Final response to referee comments on egusphere-2025-5422

First of all, we would like to thank reviewer #2 for the constructive comments, which helped to improve the manuscript. Below we give answers and clarifications to all comments made by the referee (repeated in italics).

Anonymous Referee #2

Main comments

Reviewer: *1. Motivation for replacing the Random Forest classifier*

The manuscript describes the implementation of the XGBoost classifier in detail, but the motivation for replacing the previous Random Forest (RF) classifier could be stated more explicitly. While XGBoost is introduced as efficient and potentially higher performing, it is not entirely clear what concrete limitations of the RF-based filter prompted this transition.

A short summary of the main shortcomings of the previous RF approach (e.g. overly conservative filtering in certain regimes, misclassification behaviour, computational aspects, etc.) would help frame the update more clearly in terms of scientific added value.

Authors: The replacement of the Random Forest (RF)-based quality filter by an XGBoost implementation was motivated primarily by practical limitations in model size, memory consumption, and better extensibility to episodic and sporadic events. Although the RF classifier used in v1.8 delivered robust classification performance with minimal tuning, it proved costly in operational terms: the trained model occupied 73 GB on disk, and model loading was slow and memory-intensive, restricting portability and routine use. These constraints become more severe as additional physical effects are incorporated, for example in the targeted extension of the quality filter in v2.0 to explicitly address aerosol-related contamination over bright surfaces.

In the RF framework, accounting for additional effects typically requires a larger and more diverse ensemble, which further inflates model size and eventually becomes impractical. Moreover, rare but physically consistent quality-degrading conditions are not always well represented, as the ensemble averaging inherent to RF can dilute their influence. By contrast, a carefully tuned XGBoost model achieves comparable or improved classification performance with substantially lower computational and memory requirements. Its boosting strategy places greater weight on informative but infrequent training samples, allowing later trees to focus specifically on such cases. As a result, the XGBoost model implemented in v2.0 has a file size of only 236 MB, enabling efficient operational use without loss of classification skill. XGBoost thus addresses practical and conceptual limitations of the RF approach in this specific application.

This motivation is reflected more explicitly in Section 2.2 of the revised text:

“In earlier versions of the product, quality filtering was accomplished using a Random Forest classifier . . . . For a global quality filtering task that spans a wide range of surface types, atmospheric states, and illumination conditions, achieving sufficient ensemble diversity quickly

drives up model size and memory demands, which limits extensibility to additional effects. In particular, infrequent but physically consistent quality-degrading conditions, such as specific aerosol events over bright surfaces targeted in the updated product version, may not be optimally captured by a Random Forest classifier, as its ensemble-averaging strategy risks diluting such sporadic signals.

To address this limitation, the latest product TROPOMI/WFMD v2.0 uses Extreme Gradient Boosting (XGBoost), providing a more compact and computationally efficient framework for comprehensive quality filtering. . . . "

Reviewer: *2. Choice of classification threshold ($p_0 = 0.5$)*

The decision threshold is set to $p_0 = 0.5$. While this is a common default in binary classification, in the context of atmospheric retrieval filtering the two types of misclassification do not have equal consequences. In particular, retaining cloud-affected scenes may introduce systematic biases in the retrieved columns, whereas rejecting valid scenes mainly reduces data yield. It would therefore be helpful to clarify whether alternative thresholds were evaluated and whether the selected value was assessed with respect to downstream geophysical performance (e.g. bias and scatter relative to TCCON).

Authors: As illustrated in Figure 3, the parameter p_0 controls the trade-off between data yield and the risk of accepting contaminated scenes. Retaining cloudy scenes is undesirable, yet an excessively strict filter that rejects nearly all scenes is equally impractical. We therefore select the standard value $p_0 = 0.5$ as a neutral decision threshold, corresponding to equal costs for false positives (cloud-contaminated scenes accepted) and false negatives (good scenes rejected).

For a quality filter, one may reasonably argue that false positives should be penalised more strongly. Doing so, however, requires an explicit definition of how much higher these costs should be. In the absence of a well-defined cost function, there is no uniquely optimal choice of p_0 when only a binary quality label is provided. The selected threshold represents a balanced and transparent default that is straightforward to interpret and reproduce. On this basis, alternative thresholds were not explored in downstream performance assessments, such as the validation against TCCON. Importantly, the TCCON comparison already demonstrates good performance for the standard choice $p_0 = 0.5$.

An alternative design would be to provide the continuous value of p_0 in the product instead of a binary quality flag, thereby allowing user-defined thresholding. This approach would, however, transfer the responsibility for defining an appropriate cost function to the user on a case-by-case basis and increase the risk of inconsistent or unintended use. Experience suggests that many users prioritise data volume, which may encourage overly permissive threshold adjustments at the expense of data quality.

For these reasons, the product provides a binary quality flag derived from the fixed threshold $p_0 = 0.5$, favouring simplicity and consistent data quality over user-controlled tuning. This rationale is now stated explicitly in the revised text:

"In the current product, this threshold is fixed by design and only the resulting binary quality flag is provided, prioritising simplicity and consistent data quality over enabling user-controlled adjustment of the decision threshold."

Reviewer: *3. Representativeness of the training dataset*

The classifier is trained on 38 randomly selected days from 2020–2021. Although an independent year is used for validation, it would be helpful to provide some additional information on the representativeness of this training sample. In particular, a brief summary of:

the seasonal distribution of the selected days, their geographical coverage (including high latitudes), and whether specific regimes such as strong aerosol or dust events are included,

would help the reader better assess the generalisation capability of the model.

Authors: The training data set provides daily global coverage with uniform seasonal sampling, including at least nine days per season. This design ensures that all physically observable combinations of latitude and solar zenith angle are represented, supporting robust generalisation of the trained model to operational data under daylight conditions. A remaining potential limitation is the representation of rare and short-lived transient phenomena, which may be sparsely sampled. When such cases are present, however, the boosting approach is well suited to exploit their diagnostic value. It was also verified that the training data also include aerosol events, although their inclusion was not an explicit selection criterion.

This clarification is reflected in the revised wording:

"The XGBoost classifier was trained using data from 38 randomly selected days distributed across 2020 and 2021, with the additional requirement that all seasons are adequately represented. This sampling strategy covers all physically observable combinations of latitude and solar zenith angle and captures a broad range of atmospheric conditions, while keeping the overall data set size manageable."

Reviewer: *4. Transparency of the feature set and model generalisation*

Since the quality filter is the central methodological component of this paper, the complete list of input features should be explicitly provided in the manuscript, even if inherited from earlier versions. I understand that the feature set is described in previous publications; however, for clarity and to keep the flow of the manuscript self-contained, it would be helpful to include the complete list of input features directly in the present paper.

In this regard: some features (e.g. surface altitude, lat, lon etc.) may correlate with climatological cloudiness or regional characteristics. It would therefore be useful to briefly discuss how the model avoids learning region-specific patterns rather than physically meaningful indicators of retrieval quality. A short discussion of generalisation across regimes would strengthen confidence in the robustness of the approach. An analysis of feature importance could be useful in this context.

Authors: A new subsection has been added to describe the feature set, present an analysis of feature importance using SHAP (SHapley Additive exPlanations), and motivate the inclusion of geospatial context. This includes a dedicated sensitivity experiment in which geographical coordinates were excluded from the training, allowing a direct assessment of their impact. The analysis further identifies the regimes in which geospatial information contributes most strongly to the classification. Including geolocation improves filtering performance, increasing precision and recall for accepted scenes, both globally and in the regimes of highest impact.

These regions are not characterised by enhanced climatological cloudiness. Instead, the strongest impact of geographic context is found over areas with above-average surface bright-

ness and below-average cloudiness. In such regimes, radiance-based cloud indicators become less discriminative, as optically thick clouds are also bright in the SWIR. Additional contextual information is therefore helpful to resolve quality ambiguities, an effect that may be further amplified in the presence of aerosol-related scattering. A related discussion is also provided in the response to Referee #1.

Additional comment(s)

Reviewer: *Attribution of improvements in v2.0: Version 2.0 introduces several updates simultaneously (spectral window adjustments, hybrid vertical grid, post-processing refinements, and the new XGBoost classifier). While the validation results clearly show improvements relative to v1.8, it would be helpful to briefly clarify to what extent these gains can be attributed specifically to the updated quality filter versus the retrieval physics changes.*

Authors: As noted in the abstract, the observed improvements are mainly attributable to the refined quality filtering. The changes in the retrieval itself play only a minor role. This point is clarified further in Section 3.5 of the revised manuscript:

"The improved performance relative to the previous version is driven primarily by the updated quality filtering, whereas the additional processing changes have only a minor effect in comparison."