Notes on revision made to egusphere-2025-4477

The authors would like to thank the two anonymous reviewers for their constructive comments and suggestions that have helped improve the quality of this manuscript. The manuscript has undergone a revision according to the reviewers' comments. Please see below our responses. For the reviewers' convenience, we have highlighted significant changes in the revised manuscript in blue.

Reviewer 1

Reviewer Comment 1.1 — Automatic determination of parameter T (L204) – How can the threshold value T be determined automatically? Although the authors discuss possible approaches around L255, it would be valuable to propose a general and robust method for future applications. I suggest testing both CO2M and TROPOMI datasets to evaluate whether an optimal solution can be consistently achieved.

Reply: Thanks for the relevant question. Indeed, it would be great if there was a way to make sure we didn't have to set hyperparameters manually. However, after testing a wide array of options (e.g., amongst other things, I tried to formulate various relations that takes signal-to-noise properties and map those into a relevant set of hyperparameters), I was ultimately not satisfied with these approaches – the solutions were over-engineered but seemed rather brittle, with little guarantee that they would work well on other datasets.

Therefore, I have decided to rather expose the full parameter landscape. I have added a function into the supplementary Python code which plots the results obtained with various (combinations of-) hyperparameters, such that a user can pick the best parameters.

The good news is that, in my tests, one generally only has to find a set of parameters that works well for a handful of images, and can then apply it to *all* (e.g., in the case of TROPOMI, 513) images. In other words, the exploration of the parameter landscape can be done quickly on just a few images, and then be used as a general solution for a given dataset.

We have added the following to the discussion, "The authors believe that rather than prescribing a value of T in this paper, it should remain an open hyperparameter for end users to decide. However, in the supplementary material we demonstrate two example workflows for selecting the parameters based on a simple grid search – either by minimizing an objective function that maximizes noise reduction while minimizing signal bias, or by performing a simple grid search optimization for the remaining hyperparameters. Once optimal parameters are found, they generally work well for a full dataset."

Reviewer Comment 1.2 — Combination with mean filtering –

While the MMSE methods preserve plume structures, the mean filter produces a much clearer image that could be more practical for identifying source locations and masking plumes. It would be useful to discuss the potential of combining the mean filter and MMSE approaches, and to test such a hybrid method on representative cases.

Reply: The jMMSE method corresponds to the mean filter in the case where there is no signal, i.e., $\mathbf{C}_{dd} = \mathbf{C}_{nn}$. So the jMMSE can *always* be made equal, or tuned towards, a mean filter by making sure our noise covariance $\mathbf{C}_{nn}$ is close to, or at, the data covariance $\mathbf{C}_{dd}$. Put differently, a mean filter is essentially one possible implementation or end-member of the jMMSE.

Reviewer Comment 1.3 — L20: Please add the typical background and enhancement levels of atmospheric CO2.

Reply: Thanks for the comment; we have modified a sentence: "Achieving this for satellite observations of CO_2 is challenging, as enhancements (∼0-5 ppm) are minor compared to background levels (∼420 ppm) and retrieval uncertainties are high (Miller et al., 2007)."

Reviewer Comment 1.4 — L21: Define all acronyms in full on first use (GOSAT-GW, CO2M, TANGO).

Reply: Thanks for the comment; we have expanded their acronyms.

Reviewer Comment 1.5 — L31: Include a short paragraph explaining the practical motivation. Why is a "data-only" Bayesian estimator advantageous (e.g., independence from explicit forward models, suitability for on-board or distributed processing)?

Reply: Thanks for the comment; we have added "Compared to alternatives like a model-driven denoising approach relying on meteorological, emission, and satellite noise priors, a data-driven method requires fewer assumptions, as the empirical cross-field covariance between co-registered images is learned directly from the data."

Reviewer Comment 1.6 — L73: Avoid using uncertain wording such as "maybe" when presenting derived equations.

Reply: Thanks for the comment; we have reformulated this to say 'can be'.

Reviewer Comment 1.7 — L107: Clarify why a factor of 0.5 was chosen for min–max normalization. Have the authors tested other scaling factors, and what effect do they have on performance?

Reply: Thanks for the comment; indeed this is a hyperparameter for BM3D, and it has an effect on the performance. However, 0.5 proved to be a suitable default parameter. We have added "Note that the factor 0.5 is one of the two hyperparameters that can be modified when applying BM3D (the other being the prior observation error, σ_c, mentioned in the following)."

Reviewer Comment 1.8 — Figure 3: Add panel numbers for clarity.

Reply: Thanks for the comment; done!

Reviewer Comment 1.9 — L187: Explain how the +44.4 dB value was obtained.

Reply: Thanks for the comment; it is now included in the text that this refers to going from the 'noisy' image to the 'denoised' image, i.e., it simply follows from subtracting the two PSNR numbers, now $91.8 - 46.7$.

Reviewer Comment 1.10 — L207: Consider showing noise-removal results for the CO2M image as well.

Reply: Thanks for the comment; these results could be found in the appendix previously, and are now part of the supplementary material.

Reviewer Comment 1.11 — L215: Please explain why averaging over year's data increases noise levels.

Reply: Thanks for the comment; just to clarify, I computed the average of each individual image's estimated noise levels (i.e., $\bar{\sigma} = (1/N)\sum_{i=1}^{N} \sigma_i$) rather than somehow first averaging all the images and then computing its effective noise level which indeed would be expected to be smaller (i.e., $\bar{\sigma}' = \bar{\sigma}/\sqrt{N}$). In our case, we simply have the case that the noise level in each individual image is, on average, slightly higher than the one provided example. (also note that the $\bar{\sigma}/\sqrt{N}$ theory does not hold anyhow for our case, as satellite images taken at different times are not sampling an identical distribution). We addressed this issue by choosing the word 'considering' rather than 'averaging over' at the start of our paragraph.

Reviewer Comment 1.12 — L238: Discuss the cause of negative emissions. The authors could generate synthetic TROPOMI data to test whether these negative values reflect true signal variability or arise from the MMSE denoising process.

Reply: Thanks for the comment; to clariyfy, 'negative emissions' were already present in the original noisy data. We've added this detail in the text. We have furthermore added some speculation about the source of this negative number:

"In fact, the major deviations in Figure 7, the Newcastle Steel Works and Camden Power Station, yield *negative* emissions when using the original noisy data. Therefore, the divergence method does not provide a reliable estimate of these source emissions in these cases. This may be due to a (combination of a) variety of reasons, e.g., there was not enough data to produce a robust average; there was a temporal bias in the sampling (the divergence method needs satellite images with plumes being blown in all directions to produce a 'point' on a map; otherwise 'dipoles' with seeming positive and negative sides appear – so if our overpass consistently samples cases where the wind is blowing in one direction, we will get artifacts); the steady state assumption was invalid; our chosen effective wind is not appropriate for the terrain and/or plume."

Reviewer Comment 1.13 — The code listings are long and could be better shown by uploading them to a public repository (e.g., GitHub) and sharing key scripts or notebooks for illustration. Data could also be deposited on Zenodo for reproducibility.

Reply: Thanks for the comment; I will add the actually used code to the supplementary material instead (including the used notebooks!).

Reviewer Comment 1.14 — For example the code at L313 U.T((0,1,3,2)) is not valid NumPy syntax.

Reply: Thanks for noticing this error; We tried to make the listing as short and concise as possible, but that resulted in this stupid mistake.

Reviewer Comment 1.15 — Consider moving Appendices B–D to the Supplementary Information to make the main text more concise.

Reply: Thanks for the comment; that is excellent advice.

Reviewer 2

Reviewer Comment 2.1 — 1. Impact on emission estimates

The success of the denoising procedures is clearly demonstrated and quantified based on two performance scores (PSNR and SSIM).

However, the eventual goal is the quantification of emissions, and the crucial question is whether the proposed denoising introduces a bias that might then also bias the estimated emissions.

The authors deal with this question in Figs. 6 & 7, where they apply the divergence method to the SO2 maps and derive emissions from spatial integration.

However, the results are hard to interpret, since the true emissions are not known!

Thus, this question should be investigated based on the *synthetic* data where emissions are known:

- create an ensemble of noisy CO2 images

- apply the proposed denoising

- quantify emissions from divergence method

- whenever a clear peak is appearing in the denoised data, quantify emissions from a plume-based method (e.g. cross-sectional flux)

- compare the resulting ensemble mean emissions to the a-priori truth for the different denoising algorithms.

This may sound like a major task - however, the authors have all the tools needed at hand, and for me this is the most important question after reading this paper.

For future applications, it should not be the goal to create highest PSNR, but to get the most accurate emissions, so the latter needs to be quantified as well and added to the performance scores.

Reply: Thanks for the comment.

You are of course correct that in the end we care mostly about the emission quantification being correct; but that is a problem that is difficult even if we have noise-free images (e.g., we might use erroneous effective wind speeds when computing the fluxes, incorrectly assume steady state conditions, assume an erroneous background field, we may have temporal sampling issues, ...). All these factors make emission estimation inherently difficult, even if we have noise free images. Therefore, we cannot simply compare our results against an a priori ground truth, and necessarily expect we get better results. What we *can* do is compare our estimates against noise free estimates, and expect we get closer to the noise free estimates. All we can safely assume is that by doing emissions quantification with images with a higher SNR, we have eliminated one source of potential errors, but there are many other sources of errors that still remain.

We have run your desired test on a full year of synthetic SMARTCARB data (as also used in the main body of the paper) with a known ground truth and, importantly, a known 'noise free' field. We observed no systematic artifacts (i.e., biases). We have added this extra test into the supplementary material.

Reviewer Comment 2.2 — 2. Correlated errors

The derivation assumes independent errors (line 60), and equations become simpler by this assumption (line 77).

This is picked up in section 4 (discussion), shortly mentioning "structural noise patterns, such as stripes".

While it remains unclear, what could cause such structural noise patterns, I am more worried by systematic impacts such as ground albedo or clouds - any bias in the input data would cause correlated errors in NO2 and CO2 (or SO2). This unavoidable source of correlated errors and potential impact on the results needs to be discussed in more detail.

Reply: Thanks for the insightful question. We have added this to the manuscript, "Another violation of the i.i.d. assumption may show up in the form of correlated errors in both satellite products due to shared dependencies in the radiative transfer (such as albedo, cloud contamination or aerosol loading). These factors enter both retrievals, thereby potentially introducing a common-mode error component. Mitigation strategies could include restricting the analysis to scenes with low sources of such errors, or extending the statistical model presented here to include such correlated errors or biases. Consequently, users should interpret the denoised image with caution in regions where strong scene-dependent retrieval artifacts are expected, as these features, shared between the low and high SNR images, may be preserved or even reinforced in the denoised output."

Reviewer Comment 2.3 — 3. Datasets

The proposed method is applied to synthetic data as well as to actual satellite measurement. For the latter, details about the data are missing.

Please add a paragraph on the used satellite data, in particular the chosen products and processor versions, plus appropriate references.

In particular for SO2 this information is crucial, since the operational processor was recently switched to the COBRA algorithm with far lower noise levels.

Reply: Thanks for your comments. We have added this paragraph, "The method is also evaluated using Level-2 trace gas products from Sentinel-5P / TROPOMI, which provides near-daily global coverage at a spatial resolution of approximately 7×3.5 km^2 at nadir. We use reprocessed (RPRO) SO_2 data (processor version 02.04.01) and NO_2 data (processor version 02.04.00) for the year 2021 (European Space Agency (ESA), 2020, 2021). We perform the denoising using a qa_value ¿ 0.35 to retain some more data for the denoising methods to work with, but after denoising, we analyse and plot the results using the standard recommended threshold of qa_value ¿ 0.75. The SO_2 product used here corresponds to the SO_2 pre-Covariance-Based Retrieval Algorithm (COBRA) product. The more recent operational RPRO product is obtained using the COBRA algorithm, which has significantly lower noise and a corrected bias in the retrieval. As a consequence, the denoising performance demonstrated here for SO_2 should be interpreted in the context of this earlier processor version; results may differ for COBRA-based products due to their improved signal-to-noise ratio and potentially different error correlation structure"

Reviewer Comment 2.4 — 4. Figures

- size/clarity: Some figure (in particular Figs. 3&4) are not very clear. I would propose to use a different colormap, vary the value range, and increase the figures. Zooming in might help as well (a large area north and south from the plumes seem to be irrelevant).

- scaling / aspect ratios: please add a km scale and choose an aspect ratio such that distances in x and y are scaled same.

Reply: Thanks for your comment. We have updated figures 3 & 4 by zooming and adding a km scale to Figure 3a and the color range; the aspect ratio (as can be seen in the figures) is already equal (i.e., the SMARTCARBdata consists of pixels of roughly length 2×2 km^2, and they are indeed seen to be (nearly) square in the figures.

Reviewer Comment 2.5 — Additional comments:

Line 11: I find the 40 decibel hard to visualize and would prefer a ratio of SNR here which would be more transparent and common.

Reply: Thanks for your comment, I have added "(equivalently, an over 10^4 increase in SNR)" to the abstract.

Just as a sidenote, the 40 dB is itself the SNR; that is, we have

$$\mathsf{SNR}_{\mathsf{dB}} = 10 \log_{10} \left(\frac{S}{N} \right), \tag{1}$$

so

$$(\mathsf{SNR}_{\mathsf{dB}})_{\mathsf{after}} - (\mathsf{SNR}_{\mathsf{dB}})_{\mathsf{before}} = 40 \tag{2}$$

implies

$$\frac{S}{N}_{\mathsf{after}} = 10^4 \frac{S}{N}_{\mathsf{before}}, \tag{3}$$

thus saying a +40 dB improvement we mean a 10 000 increase in SNR.

Reviewer Comment 2.6 — Caption Figure 1: I don't understand / see the "high contrast edges" which should be visible in SO2 but not in NO2.

In Box C, I see far more plumes in NO2 than in SO2. So to which "edges" does this statement refer/what is the message of this section?

Reply: I have replaced Box C with an arrow, pointing into the corridor I intended to highlight in the figure and the caption.

Reviewer Comment 2.7 — Line 85: Please explain why hereafter it's E(M) instead of E(∼M).

Reply: Excellent catch! With $\mathbf{M}$ I intended to imply the noise-free version of the data, such that $\mathbb{E}[\mathbf{M}] = \mathbb{E}[\tilde{\mathbf{M}}]$ under the i.i.d. noise assumption. However, I had not defined $\mathbf{M}$ as such to begin with. Therefore, I have changed those occurences to $\mathbb{E}[\tilde{\mathbf{M}}]$ in the main text.

Reviewer Comment 2.8 — Line 165: Please provide information on the resolution/gridding of the SMARTCARB data.

Reply: I have added "The SMARTCARB dataset is a synthetic quasi-Level 2 product at the CO2M spatial resolution (roughly 2×2 km^2) and swath length (roughly 250 km) over primarily Germany and surrounding regions, for the year 2015."

Reviewer Comment 2.9 — Fig. 3: CO2 colorbar is given for upper row, but applies for left column. NO2 colorbar is given for lower row, but applies for right column.

Reply: Thanks for the comment. I have switched the subfigures around such that both CO2 images are now given on the top row and both NO2 images are given in the bottom row, aligned with their colorbars.

Reviewer Comment 2.10 — Line 196 / footnote 3: So what is the benefit of such a coarse AMF correction? Isn't this just an upscaling of the SO2 data? Does it make any difference for the resulting maps and noise scores if this correction is applied or not?

Reply: Thanks for your comment. In the section we're looking at, we expect that the SO2 enhancements we see on the satellite picture correspond merely to powerplant emissions (which are present close to the surface). Hence, I applied an (admittedly coarse) AMF correction which says that we expect all mass to be present close to the surface. Indeed, this is an 'upscaling' (i.e., this makes the SO2 values larger) of the data. It makes a difference in the sense that the amplitudes of the resulting maps is an order of magnitude larger as we apply this AMF correction compared to if we didn't. I have compared some of our emission estimates to reported emission estimates (e.g., for Matimba/Medupi), and using this AMF correction we get close to the reported values, whereas if we do not apply this AMF correction we are an order of magnitude off.

Reviewer Comment 2.11 — Line 200: Hard to read with multiple negations (large region of low inverse SNR - is this good or bad?).

Reply: I fully agree that inverse SNR is not as intuitive as traditional signal-to-noise ratio; but in the context of Figure 5 we can see that where the inverse SNR (so, the noise-to-signal ratio) is small we have good signal whereas if the inverse SNR is high we have bad signal; we thus see the same trends in 5(b)-5(d)-5(g)-5(h), indicating that the signal is protected where it is good. In the manuscript we already write '...regions of low inverse SNR values (indicative of a good SNR)'. So, even though it is not very intuitive, I don't quite see how to make things better.

Reviewer Comment 2.12 — Line 220: I assume the averaging of the ERA5 winds uses the GNFR-A profile as weights - please clarify.

Reply: Correct! I added "In this case, wind fields were computed by vertically averaging ERA-5 reanalysis fields using the GNFR-A emission profile as weights."

Reviewer Comment 2.13 — Fig. 5:
- which filter has been applied for the satellite measurements?

Reply: It is unclear what your question is – the figure subtitles and caption state which denoising methods where applied.

Reviewer Comment 2.14 — please use consistent units for columns (molec/cm2 as in Fig. 3 or mol/mˆ2 as in the TROPOMI products, but then also for the synthetic data).

Reply: We have modified the synthetic data to use the same units, mol/cm2. I opted for mol/cm2 for the TROPOMI images because it yielded a good kind of 'value range' with single or double digit normal numbers.

Reviewer Comment 2.15 — add the PSNR/SSIM scores to (e) and (f), as in Fig. 4.

Reply: This cannot quite be done for figure 5, because it requires knowing the ground truth which is unknown. We have included an estimate of the noise following the method of Immerkaer (1996) in the text, but that is, essentially, nothing more than an estimate.

References

European Space Agency (ESA) (2020). Copernicus sentinel-5p tropomi level 2 sulfur dioxide total column products.

European Space Agency (ESA) (2021). Copernicus sentinel-5p tropomi level 2 nitrogen dioxide total column products.

Immerkaer, J. (1996). Fast noise variance estimation. *Computer vision and image understanding*, 64(2):300–302.

Miller, C., Crisp, D., DeCola, P., Olsen, S., Randerson, J., Michalak, A., Alkhaled, A., Rayner, P., Jacob, D. J., Suntharalingam, P., et al. (2007). Precision requirements for space-based x_{CO_2} data. *Journal of Geophysical Research: Atmospheres*, 112(D10).