We thank Referee 1 for the highly appreciated review. We refer to the individual comments as follows:

General comments

Beecken et al. describes the results of a measurement campaign undertaken to evaluate the performance of different instruments for the monitoring of pollutants emitted by ships. The manuscript focuses on SO_x emissions and compares the results of the different instruments and measurement principles. The importance of proper calibration is highlighted. Generally, the manuscript is well written and fits into the scope of AMT. It can be accepted for publication after some minor clarifications.

Sect 2: Is the evaluation process the same for all instruments? Where the same evaluation programs used or where there at least common settings (for example for peak detection and ship assignment to a plume)? A brief summary of the steps of the evaluation process might be beneficial.

AC: The principle of evaluation process is similar between the groups. However, the individual implementations are also reflecting the different instrumentations. The following sentence is added for clarification: "The exact implementations of the FSC calculation vary between the different groups according to their instrumentations and corrections for instrumental cross-sensitivities to other gas species, but all follow the principle described above. A detailed description of the instrument individual data analysis can be found in SCIPPER deliverable D2.3, Section 2 (Beecken et al., 2019). " (l. 112ff).

Sect 2.1: Here the calculation of the fuel sulphur content is explained. I feel there should be a brief description of how the baseline is defined for SO_2 and CO_2. Are the baselines defined in the same way for all instruments? Are there any additional sources next to ships that could cause enhancements in SO_2 and CO_2 at the same time?

AC: A short description of the baseline definition is added in this section now: "The baseline which is used for subtracting the background was obtained from the ambient VMR levels before and after the plume appears in the sensor signals. " (l. 100f).
Regarding influence of other sources, a clarification is added in section 2.2.1: "The allocation of the measured plumes with individual vessels is achieved through simultaneous wind and AIS data recording. Measurements are discarded in cases where several sources cannot be distinguished, such as from potentially mixed plumes of two vessels passing the sampling site at the same time. "
(l. 135ff).
It is also added now in section 2.5 that "… there are no further sources in that area that interfere with the plume measurements. ".

Sect 2.5: Perhaps this section should be moved to the beginning of section 2, before the detailed description of the instruments, data evaluation and description of the uncertainties.

AC: It is preferred to put this section as 2.5 because it relates and names specific systems that are first introduced in section 2.2 *Measurement Systems*. The sections 2.3 *Calibration* and 2.4 *Uncertainty* are directly connected to section 2.2 and were therefore kept in this order

Sect 3.2: Some of the instruments are specifically used by the BSH to monitor ship emissions. Was

the negative bias already known and is this usually corrected for? Is it relevant for the identification of non-compliant ships?

AC: The negative bias affected all systems except exp.uas not only BSH's systems. However, its magnitude was first noticed in this broad study due to the comparison to the high number of references, i.e. fuel samples and bunker delivery notes. It is indeed relevant because it leads to an underestimation of the actual FSC.

Specific comments

Line 146: How are these sweet spots identified, how long does it take to find a good position and how long does the UAV need to be in this position for an accurate measurement? Are the results significantly different for measurements outside the sweet spot?

AC: The sweet spot is described in more detail now in section 2.2.2: "Typical sampling distances for these systems are in the range of 50 to 100 m from the funnel's exit and the UAVs are piloted into sweet spots within the plume. Here, the sweet spot describes a plume region, where the expected VMRs of the species of interest can be well quantified according to the sensor specs. The guiding species is CO2 with its VMR targeted to be 100 to 200 ppm above the background. Carbon dioxide is measured using a compact Non-Dispersive InfraRed (NDIR) sensor while other species such as SO2, NO, and NO2 are measured using Electro-Chemical (EC) sensors. Typical VMRs in the sweet spots are in the range of a few tens of ppb for SO2, depending on the vessels' fuel and the presence of any abatement systems, and in the order of single digit ppm for NO and several hundred ppb for NO2, respectively. Typical residence times in the sweet spots are between 30 s to several minutes. "
(l. 156ff).
The ratios do not vary significantly within the plume around this region.

Line 177: Any reason why the units are replaced after 100 hours or 1 year after their production? Is regular calibration not good enough?

AC: The EC sensors have a limited lifetime which is stress tested for the applied sensors to be at least 100 hours of operation. This is described in more detail now: "The ECse units sensors have a lifetime of at least 100 hours of operation and can be operated for least one year after production without effects of sensor deterioration that are impacting the measurements (Explicit, 2018). They are replaced accordingly." (l. 190ff).

Line 253: The stationary instruments were set up right next to each other and should always see the same plumes, but it seems that 25% of the detected plumes were only detected by one of the instruments. Any explanations on this?

AC: To distinguish a plume from the background is challenging and the measured VMRs need to show a significant difference above the respective background levels which allows them to be distinguished from noise. So, the differences of the sensitivities of the gas analysers but also the differences in the integration times have an impact on the detection of plumes. Further, the different groups used individual algorithms to detect plumes.

Line 258: Is the deviation a known issue for SO_2 calibration gas and is this regularly tested for?

AC: The calibration gases were tested in different ways by the groups to ensure their quality. And ways to ensure the quality of calibration gases and the validation of measurement

instrumentation are presented in this paper. However, the magnitude of deviation has not been experienced before.

Line 260: What kind of correction was applied and how were possible affected data identified?

AC: The calibration history was fully traceable from the calibration log which allowed to identify the affected data. A cross-comparison delivered the data needed for a mathematical correction. The sentence is now updated to: "The affected data were corrected afterwards through recalculation using an updated calibration curve. " (l. 280f)

Line 320: Was there a specific reason the UAVs measured in different distances?

AC: The two drones were operated with a safe operating distance from each other for deconfliction, please see explanation in line 230.

Technical suggestions

Add NOx = NO + NO2 where it first appeared in the introduction.

AC: This is adapted according to suggestion. (l. 35)

References

Beecken, J., Irjala, M., Weigelt, A., Conde, V., Mellqvist, J., Proud, R., … Duyzer, J. (2019). *Review of available remote systems for ship emission measurements* (No. D2.1). The SCIPPER Project (European Commission - Horizon 2020 No. 814893).
Explicit. (2018). *Airborne Monitoring of Sulphur Emissions from Ships in Danish Waters—2017 Campaign Results*. Ministry of Environment and Food of Denmark. Retrieved from https://www2.mst.dk/Udgiv/publications/2018/04/978-87-93710-00-9.pdf