Answers to Specific Comments:

Is there any specific reason for adopting June 29 as the demarcation between labeling and blind data? Please specify. There might be a significant uncertainty when choosing different demarcations. The authors need to address such uncertainties in the revised manuscript.

MS data were recorded from June 26th to July 2nd, 2018, with only 2 hours of measurement on July 2nd, hence roughly 7 days in total. For simplicity, we arbitrarily chose the data from June 26th to June 30th to create the benchmark dataset and for training and the June 30-31 data for testing, as we did not note a severe time-dependence of the composition of the 8 chosen classes during the whole measurement campaign.

In Section 3, a brief description of the sampling lines, flow rate, and residence time for all measurements would be good. Further, it would be beneficial if the authors provide more details about the calibrations for the SPMS.

The SPMS instrument is a bipolar time-of-flight mass spectrometer (ATOF-MS) with an aerodynamic lens and an optical sizing unit. Detailed descriptions of its functionality can be found in (L. Li et al., 2011) and (Zhou et al., 2016). Briefly, for velocimetric particle sizing, two continuous wave lasers with a wavelength of 532 nm, ellipsoidal mirrors, and photomultipliers are employed. The compact mass spectrometer in Z-TOF geometry (Pratt and Prather, 2012), is equipped with continuous-wave 248.3 nm KrF excimer lasers used to detect the particles prior to LDI (L. Li et al., 2011; Passig et al., 2020; Schade et al., 2019; Y. Zhou et al., 2016). This wavelength is very well suited for resonance-enhanced laser desorption/ionization (LDI) of iron and other transition metals of interest for the analysis of ship exhaust particles in ambient air (Passig et al., 2020). The optical setup was optimized to achieve a hit rate of about 50% (#mass spectra/sized particles). The lens (f = 200 mm) is brought to an off-focus position of 7 mm relative to the particle beam, resulting in a spot size of 150 x 300 µm and an intensity of 5 GW cm^{-2} at 6 mJ pulse energy (Passig et al., 2020; Schade et al., 2019). To analyse a sufficient number of particles, a multi-stage virtual impactor was used (Model 4240, MSP Corp., USA). From the 300 L min^{-1} intake airflow, particles were concentrated into 1 L min^{-1} carrier gas stream (6 x 4 mm conducting tube), from which 0.1 L min^{-1} entered the SPMS instrument after a transfer time of few seconds. Monodisperse polystyrene particles were used for the size calibration of inlet and soot particles for the mass calibration of the mass spectrometer. Particle numbers were not corrected for size-dependent or type-dependent detection efficiencies (Shen et al., 2019).

More evidence needs to be listed to support the eight coarse particle classes used in this study. Why didn't the authors choose 7 or 9 coarse particle classes instead of 8?

Of note, the abundance of particles in different classes varies greatly, for example, particles in the K-rich and OC-EC classes make up about 80% of all particles measured. In order to obtain a benchmark dataset with comparable numbers of spectra in each class (ideally the same number of particles in each class, for a balanced dataset) we viewed a number of mass spectra much larger than the 24,000 mass spectra we eventually assigned to one of 8 classes ('labeled') for the results presented in this paper. The vast majority of these mass spectra belonged to the 8 classes. The number of particles not belonging to one of these 8 classes was very small and not enough to create separate classes for them in the dataset.

Some secondary particles are classified as primary emissions, e.g., K-rich particles and OC-EC particles. I also notice that V-rich particles contain 54/56Fe+ signal.

The main purpose of this paper is to verify whether supervised learning can accomplish the task of automated classification of SPMS data. Therefore, the 8 classes in this dataset were not divided further into subclasses (primary or secondary particles). Later, we will create refined datasets of labeled data, taken into account, for example, the degree of aging of particles emitted from ships, refining e.g. the Fe-rich class into Fe-EC, Fe-Sul, Fe-Nit, Fe-Nit-EC (primary and secondary particles).

Does it mean that the resolved V-rich particles were a mixed factor? How about combining the two factors together?

Indeed, many particles within the V-rich class showed a mixture with 54/56Fe. Specifically, the aerosol particles emitted from ships using heavy fuel oil (HFO) have shown a frequent combination of V-Fe-Ni ions.

It would be better to add detailed comparisons (including time series and mass spectra) between the predictions and the results obtained in a previous study using ART-2a. Otherwise, we don't know how significant the similarity variations are.

1) The results obtained with the ART-2a and ML-based approaches are not well comparable at the level of classification accuracy. With ART-2a, depending on the vigilance parameter (range from 0 to 1) a different number of clusters is generated. In most cases, for a practically feasible classification result, manual selecting and merging clusters are required. The smaller the vigilance factor, the fewer the number of generated clusters, and the lesser post-processing needed, but the accuracy will undoubtedly decrease. In the extreme case, when the vigilance is set to 1, each mass spectrum will form a different cluster and the classification accuracy would be 100 %, because we need to manually select and merge all of the clusters. Hence, with ART-2a heavily relying on manual post-processing, to compare its performance to that of automated classification algorithms would be not fair.

2) The main purpose of this paper is to verify whether supervised learning can be applied to the fully automated classification task of SPMS data, which cannot be achieved by ART-2a. A detailed comparison with previous studies using ART-2a (although, as motivated, difficult and case-dependent) would significantly increase the length of the text and make the topic of the paper less clear. We will consider taking up this comparison in a future paper specifically devoted to it, addressing also the balance between manual post-processing workload and the classification accuracy.