Author responses to Anonymous Referee #2 are in bold below:

This paper deals with a commonly used ground-based method (OTM33A) for estimating emissions rates. Recent papers have highlighted the relevance of site-level (facility-wide) emission estimates. The authors perform tests to assess accuracy of this approach in the context of methane emissions from single sites as well as ensembles (i.e., characterize emissions distributions from a population of sites). The results are relevant due to the increasing use of the approach. I recommend publication after some minor edits/clarification. Two main points to be addressed/expanded by the authors: (1) Effect of multiple sources-distance selection (2) Determination of non-detects and potential effect of overestimation in determining fraction of sites that fall below detection limit. Additional comments:

We greatly appreciate the reviewer's careful consideration of the manuscript. We have addressed the noted issues as detailed below.

INTRODUCTION: Page 1, Line 23: "Site-level measurements are therefore necessary for improving emission estimates of the O&G production sector." This is true, might be also useful to mention importance of site-level measurements in conjunction with component-level measurements to understand source of emissions.

The text on line 23 has been changed to include component-level measurements and the study by Brandt et al., 2014, is included to help emphasize that point.

P1, L23-24 now read: "Site- and component-level measurements are therefore necessary for improving emission estimates of the O&G production sector (Brandt et al.,2014).

Page 2, Line 3: 'However, more permanent approaches are still under development and must be approved as equivalent monitoring technologies before they can replace existing EPA approved Leak Detection and Repair (LDAR) methods like optical gas imaging (OGI).' Suggest expanding discussion of difference between leak detection and leak (emissions) quantification, which is important in the context of LDAR and equivalency. One could argue that main goal of LDAR is not improving inventories, but repairing leaks. I think this idea needs to be further developed to link it to importance of site-level measurements.

Additional text has been added to page in an attempt to emphasize that LDAR does not typically generate data that can be used to improve emission inventories. While we agree that there is ample material to be discussed in terms of LDAR methods and equivalency, we believe this is beyond the scope of the current manuscript.

P2, L6-10 now read: "Annual or semi-annual LDAR programs already in place rarely quantify total emissions from a site, and the efficacy of these programs depends on many factors including employee experience, leak size, and meteorological variables like wind speed and temperature (Ravikumar et al., 2016, 2018). This makes LDAR programs an important tool for finding leaks and reducing emissions, but they often do not explicitly quantify or provide data of the actual emission rate from production sites, and this limits usefulness for improving emission inventories."

METHODS It might be useful to briefly discuss the detection limit of the method (threshold for considering a site as non-detect). This is discussed in previous papers, but might be useful to summarize here. Consequently, discuss the potential overestimation at lower emission rates with the threshold for non-detects. This is something that matters for the ensemble.

The method limit of detection has been added to the methods section.
P5, L30-31 now reads: "The estimated lower detection limit of the method is 0.01 g s^{-1} 0.036 kg h^{-1}) (Brantley et al., 2014)."

In the current study there were no "non-detects" meaning there is no bias in the Christman or METEC ensembles. In the field, careful consideration of non-detects is essential, but we feel this is best addressed in the papers covering those field deployments as methodology varies slightly from one study to another. Overall, the slight underestimation of total mass flux found in this study and the large underestimation reported by Bell et al., 2017 support OTM 33A being, if anything, slightly low for an ensemble of measurements. In general, it is not too critical what number is inserted for the low emission wells as the mean of the ensemble is dominated by higher emission sites and the uncertainty in the number of high emission sites.

Page 5, line 23: Might be good to mention that this could also affect flares (in addition to liquids unloadings).

P5, L27-29 now read: "OTM 33Astruggles to quantify plumes with a particularly high vertical velocity or buoyancy (such as manual unloadings, lit or unlit flares, or very hot emissions)."

Page 9, line 11-14. What happens with multiple sources on site? This paragraph hints at the importance of using OGI to locate source. Might be useful to expand on distance selection under various sources (i.e., based on highest emission point?)

The analysis presented here suggests that, at least for emission points that are within 6 m of each other, no selection of a specific source is necessary given that the observed error of ~10% is much smaller than other errors associated with the method. This section has been expanded to more clearly explain the relatively small impact of not knowing the exact source location on smaller sites (these are the sites typically measured via OTM 33A).

Page 10, line 9-11. 'These results also indicate OTM 33A does not drastically underestimate the total emissions for an ensemble or group of measurements, and that scaling up mean emissions measured with OTM 33A to an entire basin is a valid approach." This is an important conclusion from the paper since the ensemble is a common application of this method. Might be good idea to further highlight in the abstract.

We agree this is an important finding. We believe the statement in the abstract on Page 1 Lines 12-13 that, "an ensemble of OTM 33A measurements may have a small but

statistically insignificant low bias." makes this point without overstating what can be determined from the current study.

Figure1: It might be useful to expand caption to include label of release points (i.e., what is the source of emissions).
The caption of Figure 1 has been expanded to include descriptions of all of the pictured release points, as well as the total number of release points (11).

Figure 2: Significant figures for R parameter.
Significant figures for all R parameter values have been appropriately reduced.