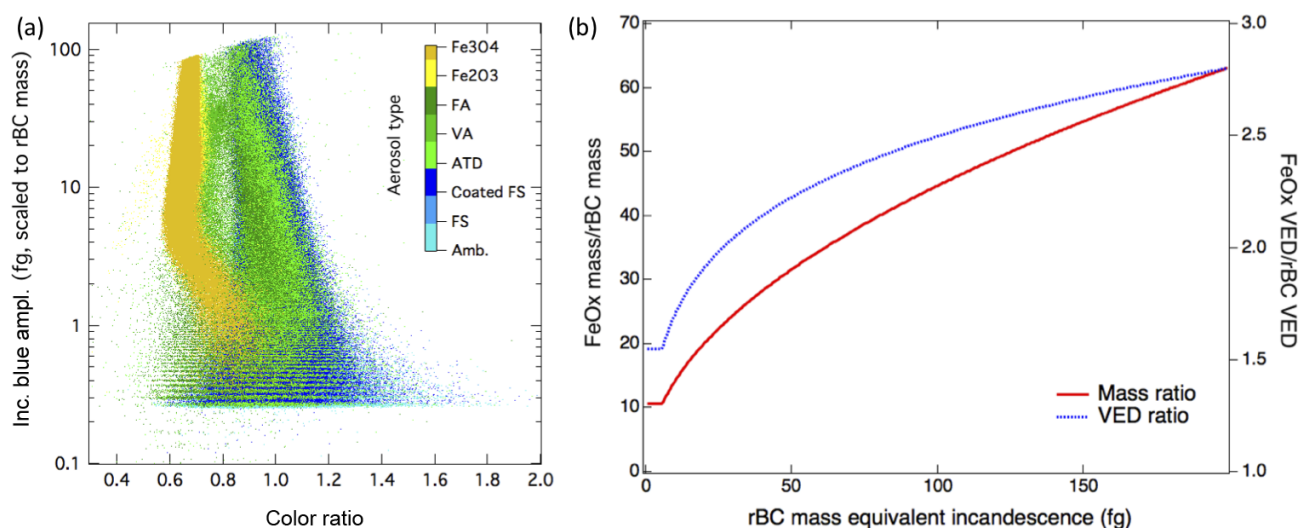
A data set was compiled from laboratory samples to simulate aerosols expected to be found in the atmosphere. Data from laboratory samples of fullerene soot (Lot #F12SO11), magnetite ( $\text{Fe}_3\text{O}_4$ ,  $< 5\ \mu\text{m}$ , Sigma Aldrich 310050), hematite ( $\text{Fe}_2\text{O}_3$ ,  $< 5\ \mu\text{m}$ , Sigma Aldrich 310069), Arizona test dust

(PTI ISO 12103-1), volcanic ash (VA) from the Eyjafjallajökull volcano (collected on the ground in Iceland), and coal fly ash (Clifty-F, referred to as FA) were measured in the laboratory. Fullerene soot is a calibration material that behaves in the SP2 similarly to ambient rBC (Kondo et al., 2011; Baumgardner et al., 2012). Arizona test dust (ATD) is a commonly used reference material for mineral dust and includes some metallic components, including 2%–5% by weight of hematite. Thickly coated rBC particles were simulated by mixing glycerol (99.5%) with fullerene soot. Each of these samples include some fraction of particles that incandesce in the SP2 laser, although ATD, VA, and FA only have a small fraction of incandescent particles relative to the particles that do not incandesce. Samples were measured with a flow rate of  $4\ \text{cc s}^{-1}$  and at low enough concentrations to avoid cases of two incandescent particles crossing the laser at the same time.

Because machine learning algorithms perform best with a large number of examples (to provide sufficient variance), we focused on acquiring a large number of measurements for each aerosol type. The total number of laboratory samples, which are subdivided between the training and test sets (discussed in the next section) are given in Table 1. The incandescent-peak-height-to-color-ratio relationship for all of the laboratory samples is shown in Fig. 2a, along with ambient rBC particles measured in the laboratory. Histograms of the color ratio for the laboratory samples are provided in Fig. S1 in the Supplement.

**Table 1.** Laboratory aerosol samples used in analysis. We test different aerosols with known incandescent components, given in the table below. The abbreviations used to refer to these materials throughout this work are given in the second column. The sample size refers to the total number of incandescing aerosols in the sample.

Material	Abbreviations	Sampling method	Sample size
Fullerene soot	FS	Nebulizer	231 101
Fullerene soot + glycerol	Coated FS	Nebulizer	162 959
Fe <sub>3</sub> O <sub>4</sub> powder (< 5 μm)	Fe <sub>3</sub> O <sub>4</sub>	Nebulizer	258 624
Fe <sub>2</sub> O <sub>3</sub> powder (< 5 μm)	Fe <sub>2</sub> O <sub>3</sub>	Nebulizer	45 609
Clifty-F fly ash	FA	Nebulizer	18 677
Arizona test dust	ATD	Nebulizer	67 102
Volcanic ash	VA	Nebulizer	33 970



**Figure 2.** (a) Incandescent-peak-height-to-color-ratio relationship for different incandescent aerosol types. Laboratory samples of different test materials show significant overlap between incandescent peak height and color ratio. Each point represents a single particle, and the abbreviations are given in Table 1, with Amb. indicating ambient particles. (b) Comparison of FeO<sub>x</sub> and rBC detection in the SP2. Relative mass and volume equivalent diameter (VED) for rBC and FeO<sub>x</sub> observed by the SP2. Because of the power law relationship for the mass to incandescent signals of FeO<sub>x</sub>, the largest particles detected by the SP2 are approximately 60× more massive than rBC with the same incandescent signal. FeO<sub>x</sub> is significantly denser than rBC, however; the volume equivalent diameter ratio is less than a factor of 3 for particles within the SP2 detection range. Inc. refers to incandescent in the figures.

## 2.4 Ambient measurements

We performed ambient sampling in Boulder, CO, from the rooftop inlet of the David Skaggs Research Center to provide samples of typical atmospheric aerosols in an urban environment and natural dust aerosols. The sampling line from the rooftop was an approximately 6 m long, 2.5 cm diameter vertical tube with a small pick-off sampling line of 0.2 cm diameter, which provides a transmission efficiency of ~ 100 % below 1 μm, and slightly enhances sampling for aerosols > 1 μm (superisokinetic). Observations of ambient aerosols were measured over a period of two days (31 October–1 November 2018). The sample flow rate of the SP2 during ambient sampling was chosen to be 4 cc s<sup>-1</sup> to match the laboratory data set.

## 2.5 Differentiation of different aerosol types

Classification of aerosols measured by the SP2 rely on differences in both the optical sizes of these aerosols and the aerosols' incandescence and evaporation in the laser beam. Previous work has noted that, to first order, FeO<sub>x</sub> can be differentiated from rBC in the SP2 because these aerosols have lower color ratios (Schwarz et al., 2006; Yoshida et al., 2016). Other properties of these aerosols impact their detection in the SP2, including a higher mass-to-incandescent relationship (Yoshida et al., 2016) and a higher void-free density for FeO<sub>x</sub> relative to rBC (5.17 g cm<sup>-3</sup> vs. 1.8 g cm<sup>-3</sup>). This significantly higher mass-to-incandescent relationship for FeO<sub>x</sub> (Yoshida et al., 2016) means that much more massive particles are detected than in the case of rBC with the same incandescent peak amplitude (varying from ~ 10× to

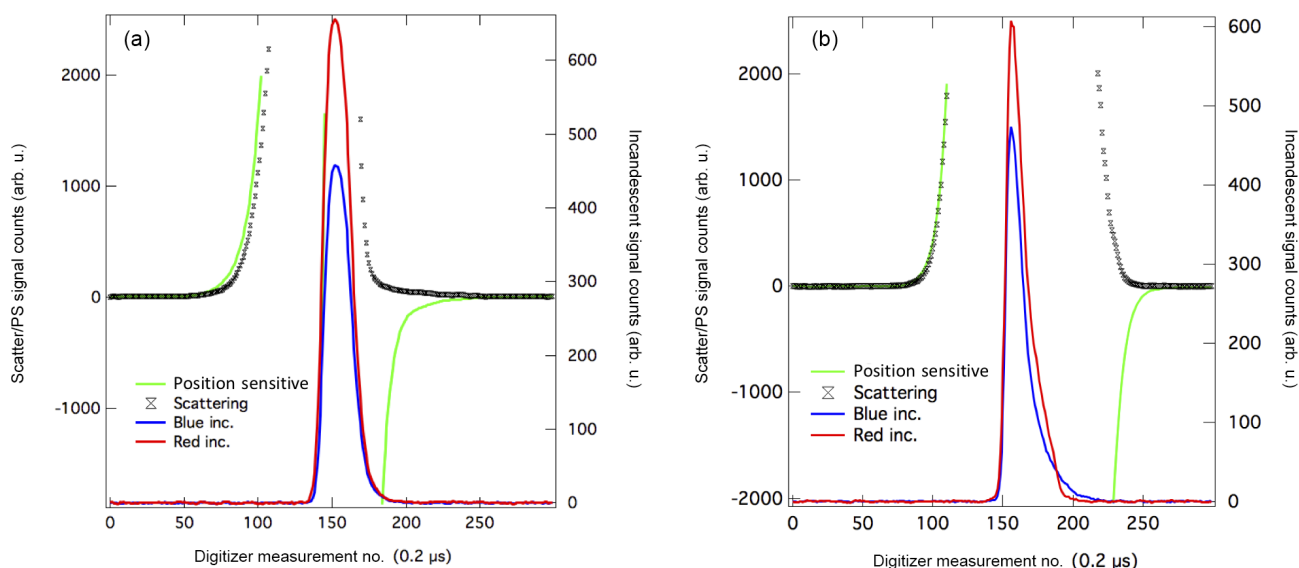
$> 60\times$  more massive for the largest particles in the SP2 detection range; see Fig. 2b). Because the density of  $\text{FeO}_x$  is nearly  $3\times$  higher, however, the volume equivalent diameter is only  $1.5\text{--}3\times$  higher. Literature values for the index of refraction of  $\text{FeO}_x$  ( $n = 2.30 + 0.46i$ , at 1000 nm) also differ significantly from rBC ( $n = 2.49 + 1.49i$ , at 1064 nm) (Huffman and Stapp, 1973; Moteki et al., 2010; Moteki et al., 2017). Yoshida et al. (2016) noted that the incandescence of hematite occurs deeper in the SP2 laser than magnetite, rBC, Taklamakan Desert dust, and Icelandic dust, likely due to a smaller imaginary part of the index of refraction for hematite than for magnetite.

As has been previously noted, color ratio alone is not sufficient to differentiate rBC and  $\text{FeO}_x$ . In practice, particles of the same type show significant statistical variability about their characteristic blackbody temperatures (see Fig. 2a), with variability in color ratio generally inversely proportional to refractory aerosol mass. The population of ambient rBC and fullerene soot particles at the smallest detectable masses demonstrated significantly greater variability in color ratio (when compared with the population of larger particles), due to the lower signal to noise on the red and blue channels. We also found that the width of the distribution of color ratios for rBC and  $\text{FeO}_x$  as a function of incandescent peak height strongly depended on the relative alignments of the PMT detectors (see Fig. S3). Because of the additional aperture in front of the red PMT in the NOAA SP2 optical head, the color ratio was particularly sensitive to the alignment of the red PMT. The incandescent peak height (mass) and a low color ratio together provide sufficient contrast to differentiate iron-oxide-containing aerosols from rBC for larger incandescent peak heights (equivalent to  $\sim 2$  fg rBC) (Yoshida et al., 2016; Liu et al., 2018; Moteki et al., 2017). When detecting mixed populations of  $\text{FeO}_x$  and rBC, the incomplete contrast between rBC and  $\text{FeO}_x$  in terms of their incandescent peak heights and color ratio limits detection at smaller sizes. The upper limit of detection for both rBC and  $\text{FeO}_x$  in the SP2 is due to the gain setting on the detectors; when the incandescent channel becomes saturated, refractory aerosol mass can no longer be quantified.

To improve contrast between rBC and  $\text{FeO}_x$  at smaller sizes (160–230 nm effective void-free diameter for  $\text{FeO}_x$ ), core scattering (amount of scattered light measured after volatile coatings have evaporated but before the refractory core has significantly evaporated) can be used as an additional parameter to differentiate rBC and  $\text{FeO}_x$ . Although the imaginary part of the index of refraction for rBC is greater than for  $\text{FeO}_x$  particles, the incandescent-peak-height-to-mass relationship for rBC is also significantly higher (Yoshida et al., 2016). Therefore,  $\text{FeO}_x$  particles with similar incandescent peak signals compared to rBC are significantly larger particles and have greater than  $4\times$  as much core scattering. However, these additional criteria also do not provide complete contrast between the two particle classes at these sizes.

As can be seen in Fig. 2, an additional complication arises in differentiating  $\text{FeO}_x$  from mixed populations of particles including other aerosols with metallic components that incandesce in the SP2, such as natural mineral dust, as the color ratios of ATD, VA, and FA overlap with  $\text{FeO}_x$  at similar incandescent peak heights. The larger variability in color ratio for ATD, volcanic ash, and coal fly ash in the laboratory samples that were tested is likely due to presence of multiple types of metallic oxides with a greater variety of characteristic blackbody temperatures than  $\text{Fe}_3\text{O}_4$  or  $\text{Fe}_2\text{O}_3$ . TEM (transmission electron microscope) images have shown that natural mineral dust is composed of numerous grains of different materials, of which iron oxides or other metallic oxides can be one component, typically embedded in quartz or feldspar (e.g., Jeong and Nousiainen, 2014). Yoshida et al. (2016) observed a greater variability in the color ratio for Icelandic dust (similar to the observed color ratios in VA sampled in this study) and speculated that these higher color ratios are due to the presence of strongly light absorbing minerals such as titanomagnetite from volcanic origins. Eyjafjallajökull ash samples were previously found to contain several metallic elements, including similar mass fractions of Fe and Al (Rocha-Lima et al., 2014). SEM-EDS (scanning electron microscopy with energy dispersive spectroscopy) characterization of fly ash demonstrated that typical samples were mainly composed of amorphous aluminosilicate spheres, with a smaller contribution of iron-rich spheres, composed of iron oxides mixed with aluminosilicate (Kutchko and Kim, 2006). On the other hand, anthropogenic iron oxide particles that have been observed in the atmosphere may be coated with organic materials and inorganic materials but have been found to be predominately metallic (Adachi et al., 2016; Li et al., 2017). (Although coal fly ash is also of anthropogenic origin, we treat these aerosols independently from anthropogenic  $\text{FeO}_x$ , as they have significantly different signals in the SP2 and generally are more similar to ATD.) Core scattering has been shown to be a useful criterion for differentiating anthropogenic iron oxide aerosols from dust-like aerosols (Yoshida et al., 2016; Moteki et al., 2017). However, this criterion can only be used on the subset of aerosols that can be optically sized in the typical configuration of the SP2. (Typically  $\text{FeO}_x$  with effective void-free diameters  $< 350$  nm, or  $< 230$  nm for the high gain setting used here).

To determine which incandescent signals were likely associated with mineral dust or anthropogenic  $\text{FeO}_x$ , we investigated whether optical scattering in the SP2 can be used to indicate nonevaporative portions of the aerosol (see Fig. 3). Given that natural dust grains typically consist of metallic inclusions surrounded by other materials, even if these portions of the aerosol incandesce in the SP2, the entire particle may not be completely vaporized. Recent measurements using the SP2 to characterize hematite content in dust particles estimated that mineral dust particles detected by the SP2 were generally  $> 500$  nm in size (Liu et al., 2018). Postincandes-



**Figure 3.** Comparison of SP2 signals. SP2 traces for two different aerosols with metallic components with similar incandescent peak heights and color ratios. (a) is from laboratory samples of  $\text{Fe}_3\text{O}_4$  and (b) is from a mineral dust particle. Black indicates the scattering signal, green is the position-sensitive detector, red is the red PMT signal, and blue is the blue PMT signal. For the particle signal shown in (a), the scattering signal disappears as the particle incandesces (around the 200th digitizer measurement point), indicating complete evaporation of the particle. For the particle signal shown in (b), scattering is still present after the incandescent signal has returned to the baseline, indicating that nonevaporative portions of the particle still remain after passing through the SP2 laser.

cent scattering, defined as the scattering amplitude of the particle measured after the incandescent signal has returned to the baseline, indicates portions of the aerosol still remain after the refractory portion of the particle has evaporated (Sedlacek III et al., 2012). Postincandescent scattering was generally nonzero for both  $\text{FeO}_x$  and the other aerosols containing metallic components (ATD, VA, FA); in the case of  $\text{FeO}_x$ , the signal was proportional to observed iron oxide mass, which may be related to previous observations that these particles appeared to be melting in the laser beam (Yoshida et al., 2018). Choosing a single threshold value for postincandescent scattering only differentiated  $\text{FeO}_x$  from other aerosols containing metallic components for  $\sim 80\%$  of the particles, however.

### 3 Supervised learning methods applied to aerosol classification with the single particle soot photometer

The inability to unambiguously classify  $\text{FeO}_x$  from anthropogenic sources from either natural mineral dust or rBC using a small number of features (the incandescent peak height, the color ratio, core scattering, and postincandescent scattering) derived from the single particle signals suggests that new analysis approaches should be explored to fully exploit this additional aspect of the SP2. The SP2's limitations in classifying different aerosols could be overcome in some cases by changes to the detection scheme, e.g., detectors with greater dynamic range for optical sizing; however, other limitations

are fundamentally linked to the LII method, as there is an overlap between the features associated with different particle types when only those four features are considered. Previous data sets (e.g., aircraft data sets) are also limited to the detection scheme described in Sect. 2.

From a mathematical perspective, the problem of classifying aerosols can be described as the search for a mapping function  $f$  that maps a feature vector  $\mathbf{x}_i$  associated with each aerosol to its correct class label  $y_i$ . Here we define a feature as an attribute (for example, the incandescent peak height) associated with a single particle  $i$ , which can be expressed as an  $n$ -dimensional feature vector  $\mathbf{x}_i \in \mathbb{R}^n$ . We would like to find a separable subspace within the  $n$ -dimensional feature space that can be used to differentiate aerosols by class. In other words, decision boundaries can be found that allow us to separate the different aerosols with minimal misclassifications. Decision boundaries are hyperplanes of dimension  $n - 1$  that subdivide this feature space such that the different classes (rBC,  $\text{FeO}_x$ , etc.) reside in distinct subspaces.

Supervised learning algorithms are a class of machine learning algorithms that map input variables ( $\mathbf{X}$ ) to a predicted output variable ( $\hat{\mathbf{Y}}$ ), after first training the algorithm using a set of input variables ( $\mathbf{X}'$ ) with known output ( $\mathbf{Y}'$ ). (Here we adopt notation to use  $\hat{y}_i$  to differentiate the predicted label from the actual label  $y_i$ . For  $n$  training examples and  $m$  features, the input vectors generalize to matrices.) Following the notation in Mohri et al. (2012), the problem that the learning algorithm attempts to solve is finding a hypothesis  $h$ , where  $h \in H$  (a subset of functions explored

by the learning algorithm), to map  $x_i$  to a predicted class label  $\hat{y}_i$ , such that the loss function  $L(\hat{y}_i, y_i)$  is small. This loss function  $L(\hat{y}_i, y_i)$  gives the cost of predicting  $\hat{y}_i$  rather than  $y_i$ . We would like to avoid hypotheses that either underfit or overfit the data. Underfitting refers to a hypothesis that does not perform well even on the initial training data set, as it does not capture the trend of the data. Overfitting occurs when a hypothesis fits the training data well but cannot generalize to new cases, because it has too closely constrained the model to the specific data set. Overfitting can be addressed by increasing the number of training examples (to provide greater instances of within class variance), while underfitting generally implies that the chosen features do not allow for enough degrees of freedom, and a more complex model is required. Since the small set of features described in Sect. 2 does not provide enough information (i.e., it underfits the data), we would like to expand the number of features that provide some information about what type of aerosol was observed. However, adding additional features creates a much larger space over which to determine appropriate decision boundaries to differentiate aerosols. Moreover, the large variation within classes (between particles of the same type with different masses or internal mixing states) makes this problem even more challenging. This kind of problem is tractable using supervised machine learning, however; and these algorithms can be readily applied using existing software libraries (e.g., Python's scikit-learn and TensorFlow libraries).

There are a variety of machine learning algorithms that can be used for classification; the choice of an algorithm depends on consideration of both the particular data set and the intended application, as there is no clear-cut superiority of performance between different algorithms. Here we focus on the application of a random forest algorithm to the SP2 observations (we initially considered other machine learning algorithms, and some additional details are provided in the Supplement). A random forest consists of an ensemble of decision trees and is described in greater detail in Sect. 3.3. This approach allows us to extend the number of features considered for an individual particle to improve classification performance. We compare two different approaches for applying this random forest algorithm to the laboratory data. In one case, we use six distinct classes (rBC, ATD, FA, VA, Fe<sub>2</sub>O<sub>3</sub>, and Fe<sub>3</sub>O<sub>4</sub>), where rBC includes both bare and coated fullerene soot in the training data set. In the second case, we use only three distinct classes: rBC (again, including coated and uncoated fullerene soot as training data), dust-like aerosols (ATD, VA, and FA), and FeO<sub>x</sub> (Fe<sub>2</sub>O<sub>3</sub> and Fe<sub>3</sub>O<sub>4</sub>).

The application of (any) supervised machine learning algorithm requires the implementation of several steps: first, data are collected and, in the case of supervised learning algorithms, labeled and randomly separated into independent training, cross-validation, and test sets. The randomly selected training data set is needed to train the model, the cross-validation set is used to determine the optimal set of hyperparameters for the algorithm, and an independent test set pro-

vides information about how the trained algorithm generalizes to new cases. Before the application of the algorithm to the data set, however, the data need to be preprocessed, which entails repairing or removing missing values and transforming variables by normalizing and scaling them. In the case of classic machine learning, features are extracted from the data set (more advanced techniques such as representation learning/deep learning operate directly on raw data to extract features, but for computational simplicity we do not explore this approach here; Goodfellow et al., 2016). The third step is training the algorithm on the training data set and optimizing its performance using the cross-validation set. The fourth step is evaluating the performance of the algorithm on the test set. Finally, the trained algorithm can be applied to new data sets. The computation steps required to train and optimize the algorithm and then apply it to new data are referred to as a machine learning pipeline. We describe the first three steps in this section and discuss the algorithm's performance on the test data set and on atmospheric measurements in Sect. 4.

### 3.1 Feature engineering from single particle signals

The typical analysis method for the NOAA SP2 reduces 80  $\mu$ s time series signals to a vector of features that can be used to determine the mass, optical size, coating state, and coating thickness of rBC given appropriate calibrations for the detectors as was described in Sect. 2. An algorithm is applied to filter out signals that may be contaminated, e.g., by multiple particles measured during the same acquisition window or other nonideal sampling conditions.

To leverage existing SP2 feature engineering and data analysis, we consider a number of features derived from the single particle signals, including features previously demonstrated to provide useful information about the particle's physiochemical properties. The features explored in this analysis are shown in Table 2 and for an example rBC particle in Fig. 4 and can be roughly divided into three categories: those associated with the incandescent channels ( $x_0$ – $x_1$ ), those associated with the scattering channels ( $x_2$ – $x_7$ ), and those derived from the timing of different signals in the beam ( $x_8$ – $x_{16}$ ). Those associated with the scattering channel are related to the optical size of the aerosol as it traverses the laser (core scattering, total maximum scattering, position-sensitive wideness, postincandescent scattering, and the optical size at a fixed point along the evaporating edge). As discussed in Sect. 2.5, postincandescent scattering ( $x_4$ ) is defined as the maximum value of the scattering signal after the blue incandescent signal has reached a peak and has returned to the baseline. Those associated with the incandescent channel relate to the mass and thermal properties of the aerosol (the blue peak amplitude and the color ratio). Those associated with timing in the beam (e.g., minimum scattering before incandescence, incandescent start position, evaporation point, incandescent used length, and incandescent total length) are related to both the size and physiochemical prop-



erties of the particle (e.g., whether it is initially coated with any volatile materials, how strongly absorbing the aerosol is, or how long it takes to evaporate in the laser); several of these features also depend on the specific laser settings. The features derived from the single particle signals in some cases will not provide values directly interpretable as a measure of the physical properties of the particles as given in Table 2 (i.e., for larger particles, the detectors may be saturated). If the systematic biases with the feature or measurement artifacts are repeatable, however, they still provide useful information to the algorithm, as these methods rely on statistical relationships between the data sets rather than an underlying physical model. To quantify detector saturation for larger particles, we have included features that indicate whether the scattering signal is high before the start of incandescence ( $x_7$ ) and how long the signal is saturated ( $x_{10}$ ). Since the application of a machine learning algorithm requires a value for every element in the feature vector, single particle signals that do not have valid values for each of the features given in Table 2 are imputed with dummy values; we discuss the details of this imputation in the next section.

These features were chosen because they generally showed some separability for different aerosol types for the laboratory samples, although no single feature or pair of features was sufficient to entirely separate different aerosol classes. In applying the machine learning algorithm for the two different cases, we initially use all features but also explore whether a reduced set of these features can provide similar classification performance.

### 3.2 Data preprocessing

Many machine learning algorithms work best if features are preprocessed so that they are normally distributed and have a mean of 0; however, one advantage of decision trees (and by extension random forests) is that they are fairly robust to feature scaling and normalization. We perform several preprocessing steps to prepare the data for use in the algorithm. (These steps are summarized in Supplement Table S1.) First, we remove data associated with particles that do not at minimum have values for both the incandescent peak height and color ratio. For features that are expected to be lognormally distributed (the blue signal amplitude, which is proportional to the particle mass, and the features associated with different optical sizes), we also take the natural logarithm in order to have normally distributed values.

The other important step in preparing data for supervised learning is figuring out what to do with particles associated with incomplete information. In certain cases (such as when the scattering detector is saturated for larger particles), the unknown number likely represents known information (e.g., that the particle is too large to optically size). For these four features (core scattering, scattering peak amplitude, evaporation scattering size, and the laser intensity at the peak of incandescence) we assume that missing values are larger than

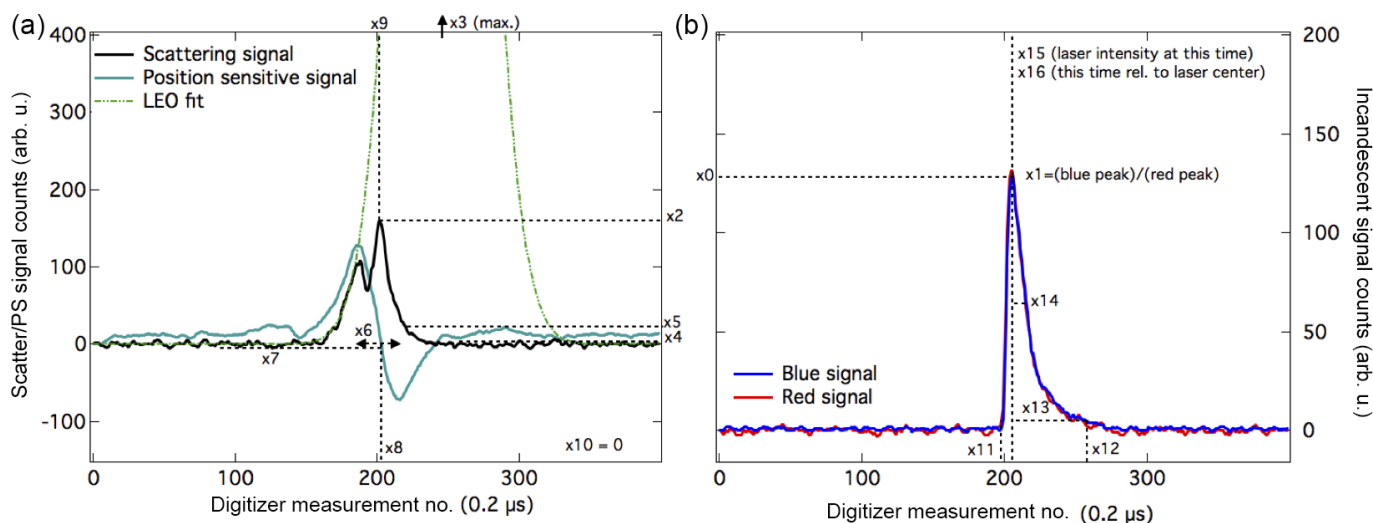
any of the values that were recorded; therefore these missing values were imputed with 110 % of the highest expected value for each of those features. For the other 13 features, we used dummy values outside the typical detection range to effectively exclude these features for that particular sample (alternatively, we tested using typical mean values for these features, which led to similar classification performance). Preprocessing could potentially be improved by implementing a learning algorithm to optimize the imputation of missing values; however, because we did not observe a significant change in performance dependent on the particular method for preprocessing, we did not explore this option.

### 3.3 Random forest

After feature selection and data preparation, we apply an algorithm that can construct a model from the data based on the samples in the training data set. In this case we use a random forest algorithm, which consists of an ensemble of decision trees. (For a schematic of a decision tree, see Fig. 5a. Additional details about decision tree classifiers are provided in Appendix A.) Random forests improve upon the performance of single decision trees by growing an ensemble of decision trees. This algorithm is more robust to overfitting than a single decision tree, as the trees rely on bagging (bootstrap aggregating): a random selection of the training samples are chosen to grow each decision tree, with replacement from the original training set. Each tree also uses a random selection of a subset of the features to define the split at each node (Breiman, 2001); we discuss additional details in Sect. 3.5. Because of this randomness, the generalization error converges for a large number of trees. The predictions of the random forest are based upon the ensemble vote from all the trees in the forest; that is, for each sample, the trained algorithm outputs a conditional probability distribution  $f(y_i) = p(y|\mathbf{x}_i, \theta)$  over the classes, with the highest probability corresponding to the most likely class of the particle  $\hat{y}_i$ , given the values of the feature vector  $\mathbf{x}_i$  and the optimized parameters of the algorithm  $\theta$ . In the implementation of the algorithm used here, the conditional probability for a new sample to belong to any particular class is determined by averaging the probability predictions from all the trees in the forest (Pedregosa et al., 2011). This algorithm has previously been used to classify single particle mass spectra (Christopoulos et al., 2018). A schematic for applying the random forest to a single particle signal from the SP2 is shown in Fig. 5b.

### 3.4 Computational resources

For this analysis, we used Python 3.6.6 with the sci-kit learn package version 0.20.0. To train the algorithm and optimize its hyperparameters, we used a remote Linux server with 24 cores (two Intel Xeon CPU E5-2695 v2 2.40 GHz proces-



**Figure 4.** SP2 signal for a single particle showing all features used in the algorithm. SP2 traces for a coated rBC particle (4 fg, 20 nm thick coating assuming  $n_{\text{core}} = 2.26 + 1.26i$  and  $n_{\text{coat}} = 1.45$ ) showing the features used in the machine learning algorithm. (a) shows the scattering and position-sensitive detectors and (b) shows the blue and red incandescent channels for the same particle. Physical interpretation and descriptions of these features are further detailed in Table 2. Since the scattering signal is not saturated at any point for this particular particle,  $x_{10} = 0$ .  $x_3$  is proportional to the maximum value of the scattering signal derived from the leading-edge only (LEO) fit (Gao et al., 2007).

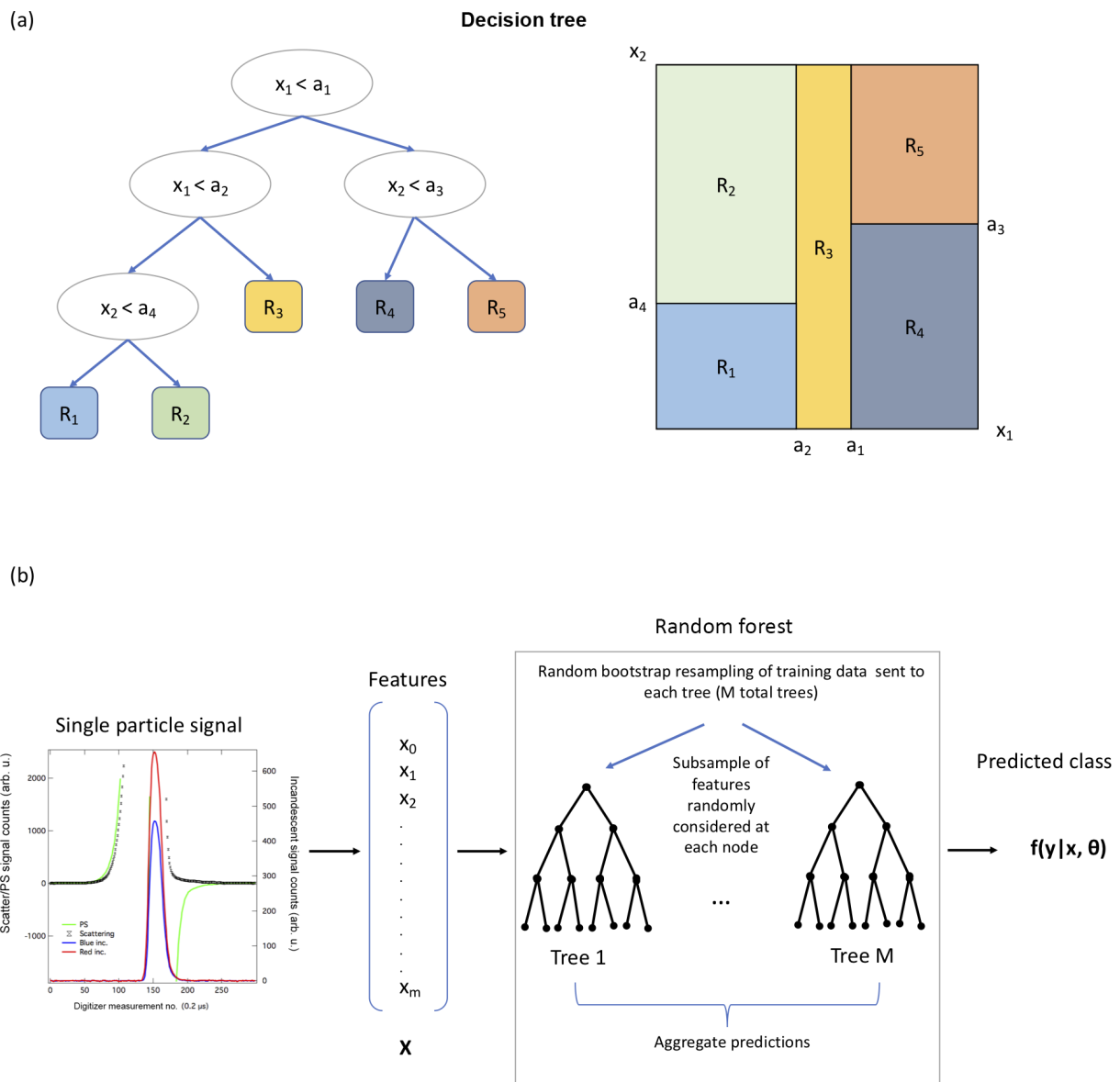
**Table 2.** Description of features from processed SP2 single particle signals. For features that correspond to times in the SP2 detection window, the time is referenced to the position-sensitive detector crossover point, which is a fixed point in space (independent of particle size). See Fig. 4 for an illustration of these features for an rBC particle measured by the SP2.

Symbol	Feature	Description/physical interpretation
$x_0$	Blue peak amplitude	Function of aerosol incandescent mass
$x_1$	Color ratio	Function of aerosol blackbody temperature
$x_2$	Core scattering	Function of optical size of core after volatile coatings have evaporated
$x_3$	Total scattering max.	Function of optical size of the aerosol including coating (estimated from LEO fitting)
$x_4$	Postincandescent scattering	Function of optical size of aerosol after refractory portion has evaporated
$x_5$	Evaporation scattering size	Function of scattering ampl. along evaporating edge when incandescent signal $\sim 1$ fg rBC
$x_6$	Position-sensitive wideness	Time difference between max. and min. of PS signal amplitude
$x_7$	Min. scattering before incandescence	Minimum value of scattering signal before start of incandescence
$x_8$	Position-sensitive trigger position	Time of PS detector crossover point (change from positive to negative)
$x_9$	Scatter peak location	Time of maximum scatter signal
$x_{10}$	Saturation width	Total time scattering signal is saturated, relative to Gaussian laser width
$x_{11}$	Incandescent start position	Time incandescent signal is first greater than min. threshold (relative to PS crossover point)
$x_{12}$	Evaporation point	Time at which aerosol has completely evaporated (relative to PS crossover point)
$x_{13}$	Incandescent total length	Time from incandescence peak location until signal has decayed to baseline
$x_{14}$	Incandescent used length	Full width at half maximum of incandescent signal
$x_{15}$	Light on laser intensity	Laser intensity calculated at blue maximum peak location
$x_{16}$	Width fraction from center	Incandescent peak location relative to center of laser, scaled by Gaussian laser width

sors with 12 cores each and two threads enabled per core), which provided  $\sim 125$  GB of RAM.

The computation time for training a random forest is directly related to the number of trees in the forest and the amount of training data used. In general, a binary decision tree has a time complexity of  $\mathcal{O}(mn \log n)$  for  $n$  training samples and  $m$  features (Pedregosa et al., 2011). One disadvantage of this method is that each decision tree needs to use all of the training data (or rather, the subset of the training

data used to train that tree) at once in order to grow the tree, requiring a significant amount of memory for large training data sets. Distributed or parallel methods can be used to improve computation efficiency for random forests, as the trees can be grown simultaneously on different processors. Depending on the complexity of the trained random forest, the decision path of the trained algorithm can also require significant memory to store. With the large training data set that we consider here, this method is computationally expensive, so



**Figure 5.** Schematic of a decision tree (a), and a random forest as applied to the SP2 signals (b). (a) A simplified example of a decision tree for a case with only two features, demonstrating how choosing threshold values of the features at each node (left) subdivides the feature space into different regions (right). The true case corresponds to the left “child” in the tree. This figure has been adapted from Mohri et al. (2012). (b) The schematic for applying supervised learning to predict the aerosol class for a particle measured by the SP2 is shown. First, single particle signals are processed and reduced to a feature vector  $x_i$ , which is then given to the trained random forest. Each decision tree within the forest makes a series of decisions based on the values of the features (which have been learned from a random subset of the training data, by considering a random subsets of features at each split) to predict the most likely class of the particle based on the observed features. The aggregated prediction from the ensemble of  $M$  decision trees predicts a probability distribution  $f(y) = p(y|x_i, \theta)$  over each class  $y$ .

although this approach serves as a proof of concept, it would be advantageous to explore other methods for improved computational efficiency.

### 3.5 Tuning hyperparameters for improved performance

In order to optimize the performance of supervised machine learning algorithms, a cross-validation data set is typically used to find the optimal set of hyperparameters for the algorithm. Hyperparameters are separate from the parameters optimized by the learning algorithm but affect its general-

ization performance. For a random forest, the hyperparameters consist of, for example, the number of trees in the forest (number of estimators), the maximum depth of each branch, and the maximum number of features (the subset of features) randomly considered at each node. Several of these hyperparameters affect the growth of individual trees, serving as alternative stopping criteria, e.g., by setting a lower limit on the number of training samples required for each split or setting an upper limit on the depth of the tree.

Random forests can use out-of-the-bag (o.o.b.) error estimates for determining hyperparameters (Breiman, 2001). Since each tree relies on only a subsample of the training data set, the performance of that tree on the subset of the training data not used in growing the tree can be used to optimize the hyperparameters. We used o.o.b error estimates to initially tune over a number of different hyperparameters to see which most impacted the classification performance. We found that the entropy criteria always performed better than using the Gini index as the metric for optimizing splitting at each node (see Appendix A for details). Increasing the number of estimators also always improved the classification accuracy, although the effect reached an asymptote after  $\sim 100$ – $120$  trees. The maximum depth of each tree showed decreasing o.o.b error for greater depths, up to  $\sim 40$  splits in the three-class case and  $\sim 30$  splits in the six-class case, with very little change for greater depths. Requiring fewer samples at each internal split and allowing fewer samples per leaf tended to decrease error. The maximum number of features randomly considered at each split demonstrated the most significant difference between the six-class and three-class cases, with the three-class case performing better with a smaller subset of features than the six-class case.

Since each hyperparameter does not independently impact the performance of the algorithm, we then used a grid search with 3-fold cross-validation over 54 different combinations of the hyperparameters to find the best set for each case (Mohri et al., 2012).  $K$ -fold cross-validation uses a subset of the training data (after the data have been randomly shuffled, and with equal relative subselections of each aerosol class) as a cross-validation data set (typically  $K \in [3, 10]$ ), while the algorithm is trained on the rest of the data. The procedure is repeated  $K$  times, using a different subset of the data each time as the cross-validation set in order to get  $K$  total values for the classification accuracy for each set of hyperparameters. For the six-class case, the optimal hyperparameters were a maximum depth of 50, 12 features considered at each split, a minimum of one sample per leaf, and a minimum of two samples per split, with 120 estimators, and using entropy as a criterion. For the three-class case, the optimal hyperparameters were a maximum depth of 50, 10 features considered at each split, a minimum of one sample per leaf, and a minimum of four samples per split, with 120 estimators, and using entropy as a criterion.

#### 4 Performance on laboratory samples and atmospheric observations

The performance of the trained and optimized random forest was tested on an independent test set of laboratory samples. Testing the trained algorithm on an independent set of labeled data allows us to determine its generalization performance (the skill of the learning algorithm to classify new particles) before applying it to the (unlabeled) atmospheric data set. We used stratified  $K$  folding (which randomly divides data with equal subselections of each aerosol class) to divide the laboratory data set into thirds, with 2/3 used for the training/cross-validation steps, and with 1/3 kept aside as a test data set.

##### 4.1 Improving the pipeline

After optimizing the algorithm, we investigate whether the machine learning pipeline can be improved. The relative importance of different features (given in Table 3) is estimated from the fraction of samples in the data set for which the decision pathway is impacted by that feature (Pedregosa et al., 2011). The importance of each feature estimated with this method only gives information over the entire population of aerosols in each class in the test set, however. For both the six-class and three-class cases, the color ratio is the most important feature, with the postincandescent scattering being the next most important feature. As both of these features were previously identified as physically meaningful for separating different aerosol types, this is not surprising.

To avoid overtraining our model, we would ideally like to use the smallest set of features possible. We remove the least important features (retaining the 11 most important features for the six-class case and the 9 most important for the three-class case; see Table 3, columns 3 and 5, for the subset of features for each case) and retrain and optimize the algorithm in each case. We chose these reductions since there was a clear break in the relative importance of different features at these points for each case (see Table 3).

Removing additional features (e.g., using only the top five most important features for the three-class case) still performed well for  $\text{FeO}_x$  and rBC but led to significantly worse classification accuracy for the dust-like aerosols, suggesting the additional features provided enough information to reduce the misclassification of dust-like aerosols from rBC by a factor of 2. The classification accuracy for these feature reductions is quantified in Tables S2 and S3. Another advantage of reducing features is that the algorithm can be trained faster (in the case of the six-class case, the training time was reduced from 92 to 52 s when using 11 rather than 17 features for each sample).

**Table 3.** Importance of different features. The relative importance of the different features for the optimal random forest for the six-class and three-class cases. All denotes that all 17 features were used in the algorithm, and reduced indicates that only a subset of features were included as input to the learning algorithm (11 features for the six-class case and 9 features for the three-class case). The top five most important features in each category are bolded.

Feature	All (six classes)	Reduced (six classes)	All (three classes)	Reduced (three classes)
Blue peak amplitude	<b>0.064</b>	<b>0.075</b>	<b>0.049</b>	<b>0.057</b>
Color ratio	<b>0.391</b>	<b>0.358</b>	<b>0.455</b>	<b>0.587</b>
Core scattering	0.032	0.044	<b>0.051</b>	0.019
Total scattering maximum	0.021	–	0.014	–
Postincandescent scattering	<b>0.148</b>	<b>0.190</b>	<b>0.183</b>	<b>0.136</b>
Evaporation scattering size	<b>0.042</b>	<b>0.053</b>	<b>0.043</b>	0.037
Position-sensitive wideness	0.018	–	0.011	–
Minimum scattering before incandescence	0.023	–	0.020	–
Position-sensitive trigger position	0.024	–	0.017	–
Scatter peak location	0.026	0.045	0.019	–
Saturation width	0.028	0.046	0.015	–
Incandescent start position	0.039	<b>0.054</b>	0.029	<b>0.057</b>
Evaporation point	0.035	0.052	0.025	<b>0.054</b>
Incandescent total length	0.030	0.036	0.023	0.028
Incandescent used length	<b>0.044</b>	0.047	0.026	0.025
Light on laser intensity	0.018	–	0.011	–
Width fraction from center	0.016	–	0.011	–

## 4.2 Confusion matrices

To quantify the performance of the algorithm for each of the cases, we visualize the true positive and false positive rates for each class using a confusion matrix (Fig. 6). Confusion matrices are useful ways to visualize how well a classifier performs. For each class, they give the number of particles of that class that are predicted to belong to that class (the true positives, along the diagonal) vs. all of the misidentifications of the particles as other classes (the false positives, the off-diagonal elements of the matrix). Since our test data set does not have the same number of particles for each class, we normalize along each horizontal row to give the fractional portion predicted for each of the class labels.

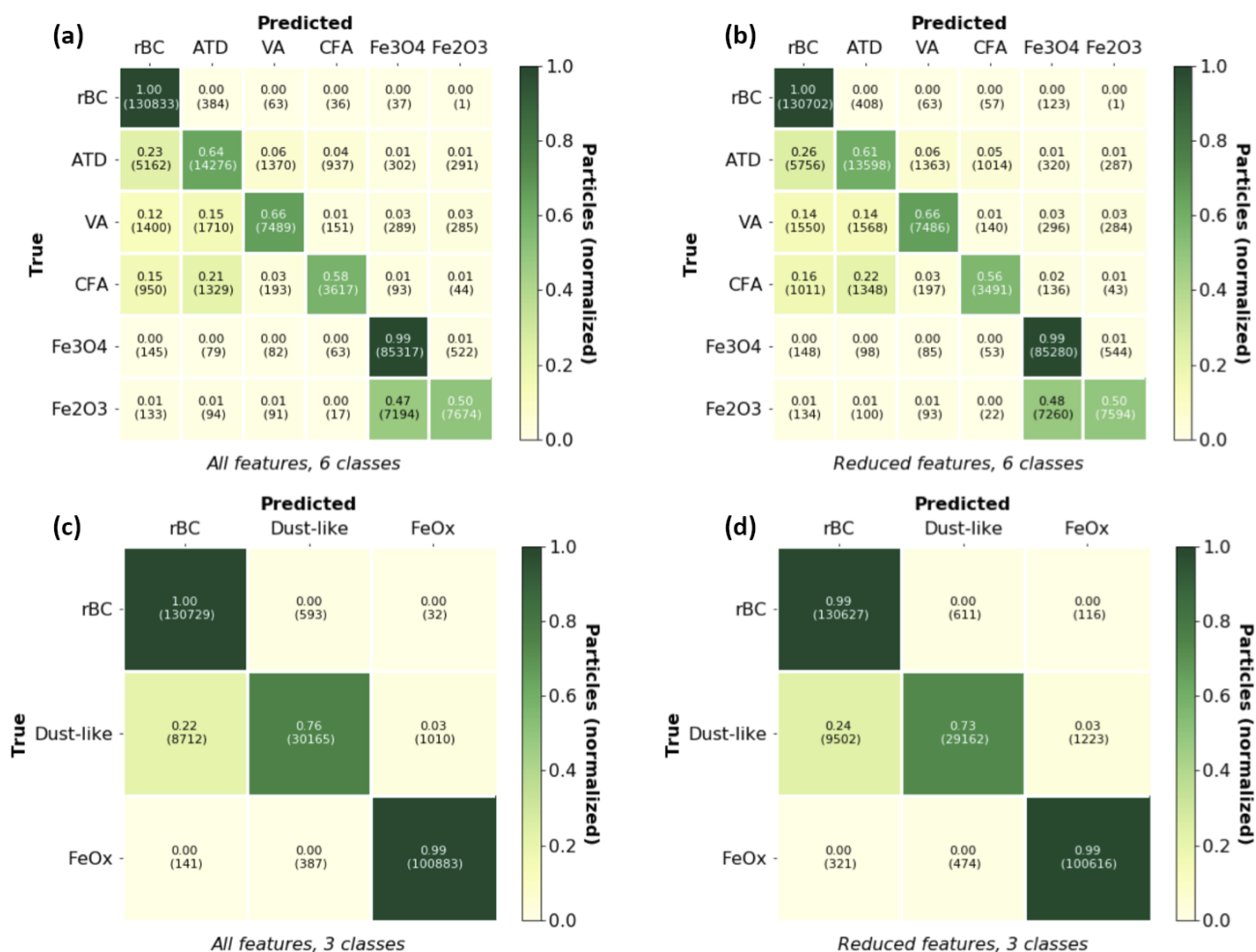
For both the six-class and three-class cases, we find the worst performance for the particles containing metallic inclusions (ATD, VA, FA – the dust-like aerosols). These particles are most likely to be misidentified as either rBC or as one another. One reason for this may be the imperfections of the laboratory data sets. Previous work has noted that there is a small fraction of rBC present in laboratory samples of ATD (Susan Kaspari, personal communication, 2019), which likely contributes to the high rate of errors between ATD and rBC. (Color ratios of particles detected by the SP2 in ATD samples were consistent with rBC and were subsequently removed after heat treating the samples – Susan Kaspari, personal communication, 2019.) A significant fraction of the color ratios for the incandescent aerosols detected in the FA laboratory samples also demonstrated a color ratio distribution more consistent with rBC, suggesting a fraction of the

incandescent aerosols detected in fly ash may also be rBC (see Fig. S1). Additionally, because of the low incandescent rate of these particles, we had acquired the fewest number of training examples for these data sets. In general, these aerosols are much less likely to be misidentified as  $\text{FeO}_x$  (1%–3% false positives) by the trained algorithm than as rBC, however.

We also find that rBC is unlikely to be misidentified as any of the other aerosol types, including  $\text{FeO}_x$ . For the six-class case,  $\text{Fe}_3\text{O}_4$  was more likely to be correctly identified than  $\text{Fe}_2\text{O}_3$ ; this could be because some portion of the  $\text{Fe}_2\text{O}_3$  is more similar to  $\text{Fe}_3\text{O}_4$  or perhaps because of the imbalance of training examples for the two particles. The confusion matrices make it clear that  $\text{Fe}_3\text{O}_4$  and  $\text{Fe}_2\text{O}_3$  are more likely to be misclassified as one another than as any of the other particle types.

### 4.2.1 Base case

To provide a basis for the comparison of the different supervised learning algorithms, we also consider a classification scheme which uses only a few features and linear decision boundaries to classify incandescent aerosols. We designate this case as the “base case” and use only the color ratio and incandescent peak height to differentiate rBC from  $\text{FeO}_x$ , as well as the additional criteria of the core scattering to differentiate anthropogenic  $\text{FeO}_x$  from mineral dust with metallic inclusions. This is based on the method that has previously been used in, e.g., Moteki et al. (2017), Ohata et al. (2018), and Yoshida et al. (2018).

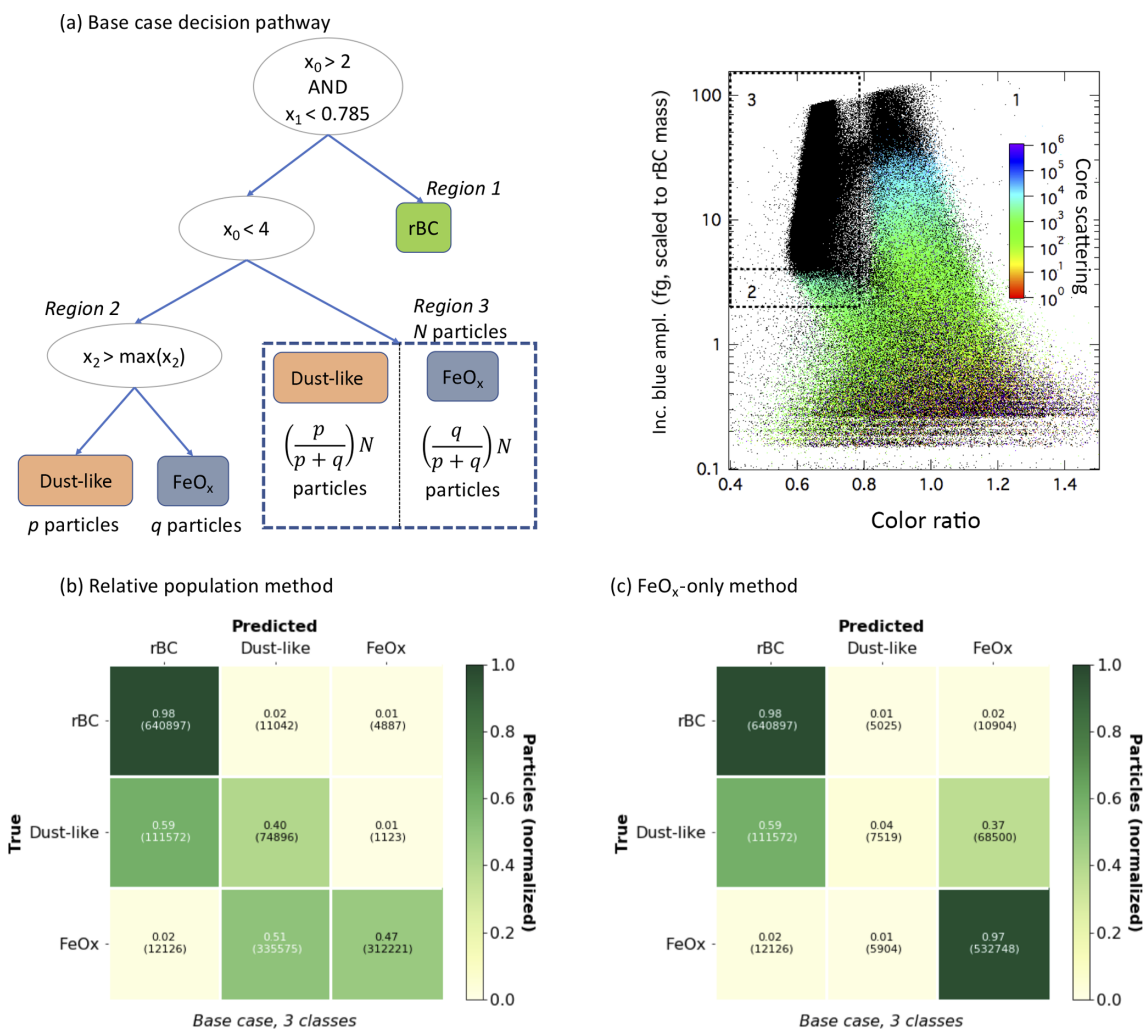


**Figure 6.** Confusion matrices for the six-class and three-class cases. A visualization of the confusion matrices using all features (a) and the reduced-feature space (b) is shown for the laboratory test data. The y axis in each figure indicates the true particle type and the x axis indicates the particle type predicted by the trained random forest on the test data set. The fraction of aerosols identified as a particular class relative to the total number of aerosols is given, with the actual number of particles shown in parentheses.

We visualize the simple pathway through feature space using this method in Fig. 7a. The threshold values for the metallic mode (including both dust-like and FeO<sub>x</sub> samples) for the incandescent peak height were chosen to be 2 fg rBC equivalent mass and the color ratio margin was chosen to be 0.785 from inspection of the modes in the data. We designate the region used for rBC as Region 1 (regions are shown in the right panel of Fig. 7a). As discussed in Sect. 2.5, we can only optically size aerosols in the metallic mode for incandescent peak height between ~2 and 4 fg rBC equivalent mass; this portion of the metallic mode is designated as Region 2. Since we cannot optically size these aerosols outside of this range (Region 3), we need an additional assumption in order to estimate classification performance. Previous work has only designated FeO<sub>x</sub> as anthropogenic or dust-like for aerosols in Region 2 and used this (and offline techniques) to provide context for interpreting atmospheric observations of

FeO<sub>x</sub> in East Asia as dominated by anthropogenic emissions (Moteki et al., 2017; Ohata et al., 2018; Yoshida et al., 2018).

To demonstrate how well the base case provides information for the entire population of aerosols measured by the SP2, we consider two possible assumptions for Region 3 using the laboratory samples; we emphasize that these results are specific to the distributions and relative numbers of particles in the laboratory samples and should be taken as illustrative only. One reasonable assumption would be to assume that the entire population of aerosols in the metallic mode would have a similar ratio of anthropogenic to dust-like FeO<sub>x</sub> as the aerosols in Region 2 (relative population assumption). However, this assumption clearly relies on the number distributions of dust-like and anthropogenic FeO<sub>x</sub> aerosols being similar over the entire range that the SP2 can measure (which would be unknown in ambient populations). Another assumption would be to take all the particles in Region 3 as



**Figure 7.** Base case classification scheme and confusion matrices. (a) We visualize the scheme for classifying aerosols using a simplistic pathway through feature space. The features are the same as in Table 2, and  $x_2 > \max(x_2)$  indicates the detector is saturated for the core scattering measurement. The true case at each node corresponds to the left “child”. We need to make an additional assumption for the particles in the metallic mode that cannot be optically sized (the right-most split, corresponding to Region 3). Saturated values for core scattering are indicated by black dots in the right figure, which shows the incandescent peak height vs. color ratio for the laboratory samples. (b) Confusion matrix for the base case assuming the same ratio of anthropogenic FeO<sub>x</sub> : dust-like in Region 3 as measured in Region 2. Since the particles in Region 3 are differentiated only on a population basis to visualize this case as a confusion matrix we have made the additional assumption that particles are correctly classified up to the total number of aerosols of that class (any additional aerosols identified as that class are assumed to be false positives). (c) Confusion matrix for the base case identifying all aerosols in Region 3 as anthropogenic FeO<sub>x</sub>, to provide an upper limit on anthropogenic FeO<sub>x</sub>.

an upper limit on anthropogenic FeO<sub>x</sub> (FeO<sub>x</sub>-only assumption).

We visualize the classification performance for the base case under these two assumptions as confusion matrices, as shown in Fig. 7b and c. Since it is not possible to identify the six individual particle types using this method, we show only the three-class case for this scheme. The relative population assumption leads to a significant underestimation of anthropogenic FeO<sub>x</sub>, as can be seen in Fig. 7b. (Since the relative population method does not classify individual aerosol particles to visualize this scheme as a confusion matrix we

assume particles are true positives up to the number of total particles of that class in our data set, and they are otherwise false positives.)

For the FeO<sub>x</sub>-only method (Fig. 7c), the true positive rate for rBC and FeO<sub>x</sub> is still quite good (98 % and 97 % respectively), but the false positives are significantly more problematic than for the supervised machine learning classification schemes. Nearly 2 % of the rBC is misidentified as FeO<sub>x</sub>, which could be problematic given that only a small fraction ( $\sim 1/250$ ) of ambient aerosols in urban areas that incandesce in the SP2 are expected to be FeO<sub>x</sub>. The misclas-

sification of the dust-like particles is even more problematic, as nearly 37 % of these particles would be identified as  $\text{FeO}_x$  for this particular data set. Given that in ambient populations we would expect to have no prior knowledge about the relative proportion of anthropogenic to dust-like aerosols, this large misclassification rate could significantly bias the interpretation of the  $\text{FeO}_x$  mass mixing ratio in certain cases.

#### 4.3 Precision and recall

The confusion matrices only provide information about how well the trained algorithm performs on each class in general. However, from a measurement perspective, we are interested in whether there might be significant bias for different classes in certain cases, such as aerosols with smaller or larger incandescent masses. To investigate how well the algorithm performs as a function of the incandescent peak height, we define two metrics, the precision and recall. The precision  $P_i$  for a particular class  $i$  is defined as

$$P_i = \frac{\# \text{ true positives}}{(\# \text{ true positives}) + (\# \text{ false positives})}, \quad (1)$$

and recall  $R_i$  for class  $i$  is defined as

$$R_i = \frac{\# \text{ true positives}}{(\# \text{ of true positives}) + (\# \text{ false negatives})}. \quad (2)$$

True positives are defined as particles of class  $i$  that are predicted to belong to class  $i$ . False positives are particles of other classes ( $j \neq i$ ) that are predicted to belong to class  $i$ . False negatives are particles of class  $i$  that are predicted to belong to a different class  $k$  ( $k \neq i$ ). Precision provides information about how accurately the algorithm identifies particles of a particular class, whereas recall provides information about how many of the relevant particles are actually identified. For  $\text{FeO}_x$ , we are most interested in maximizing the precision (as opposed to the recall), i.e., we would like to avoid falsely identifying other types of particles as  $\text{FeO}_x$ . Lower recall would translate to an underestimation of the mass mixing ratio of  $\text{FeO}_x$ , whereas having lower precision could introduce a significant systematic bias due to other aerosols being misidentified as  $\text{FeO}_x$ .

Figure 8 shows a side-by-side comparison of the performance of the six-class (reduced features) case and three-class (reduced features) case on the laboratory test set. The true and predicted labels for the sample set, shown as a function of the incandescent peak height vs. color ratio, are shown. Comparing the true labeled data set with the predicted labels indicates that the portion of the dust-like particles that overlap with the rBC population is the one that is most likely to be misidentified. As stated previously, this may be due to a fraction of rBC present in the laboratory samples of ATD and FA. The precision and recall for each of the classes is also shown, binned as a function of the incandescent peak height. In general, the performance of the classifiers are better for  $\text{FeO}_x$  at larger masses than at smaller masses; this is

likely impacted by the size distribution of the particles in our training data set (see Fig. S2 for the size distributions of the iron oxide samples).

#### 4.4 Performance on atmospheric observations

We finally apply the machine learning pipeline for the three-class, reduced-feature case to the atmospheric data sets acquired in Boulder, CO (Fig. 9). Generally we observe the three modes that we expect for rBC, dust-like aerosols, and  $\text{FeO}_x$  based upon the predictions of the random forest algorithm. These observations indicate that the laboratory samples in general have similar responses in the SP2 to the aerosols that we are observing in the urban environment. This suggests that the algorithm identifies iron oxides sourced from anthropogenic emissions with the laboratory samples of pure iron oxides. This application demonstrates that a significant feature of  $\text{FeO}_x$  is present in the atmosphere in Boulder, as has previously been observed in urban areas in East Asia (Yoshida et al., 2016; Moteki et al., 2017; Ohata et al., 2018).

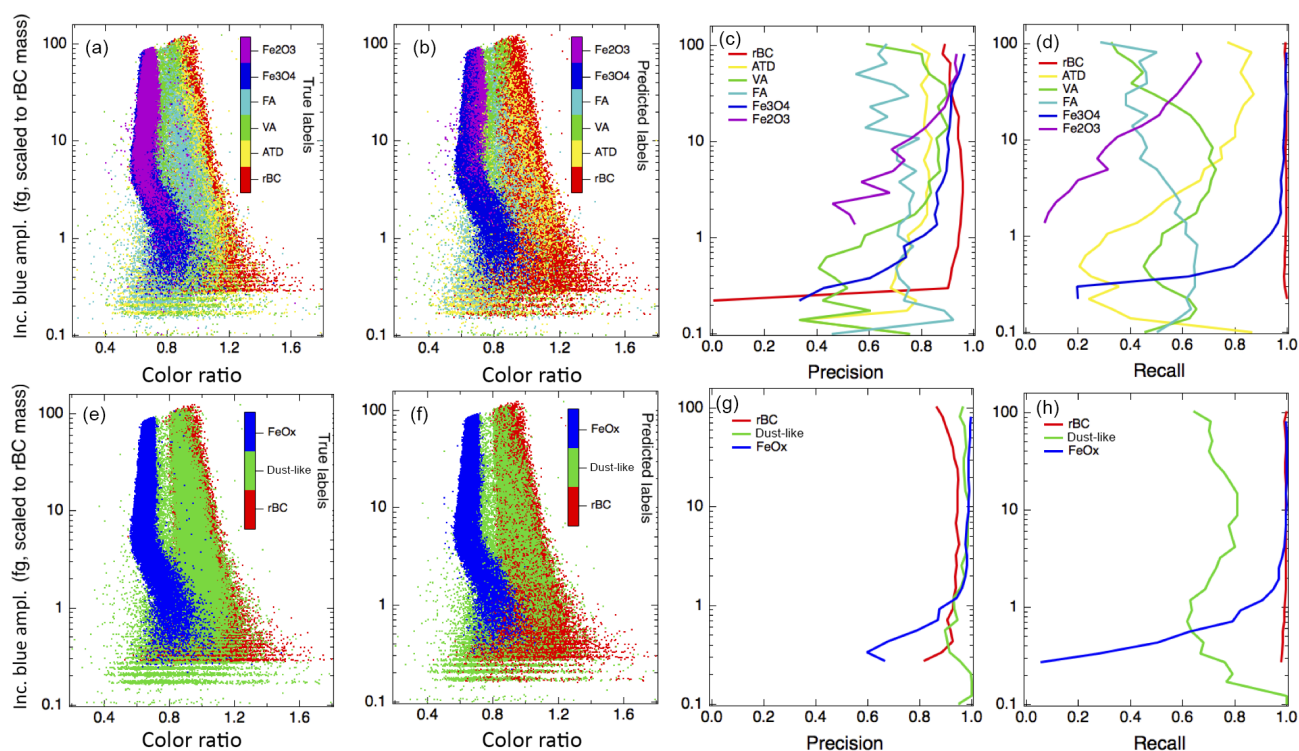
The algorithm also identifies a significant fraction of dust-like aerosols, both at cooler color temperature ratios and mixed into the population of particles in the rBC mode. At the larger masses for the rBC color ratio mode, the algorithm does appear to be misidentifying a fraction of the rBC with cooler color temperature ratios as dust-like particles (as all particles below  $\sim 0.8$  and with incandescent blue amplitudes  $> 5$  fg rBC equivalent incandescence on the shoulder of the rBC mode are identified as dust-like in Fig. 9). This misidentification is likely due to the differences between the SP2 response to ambient rBC vs. fullerene soot, as ambient rBC has a greater prevalence of particles with a lower color temperature ratio than fullerene soot (see Fig. S1). This result suggests that the algorithm has been overtrained to the specific threshold values for the color ratio in this case.

Previous studies have indicated that a fraction of rBC present in the atmosphere may be attached to other aerosols such as natural dust (attached-type rBC) rather than present in a core-shell structure (coated rBC) (Sedlacek III et al., 2012; Dahlkötter et al., 2014; Moteki et al., 2014). In the ambient measurements in Boulder, we found  $\sim 7\%$  of the aerosols in the rBC mode were predicted to be dust-like (excluding color ratios below 0.85 to eliminate any effects of the overtraining due to the threshold color ratio value), similar to previous observations of  $< 10\%$  attached-type rBC in an urban area (Tokyo) (Moteki et al., 2014). This suggests that machine learning could also potentially be used to identify attached-type rBC aerosols over a greater rBC mass range than previously developed algorithms (Moteki et al., 2014).

## 5 Conclusions and recommendations

We explored the advantages and limitations of using supervised machine learning to classify absorbing aerosols de-





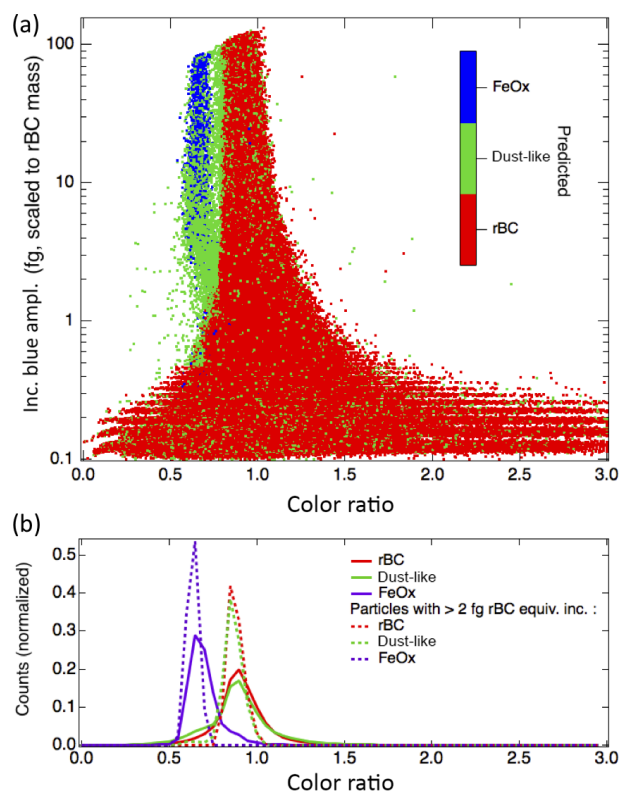
**Figure 8.** True and predicted labels, precision, and recall by incandescent peak height for the six-class (reduced features) and three-class (reduced features) cases. (a–d) show the performance of the six-class classification scheme and the (e–h) show the performance of the three-class classification scheme. (a) and (e) show the laboratory test data, color coded by its true class label (particles differ between the six-class and three-class cases because different subsets of the laboratory data were randomly chosen as test sets in each case). (b) and (f) show the predictions of the algorithms for the class labels. (c) and (g) show the precision for each class, binned by incandescent peak height. (d) and (h) show the recall, binned by incandescent peak height for each class.

tected via laser-induced incandescence. This paper serves as proof of concept that supervised machine learning is a useful technique for analyzing and classifying laser-induced incandescent signals acquired by the SP2. This method improves upon the performance of previous classification methods using only three or four features (the incandescent peak height, the color ratio, the core scattering, and the postincandescent-scattering amplitudes) derived from the single particle signals and indicates that the SP2 does provide enough information via laser-induced incandescence to identify FeO<sub>x</sub> with few misclassifications as other types of aerosols that the SP2 can detect. In past studies, decision boundaries had been based on inspection of the data rather than statistical considerations, and the method presented here provides a more statistically consistent method.

In order to use supervised learning algorithms to classify aerosols with an SP2 during aircraft and field observations, it is very important to acquire the samples for training data sets with the same instrument, optical configuration, and operating conditions as the data sets to be processed. Several of the features (in particular the color ratio) demonstrated strong dependence on detector alignment or may be affected by the specific laser power settings (see Fig. S3 for additional

details). This configuration dependence leads to a greater incidence of misclassified particles if algorithms trained with data taken with one instrument configuration are applied to data sets attained with another, as the algorithms can be over-trained. This makes the application of the algorithms to aircraft observations more challenging, as changes in pressure and flow rates during sampling may also impact some of these features. One potential solution is to take a large training data set simulating a number of different alignment configurations, although for simplicity, we have not explored this approach here.

We recommend that the three-broader-class approach be used, as this method provided clear advantages over the six-class approach. The incandescent onset position of ambient FeO<sub>x</sub> observed in East Asia was found to be between that characteristic of Fe<sub>2</sub>O<sub>3</sub> and Fe<sub>3</sub>O<sub>4</sub> in pure laboratory samples, suggesting that combustion iron oxide aerosols found in the atmosphere may be homogeneous internal mixtures of these two iron oxides (Yoshida et al., 2018). This provides additional motivation to use the three-class classification scheme, as ambient FeO<sub>x</sub> may have characteristics on a continuum between pure laboratory samples of Fe<sub>2</sub>O<sub>3</sub> and Fe<sub>3</sub>O<sub>4</sub>. When applying this method to atmospheric measure-



**Figure 9.** Application of algorithm to observations in Boulder, CO. **(a)** Ambient data acquired from a rooftop inlet demonstrate the performance of the three-class, reduced-feature implementation of the random forest algorithm after it has been trained on laboratory data. A clear feature of  $\text{FeO}_x$  is observed in the ambient data. The larger variety of color ratios at the smaller incandescent peak heights ( $< 0.5$  fg rBC equivalent mass) than observed in laboratory data is due to the greater prevalence of small rBC aerosols in the urban environment than in the nebulized fullerene soot samples. **(b)** Histograms for the color ratios of the particles identified to belong to each of the three classes are shown, both for the entire population identified and also only for particles with larger incandescent blue amplitudes.

ments in Boulder, CO,  $\sim 7\%$  of aerosols in the rBC mode were quantified as dust-like. These aerosols were associated with a greater incidence of nonvolatile particles (that did not evaporate completely in the SP2 laser) than the rBC not identified as dust-like, suggesting that this method may also be useful for identifying rBC attached to other types of aerosols, such as mineral dust. However, because the recall for dust-like aerosols in laboratory samples was only  $\sim 70\%$ , there is still significant room for uncertainty in the interpretation of these aerosols.

Using a random forest ensured good performance and demonstrated that this method in principle works for classifying aerosols detected via laser-induced incandescence; however, we have not specifically optimized this method for computational efficiency, and other supervised learning al-

gorithms may offer advantages in this respect. The cross-validation step required the greatest amount of computation time ( $\sim 1\text{--}2$  h for the 3-fold cross-validation grid search using 48 threads), as it required repeatedly training the model with different options for the hyperparameters. However, this step only needs to be performed once; after the hyperparameters have been optimized, the computation time for training the three-class, reduced-feature case with  $\sim 6 \times 10^6$  particles was approximately 45 s with the optimal set of hyperparameters, and for making predictions on the test set of  $\sim 3 \times 10^6$  particles it was  $< 5$  s. Although we have used a Linux server with multiple processors for this study, we have also deployed this method on a laptop (MacBook Pro with 3 GHz Intel Core i7 processor with four cores and 16 GB 1600 MHz DDR3 memory); in this case training time was  $\sim 300$  s and testing time was  $< 5$  s. Computation time could be reduced by using a smaller training data set, although with some trade-offs in classification accuracy.

Another approach would be to use an unsupervised learning algorithm to classify atmospheric observations, as supervised learning algorithms rely on ambient aerosols having a similar response in the SP2 as laboratory samples. Given the low relative incidence of  $\text{FeO}_x$  vs. rBC, clustering algorithms that assume consistent cluster size would likely perform poorly, however. Some unsupervised approaches, such as hierarchical agglomerative clustering analysis, which has previously been used to classify biological aerosols detected via UV-light-induced fluorescence (Robinson et al., 2013; Ruske et al., 2017, 2018; Savage and Huffman, 2018), are more appropriate for data sets where cluster size is not expected to be consistent. However, these methods are significantly more computationally intensive than the approach we explored here (e.g., with a time complexity scaling as the square of the number of samples or worse; Müllner, 2011).

Here we have taken the approach of leveraging previous feature engineering from SP2 incandescent and scattering signals. Since we are using features derived from processing the raw SP2 time series, some features, particularly those associated with the aerosols that do not incandesce with high efficiency and are internally mixed with nonvolatile materials (ATD, VA, and FA), may be biased due to detector saturation. One solution to this issue that may improve classification performance for these aerosols would be to use the raw time-resolved single particle signals from the four channels directly as features, although this would be computationally more expensive than the approach taken here. A further adaptation of this method would use representation learning/deep learning to learn features directly from the raw SP2 signals; however, these methods are generally computationally expensive (often requiring the use of GPUs for parallel processing). These algorithms also have a large number of adjustable parameters that makes their out-of-the-box application more challenging. We do not consider this approach here but suggest it may be a potentially useful direction for future research.

Several recent observations of the size distributions of ambient  $\text{FeO}_x$  in East Asia have indicated a significant number fraction of  $\text{FeO}_x$  at smaller sizes ( $< 300$  nm) (Moteki et al., 2017; Yoshida et al., 2018); however, the nebulized samples of  $\text{Fe}_2\text{O}_3$  and  $\text{Fe}_3\text{O}_4$  in our laboratory data set were predominantly between 350 and 1200 nm volume equivalent diameter (see Fig. S2). These results indicate that particular care would need to be taken when acquiring a training data set appropriate for classifying smaller iron oxide aerosols. These aerosols have been linked to neurodegenerative diseases such as Alzheimer's and have even been detected inside the human brain (Maher et al., 2016); improving their atmospheric detection is an important concern for air quality and human health. The worst classification performance was observed for smaller  $\text{FeO}_x$  particles, likely because there were fewer examples of these  $\text{FeO}_x$  samples in the training data sets than larger  $\text{FeO}_x$ . Given that even in the best-case scenarios machine learning algorithms generally do not perform with  $> 98\%$ – $99\%$  accuracy, the significantly greater presence of rBC in the atmosphere would likely lead to significant misclassifications of rBC as  $\text{FeO}_x$  at the smallest sizes even in the best scenarios, suggesting other online approaches should be explored.

*Code and data availability.* Supervised learning algorithms used in this work are available through the sklearn python package (<http://scikit-learn.org/stable/>) (Pedregosa et al., 2011). Code used in this study and laboratory data used for training/testing are available upon request from the author.

## Appendix A: Decision trees

Decision trees are classifiers that work by sequentially subdividing the training data set based on learned threshold values of the features at each node. At each node, the threshold values are determined by minimizing the impurity  $F$  over the classes associated with the subset of samples in the two resulting “children”. Typical measurements of node impurity  $F$  are the information entropy or the Gini index (Mohri et al., 2012). Information entropy provides a metric for quantifying how much information is in an event; for example, decisions that split the training samples such that a single class is represented in each child have lower entropy than splits resulting in multiple classes, since there is greater information gain. The Gini index measures the likelihood that a randomly chosen sample would be mislabeled given the values of the labels in the subset associated with each child. After a sufficient purity according to one of these metrics has been reached, the algorithm is stopped. The class associated with the majority of the training samples after the terminal split along any particular branch of the tree is associated with that leaf, and new samples that satisfy the same criteria are predicted to have that class.

Decision trees are fairly robust even for cases of features that are not normally distributed. They have the advantage of having few tunable parameters, meaning that their out-of-the-box implementation is simpler than many other machine learning algorithms. They can also directly handle multiclass classification problems such as the one we consider here. They are nonparametric machine learning algorithms; i.e., no a priori assumptions are made about the function to be learned, and the complexity of the model is a function of the training data set size (Goodfellow et al., 2016). These types of algorithms do typically require more training data and longer training times than parametric models but can generally result in more powerful models. However, a common problem for decision trees is overfitting, meaning that their generalization to new examples can be poor; even small changes in the training data set can lead to different outcomes. Using an ensemble of decision trees (a random forest) typically provides better generalization performance than a single decision tree.

*Supplement.* The supplement related to this article is available online at: <https://doi.org/10.5194/amt-12-3885-2019-supplement>.

*Author contributions.* KL conceived of the study, took the laboratory data, and did the analysis.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* The author would like to thank Joshua Schwarz for useful discussion in preparing this paper and carrying out this study. Bernadett Weinzierl is acknowledged for providing the volcanic ash samples and Karl Froyd for providing the samples of coal fly ash. Douglas Ohlhorst and Richard Tisinai are also acknowledged for their help in accessing the computational resources used to carry out this study.

*Financial support.* This research has been supported by the NASA Tropospheric Composition Program, the NASA Radiation Sciences Program, and the NASA Upper Atmosphere Research Program.

*Review statement.* This paper was edited by Mingjin Tang and reviewed by two anonymous referees.

## References

- Adachi, K., Moteki, N., Kondo, Y., and Igarashi, Y.: Mixing states of light-absorbing particles measured using a transmission electron microscope and a single-particle soot photometer in Tokyo, Japan, *J. Geophys. Res.-Atmos.*, 121, 9153–9164, 2016.
- Baumgardner, D., Popovicheva, O., Allan, J., Bernardoni, V., Cao, J., Cavalli, F., Cozic, J., Diapouli, E., Eleftheriadis, K., Genberg, P. J., Gonzalez, C., Gysel, M., John, A., Kirchstetter, T. W., Kuhlbusch, T. A. J., Laborde, M., Lack, D., Müller, T., Niessner, R., Petzold, A., Piazzalunga, A., Putaud, J. P., Schwarz, J., Sheridan, P., Subramanian, R., Swietlicki, E., Valli, G., Vecchi, R., and Viana, M.: Soot reference materials for instrument calibration and intercomparisons: a workshop summary with recommendations, *Atmos. Meas. Tech.*, 5, 1869–1887, <https://doi.org/10.5194/amt-5-1869-2012>, 2012.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, 2001.
- Christopoulos, C. D., Garimella, S., Zawadowicz, M. A., Möhler, O., and Cziczo, D. J.: A machine learning approach to aerosol classification for single-particle mass spectrometry, *Atmos. Meas. Tech.*, 11, 5687–5699, <https://doi.org/10.5194/amt-11-5687-2018>, 2018.
- Dahlkötter, F., Gysel, M., Sauer, D., Minikin, A., Baumann, R., Seifert, P., Ansmann, A., Fromm, M., Voigt, C., and Weinzierl, B.: The Pagami Creek smoke plume after long-range transport to the upper troposphere over Europe – aerosol properties and black carbon mixing state, *Atmos. Chem. Phys.*, 14, 6111–6137, <https://doi.org/10.5194/acp-14-6111-2014>, 2014.
- Dall’Osto, M., Beddows, D., Harrison, R. M., and Onat, B.: Fine iron aerosols are internally mixed with nitrate in the urban European atmosphere, *Environ. Sci. Technol.*, 50, 4212–4220, 2016.
- Freund, Y. and Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.*, 55, 119–139, 1997.
- Gao, R., Schwarz, J., Kelly, K., Fahey, D., Watts, L., Thompson, T., Spackman, J., Slowik, J., Cross, E., Han, J.-H., Davidovits, P., Onasch, T. B., and Worsnop, D. R.: A novel method for estimating light-scattering properties of soot aerosols using a modified single-particle soot photometer, *Aerosol Sci. Technol.*, 41, 125–135, 2007.
- Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, <http://www.deeplearningbook.org> (last access: 15 March 2019), 2016.
- Gysel, M., Laborde, M., Mensah, A. A., Corbin, J. C., Keller, A., Kim, J., Petzold, A., and Sierau, B.: Technical Note: The single particle soot photometer fails to reliably detect PALAS soot nanoparticles, *Atmos. Meas. Tech.*, 5, 3099–3107, <https://doi.org/10.5194/amt-5-3099-2012>, 2012.
- Heimerl, K., Weinzierl, B., Gysel, M., Baumgardner, D., Kok, G., Linke, C., Schnaiter, M., Schwarz, J., Sheridan, P., Subramanian, R., and Walker, J.: Using a Single Particle Soot Photometer to detect and distinguish different absorbing aerosol types, *European Aerosol Conference*, 44, 663–675, 2012.
- Huffman, D. R. and Stapp, J. L.: Optical measurements on solids of possible interstellar importance, in: *Interstellar dust and related topics*, Springer, 297–301, 1973.
- Ito, A., Lin, G., and Penner, J. E.: Radiative forcing by light-absorbing aerosols of pyrogenetic iron oxides, *Sci. Rep.*, 8, 1–11, 2018.
- Jeong, G. Y. and Nousiainen, T.: TEM analysis of the internal structures and mineralogy of Asian dust particles and the implications for optical modeling, *Atmos. Chem. Phys.*, 14, 7233–7254, <https://doi.org/10.5194/acp-14-7233-2014>, 2014.
- Kondo, Y., Sahu, L., Moteki, N., Khan, F., Takegawa, N., Liu, X., Koike, M., and Miyakawa, T.: Consistency and traceability of black carbon measurements made by laser-induced incandescence, thermal-optical transmittance, and filter-based photoabsorption techniques, *Aerosol Sci. Technol.*, 45, 295–312, 2011.
- Kutchko, B. G. and Kim, A. G.: Fly ash characterization by SEM-EDS, *Fuel*, 85, 2537–2544, 2006.
- Li, W., Xu, L., Liu, X., Zhang, J., Lin, Y., Yao, X., Gao, H., Zhang, D., Chen, J., Wang, W., Harrison, R. M., Zhang, X., Shao, L., Fu, P., Nenes, A., and Shi, Z.: Air pollution–aerosol interactions produce more bioavailable iron for ocean ecosystems, *Sci. Adv.*, 3, e1601749, <https://doi.org/10.1126/sciadv.1601749>, 2017.
- Liati, A., Pandurangi, S. S., Boulouchos, K., Schreiber, D., and Dasilva, Y. A. R.: Metal nanoparticles in diesel exhaust derived by in-cylinder melting of detached engine fragments, *Atmos. Environ.*, 101, 34–40, 2015.
- Liu, D., Taylor, J. W., Crosier, J., Marsden, N., Bower, K. N., Lloyd, G., Ryder, C. L., Brooke, J. K., Cotton, R., Marengo, F., Blyth, A., Cui, Z., Estelles, V., Gallagher, M., Coe, H., and Choularton, T. W.: Aircraft and ground measurements of dust aerosols over the west African coast in summer 2015 during ICE-D and AER-D, *Atmos. Chem. Phys.*, 18, 3817–3838, <https://doi.org/10.5194/acp-18-3817-2018>, 2018.

- Maher, B. A., Ahmed, I. A., Karloukovski, V., MacLaren, D. A., Foulds, P. G., Allsop, D., Mann, D. M., Torres-Jardón, R., and Calderon-Garciduenas, L.: Magnetite pollution nanoparticles in the human brain, *P. Natl. Acad. Sci. USA*, 113, 10797–10801, 2016.
- Matsui, H., Mahowald, N. M., Moteki, N., Hamilton, D. S., Ohata, S., Yoshida, A., Koike, M., Scanza, R. A., and Flanner, M. G.: Anthropogenic combustion iron as a complex climate forcer, *Nat. Commun.*, 9, 1593, <https://doi.org/10.1038/s41467-018-03997-0>, 2018.
- Mohri, M., Talwalkar, A., and Rostamizadeh, A.: *Foundations of machine learning (adaptive computation and machine learning series)*, Mit Press Cambridge, MA, 2012.
- Moteki, N. and Kondo, Y.: Dependence of laser-induced incandescence on physical properties of black carbon aerosols: Measurements and theoretical interpretation, *Aerosol Sci. Technol.*, 44, 663–675, 2010.
- Moteki, N., Kondo, Y., and Nakamura, S.-I.: Method to measure refractive indices of small nonspherical particles: Application to black carbon particles, *J. Aerosol Sci.*, 41, 513–521, 2010.
- Moteki, N., Kondo, Y., and Adachi, K.: Identification by single-particle soot photometer of black carbon particles attached to other particles: Laboratory experiments and ground observations in Tokyo, *J. Geophys. Res.-Atmos.*, 119, 1031–1043, 2014.
- Moteki, N., Adachi, K., Ohata, S., Yoshida, A., Harigaya, T., Koike, M., and Kondo, Y.: Anthropogenic iron oxide aerosols enhance atmospheric heating, *Nat. Commun.*, 8, 15329, <https://doi.org/10.1038/ncomms15329>, 2017.
- Müllner, D.: Modern hierarchical, agglomerative clustering algorithms, *arXiv preprint arXiv:1109.2378*, 1–29, 2011.
- Ohata, S., Yoshida, A., Moteki, N., Adachi, K., Takahashi, Y., Kurisu, M., and Koike, M.: Abundance of Light-Absorbing Anthropogenic Iron Oxide Aerosols in the Urban Atmosphere and Their Emission Sources, *J. Geophys. Res.-Atmos.*, 123, 8115–8134, 2018.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 12, 2825–2830, 2011.
- Rocha-Lima, A., Martins, J. V., Remer, L. A., Krotkov, N. A., Tabacniks, M. H., Ben-Ami, Y., and Artaxo, P.: Optical, microphysical and compositional properties of the Eyjafjallajökull volcanic ash, *Atmos. Chem. Phys.*, 14, 10649–10661, <https://doi.org/10.5194/acp-14-10649-2014>, 2014.
- Robinson, N. H., Allan, J. D., Huffman, J. A., Kaye, P. H., Foot, V. E., and Gallagher, M.: Cluster analysis of WIBS single-particle bioaerosol data, *Atmos. Meas. Tech.*, 6, 337–347, <https://doi.org/10.5194/amt-6-337-2013>, 2013.
- Ruske, S., Topping, D. O., Foot, V. E., Kaye, P. H., Stanley, W. R., Crawford, I., Morse, A. P., and Gallagher, M. W.: Evaluation of machine learning algorithms for classification of primary biological aerosol using a new UV-LIF spectrometer, *Atmos. Meas. Tech.*, 10, 695–708, <https://doi.org/10.5194/amt-10-695-2017>, 2017.
- Ruske, S., Topping, D. O., Foot, V. E., Morse, A. P., and Gallagher, M. W.: Machine learning for improved data analysis of biological aerosol using the WIBS, *Atmos. Meas. Tech.*, 11, 6203–6230, <https://doi.org/10.5194/amt-11-6203-2018>, 2018.
- Savage, N. J. and Huffman, J. A.: Evaluation of a hierarchical agglomerative clustering method applied to WIBS laboratory data for improved discrimination of biological particles by comparing data preparation techniques, *Atmos. Meas. Tech.*, 11, 4929–4942, <https://doi.org/10.5194/amt-11-4929-2018>, 2018.
- Schwarz, J., Gao, R., Fahey, D., Thomson, D., Watts, L., Wilson, J., Reeves, J., Darbeheshti, M., Baumgardner, D., Kok, G., Chung, S. H., Schulz, M., Hendricks, J., Lauer, A., Kärcher, B., Slowik, J. G., Rosenlof, K. H., Thompson, T. L., Langford, A. O., Loewenstein, M., and Aikin, K. C.: Single-particle measurements of midlatitude black carbon and light-scattering aerosols from the boundary layer to the lower stratosphere, *J. Geophys. Res.-Atmos.*, 111, 1–15, 2006.
- Schwarz, J., Spackman, J., Gao, R., Perring, A., Cross, E., Onasch, T., Ahern, A., Wrobel, W., Davidovits, P., Olfert, J., Dubey, M. K., Mazzoleni, C., and Fahey, D. W.: The detection efficiency of the single particle soot photometer, *Aerosol Sci. Technol.*, 44, 612–628, 2010.
- Sedlacek III, A. J., Lewis, E. R., Kleinman, L., Xu, J., and Zhang, Q.: Determination of and evidence for non-core-shell structure of particles containing black carbon using the Single-Particle Soot Photometer (SP2), *Geophys. Res. Lett.*, 39, L06802, <https://doi.org/10.1029/2012GL050905>, 2012.
- Stephens, M., Turner, N., and Sandberg, J.: Particle identification by laser-induced incandescence in a solid-state laser cavity, *Appl. Optics*, 42, 3726–3736, 2003.
- Yoshida, A., Moteki, N., Ohata, S., Mori, T., Tada, R., Dagsson-Waldhauserová, P., and Kondo, Y.: Detection of light-absorbing iron oxide particles using a modified single-particle soot photometer, *Aerosol Sci. Technol.*, 50, 1–4, 2016.
- Yoshida, A., Ohata, S., Moteki, N., Adachi, K., Mori, T., Koike, M., and Takami, A.: Abundance and emission flux of the anthropogenic iron oxide aerosols from the East Asian continental outflow, *J. Geophys. Res.-Atmos.*, 123, 11194–11209, <https://doi.org/10.1029/2018JD028665>, 2018.
- Zawadowicz, M. A., Froyd, K. D., Murphy, D. M., and Cziczo, D. J.: Improved identification of primary biological aerosol particles using single-particle mass spectrometry, *Atmos. Chem. Phys.*, 17, 7193–7212, <https://doi.org/10.5194/acp-17-7193-2017>, 2017.