Authors' reply to RC2 by Anonymous Referee #2

We would like to thank Anonymous Referee 2 for the constructive and encouraging review of our manuscript. Please find our point-by-point replies below. The Referee's comments are printed in blue, and our replies in black.

Review of Kuhn et al., "NitroNet- A deep-learning NO2 profile retrieval prototype for the TROPOMI satellite instrument"

Reviewer suggestion: minor revisions.

This paper presents a new NO2 retrieval model to produce vertical profiles from satellite observations, using a machine learning approach. In my opinion, this is an impressive piece of work, thoroughly explained, well executed, and producing impressive results. The results of NitroNet comparisons to vertical columns and surface values, within and outside of the training times/regions, shows very good promise.

I think that the main weakness of the paper lies in the challenge of verifying the NO2 vertical profiles, not just columns and surface values. This is inherent to the point of the paper of course, i.e., that NO2 vertical profile measurements are sparse. The authors tackle this by comparison to the FRM4DOAS MAX-DOAS dataset, and results are promising. I think it would improve the paper to include comparison to more MAX-DOAS datasets if possible, outside the European domain and over more seasons. Perhaps this could be achieved by looking at a few discrete layers in the profile, not necessarily full profile comparison plots. I also think the authors should consider whether verification against cloud-sliced NO2 data, or aircraft campaign NOx measurements, are an option to demonstrate the capability of NitroNet to provide information on free- and upper-tropospheric NO2 tropospheric profile.

We acknowledge, that the validation against NO_2 profile observations is of high importance. However, the required reference data are hard to obtain. For example:

- NO_2 profiles from cloud slicing are usually reported on coarse spatio-temporal grids due to averaging (e.g. seasonal means with 1° x 1° horizontal resolution and 5 tropospheric layers, see Marais et al., 2021).

- A reasonable validation against aircraft measurements would require data recorded on the central European domain after TROPOMI went operational (2017). There are datasets which meet these criteria, (see e.g. Riess et al., 2023; Brenninkmeijer et al., 2007), but these are just as sparse as the FRM_4DOAS measurements used in our manuscript.

- NO_2 sonde measurements are equally sparse, and should be considered immature compared to other measurement methods.

The possibility of using such reference data for validations in the future was added to the outlook. The following sentence was added to sect. 5: *NO_2 profile observations from cloud-slicing (see e.g. Marais (2021)) or aircraft measurements (see e.g. Riess (2023); Brenninkmeijer (2007)) may be used for further validation of NitroNet at various altitudes.*

Based on the Referee's suggestions, the validation against FRM_4DOAS profiles was extended. More details are given further below in the reply to the Referees comment referring to l. 425 at the bottom.

I have listed some specific minor revisions below.

Introduction:

It is worth mentioning that there are methods of determining some vertically-resolved NO2 information from satellite observations, e.g. cloud-slicing, and also there are aircraft campaigns providing vertically-resolved NOx information.

See the answer above. Cloud slicing was added to the introduction. Aircraft measurements were already

mentioned there.

The following paragraph was added to sect. 1: *Although further measuring platforms (e.g. sondes, aircraft) and methods (e.g. Light Detection and Ranging instruments (LIDAR), or "cloud-slicing") exist, these are not routinely deployed (see e.g. Sluis et al. (2010); Bourgeois et al. (2022); Lange et al. (2023); Riess et al. (2023); Volten et al. (2009); Berkhout et al. (2018); Su et al. (2021), Marais et al.(2021)). Particularly aircraft measurements and cloud slicing are appreciated for resolving along the vertical axis, although at lower spatio-temporal resolutions (e.g. cloud slicing: seasonal means with 1° × 1° horizontal resolution and 5 tropospheric layers, see Marais et al. (2021)) or sparse spatio-temporal coverage (aircraft measurements).*

You mention that TROPOMI NO2 relies on a priori profiles, but it is also worth noting in your initial comments that the same is true for MAX-DOAS NO2 vertical profiles.

Please note, that only MMF (not MAPA) depends directly on an a priori profile. MAPA depends implicitly on a priori assumptions, e.g. in the form of the profile parametrization.

The following sentence was added to sect. 1: *Additionally, the commonly used retrieval algorithms suffer from significantly reduced sensitivity at higher altitudes (> 2 km), and depend on a priori assumptions.*

Line 89: 'cannot' rather than 'can not', and later in the sentence I think you mean 'inherent to the training data' not ,immanent…'

We have replaced this occurrence (and several others) of „can not" with „cannot". Also, „immanent" was replaced with „inherent".

Line 110: it would be helpful to the reader to include a brief comment on why the O3 VCDs are included in NitroNet.

The following sentence was added to sect. 2.2: *Additionally, although much less influential, total O_3 VCDs are used, assuming they are also informative of the tropospheric O_3 column, and thus of the tropospheric NO_x photochemistry.*

Line 137: MAX-DOAS measurements are strongly influenced by clouds. You mention the filtering of clouds by virtue of the selected TROPOMI QA flag: is there a similar filtering for cloudy results for FRM4DOAS MAX-DOAS results?

According to Beirle et al. (2019), MAPA does not provide automatic cloud flagging yet. However, MAPA provides three quality flags („valid", „warning", „error"), which were also shown to be sensitive to cloud effects. In our analysis, we removed all MAPA profiles flagged with „error". No other filter criteria were used.

The following sentence was added to sect. 2.3: *All profiles flagged as "erroneous" by MAPA were discarded. Note, that although MAPA does not support automatic cloud filtering yet, the described „error" flagging was shown to be sensitive to cloud effects, as well (see Beirle et al. (2019)).*

Line 176: A reference for Shapley scores would be good here.

We have added the following reference:

Štrumbelj, E., Kononenko, I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 647–665 (2014). https://doi.org/10.1007/s10115-013-0679-x.

Line 191: This statement is a little unclear to me: 'The learning rate was halved whenever training progress had stalled over several epochs'. Perhaps you could clarify?

Close to the loss minima, training can stall if the learning rate is chosen too large. This is because the parameter updates can overshoot, thereby missing the ideal solution to the optimization problem. This can be solved by using a learning rate scheduler, which decreases the learning rate upon stagnation of the loss. In our routine, the learning rate was halved, whenever the validation loss had not decreased over 20

epochs. More information on learning rate schedulers can be found e.g. in the torch documentation (`https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`).

The following paragraph was added to sect. 3.2: *In order to reduce early stagnation of the training process as a result of too large learning rates, a simple learning rate scheduler was used (`ReduceLROnPlateau`, see Paszke et al. (2019)). The learning rate was halved whenever the training progress, as measured by the validation loss, had stalled over several epochs (meaning full iterations over the training set).*

Line 211: Is the low bias you mentioned improved or worsened if the filtering criteria are relaxed from the tuned DVCD and DPBLH?

The bias is given in reference to training the neural network on un-filtered data. Relaxation of the filter criteria would lead to a reduced bias. Note, that the bias can be immediately corrected for, because it is already quantified during training (before „prediction time"), which is already mentioned in the manuscript.

Line 251: high NO2 in the upper troposphere is also linked to long lifetime of NOx reservoirs, lightning and subsidence from the stratosphere.

The following sentence was added to sect. 4.1: *(…) which could be linked e.g. to aircraft emissions, decay of NO_x reservoirs, lightning, or stratosphere-troposphere exchange.*

Line 254: Could you provide a brief comment on why the model performs better at high NO2 concentration than low? Is this largely due to the better agreement in the lower troposphere/more polluted layers?

The following paragraph was added to sect. 4.1: *The relative prediction errors are smaller at higher NO_2 concentrations. This is because the high NO_2 concentrations at the surface are more strongly correlated to the NO_2 VCD, which is the main model input. Vice versa, the correlation is weaker in higher layers, where the concentration tends to be lower. Therefore, the combined input variables are more descriptive of the lower, more polluted layers, and allow the neural network to make a more precise prediction.*

Figure 5: I presume that the WRF-Chem comparison to Airbase is achieved with in-situ bias correction (F factor) calculated by WRF-Chem, and that the NitroNet comparison is achieved with F calculated by NitroNet? How well do the F factors agree between WRF-Chem and NitroNet? Could any discrepancies in F factor help explain any of the observed in-situ NO2 biases in Fig 5?

The comparison is made as described by the Referee. When training NitroNet on the *F* targets, a relative test error (note: not a bias) of 5 % was determined (this was already mentioned in sect. 3.5). Therefore, discrepancies in *F* are most likely not the reason for the observed biases.

The following paragraph was added to sect. 4.1: *As mentioned before, NitroNet is able to reproduce the NO_z correction factors of WRF-Chem with a relative precision of $\pm$ 5 % and no bias. Due to the good agreement between WRF-Chem and NitroNet in this regard, the prediction of the NO_z correction factors cannot explain the low-biases observed in Fig. 5.*

Figure 9: Is it possible to show the standard deviation of the mean monthly profiles for each technique? It would be interesting to know how significant the profile differences are given the in relation to the variability across the month. Just to clarify, have you only taken MAX-DOAS profiles from FRM4DOAS at the TROPOMI overpass time?

The standard deviations were added to Fig. 9 as requested by the Referee, and an explanatory sentence was added to its caption. The updated Fig. 9 is also shown below.

The following sentence was added to sect. 4.2.2: *A temporal threshold of 60 minutes is used, meaning that each NitroNet NO_2 profile is associated with the average over all colocated MAX-DOAS profiles recorded within 60 minutes of the corresponding satellite overpass.*

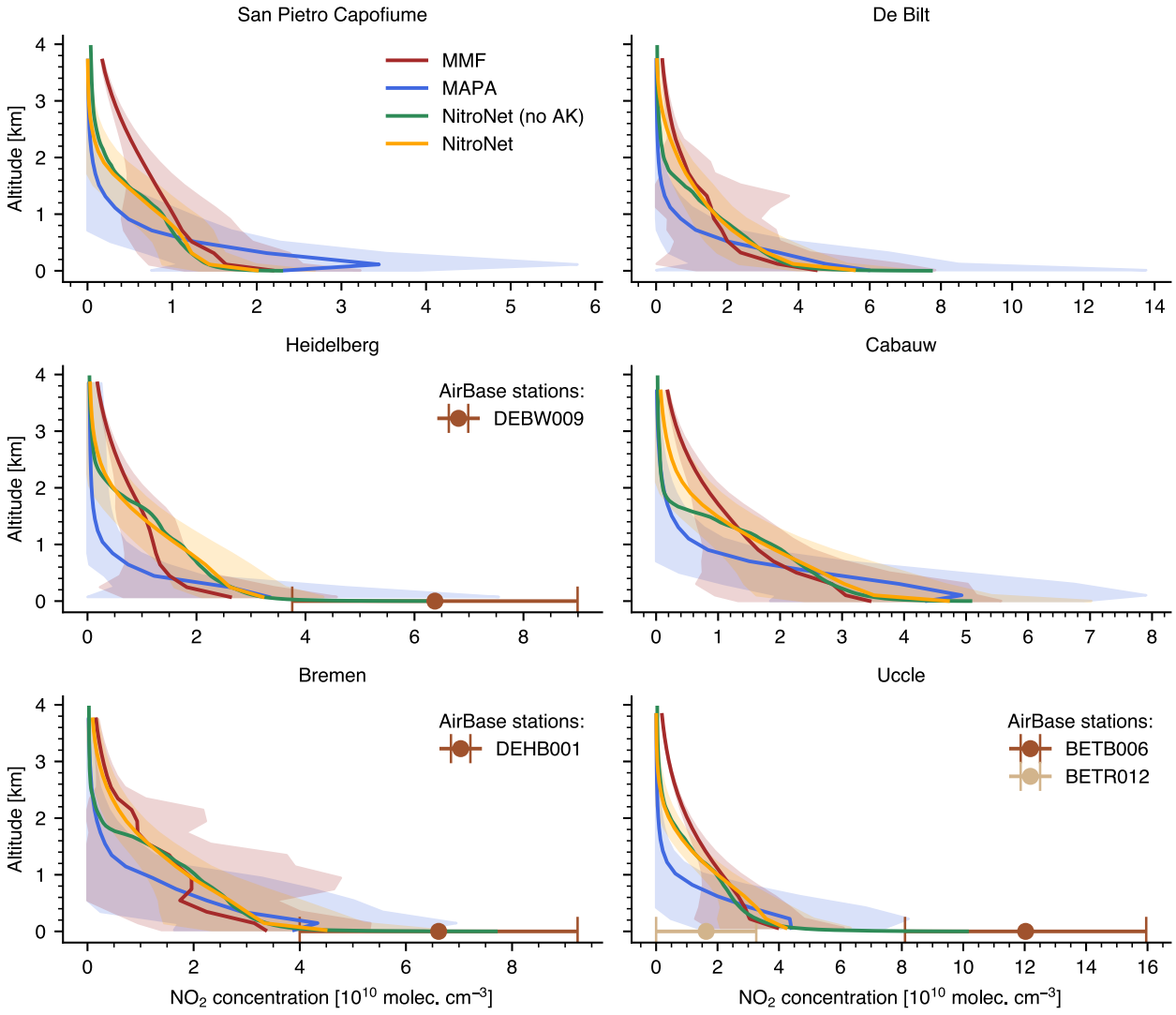

Figure 9. Comparison of monthly-mean FRM4DOAS NO_2 profiles against NitroNet profiles (May 2022). The monthly standard deviations of the profiles are drawn as shaded regions in the background. Where available, colocated AirBase measurements of the surface NO_2 concentration within a radius of 5 km were drawn at 0 m altitude.

Line 375: You say in relation to profiles with elevated layers of NO2 that 'NitroNet is unable to reproduce this profile type, most likely because the training dataset contains very few corresponding examples'. Is this something that can be rectified? In principal, or even better if you're able to show it, is it possible to provide more elevated layer examples in the synthetic training data to address this problem?

At the moment, we cannot overcome this obstacle. As described in our preceding publication (Kuhn et al., 2024), the WRF-Chem model struggles to reproduce such elevated layers in many cases (except close to very strong point sources, e.g. power plants). Because NitroNet is trained on WRF-Chem data, it suffers from the same limitations.

In the future, we can attempt to improve NitroNet's profile diversity by

- producing WRF-Chem training data on significantly higher resolution
- attempting to train NitroNet with a weighted training set, where profiles with elevated layers are given larger weights

The following sentence was added to sect. 4.2.2: *As shown in Kuhn et al. (2024), the WRF-Chem model, which provides NitroNet's training data, also struggles to reproduce elevated layers in some locations.*

The following sentence was added to sect. 5: *Similarly, NitroNet could benefit from training data of higher horizontal resolution, which might improve its ability to reproduce more complex NO_2 profile shapes, e.g. with elevated layers.*

Line 400-401: There are a number of outstanding research questions related to NOx over the oceans, for example the contribution of ship emissions in the lower troposphere, and the role of lightning in upper tropospheric NOx over the ocean. Is your hypothesis here that NitroNet performs worse over the oceans because the model gets ship NOx emissions wrong, biasing your training set? Rather than state that the oceanic regions are less relevant, it would be good to understand your thoughts on how NitroNet could be improved over the oceans.

We acknowledge the importance of NO_2 retrievals over water in relation to the research questions mentioned by the Referee. The relevance of oceanic regions was emphasized in the outlook and the paragraph in question was changed accordingly.

To clarify, our hypothesis is the following: Fig. 10b and Fig. 12a show an abrupt jump in NitroNet's prediction errors at the land-water boundaries. However, this is not observed in Fig. 4. If NitroNet's predictions work well on the training domain, but poorly on foreign domains, it indicates that the foreign domains are characterized by qualitative differences in the combination of features (inputs) and targets (outputs). Such differences could be caused, e.g. by an unrepresentative amount of shipping routes on the training domain.

The following paragraph was added to sect. 4.3: *The most likely explanation is that the training dataset does not contain enough representative examples of NO_2 profiles over water. The water regions of the training set must be assumed less representative, e.g. because they are pervaded by unusually many shipping routes, which may lead NitroNet to overestimate NO_2 over more remote water bodies.*

The following paragraph was added to sect. 5: *In particular, it might also help to resolve the prediction errors over water, which could be useful in addressing some of the outstanding research questions related to NO_2 over the oceans (e.g. the contribution of ship emissions and lightning to the lower / upper troposphere).*

Line 425: In terms of seasonal performance of the vertical profiling capability, it would be really valuable to assess NitroNet against the FRM4DOAS network over seasonal timescales. Seasonal comparison at a few specific altitudes, e.g. 0, 1, 3 km, would give an indicator of whether NitroNet consistently achieves its aim of providing NO2 vertical profiles.

We reply to this comment together with the Referee's suggestions in the initial general comment above.

We acknowledge, that the comparison to FRM_4DOAS data should be more detailed, with specific focus on the assessment of NitroNet's ability to predict realistic NO_2 profiles. As suggested by the Referee, an evaluation at individual altitudes over the period of one year was added to the manuscript, and is also shown below in Fig. 14. The altitude ranges (0 - 200 m, 200 - 400 m, 400 - 600 m, 600 - 1000 m, 1000 m - 2000 m) were chosen based on the limited vertical sensitivity of the MAX-DOAS retrievals beyond.

The following paragraphs were added to sect. 4.3:

Figure 14 shows a full-year evaluation of NitroNet against NO_2 concentrations from FRM_4DOAS in selected altitude ranges. For this analysis, NitroNet's average bias (left panel) and absolute error (right panel) over all previously shown FRM_4DOAS instruments were computed for a full year of data, with either MMF or MAPA used as reference. Each subplot of Fig. 14 is restricted to a specific altitude range (0 - 200 m, 200 - 400 m, 400 - 600 m, 600 - 1000 m, 1000 - 2000 m). In the lowest evaluation layer, at 0 - 200 m, there is particularly good agreement between MAPA and MMF, with NitroNet biases between -70 % and +20 % over the course of the year. Here, a similar tendency as in Fig. 13 can be observed, with low biases occurring during winter, and high biases during summer. The summertime high biases are of similar magnitude than in the comparison to TROPOMI VCDs and AirBase surface measurements (approximately + 15 % vs. + 23 %, and

+ 10 %, respectively). Particularly in the higher layers, the validation against MMF yields far lower mean biases, mostly in the range from -30 % to + 30 %, while the validation against MAPA result in larger biases of ~ 100 % at 600 - 1000 m, and ~ 200 % at 1000 - 2000 m. This owes to the steeper vertical concentration gradients of the MAPA profiles due to their assumed profile shape, and aligns well with the profiles shown in Fig. 9. The large relative biases of NitroNet in relation to MAPA might appear concerning at first, and should be put into perspective based on the following considerations:

First, it is hard to assess, which of the two retrieval algorithms yields more trustworthy results. Although conceptionally different, MAPA and MMF both suffer from increasingly poor sensitivity at higher altitudes. This is also the case here, as exemplified by the MMF averaging kernels shown in Fig. C6, which indicate an effective vertical sensitivity of up to 1.5 km in Heidelberg, May 2022. In consequence, the retrieval results are considerably affected by a priori assumptions. In the case of MMF, an a priori profile is taken from a WRF-Chem simulation over Mexico (see Friedrich et al. (2019)), which might be entirely unrepresentative of the central European domain investigated here. Parametrized retrievals such as MAPA do not require a priori profiles, which is an advantage in this context. Nonetheless, MAPA still depends on other a priori assumptions, e.g. in the form of the assumed profile shape by the choice of parametrization. In particular, the exponential tail of the MAPA profiles towards higher altitudes, which is the dominant characteristic here, is prescribed.

Second, computing the relative biases of NitroNet involves division of the absolute errors by the NO_2 concentrations of MMF, and MAPA, respectively. In the case of MAPA, these can be considerably small (e.g. ~ $0.1 \cdot 10^{10}$ molec. cm^{-3} for 1000 - 2000 m, see Fig. 9 for reference), for the reasons discussed above. Thereby, even moderate absolute errors (see right-side panel of Fig. 14) can result in large relative biases. Thus, the assessment of model performance by means of the prediction biases is informative in the lowest 3 evaluation layers (up to 600 m), but not beyond.

Another important finding of Fig. 14 is that the seasonal trends observed in Fig. 13 are represented in the lowest layer (0 - 200 m), but not the higher ones. This indicates, that the seasonal biases of NitroNet (and the underlying WRF-Chem training data) might be rooted in the lower regions of the troposphere.

We appreciate the reviewer's suggestion to include FRM_4DOAS data from other regions of the world as well. However, the remaining FRM_4DOAS instruments are located far away from the central European training domain of NitroNet (e.g. in Ny-Alesund, Norway). As shown in the manuscript, NitroNet's prediction quality can vary under such conditions. Many of these instruments are also operated in remote locations and / or have no colocated AirBase measurements. A validations against these instruments would require considerable additional efforts, introduce new uncertainties, and most likely contribute little to the overall assessment of NitroNet's performance.

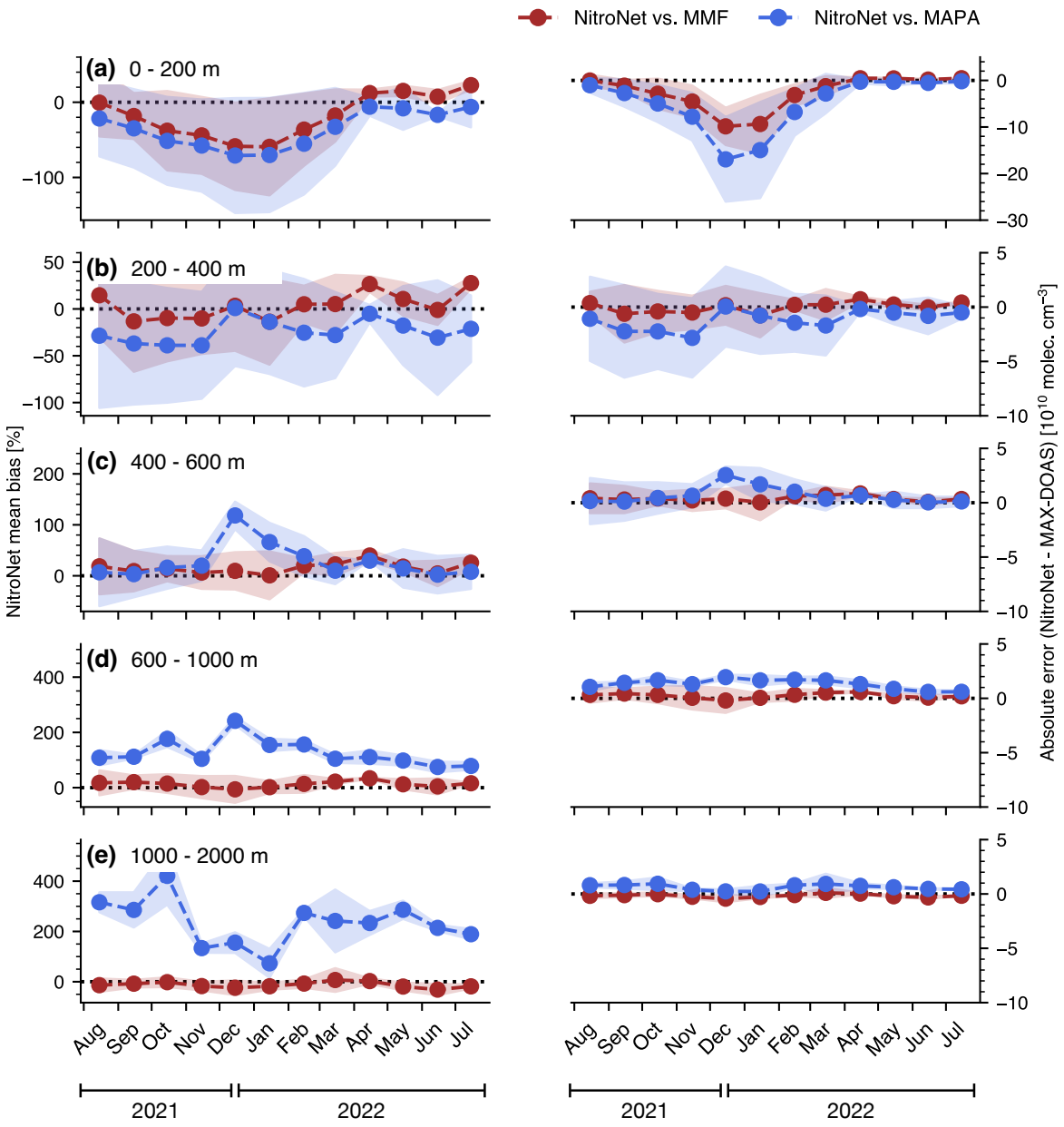

Figure 14. Seasonal evaluation of NitroNet against NO_2 concentrations from the FRM_4DOAS dataset. The left panel shows NitroNet's monthly-mean biases averaged over all available MAX-DOAS instruments. The right panel shows the corresponding absolute concentration errors. Each subplot refers to a specific altitude range, namely **(a)** 0 - 200 m, **(b)** 200 - 400 m, **(c)** 400 - 600 m, **(d)** 600 - 1000 m, and **(e)** 1000 - 2000 m.

Figure 13: I may be missing something here, but I'm unsure how the monthly mean correlation coefficients can be almost all above the daily mean correlation coefficients, and the monthly mean RMSE can often be below all the daily RMSE values for a given month (e.g. Apr-Jul 2022)?

In this context, it is important to distinguish between the monthly mean of a diagnostic, and a diagnostic computed on monthly means. Here, „diagnostic" refers to bias, RMSE, or correlation coefficient.

Figure 13 shows the diagnostics computed on monthly-mean data. It does not show the monthly means of the diagnostics computed on daily data. Monthly-mean data has significantly reduced noise compared to unaveraged data, resulting in larger correlation coefficients and lower RMSE. The bias is unaffected by this, because it is insensitive to (centered) noise.

We wish to keep the evaluation this way, and refer to our reply to RC 1. Please note, that the monthly means of the diagnostics computed on daily data can be directly obtained from Fig. 13 as shown. This is not the case for the diagnostics computed on monthly means.

References:

Marais, E. A., Roberts, J. F., Ryan, R. G., Eskes, H., Boersma, K. F., Choi, S., Joiner, J., Abuhassan, N., Redondas, A., Grutter, M., Cede, A., Gomez, L., and Navarro-Comas, M.: New observations of NO2 in the upper troposphere from TROPOMI, Atmos. Meas. Tech., 14, 2389–2408, https://doi.org/10.5194/amt-14-2389-2021, 2021.

Riess, T. C. V. W., Boersma, K. F., Van Roy, W., de Laat, J., Dammers, E., and van Vliet, J.: To new heights by flying low: comparison of aircraft vertical NO2 profiles to model simulations and implications for TROPOMI NO2 retrievals, Atmos. Meas. Tech., 16, 5287–5304, https://doi.org/10.5194/amt-16-5287-2023, 2023.

Brenninkmeijer, C. A. M., Crutzen, P., Boumard, F., Dauer, T., Dix, B., Ebinghaus, R., Filippi, D., Fischer, H., Franke, H., Frieß, U., Heintzenberg, J., Helleis, F., Hermann, M., Kock, H. H., Koeppel, C., Lelieveld, J., Leuenberger, M., Martinsson, B. G., Miemczyk, S., Moret, H. P., Nguyen, H. N., Nyfeler, P., Oram, D., O'Sullivan, D., Penkett, S., Platt, U., Pupek, M., Ramonet, M., Randa, B., Reichelt, M., Rhee, T. S., Rohwer, J., Rosenfeld, K., Scharffe, D., Schlager, H., Schumann, U., Slemr, F., Sprung, D., Stock, P., Thaler, R., Valentino, F., van Velthoven, P., Waibel, A., Wandel, A., Waschitschek, K., Wiedensohler, A., Xueref-Remy, I., Zahn, A., Zech, U., and Ziereis, H.: Civil Aircraft for the regular investigation of the atmosphere based on an instrumented container: The new CARIBIC system, Atmos. Chem. Phys., 7, 4953–4976, https://doi.org/10.5194/acp-7-4953-2007, 2007.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y., and Wagner, T.: The Mainz profile algorithm (MAPA), Atmos. Meas. Tech., 12, 1785–1806, https://doi.org/10.5194/amt-12-1785-2019, 2019.

Kuhn, L., Beirle, S., Kumar, V., Osipov, S., Pozzer, A., Bösch, T., Kumar, R., and Wagner, T.: On the influence of vertical mixing, boundary layer schemes, and temporal emission profiles on tropospheric NO2 in WRF-Chem – comparisons to in situ, satellite, and MAX-DOAS observations, Atmos. Chem. Phys., 24, 185–217, https://doi.org/10.5194/acp-24-185-2024, 2024.