Answer to Referee #1

We would like to thank Referee #1 for his/her positive and constructive comments and suggestions. We have studied comments carefully and made corrections, which we hope meet with approval. Comments and responses are listed as follows. In order to facilitate the reference to the questions and proposed changes, we use the following color coding:

Color coding:

Referee comment

Our answer

Proposed change in manuscript

This is an interesting paper on an important issue. Sulphur compliance monitoring at sea. The method using drones is rather new and the paper shows details that could help other researchers in this area. Especially the verification measurements presented in the paper and the experiences with the on-board verification attempts are important. Worldwide readers will be interested in the results of the measurements but also in the experiences with on-board inspections. The paper reads well and although it is only a small contribution, I think it deserves publication paying attention to some of the points below. I would like to see a more detailed and perhaps more quantitative treatment of the way the measured data are handled and converted to FSC and when they are rejected. I have only a few small comments on the English (see below).

Thank you for the comments, we are very encouraged.

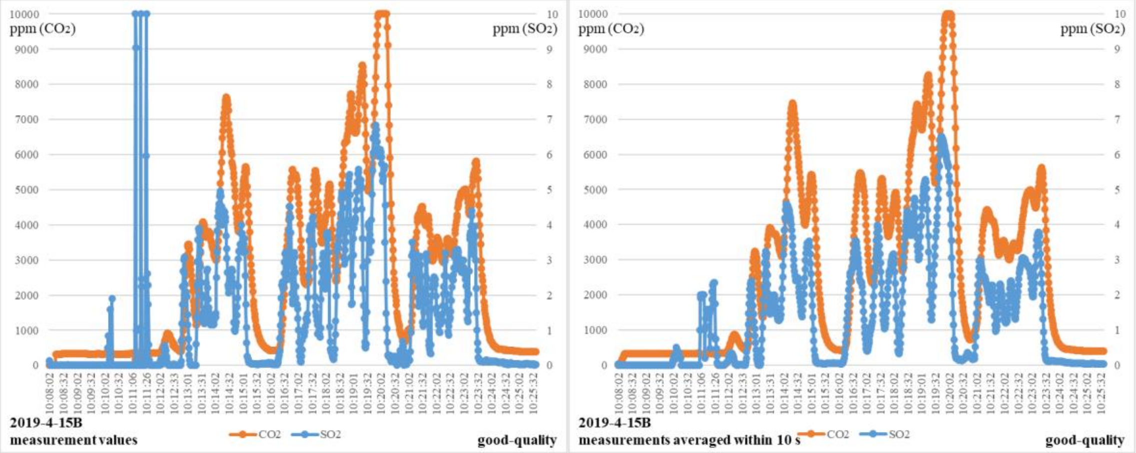
Other more general comments:

- The qualifications poor and good are recognizable. We use that as well in our monitoring. Yet I would like to challenge the authors to come up with a more objective assessment of the quality of each measurement or at least some description as to why some measurements are considered poor. It could be difficult to find objective measures, but it seems needed before the method will become a true enforcement tool.

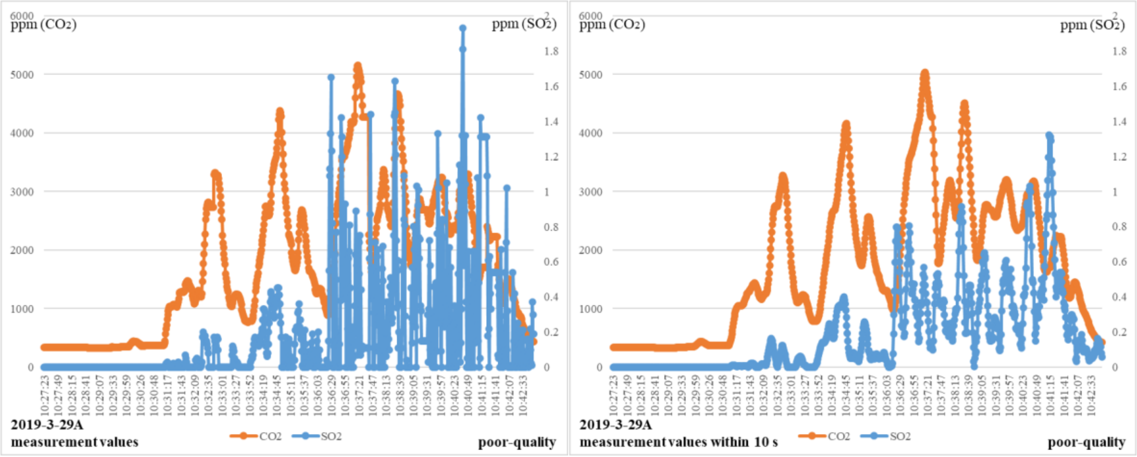
How to objectively and correctly evaluate the quality of plume data is indeed a very important scientific issue, which is also the key to the effective application of the measurement results for maritime law enforcement. I think it would be ideal to design a computational model with input values for plume measurements, output values for FSC results and corresponding confidence levels (or a score that represents the quality of plume data).

This model requires two aspects of work: 1. Adequate an adequate data sets, including gas measurements and the true value of FSC. However, in many cases the true value of the FSC is not available, especially sailing ships. 2. Design and implement the model based on the data sets.

Therefore, there needs a lot of work to implement this model. In fact, that's what we're working on right now. In our current application, however, we assess data quality primarily on the basis of experience. To illustrate how to evaluate data quality, I have supplemented the manuscript with typical good- and poor- data.



(a) (b)



(c) (d)

Figure 5. Typical measurement data for SO_2 and CO_2 concentrations, and their corresponding average values within 10 s. (a) and (b) good-quality data from plume ID 2019-4-15B. (c) and (d) poor-quality data from plume ID 2019-3-29A. There are some errors in the measurements from 10:11:06 to 10:12:02 in (a), which may have been caused by sensor uncertainty. These data were ruled out and did not affect the calculation results.

- The S-content is now derived from the “fluctuating” signal presented in figure 5. It is not entirely clear why the 10 s averaged data would lead to a better result. I can hardly see the difference. Please show why this is better.

I have added description and discussion in the manuscript as bellow.

The average gas concentration within 10 s was chosen for the FSC calculations; however, this does not mean that 9 s or 11 s could not have been selected. To demonstrate this, a comparison calculation was carried out using both 9 s and 11 s, which showed that these led to very little differences in the results. However, it is necessary to ensure that the gradient of the gas measurements is stable within the sampling time (the interval length of the integral). Moreover, the interval length cannot be too short (e.g., 2 s) or too long (e.g., 20 s). If the time is too short, it is difficult to determine whether the measurements are stable and undisturbed over time. Similarly, if the time is too long, it is also difficult to ensure that all of the measurements in the integral interval are stable and undisturbed. In addition, during the flight of the UAV in this study, the time available for measuring the plume was ~5 minutes. As both the ship and the UAV were moving at this time, it was virtually impossible to ensure that the UAV was flying consistently within the plume and obtaining stable measurements. Accordingly, 10 s is also a relatively appropriate value for the measurement process.

– It should be noted that the signal is not noisy but simply reflects the incomplete mixing of the exhaust gas with clean air. The peaks represent the air that is exhausted from the funnel i.e. only in the peaks you will find the ratio between SO2 and CO2 that is a direct measure of the fuel composition. The highest peak is probably the best choice but could still be a result of mixing of clean air with exhaust air and lead to bias in the result. Or is this negligible? I would welcome a discussion showing that the peak height is a good measure.

According to my understanding, the key problem is how to ensure that the gas being measured is evenly mixed. Or, the selected peaks of SO_2 and CO_2 are measured from the complete mixing of the exhaust gas with clean air. And then the ratio between SO_2 and CO_2 can be used to calculate the fuel composition. Assuming that the gas is mixed complete, the variation trend of SO_2 and CO_2 measurements should be the same (given that the corresponding time of the sensor is not consistent, there may be some deviation), and this trend can be easily identified in the peak area.

But not all peaks can be used to calculate the FSC. Therefore, in the previous study, we developed a selection process. The measurements from incomplete mixing have been ruled out. Meanwhile, the measurements with errors have also been ruled out. The maximum values are likely to have been measured in the center of the ship’s plume. At that location, the measurement value is relatively stable, and the probability of interference from other factors is lower. At the same time, the higher the peak value, the greater the proportion of exhaust gas, so the impact from the incomplete mixing of the exhaust gas with clean air is smaller.

To sum up, the obvious and stable maximum peak appears in the measured value over 10 s periods as the calculated value is a more appropriate choice. There are, of course, cases where multiple similar peaks can occur simultaneously. At this time, their calculations may be very similar, in which case, the results obtained by the calculation of the highest peak should have high credibility. However, it is difficult to describe the problem quantitatively and further research is needed. The ideal model that I mentioned above need a lot of work to do. I have added description and discussion in the manuscript as bellow.

Nevertheless, there is also some uncertainty associated with choosing the peak values. After ruling out the peak values across the full range as well as those corresponding to dramatic changes, the global maximum values were selected as the peak values to calculate the FSC. The maximum values probably correspond to the measurements taken in the center of the ship's plume. At that location, the measurement values were relatively stable, and the probability of interference from other factors was lower. Furthermore, the higher the peak value is, the greater the proportion of exhaust gas is; hence, the impact from the incomplete mixing of the exhaust gas with clean air is relatively small.

In summary, the obvious and stable maximum values are selected as peak values to calculate the FSC. There are, of course, situations where multiple similar peaks can occur simultaneously. In this case, their calculated FSCs may be very similar, and the results obtained by the calculation of the highest peak should have high credibility, for instance, the measurements of plume 2019-4-15B.

I also added the selection of peak values of Figure 5 in the Table 2 as bellow.

Table 2: All peak values and their corresponding FSC results. The background values of plume 2019-4-15B were 0 ppm and 310 ppm for SO_2 and CO_2, respectively. The background values of plume 2019-3-29A were 0 ppm and 329 ppm for SO_2 and CO_2, respectively. The remarks indicate the reason for choosing or not choosing the peak. It can be seen that the peak value of plume 2019-4-15B was more obvious and that the results obtained by multiple alternative peaks were similar. The peak of plume 2019-3-29A was less obvious and there were fewer alternative peaks. This was also the basis for distinguishing data as being of a "good"/"poor" quality. The FSC result of selected peak values are marked as "√".

Plume ID	Time point	Peak value of SO_2 and CO_2 (ppm)	Estimated value of FSC (% (m/m))	True value of FSC (% (m/m))	Remark
2019-4-15B	10:12:52	2.406, 2020	0.326	0.168	Reject; less obvious peak values
	10.13.23	3.235, 2372	0.364		
	10.14.07	4.594, 4665	0.245		Non-maximum peaks of alternative peak values
	10.14.57	3.529, 4872	0.179		
	10.16.39	3.549, 4444	0.199		
	10:17:27	3.989, 3911	0.257		
	10:18:01	3.159, 4607	0.171		
	10:18:47	4.757, 6895	0.168		
	10:19:11	5.287, 7634	0.167 (√)		Maximum peak of the alternative peak value
	10:19:46	6.515, 8100	0.194		Reject; measurements exceeded the range
2019-3-29A	10:34:41	0.399, 3880	0.026	0.035	Reject, less obvious peak values
	10:35:19	0.258, 2011	0.036		Non-maximum peaks of the alternative peak values
	10:37:15	0.567, 4994	0.028		Reject; less obvious peak values
	10:38:27	0.913, 4022	0.057 (√)		Maximum peak of the alternative peak value
	10:40:37	1.031, 2996	0.090		Reject; error in the measurement data
	10:41:13	1.321, 1700	0.224		

According to my understanding, the above comments are all about the 2.4 Calculation and 2.5 Uncertainties. I have rewritten these parts.

- Why not convert table 1 and table 2 into x-y graphs? Perhaps if combined?

Please note that the estimated (UAV) and true (sampled fuel) values of the FSC from 11 berthing ships are list in table 1. But there are only estimated (UAV) values of the FSC from 27 sailing ships are list in table 2. It seems that table 2 cannot convert into x-y graphs and the table 1 and table 2 cannot be combined.

- Our enforcements contacts tell us that ship owners will normally use fuel with a Sulphur content just below the limit. If I look at all the individual samples, I don't see that. Is there an uncertainty that is missed or are the vessels changing from one fuel to the other at the time of the sampling?

I think it's different in different areas, especially when the regulations are different.
In the DECA of China, the regulations are as follow:
Starting January 1, 2017, the FSC cannot exceed 0.5% (m/m) during berthing, excluding the first hour after arrival and the last hour before departure.
Starting January 1, 2018, the FSC cannot exceed 0.5% (m/m) during berthing.
Starting January 1, 2019, the FSC cannot exceed 0.5% (m/m) for both sailing and berthing ships.
Overall, our FSC monitoring results are shown in Figure 6. It shows that the FSCs of the sailing ships were considerably higher than those of the berthing ships. The FSCs of berthing ships are measured in the year of 2018 and 2019. The FSCs of sailing ships are measured in the year of 2019. On July 15, 2019, it was the first time that a sailing ship had been caught for having failed the FSC regulations in China.
I believe that with the further implementation of the policy, the ship owners will normally use fuel with a Sulphur content just below the limit in the DECA of China.

More specific comments: Abstract: Line 12 Emissions of CO2 and SO2 are not measured if I am correct. S content (S%) is measured.

Yes, this sentence "measure the sulfur dioxide and carbon dioxide emissions from sailing ships" is not appropriate. I guess the S content (S%) is calculated from the concentrations of SO_2 and CO_2. This sentence has been modified as follow.

The present study adopts a monitoring method involving an unmanned aerial vehicle (UAV) that takes off from a patrol boat to measure the concentrations of SO_2 and CO_2 within the plumes of sailing ships.

Line 13: I don't think the costs of this method are presented or discussed explicitly in this paper. The cost of a vessel capable of operating in open sea seems neglected. In our country that is not low cost. This is also rather costly.

Yes, this is rather costly. The patrol boats of the maritime department should cruise in the area regularly (once a week). To keep costs down, we measure the ships plume at the same time. Therefore, the cost is lower compared to aircraft and applicable to maritime department. I have added the discussion in the manuscript.

The method proposed in this study can be used to monitor ship emissions at a comparatively low cost to understand the FSCs of sailing ships in open waters. Although the cost of using patrol boats is not negligible, it is convenient and lower cost for maritime authorities compared with small aircraft.

Line 17 According to the monitoring results: I suggest changing to: Based upon the online monitoring results ....
OK, this sentence has been modified.

Line 69: low cost but doesn't include the cost of sailing

As mentioned above, I have added discussion in the manuscript.

Line 90 precision of 5 % at full range (is 10 ppm) or 0.5 ppm. Is that correct? Please mention. And how high is that compared to the observed values? What is the Sulphur content of the example presented in figure 5? Please add.

Yes, it's relative to the range (0.5 ppm is 5% of 10 ppm full range). According to the comments of Referee #2, I have supplemented the detailed parameter information of the UAS. In the original manuscript, the range of SO_2 sensor and accuracy of CO_2 sensor were wrong. I have modified and cross-checked to make sure the product information was right.

There are different range models of sensors, such as 5 ppm, 10 ppm, 100 ppm and so on. Precision generally depends on the size of the range. For example, precision of 0.25 ppm for 5 ppm, 0.5 ppm for 10 ppm, and 5 ppm for 100 ppm. It needs to make sure that observed values do not exceed the range in most cases, and the precision should not be too low. Therefore, observed values of SO_2 are generally in the range of 0-10 ppm. CO_2 are 400-5000 ppm range.
As mentioned above, I have added Table 2.

Line 90 etc. Could some details or results of the calibration procedure be presented as well.

I have added the details in the manuscript.

These sensor characteristics were provided by the instrument manufacturer and were ensured to be within the tolerances by calibration. The zero and full scales are usually calibrated by a standard mixed gas when the equipment is used on a daily basis. The major parameters of the UAS are listed in Table 1.

Line 107 "measure the concentration of SO2 and CO2". Change to: Measure the concentration of SO2 and CO2 in the plume

OK, this sentence has been modified.

Line 124 EF has no unit or? I wonder why equation 1 is mentioned. I think it is a bit confusing.

EF has unit, g_{SO2}/kg_{fuel}, in g emitted per kg fuel. I have added it in manuscript.
Equation 1 is a description of the measurement principle. We use the parameter of 10s to derive equation 2. If don't discuss equation 1, just list equation 2. It seems too sudden to the reader.

Line 132: What is sampling rate? Electronically? And why are the SO2 and CO2 sensors not synchronized? In the graphs it looks like a delay. And you could just shift them a little. Why is that not done? And why is the 10 s averaged data better. I don't see that. Figure 5 Please use equal y-scales in the right and left panel. This is confusing.

The sampling rate is 1s.
This is due to inconsistent response times of different sensors. In the vast majority of cases, the response time of SO_2 are faster, and the CO_2 looks a little bit behind SO_2.
Of course, we can adjust the timing artificially. For example, adjust the peak time of CO_2 to the peak time of SO_2.
I've done a comparative calculation. The data set is of the estimated (UAV) and true (sampled fuel) values of the FSC

from 11 berthing ships (using second-generation pod). On the whole, the difference is not great and the accuracy was slightly reduced. Therefore, we chose the method described in the manuscript to do the calculation.
The average gas concentration over 10 s was chosen for the FSC calculations; however, this does not mean that 9 s or 11 s could not have been selected. To demonstrate this, a comparison calculation was carried out using both 9 s and 11 s, which showed that these led to very little differences in the results. However, it is necessary to ensure that the gradient of the gas measurements is stable over the sampling time (the interval length of the integral). Moreover, the interval length cannot be too short (e.g., 2 s) or too long (e.g., 20 s). If the time is too short, it is difficult to determine whether the measurements are stable and undisturbed over time. Similarly, if the time is too long, it is also difficult to ensure that all of the measurements in the integral interval are stable and undisturbed. In addition, during the flight of the UAV in this study, the time available for measuring the plume was ~5 minutes. As both the ship and the UAV were moving at this time, it was virtually impossible to ensure that the UAV was flying consistently within the plume and obtaining stable measurements. Accordingly, 10 s is also a relatively appropriate value for the measurement process.
Therefore, 10 s is an appropriate parameter.
I have added the details in the manuscript as mentioned above.

OK, I have added equal y-scales in the right and left panel.

Line 137: The 10 sec averages hardly differ from 1 sec data. How could that happen?

Does this refer to the data in Figure 5?
The data in Figure 5 (b) is relatively smooth, making it easy to select peak values.
The data in Figure 5 is of good-quality and the difference is not significant. This is even more pronounced if the data is of poor-quality. I've added good- and poor- typical figures as mentioned above.

Line 140 What is meant by the calculated function? Not just the average measurement value?

It is “calculate the average of the data within 10 s”. This sentence has been modified.

Line 160: What would be objective criteria to tell whether it’s a poor-quality plume?

In the previous study, we developed a selection process. If most of the data in the measurement dataset is ruled out, then it is a poor-quality plume. If multiple peaks are preserved and their FSC result are similar, then it is a good-quality plume. But, strictly speaking, we do not have a completely objective evaluation criterion. This requires the computational model I mentioned above. I have added tow typical example in the Figure 5 and Table 2.

Line 167 Is this correct 20% (m/m)? I would expect 20% uncertainty with no units. So I suggest to leave (m/m) out.

Yes, this sentence has been modified.

Line 168 and 169. Unclear what is meant here (after therefore)? It could be interesting for the reader interested in enforcement to mention (even in the abstract) how many of the Non Compliants were detected and how many were missed (i.e. ships that were not identified as Non Compliants but had an FSC above the limit.

Please note that boarding inspection is a complicated process. The target ship cannot stop immediately in the channel for inspection and have to sail to the anchorage. Patrol boats need to follow to the anchorage for boarding inspection. In the process, to avoid punishment, crew will take various measures to drain the high-sulfur fuel in the main engine fuel oil pipeline.

The whole process of boarding inspection requires more than half a day and the work of more than a dozen law enforcement officers (such as sailors, driver, inspectors, VTS watchmen).

Therefore, in order to ensure that can accurately tracked down the offending ship. Law enforcement officers of the Pudong maritime safety administration only intercepted four sailing ships for which the UAV FSC results were of a good-quality and all exceeded 2% (m/m). The chemical FSC results of the four ships were: 0.534% (m/m), 0.744% (m/m), 0.813% (m/m), and 1.991% (m/m).

If the UAV FSC results is just over 0.5%, and then we board the ship for inspection. I believe the Non Compliants and Missed will happen. However, the factor of switching fuel cannot be ignored. Intercept and boarding a sailing ship are not just an experiment.

Line 189: perhaps the word optimistic is not the right word. Perhaps it should be: The uncertainty in the assessment is not small but the results so far, do not lead to optimism with respect to the FSC used by ships sailing in the area.

OK, this sentence has been modified.

Line 210: It Is only a small sample isn't it, but still convincing looking at figure 6. Please mention that.

OK, I have added the this in the manuscript.

Page 15: Number of digits in the given numbers are large (such as 40913 ppm, I suggest changing that to 4.1 %)

OK, these words have been modified.

Page 15 Why are the results of the sampling by the maritime authority not given in the table.

OK, I have added this.

English:

Line 15 ships → vessels

The word "ships" or "vessels" seems all appropriate. For consistency, "ships" is used all over the text.

Line 46 Several studies have suggested monitoring methods or similar. Otherwise I understood wrong but then the sentence is not very clear.

I have changed the sentence "Several studies have suggested monitoring ship emissions to estimate the FSC of the target ship" as "Several studies have suggested estimating FSC by measuring ship plumes".

Line 45 and 66 supervise or supervision→ enforce and enforcement; Line 54 and 65 navigation → navigating (?); Line 60: airplace→ aircraft; Line 65 inaccurate → non representative; Line 67 Suggestion: leave therefore out Line 85 extracts gas? → draws air; Line 108: approximately a few hundred meters. This is double: a few hundred meters is already an expression showing that it is an approximate value. Suggest leaving the word approximately out.; Page 13 legends to figure 4: the enlarged UAV is shown in the top left corner (and it is in the right corner: a detail)

The above English grammar problems have been revised. Thank you very much for your earnest help in pointing out the English problems.

NB: I have two versions of the text. Also, one with an appendix including two figures: Figure A2 The Chinese text could be difficult to read for non-Chinese readers. Perhaps add some explanation of the Chinese text.

As the suggestion of the Associate Editor, "remove the figures from the appendix as they do not contribute to the discussion of the measurement techniques”. I have removed the Figure A1 and A2.