author comments on the manuscript "Monitoring CO emissions of the metropolis Mexico City using TROPOMI CO observations", reviewer 2

We would like to thank the reviewer for the constructive comments that aided us to improve our manuscript. In this document we provide our replies to the reviewer's comments. The original comments made by the reviewer are numbered and typeset in italic and bold face font. Following every comment we give our reply. Here line numbers, page numbers and figure numbers refer to the original version of the manuscript, if not stated differently. Additionally, the revised version of the manuscript is added.

1 General Comments

1. ***two-step minimization: The optimization of the emissions requires two steps. The first step (pre-fit) filters out a significant fraction of the TROPOMI data based on several criteria, while the second step (final fit) includes the entire data set. Some of these criteria require more justification, especially when threshold values are bieng used without any justification.***

   adjusted

   We relaxed the filter criteria for the Pre-fit retrieval and by that were removing the criteria criticized by the referee. The changes in our results are insignificant (please see the attached Fig.1 for a comparison). Accordingly, we recalculated everything and changed all the resulting numbers in the manuscript. The description of the filter criteria (p6,l23- p7l5) is changed from:
   " (3) The quality of the forward model depends on the meteorological situation, where we consider model simulations for low wind speeds more reliable. This considerations led to the criteria of the data filtering for the pre-fit. Thus, we only select overpasses which meet all of following filter criteria:

   - 70 % of the data domain is covered by TROPOMI observations
   - for all observations the across track pixel size is < 15 km.
   - the average wind speed of the scene is < 4 m/s.
   - The fit residuum $\langle\delta\rangle$ < 8ppb, and the standard deviation $\sigma(\delta)$ < 8ppb to limit the effect of too large forward model errors.
   - $\sigma(\delta)/\delta(\mathbf{y}_{\mathrm{meas}})$ < 0.65 to ensure that the forward model can explain the variability of the measured CO field.
   - $\sigma(\mathbf{F}(\mathbf{x}))$ > 4 ppb to ensure that the model data contain a clear pollution signatures.
   - the Pearson correlation coefficient r > 0.3 between $CO_{TROPOMI}$ and CO_{WRF}.

   The filter criteria reduce the original set of 551 overpasses to 148, which we consider to be sufficient to estimate the overall average emission rate per district, yielding the prior state vector $\mathbf{x}_a$. "
   to
   " These considerations led to the criteria of the data filtering to determine the mean emission for each district. We only select overpasses which meet both filter criteria:

   - 70 % of the data domain is covered by TROPOMI observations
   - for all observations the across track pixel size is < 15 km.

   The filter criteria reduce the original set of 551 overpasses to 199, which we consider to be sufficient to estimate the overall average emission rate per district, yielding $\bar{x}$. "

   The statement of the referee about the optimization is addressed in the following comment.

2. ***But more importantly, this two-step process implies that the same measurements have been used twice. This approach seems to be a solution to the noise affecting the model-data residuals, hence limiting the convergence of the optimization system. In figure 5, there is almost no difference between the pre-fit and the final fit. The results are already constrained after the pre-fit step, except that the error bars further decrease, which seems artificial without assimilating additional measurements. To be clear, data should never be used twice in the optimization as it artificially in- creases the information content from the data set. Instead, the emissions should be produced by a single optimization procedure, iterative or at once, but extracting infor- mation only one time. If the noise is the inherent problem here, it***

should be treated by filtering out the noise in the data or in the model. Smoothing data signals, decom- posing the signals into frequencies, or averaging over time (both model and data) will help extract the information from noisy model-data residuals. Other approaches like Wasserstein distance or other machine learning techniques will help remove the noise. The optimization procedure needs to be revised to produce robust emissions and un- certainties. As it stands, the selection of data is arbitrary and the optimization uses the sae data multiple times.

adjusted

We completely rewrote section 3.2 about the inversion methodology (please see the new version of the manuscript). Our aim is to show that our two-step inversion scheme is extracting orthogonal information from the measurements and by that not the same information is extracted twice. "

3. *Background determination: The determination of the background CO values is never ex- plained in details. The first paragraph of Section 4 describes very briefly that CO background has been fitted. Which domain has been used? The entire State of Mex- ico? Considering the topography, the gradients over the domain, and the potential contamination by other sources (such as fires, but also cities, industries, shipping, . . .), the uncertainty associated with the background determination needs to be quanti- fied and included in the optimization. The uncertainty associated with the background seems to be equal to zero in the final fit (sec- tion 3.4). How is the uncertainty defined in the optimization? Zero values appear in the prior error covariance matrix (P7-L14). It suggests that the background is pre-defined (be- fore the final optimization). A section should describe precisely how the background values are determined in (or before) the optimization. Assuming you provide a coherent one-step minimization procedure, and a robust de- termination of th background values for each day, I recommend that you include some pseudo-data experiments to evaluate the potential of your optimization to constrain the city emissions. Simple perturbations should be added to urban districts to determine the actual constraint from TROPOMI data.*

adjusted

In our approach we chose to infer information about the background CO directly from the TROPOMI CO measurements by simultaneously fitting the parameters $(\alpha_{\mathrm{bg}}, \alpha_{\mathrm{bg}})$ of Eq. 5) together with the emission estimates of the different city districts in the same domain. This is advantageous because the inversion can decide itself which part of the measured signal is classified as background and which as pollution from the city districts which minimized the change to introduces biases in the emission estimates. For the same reason we are fitting the background parameters always without imposing a regularization for each TROPOMI orbit. Here, the parameter α_{bg} models a background CO concentration and α_{bg} a gradient of the vertical CO concentration dependent on the orography of the region.
Please see our answer to the third major comment of referee 1 and the corresponding changes to the manuscript.
The zeros in the regularization matrix ensure that the background parameters are not regularized. To make this more clear We change the sentence p7,l14:
from
" such that the elements of the state vector $\alpha_{\mathrm{b}g}$ and α_{elv} are not regularized. "
to
" . . . where the zeros ensure that the elements of the state vector α_{bg} and α_{elv} are not regularized. Obviously, the . . . "
Furthermore, the errors of the background are shown in Fig.2 for each point. They are small because we have a strong signal about the background in the TROPOMI data.

4. *As a last comment, the selection of data with low wind speed conditions will increase the model errors. If the aboslute wind speed is 2 m/s, an error of 2 m/s in wind speed corre- sponds to an error of 100% on the emissions. It maximizes the local enhance- ments which helps with the large noise, but this filter is typically the opposite in most studies using satel- lite data (minimum wind speed). Removing noise will help removing that threshold which seems to be a simple but risky solution to reducing the noise in model-data residuals.*

adjusted

Please see our answer to the first mayor comment of the referee. In the new version of the manuscript we will not filter on the wind speed anymore. We found removing these filter criteria have no significant

effect on our results (see the attached Fig. 1).

2 Technical Comments

1. **P2 L1: and its transport in the atmosphere : Unclear.**

   adjusted

   We changed the sentence p2,l1 from:
   " These characteristics established CO as a tracer for air pollution and its transport in the atmosphere . . . "
   to
   " These characteristics established CO as a tracer for air pollution and transport processes in the atmosphere "

2. **P2 L20: Add references for the error sources**

   adjusted We added a reference to (Borsdorff et al., 2019) where we discussed the errors on the example of pollution transport from Tehran.

3. **P2 L30: Here the emission estimation changed by 0.42 Tg/yr 30 in only 2 years from 2014 to 2016, due to a change in the mobile emission model from mobile to moves. This means that the emission model changed and not the emissions. Has this change been confirmed or validated by other data?**

   adjusted

   We haven't found a validation or confirmation of this. However, it shows that the error bars on the emissions of the INEM inventory are big. To prevent confusions, we removed the sentence.

4. **P4 L5: The inventory is time dependent and accounts for the diurnal, week-to-week and monthly variations of the emissions How accurate are these cycles? Considering the overpass time is fixed, the mismatch can be explained by a difference in diurnal cycles in and out of the city. How was the inventory constructed? Does it include traffic counts? Are the other sectors using temperature-dependent relationship?**

   adjusted

   There are no sectors using temperature-dependent relationships in the inventory. We are adding the following sentence at p4, l5:
   " The weekly and daily mobile temporal profiles are derived from traffic counts in Mexico. The emissions inventory is described in Garcia et al. (2018). "
   We added the following sentence at p11,l10:
   " Another potential error source of our method are the accuracy of the week-to-week and monthly variations of the emissions in the INEM inventory considering the fixed overpass time of TROPOMI."

5. **P4 L29: Here, z is the mean elevation in the TROPOMI CO ground pixels and zref = 2240 m the reference altitude which is set to the elevation of Mexico City. This correction is unclear. The altitude used by TROPOMI is defined as a surface pressure. The altitude error depends on the difference between the WRF surface pressure and the TROPOMI surface pressure. Why using an average altitude of Mexico City as a reference?**

   adjusted

   Actually, the reference altitude can be arbitrary chosen. However, we selected a altitude that is in the range of our region of interest. We change the sentence at p4, l29 from:
   " Here, z is the mean elevation in the TROPOMI CO ground pixels and $z_{\mathrm{ref}} = 2240$ m the reference altitude which is set to the elevation of Mexico City. "
   to
   " Here, z is the respective elevation of the TROPOMI CO ground pixels and $z_{\mathrm{ref}} = 2240$ m is an arbitrary reference altitude set to the elevation of Mexico City, . . . "

6. **P5 - L4: local enhancements of CO are due to emissions of the city districts of the same day Have you tested that assumption? Basin cities are often problematic with low wind speed for days, which can accumulate signals from more than one day in the basins (example: Los Angeles during Winter). An averaged wind speed or residence time of tracers would help justify this assumption.**

   adjusted

   We added the following sentence at p11,l10: " Furthermore, basin cities can be problematic with low wind speed for days, which could lead to accumulate signals from more than one day in the basins which is not yet covered by our approach. To account for this effect in our inversion needs major adjustments, which will be investigated in follow up studies. "

7. **P5: Use the current notation for multivariate regression used in most publications (observation operator H, state vector x, prior error cov B, Obs error cov R, observations y, Kalman gain K).**

   not adjusted

   We are using the notation of Rodgers (2000) that is commonly used in our field.

8. **P6 - L17: with little forward model errors. Unclear. Re-phrase.**

   adjusted

   We changed the sentence p6,l17 from:
   " On one hand, it should be large enough to estimate mean emissions for the period of TROPOMI observations, and the other hand strict data filtering is required to get a stable inversion with little forward model errors"
   to
   " On the one hand, the ensemble should be large enough to estimate mean emissions for the considered time period, but on the other hand it should be strictly filtered for cases where the forward model is in good agreement with the measurement such that a stable inversion of the all emissions is possible. "

9. **P6 L24: for low wind speeds more reliable. The model errors are critical during low wind speed conditions when a slight change in wind speed can affect the magnitude of the observed enhancements. Typically, high wind speeds should be avoided because local enhancements are weak while low wind speeds should also be avoided when a small change in the wind speed can significantly change the local enhancements.**

   adjusted

   Please see our answer to the first mayor comment of the referee. In the new version of the manuscript we will not filter on the wind speed anymore. We found removing this filter criteria has no significant effect on our results (see the attached Fig. 1).

10. **P6 L30: fit residuum ¡ 8ppb, and the standard deviation ¡ 8ppb to limit the effect of too large forward model errors. TROPOMI is an averaged enhancement over a grid cell. Point sources will be under-estimated in the data as the plume will not be mixed over the entire grid cell. This bias has been presented by the TROPOMI team. Can you conform the relationship between point sources over the domain and the location of these high model-data differences?**
    In addition, removing noisy pixels will artificially decrease the uncertainty by removing undesirable pixels. Some of these large model-data differences might be real transport errors or observation noise.

    adjusted

    The under-estimation reported by the TROPOMI was analyzed for point-sources smaller than a TROPOMI pixel deploying the mass balance method. This does not hold here, we are looking at city districts that in general extend over multiple TROPOMI ground pixel (see Fig. 4 of the old manuscript).
    The filter criteria criticised by the referee is not applied anymore in the new version of the manuscript. Please have a look at our answer to the first major comment of the reviewer. Removing this criteria was not changing our results significantly (see the attached Fig. 1).

11. **P7 L1: ...(ymeas) ¡ 0.65 to ensure that the forward model can explain the vari- ability of the measured CO field. This value seems arbitrary. How did you define it?**

   adjusted

   Please see our answer to the first mayor comment of the referee. In the new version of the manuscript we will not apply this filter anymore. We found removing this filter criteria has no significant effect on our results (see the attached Fig. 1).

12. **P7 L4: What I the impact on the seasonal distribution? Does it remove data evenly over the year? This filters are likely to bias your results over specific season. A figure showing the time depencen of the filtered data is needed (or statistics)**

   adjusted

   Please see the major comment one of this referee. We removed many of this filters and our results are still stable. However, an even sampling over all seasons is not use full here. We described in p8l1-9 that in particular measurements during the biomass burning season are too difficult to interpret with our method and should be rejected to ensure data quality.

13. **P7 L15: Kbg and Kelv are not regularized How can you optimize the emissions without regularizing the background values? Are they pre-determined? How were they defined?**

   adjusted

   The signal about the background parameters is very strong. The parameters are fitted without regularization for each overpass of TROPOMI. Please see the major comment of referee 1.

14. **P7 15-20: The balance between prior information and data constraint is usually computed with the Chi2 normalized distance. A value near one will define the optimal balance between the two. Michalak, A., Hirsch, A., Bruhwiler, L., Gurney, K., Pe- ters, W., and Tans, P.: Maximum likelihood estimation of co- variance parameters for Bayesian atmospheric trace gas sur- face flux inversions, J. Geophys. Res.-Atmos., 110, D24107, https://doi.org/10.1029/2005JD005970, 2005.**

   not adjusted

   A major problem we are facing in this study is that the priori information from the INEM inventory seems to be biased. Hence, we don't want to constrain our inversion to much by it.

15. **P7 L21-24: This approach weighs toward pixels that are co-located with the sources. In other words, it selects preferentially the pixels above the city. In general, it should work but an evaluation perido would be helpful (with and without the filter) to measure the impact of the selection. This approach might bias the results if the model under- /over-estimate urban pixels.**

   not adjusted

   When we don't apply the regularisation the inversion depends on forward model errors what we showed in this study.

16. **P7 L23: temporal variation of the INEM emissions to be about 40%**

   . . .

   vary with 60% around their average How do you define the 60%? The link between 4060% is not explained. In addition, temporal variations and mean emission errors are not supposed to scale together. This part needs to be described more carefully. The emission errors should depend on the emissions alone instead of their temporal vari- ability.

   adjusted

   The value 60% was chosen to not over regularize the inversion. Hence, we could of course enforce 40% here but this choice is a balance between propagation of errors and extraction of information content from the measurement. To make this more clear, we change the paragraph at p7,l22 from:

5

" Considering the temporal variation of the INEM emissions to be about 40%, we adjusted the regularization parameter $\gamma_1, \cdots, \gamma_{10}$ such that the retrieved emissions vary with 60% around their average. Hence, our regularization is not enforcing that the retrievals show the same variability as the emissions of the inventory. The value 60% is an empirical parameter which the retrieval more freedom to balance information content and noise propagation. This puts a moderate constraints on the inversion ensuring on one hand a stable inversion and on the other hand a realistic variation of the retrieved emissions around the priori. "

to

" Considering the temporal variation of the INEM emissions to be about 40%, we adjusted the regularization parameter $\gamma_1, \cdots, \gamma_{10}$ such that the retrieved emissions vary within 60% around their average. The value 60% is empirical chosen to balance information content against noise propagation. It puts a moderate constraints on the inversion ensuring on the one hand a stable inversion and on the other hand a realistic variation of the retrieved emissions around the priori. "

17. *Figure 4: This figure provides illustrations but is not very helpful to prove that the WRF model is reliable or good enough. Instead, model-data mismatches should be pre- sented in a synthetic figure, for different times of year, using whisker boxes. Snapshots for four days out of 160 is too few to convince the readers. This figure should be replaced.*

   adjusted

   We disagree. We think that this examples clearly show the advantages and limitations of our our approach as we also discuss it in the manuscript.

18. *P8 L18: This clearly shows that regional models like WRF have a great potential for the interpretation and analysis of TROPOMI data. No, this does not demonstrate the model capabilities nor the ability fo the model to extract emissions. Re-phrase.*

   adjusted

   We changed the sentence p8,l18 from:
   "This clearly shows that regional models like WRF-chem have a great potential for the interpretation and analysis of TROPOMI data. "
   to
   " This clearly shows that regional models like WRF-chem have a great potential to reproduce the large-scale patterns seen by the TROPOMI instrument. "

19. *P8 L21: For atmospheric conditions under high wind speeds the WRF simulations can deviate more from the TROPOMI measurements as shown in Fig. 4 (c). This single day is too limited to conclude anything. More statistics on windy days are needed to prove your point is valid here.*

   adjusted

   The aim of the study is not to analyze model error under high wind speeds. Please have a look at our answer of the major comment of referee 1. We removed the filtering on wind speed and our results were not changing significantly (see the attached Fig.1)

20. *Figure 5: The modeled and observed XCO should be presented first, summarized for the days available before and after filtering. Do the residuals show a seasonality? The results of the optimization are difficult to interpret without the evaluation of the initial model results.*

   not adjusted

   The comment is not well formulated and we don't understand it.

21. *P9 L1-6: Large model-data mismatches are expected from observations, model er- rors, and prior errors. If observations are being used twice (if I understood correctly), mismatches will decrease automatically. Optimization should never be performed a second time with the same data. Unless I misunderstood the approach (different data are being used between these two steps), only one step should be performed. Other- wise the constraint from the data is over-estimated.*

adjusted

Please see our answer to the major comment two of the referee.

22. *P9 L7: Furthermore, non-uniform variation of the background CO concentration can be a additional reason for this scatter (as shown in Fig. 3). How did you determine the background? How do you separate the contribution from the city emissions? Is it all performed within the inversion? If so, how do you define background uncertainties? Some additional tests should be performed. If you introduce a background in the bias, is your optimization system able to recover that bias? Figure 6: statistics should be presented for the entire data set and not only for four days.*

adjusted

The background parameters are retrieved together with the emission estimates from the measurements of each TROPOMI overpass. Please have a look at the major comment of referee 1 were we adapted the manuscript accordingly. Fig. 2 of the old manuscript shows for each TROPOMI measurement the inverted background values together with the retrieval errors represented as error bars. However, it is for sure also important to analyze individual cases like shown in Fig.5 because not every issue becomes visible when looking at statistics. We tested our inversion routines and they are working. Hence, introducing and retrieving an artificial background is not helping further here.

We added a sub-figure in Fig.6 showing the goodness of the fit between TROPOMI and WRF for the priori emissions, the ones of the Pre-fit and Final. We changed the paragraph p6, l3-8 from:

" This yields the mean

$$\langle \delta \rangle = \frac{1}{J} \sum_{j=1}^{J} \delta_j \tag{13}$$

and the standard deviations of the residuals

$$\sigma(\delta) = \frac{1}{J-1} \sum_{j=1}^{J} (\delta_i - \langle \delta \rangle)^2 \tag{14}$$

The standard deviation $\sigma(\mathbf{y}_{\mathrm{meas}})$ and $\sigma(\mathbf{F}(\mathbf{x}))$ of the TROPOMI CO field and the corresponding WRF-chem forward simulation completes our set of diagnostics.
"
to

" To evaluate the fit quality for each overpass, we consider the fit residuals $\delta_i = y_i - \mathbf{K}_i x_{\mathrm{est},i}$. Additionally, we evaluate the goodness of the fit described by the reduced chi squared value,

$$\chi_i^2 = \frac{1}{\nu_i} \sum_{l=1}^{L} (\delta_i, l/y_{\mathrm{err},ik})^2 \ . \tag{22}$$

Here L is the number of observation of a single overpass, $y_{\mathrm{err},i}$ the retrieval error, and $\nu_i = L - \mathrm{DFS}_i$. "
We changed the caption of Fig.6 from:
" ... as a robust estimation of the standard deviation and the number of collocations (c). The number of collocation of the Pre-fit is the same for all tracer domains (blue line) but in the Final-fit it is changing due to the information content filtering. Here, a collocation corresponds to a specific day because TROPOMI overpasses the region only once.
"
to
" ... as a robust estimation of the standard deviation, (c) the median of the goodness of the fit (χ^2), and the number of collocations (d). The number of collocations and the χ^2 values of the apriori simulation and Pre-fit are the same for all tracer domains (blue and grey line) but in the Final-fit it is changing due to the information content filtering. Here, a collocation corresponds to a specific day because TROPOMI overpasses the region only once. "
We added the discussion at p9,l9:
" The χ^2 values in Fig. 6 clearly show that the agreement between TROPOMI and WRF can be improved by fitting the emissions of the different city districts (blue line) instead of using the INEM inventory (grey

line). The regularization approach increases the χ^2 values (green bars) because the inversion can less compensate differences between TROPOMI and WRF by choosing unrealistic emissions. However, the χ^2 values of the Final-Fit are still lower than the ones for the prior INEM emissions (grey line). Overall, the χ^2 values exceeds 1 which indicates that the difference between TROPOMI and WRF is dominated by systematic errors in the WRF simulation. "

23. ***P9 L13-14: the averaging kernel shows that the Final-fit inversion is insensitive to deviations of the Tulancingo emission from the prior estimate. Whereas the Pre-fit inversion estimates very small emissions for this district, the subsequent regularization changes the emission only marginally. This is the direct consequence of performing an optimization with the same data twice. Some of the constraint has already been introduced in the emissions.***

    adjusted

    Please also see our answer to the major comments of the referee. The idea of the Pre-fit is to estimate mean emissions for all regions to prevent biases due to imposing a regularization. Hence, we want that the emissions of the Pre-fit and Final are the same.

    We change the sentence at p9, l14 from:
    " Moreover, the averaging kernel shows that the Final-fit inversion is insensitive to deviations of the Tulancingo emission from the prior estimate. The reason for this is that the Pre-fit inversion was only estimating very small emissions for this district and the regularization of the Final-fit is changing this only marginally. "
    to
    " It indicates that TROPOMI measurements can be used to distinguish emissions of the different urban districts of Mexico, with the exception of the emissions of district Tulancingo. Due to the small mean emission, the averaging kernel indicates a low sensitive of the data product. "

3 Special Comments

1. ***Specific comments: P6 L15: depends crucially = highly depends on***

    corrected

References

Borsdorff, T., aan de Brugh, J., Pandey, S., Hasekamp, O., Aben, I., Houweling, S., and Landgraf, J.: Carbon monoxide air pollution on sub-city scales and along arterial roads detected by the Tropospheric Monitoring Instrument, Atmospheric Chemistry and Physics, 19, 3579–3588, https://doi.org/10.5194/acp-19-3579-2019, URL https://www.atmos-chem-phys.net/19/3579/2019/, 2019.

Rodgers, C. D.: Inverse methods for atmospheric sounding: theory and practice, vol. 2 of *Series on atmospheric, oceanic and planetary physics*, World Scientific, Singapore, River Edge, N.J., reprinted : 2004, 2008, 2000.
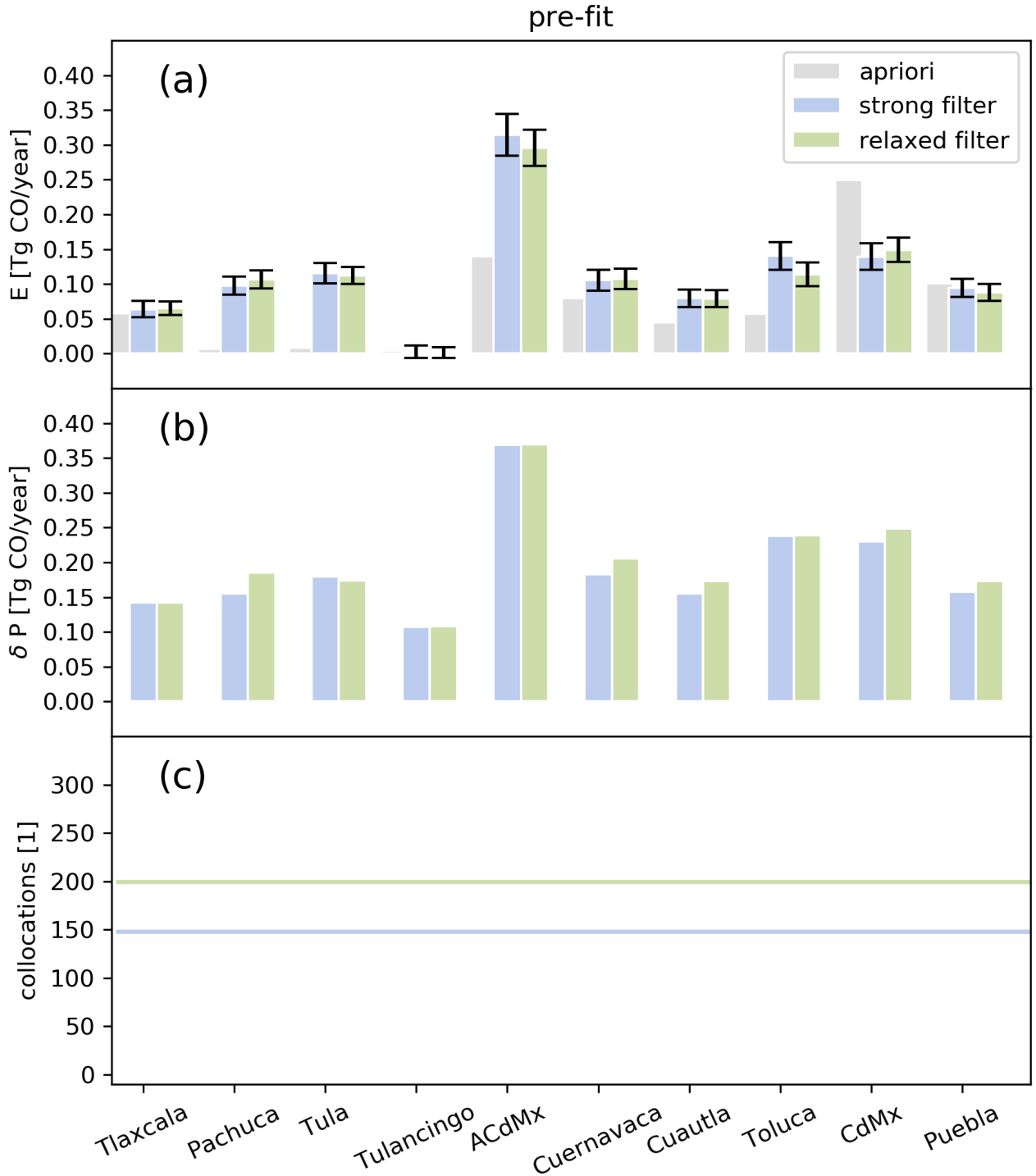
Figure 1: Statistics of CO emissions averaged from the 9th of November 2017 to the 25th of August 2019 for the tracer domains shown in Fig.1. (a) Median of the priori emissions (adapted INEM inventory) used for the WRF-chem simulation (grey) and retrieved from the TROPOMI data (Pre-fit in blue, Prefit-fit with relaxed filtering in green). The error bars indicate the standard error of the mean calculated from the delta percentiles (b) used as a robust estimation of the standard deviation and the number of collocations (c). The number of collocation of the Pre-fit is the same for all tracer domains (blue and green line). Here, a collocation corresponds to a specific day because TROPOMI overpasses the region only once.