General response to all reviewers

We thank the reviewers for their positive assessment of our work and for providing constructive and thoughtful feedback, which has helped us to improve the paper. We have carefully considered each comment and revised the manuscript accordingly.

Below, we respond to each reviewer and each comment individually. For context, the original reviewer comment is shown in *gray italic text*, followed by our response in black text, and a summary of the corresponding changes to the manuscript in blue text. Unless stated otherwise, all line numbers mentioned in our responses refer to the revised manuscript with tracked changes.

Important note on changes to the results section

A comment by reviewer #2 has helped us identify a mistake in our code that fits the PCA dimensionality reduction of the training targets before neural network training. We have corrected this, retrained the neural networks, rerun all analysis scripts, and updated the results section accordingly. The correction improved the performance of the base emulator and brought the hybrid MLP-PCA performance closer to the hybrid CNN. The revised results give a more accurate picture of the relative merits of the different emulator designs and support the recommendation of both reviewers to weaken our claims that a hybrid approach is strictly necessary. We have adjusted the corresponding statements in the abstract, discussion, and conclusions. Our key findings and conclusions remain unchanged: the hybrid approach is more accurate, requires less training data, and generalizes more robustly than the purely data-driven base emulator.

Responses to reviewer #1

Specific comments

Line 10: I suggest expressing the emulator-induced XCH4 and XCO2 errors also, or perhaps preferably, as relative errors in percent, as these may be easier for readers to assess than the absolute values in ppb and ppm. In addition, reporting values such as 7.22 ppb and 4.25 ppm with three significant digits may suggest a level of precision that is not warranted for these summary error metrics; fewer significant digits might be more appropriate.

Response:

We agree that reporting relative errors will help readers who are unfamiliar with specific XCH4 and XCO2 accuracy requirements to assess our results. We therefore now report emulator-induced errors both as absolute values and as relative errors in percent (relative to the per-sample simulated XCH4/XCO2).

Regarding the number of significant digits, we agree that some of our reported numbers did suggest an unwarranted level of precision. To determine an appropriate number of decimals, we have computed bootstrapping uncertainties on the retrieval test set for all summary error metrics. Based on this analysis, we now decreased the number of decimals to one for all reported absolute XCH4 errors. For absolute XCO2 errors and all relative errors in percent we keep two decimals as the bootstrapping uncertainties are on the order of 0.05 ppm and 0.01% for most models. For correlation coefficients reported in Figure 5a (previously Figure 4a), we have increased the number of decimals to four since the bootstrapping uncertainties (between 0.0001 and 0.008) show that these differences are meaningful and are otherwise hidden by rounding. For comparability, we use the same number of decimals across all models, even though non-scattering and Base MLP-PCA error uncertainties are typically higher than those of the full-physics model and the hybrid emulators.

Changes to the manuscript:

1. Added relative errors in the abstract (Lines 11-15)
2. Adjusted the number of decimals of absolute XCH4 errors to one in the abstract (Lines 11-12 and 14-15)
3. Adjusted the number of decimals of absolute XCH4 errors to one in the results section (Lines 482-484 and Figure 5a)

Lines 155–160: The atmosphere is represented by only 20 homogeneous layers that are equidistant in pressure. This may well be sufficient for the present retrieval setup, but it would be useful to briefly justify this choice, for example by referring to previous sensitivity studies or established RemoTeC/LINTRAN configurations showing that this vertical resolution is adequate for XCH4 and XCO2 retrievals.

Schepers et al. (2014) found that LINTRAN v2.0 reaches 0.5 % accuracy on the observable for a homogeneous atmosphere of unity optical thickness with as few as 10 model layers. We use twice that number to further increase the accuracy and

account for deviations from this idealized scenario. Our choice of K=20 layers lies within the range used in previous RemoTeC retrieval studies: Lu et al. (2022) use 15 layers, whereas Hu et al. (2016) use 36 layers.

Changes to the manuscript:

1. Added justification for choosing K=20 to Section 2.3 (Lines 158-160).

Lines 165–170: The aerosol representation involves several simplifying assumptions, including spherical particles, fixed refractive indices, and a prescribed vertical distribution. These assumptions may be well established for the RemoTeC retrieval framework, but it would be helpful to briefly motivate them here or refer more explicitly to previous studies demonstrating that this simplified aerosol representation is adequate for the present XCH4/XCO2 retrieval application.

These are indeed well-established assumptions used in all RemoTeC retrieval applications. In a simulation study, Butz et al. (2009) found that more complex aerosol models lead to convergence issues, because spectrometer measurements alone do not contain enough information to retrieve a large number of aerosol state vector elements. The proposed approach with three effective aerosol parameters was found to yield a good trade-off between retrieval stability and retrieval error. Hu et al. (2016) arrived at the same conclusion for operational TROPOMI XCH4 retrievals.

Changes to the manuscript:

1. Added motivation for the simplified aerosol model and references to the original papers introducing this model as well as to more recent papers employing it for the GOSAT, GOSAT-2 and TROPOMI missions (Lines 172-175).

Lines 207–219: The logarithmic transformation of the radiances and the PCA-based dimensionality reduction appear to be important and well-motivated design choices for reducing the complexity of the learning problem. Could the authors briefly comment on how important these steps are in practice? For example, did tests without the logarithmic transformation or PCA result in noticeably less stable training or reduced emulator accuracy? A short qualitative statement based on their model-development experience would already be helpful.

During model development, we found that without the logarithmic transformation of training targets, the spectral RMSE increased by up to an order of magnitude. The PCA-based dimensionality reduction is particularly useful for smaller training datasets as it helps the model focus on the relevant spectral features in the high-dimensional output space. For larger datasets, a pure MLP without pre-computed dimensionality reduction can in principle learn the same function and thus achieve the same accuracy. However, we find that it still helps stabilize the training and reduce sensitivity to hyperparameter choices and weight initialization.

Changes to the manuscript:

1. Added statement on the effect of the logarithmic transformation of training targets to Section 2.5.1 (Lines 216-217).

2. Added a qualifying statement on the practical importance of using the PCA dimensionality reduction of training targets to Section 2.5.1 (Lines 236-237).

Lines 305–310: Reducing the 20-layer temperature profiles to only two principal components is an efficient way to limit the dimensionality of the neural network input. At the same time, this may remove finer-scale vertical temperature structure. Could the authors briefly comment on whether such information is expected to be relevant for the Sentinel-5 retrieval geometry considered here, and whether two principal components were found to be sufficient in sensitivity tests or previous retrieval studies?

We agree that using only two components will remove some of the finer-scale vertical features of temperature profiles, however the first two components already explain more than 92% of the variance and we have found during model development that using more than 2 components did not yield any significant improvement in emulator accuracy on the validation set. We have included the results of this sensitivity study in the appendix and refer to it in the main text.

Changes to the manuscript:

1. Included information on the explained variance of two principal components in Section 2.6.2 (Line 317).
2. Added a brief summary of our sensitivity study to Section 2.6.2 and added a reference to the corresponding results presented in Figure A1 in Appendix A (Lines 319-320).

Lines 430–435: Since the Jacobian evaluation is an important part of demonstrating that the emulators can be used reliably within the iterative retrieval, I suggest moving one representative Jacobian figure from Appendix A into the main text. In addition, Figures A1–A3 show noticeably larger deviations for the aerosol central height (z_aer). The manuscript notes that this parameter also has the lowest instrument sensitivity; could the authors briefly discuss whether the reduced Jacobian agreement is mainly a consequence of this weak sensitivity and whether it has any noticeable impact on the retrieval performance?

We agree and have moved the Jacobian analysis for the Hybrid MLP-PCA emulator as a representative figure to the main text.

The reduced Jacobian agreement for z_aer can indeed be attributed to the low instrument sensitivity to this parameter, which makes it harder for the neural networks to learn the subtle effects of changes in aerosol central height when trained on radiance spectra alone. Importantly, the same weak sensitivity makes parameters like aerosol central height difficult to retrieve accurately, even when using the full-physics forward model (see Figure C1 in Appendix C), because it is poorly constrained by the measurements and thus strongly dependent on regularization. Errors in the associated derivates are therefore expected to have only small effects on the emulator-induced retrieval errors. To get an upper bound on this effect, we have performed additional retrieval experiments where we use the Base MLP-PCA emulator as forward model, but replace the associated network Jacobians with ground-truth full-physics Jacobians. This leads to only a moderate RMSE improvement of 1.0 km for z_aer, and 0.93 ppb for XCH4. Since these are improvements due to the combined Jacobian errors of all state vector elements, the effect of errors in z_aer Jacobians alone is expected to be much smaller than that. Nevertheless, we do acknowledge that there is room to further improve emulator-

based retrievals by improving their Jacobians, e.g. in a finetuning step where Jacobian errors are included in the loss function during training.

Changes to the manuscript:

1. Moved the Jacobian evaluation figure for the Hybrid MLP-PCA emulator to the main text (now Figure 4) and adjusted the Figure references in Section 3.1.3 accordingly (Lines 453-454).
2. Added an explanation for the low derivative agreement for parameters with low instrument sensitivity and a discussion of the effect this has on retrieval performance (Lines 456-464).
3. Added the low derivative agreement for z_aer as a motivation for the discussion on including an additional Jacobian loss term to further improve emulator performance in the future (Lines 595-600).

Line 470: The runtime comparison is based on measurements on a single CPU core. It may be useful to briefly note that the reported relative speed-ups are implementation- and hardware-dependent, in particular since neural network inference may benefit considerably from GPU acceleration or batched processing.

Good point. Since we already discuss the potential of GPU acceleration, and more efficient implementations of neural network Jacobians and non-scattering calculations at the end of Section 3.2.2, we have modified this paragraph slightly to include the potential of batched processing, and to make it explicit that all reported speed-ups depend on both hardware and specific implementation.

Changes to the manuscript:

1. Added an explicit statement highlighting that all reported speed-ups are hardware- and implementation-dependent (Lines 524-525).
2. Added batched processing to the list of potential improvements leading to further acceleration of neural network emulators (Line 525).

Lines 538–540: The suggestion that emulator-based retrievals could ultimately exceed the accuracy of operational full-physics retrievals is interesting, but appears somewhat speculative based on the present results. The authors should qualify the statement as a possible future perspective.

We have adjusted the phrasing accordingly.

Changes to the manuscript:

1. Changed the sentence opening to reflect a possible future perspective ("In the future, this may open up opportunities […]"), rather than a proven statement (Lines 574-575).

Lines 555–560: Given that the present evaluation is based entirely on synthetic measurements, I suggest slightly qualifying the statement that the hybrid emulators are candidates for implementation in operational Sentinel-5 processing. The authors appropriately discuss the need for evaluation with real measurements later in this section, so a small wording adjustment here would make the distinction between demonstrated

feasibility and operational readiness clearer. Also, "this emulators" should read "these emulators".

We have adjusted our phrasing in this paragraph to weaken our claim that the hybrid emulators are ready for implementation in operational processing and instead highlight already in this paragraph that so far our emulators have only been evaluated in model-consistent synthetic experiments.

Changes to the manuscript:

1. Changed the phrasing "are promising candidates for implementation in operational Sentinel-5 processing pipelines" to "are promising candidates to consider for future implementation in operational Sentinel-5 processing" (Line 602).
2. Made it explicit that the emulator-induced retrieval errors discussed here are based on model-consistent synthetic experiments (Lines 604-606).
3. Changed "this emulators" to "the hybrid MLP-PCA emulator" (Lines 604-605).

Technical corrections

Line 160: "reach the required accuracy required": please remove the duplicated "required".

Thank you for pointing out this issue.
We have corrected this in the revised manuscript (Line 165).

Line 525: A full stop is missing after "training distribution".

Thank you for pointing out this issue.
We have corrected this in the revised manuscript (Line 561).

Figure B1 caption: "aerosol layer hight" should read "aerosol layer height".

Thank you for pointing out this issue.
We have corrected this in the revised manuscript (now Figure C1 caption).

Responses to reviewer #2

1. The main drawback is that the tests (although performed in the retrieval framework) are based on synthetic data. There are several simplifications adopted by the authors (known noise model, aerosol parameters, cloud free cases, Lambertian albedo etc). In practice, we are dealing with much larger uncertainties. I do not think that an evaluation with real S5p data is a must for this paper, but at least it should be seen in the title and throughout the paper that we are dealing with "a model-consistent synthetic experiment". Also claims about operational implementation should be formulated more carefully.

We deliberately chose a model-consistent synthetic setup, because the goal of this study is to assess the emulators as function approximators of the forward model. Applying an operational retrieval algorithm to synthetic data allows us to cleanly isolate emulator-induced retrieval errors from all other error sources (as already explained in the Introduction, Lines 86-88). However, we agree that the simplifying assumptions of this setup should be made more explicit throughout our manuscript, and that evaluating the emulators on real measurements is an essential next step (as already discussed in the Discussion, Lines 620-626). We have therefore revised the abstract and discussion to make it very clear that we are performing model-consistent experiments and have reformulated our statements on operational implementation as future prospects. We would like to keep the title as is to highlight the approach and its potential rather than focusing on the specifics of our experimental setup, which is now explicitly described in the abstract.

Changes to the manuscript:

1. Added "in a model-consistent experiment" to the first sentence in the abstract summarizing our study (Line 4-6).
2. Added "from synthetic measurements" to the sentence explaining our model evaluation in the abstract (Line 9).
3. Changed "the application to operational Sentinel-5 retrievals" to "the potential for future applications to operational Sentinel-5 retrievals" in the opening paragraph of the Discussion (Line 552).
4. Changed the phrasing "are promising candidates for implementation in operational Sentinel-5 processing pipelines" to "are promising candidates to consider for future implementation in operational Sentinel-5 processing" (Line 602).
5. Made it explicit that the emulator-induced retrieval errors discussed here are based on model-consistent synthetic experiments (Lines 604-606).

2. We see that the base emulator reaches a plateau at large training-set sizes. And it is interpreted as a saturation of the network capacity (line 415). If the base emulator capacity is limited, part of the large difference between the base and hybrid approaches could result from the chosen architecture rather than the hybrid formulation itself. So far we see that the hybrid representation makes the learning problem much easier (what we can expect), but the stronger conclusion that a base approach cannot reach the required accuracy cannot be made. Therefore I suggest adding a control experiment with a substantially larger base MLP, at least for the full training set in order to check whether the remaining performance

gap persists when network capacity is increased. Also the conclusions regarding the fundamental limitations of "end-to-end" emulation should be softened.

Thank you for pointing out this potential explanation for the relatively poor performance of the base emulator. Through experiments with larger neural networks, we have identified a mistake in our code that fits the PCA dimensionality reduction of the training targets before neural network training (see also the general note to all reviewers at the beginning of this document). This bug has led to a sub-optimal constraint in the MLP-PCA training and caused the plateau in the data-efficiency plot. After fixing this, the RMSE of the base emulator now continues to decrease with larger training datasets just like for the hybrid emulators, leaving the constant offset and the performance for the smallest dataset size as the key differences between base and hybrid emulators. Based on these updated results, we agree that the claims about the fundamental limitations of end-to-end emulators need to be softened, and we have revised the abstract, results and discussion section accordingly.

Changes to the manuscript:

1. Changed "a hybrid approach is needed" to "a hybrid approach is much better suited" in the abstract (Line 10).
2. Adjusted the data-efficiency results in Section 3.1.2 to reflect the fixed behavior of RMSE and absolute bias as a function of training data size (Line 423-446).
3. Changed "the application to operational Sentinel-5 retrievals" to "the potential for future applications to operational Sentinel-5 retrievals" in the opening paragraph of the Discussion (Line 552).
4. Modified the first sentence of Section 4.1 to include the qualifying statement that the end-to-end emulators require large training datasets to achieve reasonable accuracy (Line 557).
5. Changed "from training examples alone" to "from a limited number of training examples" to reflect the fact that the end-to-end emulator is not unable to reach the required accuracy but simply needs more training data to do so (Lines 559-560).
6. Changed "This highlights the need for more advanced hybrid approaches" to "This highlights the potential of more advanced hybrid approaches" (Line 564).
7. Adjusted the paragraph on the accuracy-speed trade-off to consider when choosing the emulator design to reflect the updated results (Lines 578-590).
8. Removed "becomes feasible when physical knowledge is incorporated into the emulator design through a hybrid approach" in the opening sentence of the final conclusions (Lines 626-627). Instead we start with the broad statement that accelerating GHG retrieals with neural network emulators is challenging but feasible, followed by a discussion of the value of our hybrid approach compared to the end-to-end approach.

3. The fact that usable Jacobians emerge from networks trained only on radiances is an interesting result. At present, however, the operational implementation evaluates the derivatives by finite differences because automatic differentiation is not available through FTorch. The finite-difference steps were selected by comparison with PyTorch derivatives. I suggest providing a quantitative measure of the finite-difference error relative to the

automatic-differentiation result, rather than stating only that it is negligible. Also, correlation coefficients alone can be misleading for derivatives with small amplitudes.

Good point, we have now added more detailed information on the accuracy of the finite difference approximation and use the bias and RMSE relative to the absolute magnitude of analytical derivatives in addition to the correlation coefficient to evaluate the accuracy of emulator Jacobians.
We find that for all state vector elements, finite difference approximations result in errors well below 0.1%, which are generally about an order of magnitude smaller than errors between the finite difference approximation and full-physics Jacobians.

Changes to the manuscript:

1. Included the exact step sizes and associated finite difference RMSE errors in Table B1 of Appendix B.
2. Adjusted the sentence on numerical approximation errors in Section 2.8.2 to provide more specific information on the magnitude of these errors (Line 394).
3. Added bias and RMSE relative to the absolute magnitude of analytical derivatives as additional evaluation metrics to all Jacobian figures (Figure 4 and Figures B1-B2).
4. Included information on bias and RMSE to Section 3.1.3 in addition to correlation coefficients (Lines 466-467).

4. As a continuation of point 3, a discussion about the possibility of adding into a loss term an L2 discrepancy of Jacobians would be very beneficial.

As we have already mentioned the potential of further fine-tuning the neural network emulators by jointly optimizing radiance errors and associated derivatives in Section 4.1, we have reformulated this paragraph to improve clarity and make the possibility of adding a Jacobian loss term more explicit.

Changes to the manuscript:

1. Added the low derivative agreement for z_aer as a motivation for the discussion on including an additional Jacobian loss term to further improve emulator performance in the future (Lines 595-597).
2. Modified the phrasing "a fine-tuning step in which the trained networks are further optimized to reduce errors in both radiances and their associated derivatives" to explicitly mention the addition of a loss term penalizing the discrepancy between emulator and full-physics Jacobians (Lines 597-600).

5. The high-emission experiment scales the entire CH4 reference profile. Since the paper motivates the method partly through rapid detection of emission hotspots, it would be useful to mention that real enhancements can be vertically localized and do not necessarily correspond to uniform scaling of the reference profile. I do not consider a new experiment essential, but this limitation should be acknowledged.

This indeed presents a limitation of our extrapolation experiment and we agree that it should be made explicit that our synthetic scenes are not representative of real emission plumes.

Changes to the manuscript:

1. Added a clarifying sentence to Section 2.6.1 that explicitly highlights the limitations of the extrapolation experiments due to the difference to typically vertically localized real plumes (Lines 309-310).

6. In introduction authors mention principal component analysis (PCA) based compression schemes. In the given references two different concepts of PCA models are presented (one is for output radiances, another is for input parameters), hence it is not correct to unite them into a single concept of PCA.

Thank you for pointing out this inconsistency.
We have removed the Efremenko reference to only focus on output dimensionality reduction, which is most relevant for our study (Line 49).

7. Some justification of the chosen architecture of NN should be given.

In addition to the more general motivation for our neural network architectures that is already given in Sections 2.5.1 and 2.5.2, we have now added more detailed results from our hyperparameter search (as explained in Section 2.7.2) to the Appendix, which demonstrates how the emulator performance changes as hyperparameters are varied and provides evidence for our final architectural choices.

Changes to the manuscript:

1. Added Figures A1 and A2 to the Appendix, which show (a selection of) our hyperparameter search for the base MLP-PCA and the hybrid CNN. These results provide the bases for our final choices on network architectures and the peak learning rate used for training.
2. Added a reference to these results in Section 2.7.2 (Lines 358-560).

8. The idea of using an approximate model without scattering as a precursor is nice. I suggest to add a comparison of the approximate model with full RT simulations to see the amplitude of differences between them and the effect of the NN correction for the spectral radiances.

Good point.
We have extended Figure 2 to also include the radiance residuals for the non-scattering approximation.