Reply to Anonymous Referee #1

We thank anonymous referee #1 for the very positive overall evaluation of our manuscript and the very constructive comments.

In the following we repeat the specific comments of the referee in italics and add our replies in regular fonts.

How do the authors plan to use this system in the future? Can the system easily be used to estimate methane emissions over a longer timescale, instead of just 2018? If so, how far back can you go (e.g., availability of observations)? How quickly can it be applied to recent years?

It is planned to use FLEXVAR for EMPA's quasi-operational system to estimate Switzerland's CH_4 emissions annually as contribution to the Swiss National Inventory Reporting. Meteorological fields from the COSMO model at horizontal resolution of 7 km × 7 km are available for the years 2002 to 2021. Therefore, the FLEXPART-COSMO back trajectories could be generated for that period with consistent meteorological fields. For analysis periods after 2021, the use of different meteorological input fields could be considered, such as e.g., the COSMO meteo data at horizontal resolution of 1 km × 1 km or analysis (or reanalysis data) from ECMWF IFS. We will add the information about the availability of the 7 km × 7 km COSMO meteo data in the revised manuscript.

Regarding the availability of observations: Various atmospheric data sets are available also for the period before 2018 (our analysis year), however the number of European stations is smaller in previous years. Nevertheless, various analyses have been performed for previous years, e.g., within the VERIFY project for the period 2005-2017.

Will this system be incorporated in existing "emission verification" efforts?

See our reply to previous point

The authors have focused their uncertainty analysis on one year. If the framework would be extended to multiple years, it is likely that part of the differences between the inverse set-ups will be constant between years (i.e., systematic), and will therefore not impact a trend analysis. If the authors agree, it would be good to add some discussion on this for context.

We have focussed our analysis on a single year (2018) since the main objective was to demonstrate the use of the new inverse modelling system and to characterize this system in some detail. While the application of FLEXVAR to other years is in principle straightforward (see our reply to first point), the analysis of the uncertainties in derived trends is rather difficult. Indeed, often the argument is made that potential systematic errors of the inverse modelling system (especially of the transport modelling) should be constant and should therefore cancel out when looking on trends. However, meteorological analyses could have e.g., some time-dependent systematic errors (biases) which might be difficult to diagnose. The analysis of trends becomes even more difficult, if the observational data coverage is changing over time.

We agree that the analysis of trends and their uncertainties is an important research question. However, such an analysis is outside the scope of the present paper.

In general, the inverse emission estimates suggest some differences between top-down and bottom-up, but uncertainties overlap. Given the wide range of sensitivity tests, can the authors provide more insights into the way forward towards reducing the top-down uncertainties such that these differences can be better understood? To what degree can more modeling efforts help, and to what degree do we need a denser observational network? Related to the previous point, will a trend be better constrained than an absolute emission estimate?

In order to better understand the differences between top-down and bottom-up estimates it would be very useful to get independent estimates on smaller regional scales, e.g., using aircraft measurements, and measurement campaigns closer to larger sources, which should help to bridge the gap between top-down and bottom-up estimates.

Of course, further increasing the observational network will improve the top-down estimates. At the same time, however, further improvement of the atmospheric transport models (especially regarding boundary layer height dynamics and vertical transport) as well as improved independent validation (using e.g., ^{222}Rn and boundary layer height measurements) will be essential to improve the inverse modelling and to better characterize their uncertainties.

We will add a short discussion of this in the revised manuscript.

Even if total emissions were much more strongly constrained, then still the significant uncertainty in natural emissions will sustain a large posterior uncertainty. Do the authors see any way to address this challenge from a top-down perspective?

With the current European observational network, it remains very difficult to disentangle natural and anthropogenic sources. In order to better quantify the natural emissions, dedicated measurements closer to natural sources should be performed.

Other comments:

I suggest splitting up the lengthy L60-L108 paragraph.

We will split the L60-L108 paragraph.

From Table 3 I understand that the standard spatial correlation is 100km in the FLEXVAR inversions. However, in the FLEXKf and TM5-4DVAR inversions 200km is used (L300 & L319). Why this difference? Could this partly explain why the FLEXVAR inversions reproduce the observations best (i.e., a less stiff state)?

For the FLExKF and TM5-4DVAR inversions we had chosen the default spatial correlation used in these systems. For FLEXVAR, we will investigate the impact of the spatial correlation on the achieved correlation.

L368-374: If I understand correctly, only in-situ observations in the optimal threehourly window are selected, then these are averaged to one daily value per site. I understand this choice, but do the authors consider that any valuable information is lost in averaging out the high-frequency signal, or in the data that are not in this threehour window? Additionally, how is the data from discrete air sampling treated, since there is not the same choice for selecting a time window?

The high-frequency variations of CH_4 mole fractions (e.g., on time scale of minutes) remains very difficult (if not impossible) to simulate with the current 7 km × 7 km COSMO meteo data. Simulation of the high-frequency variations would require much higher spatial and temporal resolution of the meteo data as well as higher resolution of the applied emission inventories.

The given 3-hour time windows were chosen since measurements and model simulations are considered most representative during these 3-hour time windows. Nevertheless, it could be certainly interesting to explore the use of the whole diurnal cycle of the measurements. In this context it is quite encouraging that FLEXVAR simulates in general the diurnal cycles relatively well at most sites (see Figure S7). Nevertheless, the use of the whole diurnal cycle in the inversion would require some more detailed investigations (including the application of appropriate temporal correlations for the observational data).

Discrete air samples were taken depending on their availability (i.e., without application of an additional time window). We will add this information in the revised manuscript.

The authors point out that an advantage of the 4DVAR approach is that, for optimization, the emission grid does not need to be aggregated. However, there is still limited information in the CH4 observations, so that correlation lengths in space and time need to be applied in the optimization. It seems useful to add the effective degrees of freedom that these correlation lengths result in, in addition to the total number of state elements, to compare to the other inverse systems (e.g., near L175).

Clearly the application of spatial and temporal correlation lengths leads to a reduction of the "effective degrees of freedom". However, it is difficult to quantify this. An option could be an analysis of the spectrum of the Eigenvalues (e.g., quantification of number of leading Eigenvectors). However, such an analysis would require substantial additional work and it is not straightforward to apply this to the different inversion techniques (4DVAR vs. Extended Kalman Filter inverse modelling system).

The authors calculate posterior uncertainties in inversions different from the reference inversions. Most importantly, the alternative inversions allow for negative emissions. For this to be a valid strategy, the alternative inversions should converge to similar emissions and reproduce observations similarly as the standard inversions. I could not find any confirmation that this is the case. I understand that qualitatively this approach makes sense and the results seem plausible, but I would like to see some evaluation of this aspect in the manuscript (or supplements).

The additional inversions used to calculate posterior uncertainties (based on the conjugate gradient algorithm) show in general very similar spatial patterns of the inversion increments as the regular inversions (using the semi-lognormal probability density function and the m1qn3 algorithm). Aggregated annual total CH_4 emissions agree on average (over all sensitivity inversions) within -1.0% to 2.1% for the 4 country regions discussed in the paper (Germany, France, BENELUX, UK+Ireland) and for individual inversions within -6.4% and 6.5%. We will add this information in the updated version of the manuscript.

Fig. 5 contains a lot of information, and takes a while to fully take in. Having everything differentiated with only color does not help this process. Perhaps it could help to use different markers/linestyles for inversions that start from different priors (e.g., circle for E1, square for E2, triangle for E3)? Then,

statements as in L562-L564 can be more easily seen. Other ways to adjust the colors, or reduce the number of colors, would be desirable: the light-yellow (INV-E1-O1-S3.1) is hardly visible and some of the greens are indistinguishable.

We will update the figure taking into account the suggestions of the reviewer.

I would be interested to see a rough comparison of the computational cost of the different inverse systems. In principle a global TM5-4DVAR inversion is needed for the baseline of the FLEXPART inversions, but this global inversion (as I understand it) only needs to be done once. Once the baseline is determined, are the FLEXPART inversions much faster than the TM5-4DVAR inversions? I expect this to be an important additional advantage of the Lagrangian approach.

Indeed, the FLEXVAR inversions are much faster than a full TM5-4DVAR inversion (with European 1^{o} x 1^{o} zoom), roughly about a factor of 7-10 for the current settings (e.g., including 2 outer loops of the FLEXVAR inversion). However, the computation of the FLEXPART-COSMO back trajectories also requires significant computational resources. Since the FLEXPART-COSMO back trajectories were computed on a different computing platform than the FLEXVAR and TM5-4DVAR inversions, however, a quantitative comparison is difficult. Furthermore, the comparison of the total computational costs will depend on the number of FLEXVAR inversions to be performed (since baselines and FLEXPART-COSMO back trajectories need to be computed only once).