Foulds et al. measured methane (CH$_4$) from aircraft and determined emissions from 21 oil and gas facilities in the North Sea oil region of Norway. They found emissions agreed with self-reported emissions from oil and gas companies, but were higher than a downscaled CH$_4$ global inventory from three years prior to the study.

Overall, I thought this was a good study worthy of publication in this journal. The study is scientifically significant, especially when considering the greenhouse gas footprint of fuel choices. The scientific methods are valid and have been used frequently in the past. I only have a few comments I’d like to see addressed before I think this paper is ready for publication.

Specific comments:

Introduction: by the end of the manuscript, the authors spent a lot of time referring to Tier 1 and Tier 3 inventories. I think this warrants a paragraph in the introduction defining and discussing the different tiers. How are they constructed? When are they used?

Section 4.1: Based on Figure B1a, it seems like the wind speed perpendicular to the flight track would be highly uncertain, since several transects were nearly parallel with the wind. Please explain how the uncertainties for C193 do not dominate the uncertainty calculation. Also, what is the estimated uncertainty for the wind direction measurement?

line 387: this seems like the lower plume extent uncertainty would be greater for the
Scientific Aviation flights than for the FAAM flights. If you already assume the plume is well mixed at the top of the boundary layer for the FAAM flights, it must also be well mixed below the lowest altitude of the aircraft.

line 443: I disagree with the authors conclusions here. Rather than claiming it is not of value to compare measurements to a spatially downscaled inventory, I think you can use measurements to instead question the value of the inventory itself. What the authors have found may in fact be evidence that a national inventory should not be spatially downscaled in this way, but that is different than saying a comparison to measurements have limited value.

Figure 4: since the authors are simply fitting y vs. x, there is no reason the data plotted on the y-axis need to be upscaled to t/yr. Why not just keep them as kg/s or whatever the units of the flux measurement are?

Figure A1: it seems the color legend is missing from this graph

Figure A2: Is the FLEXPART footprint taken from a spot on the flight track? Because it looks like it is offset further downwind from the flight track.

line 393: the phrase “clearly not a robust comparison” seems subjective unless the authors show day-to-day variability in emission, which at this point in the paper, they haven’t. I suggest “likely is not” or something like that.

line 647: I prefer the sentence in the abstract that mentions “all 21 facilities”. I would add that to the conclusions section as well.

line 652: is this across “fields” or across all facilities?

Technical comments:

line 69: add “the” before “UK”

line 71: remove comma after “et al.”
line 81: add period after “et al”

line 114: don’t hyphenate “in-flight”

line 267–268: annual is mentioned twice in this sentence

line 300 and elsewhere: suggest “data ... were”

line 386: suggest starting sentence with “However, there is also some additional uncertainty...”

line 660: is “compiled” a typo? It doesn’t make sense to me.