Final response to referee comments on egusphere-2025-5422

First of all, we would like to thank reviewer #1 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 #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"