In response to Anonymous Referee #2 comments from December the 21st, 2020.

The manuscript describes a study to estimate methane emissions from the Upper Silesian coal mining area using an aircraft campaign. Two flights encircling the region have been conducted that measured methane concentrations up- and downwind in the plume. High resolution model simulations were used to relate the observations to possible emission locations, and in this way estimate the emissions for the chosen campaign day. Studies like this are highly relevant for climate research since the help to obtain insight in an important source of atmospheric methane, the second most important greenhouse gas after carbon dioxide. The subject is therefore well within the scope of ACP, and could be published after some minor changes.

The used method is rigorous and robust. The description of the aircraft flights is very informative, including the reasoning behind the flight plan. The simulations are based on combination of a trajectory model (FLEXPART) driven by a high-resolution meteorological model (WRF), where the later is also guided by wind lidar profiles taken during the same campaign. This gives trust that the simulations are as close to reality as can be expected from such simulations. Also the estimate of the emissions is done rigorous and seems hardly guided by a priori assumptions.

The results show that using such campaigns it is in principle possible to make a rather accurate estimate of actual emissions in a region with a large number of small sources that together aggregate into a major plume. Since the campaign is not more than a snap shot, it cannot be expected that these results can be extrapolated to for example a yearly total. However, could the authors give some reflection on how to use this kind method to do so? Could these campaigns be regularly repeated, or would for example regularly performed soundings (with FTIR or AirCore) be an alternative?

Dear Referee,
Thank you very much for your kind and helpful comments on the analysis presented in the manuscript. We absolutely agree, that extrapolation to yearly totals requires not only snapshot observations, but continuous measurements at various meteorological conditions. This manuscript is intended as a further step towards this goal. Achieving this goal will benefit substantially from concurrent FTIR soundings, aswell as active AirCores, as the method presented herein is not restricted to in situ data but can also be adapted to total column measurements. Total column measurements, especially mobile FTIRs, are in fact complementary to in situ observations and have already been used in explorative studies similar to the one presented here (Luther et al., 2019).

1. **Lines 80-86: Is there information available on how the EDGAR inventory estimated these CH4 emissions? In section 4.4 a comparison is made with E-PRTR. Since the later is more "specific", one could argue that it is more accurate. Aren't these numbers then not used in EDGAR?**
EDGAR v4.3.2 includes data from EPRTRv4.2 for the industrial sector, according to Janssens-Maenhout et al., 2017. However, the same reference also states: "A point-source database, such as the European Pollutant Release and Transfer Register (EPRTR18), allows a more homogeneous input for an inventory compiled under such a facility-based approach. The European study of Theloke et al. (2011), which aimed to complement EPRTR point sources with information on diffusive sources per country that together match national totals, revealed large inconsistencies, which prevented closing the two approaches in a satisfying way." It seems like discrepancies between the two inventories are still large.

2. **Figure 5: Temperature is 2-4 degrees biased ($\sim$ 1 %). Is that problematic for computing air densities etc? Think it should be related to the error estimates in section 4.3. Observed temperature is used in eq. 10 for the in-situ observations, should assimilated temperatures then be used in the simulations?**
The in-situ concentration measurements use the actually measured temperatures, so the bias does not affect the measurements themselves but only the comparison between model and simulation. It is de facto related to the uncertainty analysis. From our earlier studies we do know, that a systematic error in static air temperature manifests approximately an order of magnitude lower in emission estimates compared to errors in wind speed and direction. Nevertheless we included an error in sensed mole fractions of $\sigma_c = \pm 10 ppb$ in the uncertainty analysis in Sect. 4.6. In theory - yes - it would be good to use assimilated temperatures in the simulations. Using data from the aircraft alone for the assimilation, however, does not improve on this bias in the present case. We tried different setups but did not succeed in reducing this

bias to a satisfactory level. Unfortunately the Doppler lidars can only report wind speed, direction at specific altitudes and derived quantities. It would certainly be great if these devices would be able to remotely sound air temperature in the future, in order to have a solid basis for temperature assimilation.

3. **Line 190: Where is the source index i used in the emission rate φ_i ? I guess that both emitted mass and emission time are source dependend.**
   That is correct. The relevant sentence has been revised to: "[...] A scaling coefficient x_i is assigned to each of the n sources $\varphi_i = m_{e,i}\tau_{e,i}^{-1}$, with the total emission time $\tau_{e,i}$ in seconds and the total mass emitted $m_{e,i}$ in kg for each simulated source. [...]"

4. **Line 209: What do we learn from d_s? Much more observations than sources, so in theory overdetermined problem. If d_s equals number of sources, is the estimate than exact?**
   The problem is mathematically overdetermined, hence no exact solution exists. An approximate solution has to be found via e.g. a least-squares approach. Here we did not use normal least-squares but further included known information on the measurement uncertainties and a-priori knowledge on the sources. The degrees of freedom for signal d_s describe the reduction in the normalized error on $\mathbf{x}$ introduced by the available observations and hence provides a measure for the improvement in knowledge of $\mathbf{x}$, relative to the a-priori, due to the observations. If d_s equals the number of sources (unknowns in the system of equations), the a-priori has no influence on the retrieval. In that case all "information" originates from the measurements alone. However, it does not mean the estimate is exact, as the system of equations used may (and will) only partially reflect the "true" atmospheric "system of equations" (the truth) and since measurement errors will propagate heavily into the estimate.

5. **Lines 239-240: This is a correct description of the Jacobian, but has that been used in the error estimates described in this section?**
   The Jacobian is not relevant in this section of the manuscript, yet is used in Sect. 4.6. The intention was to introduce the quantities needed later to allow for easier reading. As it is not relevant in this section we removed the sentence in a revised version of the manuscript.

6. **Figure 7, right panel: what is the background color, emissions from EDGAR ?**
   The background colors from the right panels in Fig. 7 and Fig. 8 correspond to the topography in the USCB region. To the south the terrain becomes elevated as approaching the Tatra mountains. The CH_4 plumes are nicely advected into the Moravian gate. We revised both figure captions to include this info: "[...] Right: Top-down view on the model output and the downwind wall observations ρ at $750\,\mathrm{m\,a.g.l}$ with underlaid topography. [...]"

7. **Line 280: Why not Fig 6 and Fig 9 as two panels next to each other?**
   Our intention was to treat both flights separately, before discussing the combined flights in Sect. 4.5. We do however agree, that the figures are similar and could be merged into a 2-panel figure for a better flow. We merged Fig. 6 and Fig. 9 in a revised version of this manuscript.

8. **Line 297. A scatter plot with x for afternoon vs morning would be useful to see if the algorithm estimates emissions to be present from the same locations, and how different these are.**
   Subregional emission estimates require at least two flights associated with different wind conditions. During single flights several emission sources are masked by sources closer to the point of measurement. In the present scenario, the large number of sources enables regional estimates for single flights, albeit subregional estimates will vary significantly and are not reliable for the reason mentioned above.

9. **Line 298. What is meant with "... neither flight can be used on its own ..." Here a remark could be placed that in the next section the morning and afternoon flights are analyzed together.**
   The degrees of freedom for signal of both flights are very similar (43 for the morning flight and 44 for the afternoon flight). A value of $d_s \sim 40$ indicates, that not all "information" stems from observations, or synonymously - the emission estimate is to some extent influenced by the a-priori. This is due to some sources being "obscured" by sources closer to

the point of measurement or sparse data availability. Enhancing the degrees of freedom for signal requires measuring downwind at different wind directions. Of course it would be convenient to sample at very different advection scenarios. This is however hardly possible, as one of the principal assumptions - constant emission rate over the timespan of the flight(s) - would be violated due to daily changes in ventilation volume of the mines. As a compromise, two flights on the same day with only slightly different wind conditions have been used in an effort to maximize likelihood of a constant emission rate. We rephrased and added the following clarifying sentence to the revised manuscript: "[...] Both flights yield similar d_s values, indicating that not all information stems from observations alone. Hence, neither flight can be used alone to retrieve all modeled sources. In an effort to minimize the dependency on the a-priori, both flights will be analyzed together in the next section. [...]"

10. **Section 4.6 The estimated emission totals have an uncertainty of 16 % (std.dev.) Could this value be related in someway to a yearly total? For example, how many campaigns would be needed to come within an accuracy of say 2 % over the year?**
As you mentioned in the general comments above it would be possible to repeat these measurements under different meteorological, most importantly wind conditions, in order to minimize statistical uncertainty. This would however not improve on the systematic uncertainty. Consequently it is hard to come up with an answer on how many campaigns would be needed to reach an uncertainty of 2%.

11. **Spell and grammar**
Spell and grammar has been corrected.