Bremen, March 26, 2026

Letter to the Editor of egusphere-2025-5422

Dear Zhao-Cheng Zeng,

On behalf of all co-authors I have prepared this document, which provides the point-by-point responses to the reviews. The corresponding changes in the revised manuscript are highlighted in the attached track-changes file. In particular, there is a new subsection about the feature set and their importances, including two new figures.

Best regards,

Oliver

Final response to referee comments on egusphere-2025-5422

First of all, we would like to thank both reviewers for their constructive comments, which helped to improve the manuscript. Below we give answers and clarifications to all comments made by the referees (repeated in italics).

Anonymous Referee #1

Major comment

Reviewer: *Information and analysis regarding the features used as input to the XGBoost algorithm are missing, even though this is a key component of the work. The manuscript refers to Schneising et al. (2023) for the feature set, which in turn references Schneising et al. (2019), where a list of 25 features is provided in Section 2.5.2 Quality filter.*

I understand that in previous studies the filtering was one of several methodological improvements and therefore described in less detail. However, in the present manuscript the quality filter is the central focus, and the specific input features should be explicitly listed and described to ensure reproducibility.

The manuscript should include a list of the features used as the input to the algorithm, along with their definitions and a brief justification of how each feature relates to cloud-contaminated scenes.

It would also be very helpful to include an analysis of feature importance (e.g., a plot analogous to Figure 2 of Keely et al., 2023). One option is to use the SHAP (SHapley Additive exPlanations) package to quantify the marginal contribution of each feature in the XGBoost model (see the SHAP documentation: https://shap.readthedocs.io/en/latest/).

Authors: We agree with the reviewer. The revised manuscript now includes a dedicated subsection describing the feature set and assessing feature importance using SHAP analysis.

Minor comments

Reviewer: *On line 139, the manuscript states that 26 features are used. Is this a typo? In Schneising et al. (2019), a total of 25 features are listed.*

Authors: In v1.5, surface roughness was introduced as an additional feature (see Table 1 of Schneising et al. (2023)), increasing the total number of features to 26.

Reviewer: *Assuming that the features are exactly the same as those used in Schneising et al. (2019), many of them appear to be reasonable proxies for cloud-related information. However, it is unclear why latitude, longitude, and altitude are included. These variables could bias the algorithm toward regions that are climatologically cloudy, rather than toward features that directly indicate cloud-contaminated retrievals. If these variables are still included in the current feature set, please justify their use.*

Authors: Latitude, longitude, and altitude are not used as climatological cloud masks. Instead, they serve as contextual information that helps the classifier interpret physically motivated cloud indicators in a more flexible manner. Their inclusion relaxes the assumption of globally fixed decision thresholds and allows the model to adapt its decision boundaries. Most of this adaptation is already driven by physical features describing surface properties, atmospheric state, retrieval diagnostics, and viewing geometry. Geospatial context can further refine the decision in special cases where the other parameters alone do not unambigiously capture the relationship between cloud proxies and actual cloud contamination. Altitude is interpreted as a surface property rather than as geospatial context, as it directly affects surface-atmosphere interactions, radiative transfer, and cloud occurrence. It is also important to note that although explicit cloud and aerosol information is used to define the quality labels during training, such information is not available during operational processing. Geolocation therefore offers always-available contextual input that, in combination with other features, can partly compensate for this missing information, without replacing the physical cloud proxies.

That geolocation acts as secondary context rather than as a driver of climatological cloud filtering is supported by two complementary analyses added in the revised manuscript: a SHAP-based feature importance assessment and a sensitivity experiment in which geographic coordinates were excluded from model training, while everything else, including the validation data set, was retained.

The SHAP analysis shows that latitude, longitude, and altitude are of minor importance compared to the direct cloud proxies, ranking 11th, 18th, and 21st among all features. Latitude and longitude together account for about 6% of the total mean absolute SHAP importance. The contributions are dominated by cloud proxies (51%), followed by surface properties (17%, including 1% from altitude), atmospheric state (11%), retrieval diagnostics (10%), and viewing geometry (5%). A corresponding discussion and figure is added to the revised version of the manuscript. Even in regions where geolocation has its strongest impact, physical cloud proxies remain dominant (45%), while the contribution of geographic context stays near 6%, confirming its role as auxiliary information.

These **R**egions of **H**igh **I**mpact (RHI) are identified by retraining the XGBoost model without latitude and longitude as features and quantifying changes in scene acceptance on the validation data set. Spatially connected clusters in the upper and lower tails of the acceptance-change distribution are denoted RHI^+ and RHI^-, indicating regions where geospatial context leads to the largest net increase or decrease in accepted scenes. Their union defines the RHI subset, comprising approximately 3% of all scenes. Including geolocation improves global filtering performance, increasing precision and recall for accepted scenes by about 0.8 and 0.4 percentage points, respectively. These gains are concentrated within the RHI, where precision and recall improve by approximately 1.7 and 2.6 percentage points. At the same time, the inclusion of geolocation features reduces the class-0 precision contrast between RHI^+ and RHI^- by 66%, yielding a more spatially uniform purity of accepted data and improved regional consistency. A map showing changes in scene acceptance and highlighting RHI^+ and RHI^- is added as an additional figure to the revised version.

The RHI are characterised by lower mean cloudiness and higher surface brightness than non-RHI regions ($\Delta \mathrm{H_2O}$: 0.2 versus 1.2, cloud parameter: 1.2 versus 2.0, apparent albedo: 0.33 versus 0.19). This pattern demonstrates that the classifier does not learn a cloud climatology from geographic coordinates. Instead, geographic information matters most for predomi-

nantly clear-sky scenes over bright surfaces, where radiance-based cloud indicators are less discriminative because optically thick clouds are also bright in the SWIR. In such regimes, geospatial context acts as a disambiguating modifier that, in combination with other features, complements the physically motivated cloud proxies rather than serving as a shortcut for climatological cloud masking. This entire discussion is added to the revised manuscript.

Reviewer: *Did the authors consider eliminating existing features or introducing new features to improve the filtering performance of the updated algorithm? If so, please describe this process and its impact on performance. If not, could this be an area of improvement for the next version which could be discussed in the conclusion?*

Authors: For tree-based ensemble methods such as XGBoost, performance gains from removing input variables are generally uncommon, as features that carry little information are rarely selected for tree splits. As a result, adding weakly informative variables typically leaves performance unchanged or leads to small improvements, unless a feature introduces systematic bias or promotes overfitting. Both risks are mitigated here through representative training data and independent validation.

This behaviour is confirmed by the sensitivity experiment described above and added to the revised version, in which latitude and longitude were removed from the feature set. Their exclusion leads to a consistent reduction in both precision and recall, observed globally and within the subset of regions where filtering decisions differ most strongly.

Features that are not beneficial tend to be neutral rather than detrimental. Given the high reliability of the externally sourced reference labels used for training, the current XGBoost model therefore includes all upstream variables that are potentially informative, consistently available at inference time, and not strongly redundant. Further improvements in the future may come from introducing additional quality indicators, such as metrics derived from simultaneous sub-scene variability in surface albedo and surface elevation. Implementing such features would, however, require a high-resolution SWIR surface reflectance climatology, which was not available for the present product version.

Reviewer: *The filtering approach focuses specifically on cloud contamination, are there other retrieval limitations that could introduce biases and should therefore be considered in the filtering process? For example, could sub-footprint altitude variability impact retrieval quality? Accurately modeling surface pressure in the presence of significant altitude variations within a footprint may be particularly challenging and could affect the reliability of the retrieval.*

Authors: The current XGBoost-based quality filter is designed primarily to identify cloud and aerosol contamination. Other sources of retrieval degradation can be envisaged, but incorporating them into a machine-learning quality filter would require a well-defined reference truth for the corresponding binary good/bad classification, which is often difficult to establish.

With respect to sub-scene altitude variability, the retrieval algorithm already provides a scene-level surface roughness metric derived from a high-resolution digital elevation model. This parameter is included as an input feature to the quality filter. Its influence is currently limited, reflecting the absence of a robust and systematic criterion that links surface roughness to degraded retrieval quality. A conservatively chosen simple threshold would risk rejecting otherwise usable scenes. A combined use of surface roughness and sub-scene albedo variability, as suggested above, may offer a promising systematic route for future extensions of quality

filtering. However, such an approach is challenging in practice, as albedo exhibits seasonal variability and would need to be available at high spatial resolution.

Similar considerations apply to other potential retrieval limitations. This motivated the addition of residual-based and spatial-consistency filtering applied after the machine-learning quality prediction. This extra step in quality control allows the exclusion of certain anomalous scenes without requiring explicit knowledge and modelling of the underlying physical causes.

Reviewer: *The absolute biases at North American TCCON sites are larger than at most other sites (e.g., Eureka ∼13 ppb, ETL ∼5 ppb, Park Falls ∼8 ppb, Lamont ∼7 ppb, Edwards ∼2 ppb, and Caltech ∼4 ppb). In contrast, most European sites show absolute biases below ∼2 ppb, with Garmisch being the exception. Can the authors comment on the reasons for this discrepancy?*

Authors: One possible explanation for the discrepancy between North American and European sites could be representation error. Variations in spatial heterogeneity and horizontal gradients may cause satellite and TCCON observations to sample different air masses, even when nominally collocated, since the applied collocation radius encompasses a finite non-negligible area. Compared to many European TCCON sites, North American sites are more often located in regions with strong real gradients driven by localised sources. These include wildfire events in East Trout Lake, Park Falls, and Caltech. Caltech is also exposed to urban emissions, while Lamont is sometimes downwind of emissions from oil and gas infrastructure. In addition, some North American TCCON sites are situated in complex terrain, which could further enhance variability within the collocation region. This holds true for Caltech and, in particular, for Eureka, where the influence of the polar vortex is also a factor. In Europe, it is mainly Garmisch that fits into the category of complex terrain. Consequently, the observed differences may, at least in part, reflect genuine atmospheric variability on spatial scales that are not fully resolved in the satellite-TCCON comparison, rather than issues in either measurement system.

Reviewer: *Some TCCON sites are impacted by wildfires in the summer months (can be clearly seen at ETL, Park Falls etc.). Even with the tight coincident criteria there is a good chance one instrument might be impacted by wildfire emissions and the other not. Maybe some additional filtering can be done to mitigate impacts, perhaps filtering of some summer months during years where fires were particularly strong.*

Authors: Potential representation errors associated with wildfires and polar vortex boundaries are already acknowledged in the manuscript. Introducing additional filtering to address these cases is considered impractical, as it would remove substantial portions of the validation data, including summer periods affected by wildfires and winter periods influenced by the polar vortex. Instead, the revised version clarifies the interpretation of the derived performance metrics, explicitly treating them as upper limits that reflect both representation errors and uncertainties in the TCCON reference. This clarification has been added to Section 3.5:

"The reported values should be interpreted as upper limits, reflecting uncertainties in the TCCON reference as well as potential representation errors arising from the non-zero collocation radius, particularly when one instrument is affected by local events such as wildfires or the polar vortex and the other is not."

Reviewer: *Typo: line 402 measurements from the TCCON should be measurements from TCCON*

Authors: We think both are correct, as 'the' refers to 'Network': "measurements from the Total Carbon Column Observing Network" → "measurements from the TCCON"

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 brightness 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."