“Evaluation of simulated CO\textsubscript{2} power plant plumes from six high resolution atmospheric transport models” by Brunner et al. assesses model simulation capabilities to understand their applicability for use in power plant emission estimation from aircraft and satellite observations. The study uses real aircraft in situ and remote sensing observations from the CoMet aircraft campaign of 2018 over the Bełchatów and Jänschwalde power plants and simulated CO\textsubscript{2}M observations (with 2 noise levels) with six models including Eulerian, Lagrangian and Large Eddy Simulation (LES) models. This is a useful and informative study that enhances our understanding of model capabilities and limitations, but the authors correctly warn that a complete comparative evaluation of the models cannot be carried out based on just a few overflights. The quality of the study is high and the interpretation of the results is generally sound, as the authors are careful not to over-interpret the relatively small number of examples used in the study.

I would have liked the study to go the additional step of actually reporting the estimated emissions of the two power plants based on the real aircraft observations and not only the simulated CO\textsubscript{2}M data (Table 5). The spread in derived emission estimates that would be obtained with the different models would be informative, although not essential for this study.

With or without the above suggestion, I only recommend some minor specific changes before I would deem the study acceptable for publication in Atmospheric Chemistry and Physics.

**Specific Points**

Line 1: “dominated” is too strong of a term, since there are other major sources like urban CO\textsubscript{2} emissions from which a large fraction is from transportation or residential heating, rather than facilities. Based on the introduction, the contribution from facilities is about 58%.
Line 10: The description should clarify “NWP models extended for atmospheric tracer transport” or something like this rather than just calling them NWP models.

Line 36: The actual prevalence of stack monitors is somewhat uncertain. Recommend changing “often measured” to “sometimes measured”.

Line 46: Recommend including OCO-3 (Nassar et al., 2022 https://www.frontiersin.org/articles/10.3389/frsen.2022.1028240/full), which has enhanced capabilities relative to OCO-2 for locations of interest within the latitude range covered (up to ~52°N), and furthermore is highly relevant to the Bełchatów examples.


Line 115: The authors made an appropriate decision to optimize the location used in approximating the power plant as a point source based on actual stack locations.

Line 277: The authors need to double-check the stated sunrise time of 3:33 CET on June 6, which seems too early, as they are likely reporting the onset of twilight rather than the actual sunrise. I am unsure however which one (onset of twilight or actual sunrise) is more relevant for the ABL height. At minimum, they need to be more careful with wording.

Figures 2-4: The comparisons are very interesting and informative regarding the model spread.

Figure 4: Variation in vertical dimension between the models strongly suggests that this will be less of an issue with satellite column data and this is confirmed by comparisons is 3.3 (Fig 9 – 10).

Figures 2-11: If the authors can make the model name label on the Figures more prominent in comparison to other text, it would significantly improve clarity for the reader.

Figure 14: Units should be specified for colour scale of the figure.
Finally, the conclusion ends somewhat abruptly. I think the manuscript would benefit with an additional paragraph dealing with the bigger picture, where the authors put this study in the context of the expected capabilities and limitations of power plant CO$_2$ emissions monitoring, verification and support (MVS) with CO2M.