General comments

I think the manuscript is overall well structured and describes a well conducted and thorough study. I believe that the article examines an important subject and nicely fills a gap in existing literature by systematically examining the use deposition measurements for source term estimation. Further, I think the approach of applying the method to both synthetic data and a real world case is a good way of demonstrating the capabilities and short-comings of the method.

Personally, I would have liked to see an experiment, where all the existing air concentration measurements were used, both with and without the deposition data (both for the synthetic and real data). I think it would be interesting to see if there is added value of combining the two datasets, or if deposition data is less important, when you have +1000 air concentration measurements already. I understand that adding this experiment might be outside the scope if this study, but perhaps a small discussion could be included, where the authors speculate a bit about this.

In addition, I believe that some points are a bit unclear and need further clarification. I have listed some specific comments below, which I think needs to be addressed before the manuscript is ready for publication. That said, I strongly support the publication of this manuscript, if these points are clarified.

Specific comments

1. **Line 95:** You write that only 30 are left for your analysis. I understand that 35 were discarded, because they were too close to the source, but earlier (line 95), you mentioned that there are 135 measurements in total. What happened to the remaining 70? You write that "*… only detections of activity per surface area (i.e. Bq m -2) were selected.*" What were the remaining, and why are they not useful?

2. **Line 95:** "*We follow the distinction made by Masson et al. (2019) to label 18 of these "activity concentration in rain water" and 12 "dry + wet fallout" … Though the description "rain water" may seem to imply only the collection of wet deposition, in general monitoring networks do not discriminate between dry and wet deposition. It is therefore assumed that both the rain water and fallout measurements contain dry and wet deposition …*"
   I am not sure I understand your approach here. You argue that you cannot justify making a distinction between the two datasets, and you therefore assume that both consist of dry+wet deposition. If this is your assumption, then why do you use the labeling? I find this is a bit confusing.
   I would probably have preferred one of two options **(a) make a distinction between the datasets, and then assume that "rain water" measurements consist of only wet deposition (then use that knowledge for the inversion)**, or **(b) don't make a distinction between the datasets and treat them equally (as you do), but then don't use the labeling and use only the combined dataset and get more robust statistics.**
   If you choose to keep your approach as it is, I think you should further justify why it makes sense to study the two datasets individually. Also, you should make it more clear that "rain water" is only a labeling and not related to the assumed deposition process. In the rest of the article, it is a bit unclear to me how you interpret this label yourself.

3. **Line 100:** "*The deposition data in the supplementary material of Masson et al. (2019) only provide the start and end dates of the measurements.*" I miss some information about typical durations of the measurement windows. I think this is a crucial information both for understanding the usefulness of the measurements in the first place, but also for understanding how large the potential error is on the assumed start and end times of the measurements.

4. **Line 105:** "*… the choice to extend the measurement interval by 8 h was made to increase the likelihood of capturing the relevant precipitation event that contributed to any wet deposition*." But that goes both ways. You may end up including a precipitation event that was in fact outside of the measurement window.

5. **Line 150:** Can you perhaps elaborate a bit on the Bayesian approach. When you write that you use "*… inverse gamma distribution for the combined model and observation uncertainties*", do you then mean as prior distribution for the uncertainties, which are assumed unknown? Please, specify this. Further, it would be nice to get a few details about the MCMC sampling method used.

6. **Line 220:** "*A relative measurement error of 50% was chosen.*" What do you use this for? Is it input to the inversion algorithms? If this is the case, do you assume that this represent only measurement error or combined measurement+modelling error? Finally, if the latter is the case, should you not use a different number for the synthetic vs. real case?

7. **Figure 5:** Here you show how the modelled dry+wet deposition correspond to the observations of two datasets "rain water" and "fallout". Again, I am not sure I understand the distinction, since you treat the measurements identically. You could increase the amount of data by combining the datasets, and obtain more robust estimates of statistical parameters such as correlation coefficient and fractional bias.
   One thing that could be interesting is to see how the modelled dry and wet deposition alone correspond to the observations. Can you for example see that the dataset labeled "rain water" seem to be dominated by the modelled wet deposition.
   If you do not include figures showing this, I would at least appreciate a section somewhere in the text, where you comment on the magnitude of the modelled wet vs. dry deposition values. If you can see that the wet deposition is dominating for the "rain water" dataset, I think this adds to the justification of studying the two datasets individually (cf. my comment nr. 2).

8. **Line 252:** "*The higher resolution meteorological data (0.1 ◦ , 1h) provides an increase in deposition by one order of magnitude, which is still an underestimation but an improvement over the lower resolution result.*"
   I think this is an interesting result, and I especially appreciate the following analysis, where you compare the accumulated column densities for the two different resolutions (Figure 6). However, I notice that both of these measurements are in the dataset labeled "rain water". For gaining further understanding, I think it would be relevant to know if the modelled deposition for these locations is mainly wet, as the labeling suggests (cf. my comment nr 7).
   Further, did you look at differences in plume structure? Maybe there are some significant differences in the resolved flow? Or perhaps the measurements are taken in an area with large concentration gradients; then even small differences in the in the plume position could explain large discrepancies.

9. **Figure 7:** You have not really introduced the term "residual cost" before this figure. Can you perhaps elaborate a bit on the interpretation of this. Do you interpret it as being proportional to a probability density?

10. **Figure 9 and 10:** It is not until I see these figures that I can guess what type of assumptions you have made about the source term in the two inversion methods. For the cost function based method, you have a release profile with different release rates for each day, while for the Bayesian inversion, it seems that you assume a constant release over a period (described by start time, end time and release rate)? These assumptions should be stated clearly somewhere

in the text. Since you have decided to use the “twin experiment” with two separate releases as the “truth”, it is especially relevant to mention that it is not possible to describe this with the source term discretization you chose for the Bayesian inversion.
Further, in Figure 10, I can see the prior distributions you have used for start time, end time and release rate, but I would like to read an explanation somewhere.

11. **Line 328:** “*Since both dry deposition and air concentration SRS fields are very similar (see Fig. 2), the inverse modelling results are expected to be similar. This is verified with the results as shown.*” I think this conclusion is very interesting, especially combined with the re-run as described in line 335.

12. **Line 339 and 400:** You conclude that lowering the detection limits of deposition measurement could aid source localization with these measurements. First of all, this statement is probably fair because lowering the detection limit can only improve the results.
However, I think you should be careful with concluding too much based on the idealized case with negligible model uncertainties. In “real world” applications, the model uncertainties are expected to be much larger than the measurement uncertainties, so I would not expect to see as big an impact.
It would be interesting to see it demonstrated on real data. One option could be to conduct the inversion with all the existing air measurement data (using the real data) and then artificially raise the detection limit of those to see what the impact would be. This is probably a task for a different study, but I think you should at least discuss what impact you would expect when using real measurement data.

13. **Line 358-360:** You describe this problem of the method being over-confident. And while I agree that you would easily be able to point out Mayak in the specific case, it is of course somewhat of a problem. Before I would consider using this method I would appreciate a discussion regarding the cause of the issue as well as possible solutions.
Could the problem be that the uncertainties are assumed too small? I am still not 100% sure how the uncertainties are treated in this Bayesian method, because you first mentioned the inverse gamma distribution, and then later wrote that you assume a 50% relative error on the measurements. (cf. my comments 5 and 6). So I look forward to elaborations on this.

14. **Figure 15:** The Bayesian method gives a very impressive result.

15. **Line 405:** “*The fallout measurements, however, provide a somewhat worse results for reasons that are unclear.*” To really discuss this, I still need clarification about why the dataset is split into these two subsets in the first place. If we knew that one dataset is dominated by wet deposition and the other by dry deposition, then I guess that would be the interesting part to discuss. However, if we assume that the two datasets are comparable, then one explanation could be that you have too little data. After all, the “fallout” dataset only consists of only 12 measurements. I hope that your answers to some of my previous questions can also help a bit with the understanding here.