Reviewer 1

Review on “Direct-sun versus Sky-Scan Pandora Formaldehyde Retrievals: Implications for OMI Validation in Tropical Southeast Asia”

Comment 1.1

General comment:

This manuscript investigates the differences between Direct-sun (DS) and Sky-scan (SS) Pandora HCHO retrievals and their implications for OMI validation over tropical Southeast Asia. The topic is relevant for satellite validation over the study domain; however, in its current form the manuscript lacks a clear motivation and several methodological choices are not sufficiently justified. The comparison between DS and SS observations is not rigorously addressed, and the satellite–Pandora comparison raises concerns regarding timing consistency, spatial representativeness, and robustness of the statistics. While the use of OMI is justified by its long-term data record (while only one year of OMI data is used in the analysis), many of the Pandora observations analyzed fall within the operational period of newer Geo-satellite instruments. Therefore, the use of GEMS, which was a key motivation for establishing the PAN-Asia Pandora network, would be highly recommended and more appropriate for the study region. In addition, several figures appear to contain unrealistic values, suggesting insufficient data screening and quality control. As a result, key conclusions presented in the abstract, such as the claim that Sky-scan retrievals systematically perform better than Direct-sun observations and that DS HCHO columns are strongly influenced by episodic enhancements, are not convincingly supported by the analysis. Substantial revisions are therefore required to clarify the methodology, improve data screening, and provide robust evidence for the conclusions.

Response 1.1

We thank the reviewer for this comprehensive assessment and fully agree that substantial improvements were required in the original manuscript. In response, the study has been fundamentally revised. First, we clarified the scientific motivation by explicitly framing the analysis around the roles of retrieval geometry, temporal sampling, and spatial representativeness in satellite validation over tropical environments. Second, we implemented an uncertainty-based quality control following Rawat et al. (2025), which removes unrealistic retrievals and ensures that all analyses are based on physically consistent datasets. Third, the DS–SS comparison is now performed using temporally matched pairs (±5 min) with quality-stratified evaluation, providing a more rigorous assessment of their consistency. Fourth, the satellite analysis has been redesigned using a time-based collocation framework and expanded from OMI-only to include TROPOMI and GEMS, addressing concerns regarding temporal consistency and regional relevance. These revisions resolve the issues related to data screening, representativeness, and statistical robustness. The key conclusions are now supported by the improved

methodology, with sky-scan retrievals showing more stable agreement with satellite observations, while direct-sun retrievals exhibit larger variability associated with localized variabilities.

Revised Text

Section 1 (Introduction – final paragraph)

"In this study, we present a comprehensive evaluation of Pandora HCHO observations across Southeast Asia, explicitly distinguishing between Direct-sun and Sky-scan retrievals and assessing their consistency with multiple satellite products (OMI, TROPOMI, and GEMS). By applying an uncertainty-based quality-control framework and a unified temporal collocation strategy, this work aims to quantify how retrieval geometry, temporal sampling, and spatial representativeness jointly influence satellite–ground agreement in tropical environments."

Section 2.1.1 (Quality Control)

"To improve the robustness of ground-based HCHO observations used for intercomparison and satellite validation, an uncertainty-based quality control (QC) protocol following the methodological framework of Rawat et al. (2025) was applied to contemporaneous Pandora direct-sun (DS) and sky-scan (SS) observations. DS and SS retrievals were first paired within a 5 min tolerance window. A high-quality reference subset was then defined using Pandora quality flags QF = 0 or 10 for both DS and SS retrievals, and dynamic absolute uncertainty thresholds were calculated separately for DS and SS as the mean plus three standard deviations of the uncertainty in this subset. Matched observations were retained when either both DS and SS absolute uncertainties were below these dynamic thresholds or both relative uncertainties were below 10 %. Additional filters required WRMS < 0.01 for both DS and SS retrievals and, for sky-scan observations, maximum horizontal distance (MHxD) < 20 km when available. Pandora quality flags were subsequently used to classify observations into high-quality (QF = 0, 10), medium-quality (QF = 1, 11), low-quality (QF = 2, 12), and unusable (QF ≥ 20) categories for diagnostic analysis. This procedure reduces the influence of retrieval noise, poor spectral fits, and unfavorable viewing geometry prior to satellite collocation."

Section 2.3 (Collocation Strategy)

"To evaluate the consistency between ground-based and satellite-derived HCHO columns, filtered Pandora observations were collocated with station-level OMI, TROPOMI and GEMS retrievals using a time-based matching framework designed to account for differences in temporal sampling. The analysis includes observations from OMI, TROPOMI, and GEMS over the period 2021–2024, allowing a more robust and statistically consistent evaluation of satellite–ground agreement across multiple observational platforms. The overall methodology of the study is illustrated in Figure 2. Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an

overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Section 2.2 (Satellite Data – addition of TROPOMI and GEMS)

"TROPOMI, launched in 2017, provides substantially finer spatial sampling than OMI and improved signal-to-noise performance. For the product version used here, the nominal pixel size is approximately 5.5 × 3.5 km² (De Smedt et al., 2021). The TROPOMI HCHO product (S5P OFFL HCHO) is derived using a similar DOAS framework but includes updated air-mass factor calculations and surface reflectance treatment (Su et al., 2020). Station-level TROPOMI HCHO values were extracted from pixels within a 10 km radius of each Pandora site. Quality screening followed recommended criteria, including qa_value ≥ 0.5, cloud fraction cloud_fraction_crb < 0.3, and SZA < 60° (De Smedt et al., 2021; Dimitropoulou et al., 2021). TROPOMI can be regarded as the next-generation continuation of the UV–visible trace-gas observing capability established by OMI, providing improved spatial resolution and signal-to-noise performance while maintaining similar measurement principles and orbital sampling. The temporal overlap between OMI and TROPOMI enables consistent long-term validation of satellite HCHO retrievals and facilitates assessment of algorithm evolution across successive instrument generations. The inclusion of both OMI and TROPOMI allows evaluation of retrieval consistency across successive satellite generations. While OMI provides a long-term observational baseline beginning in 2004, TROPOMI extends this record with enhanced spatial resolution and improved sensitivity to sub-pixel variability. The overlap period between the two sensors enables assessment of temporal continuity in satellite HCHO products and supports robust validation of long-term atmospheric composition trends.

Satellite observations from the Geostationary Environment Monitoring Spectrometer (GEMS) onboard the GEO-KOMPSAT-2B platform were additionally used to complement polar-orbiting measurements. GEMS provides hourly hyperspectral observations over East and Southeast Asia, enabling improved characterization of diurnal variability in tropospheric formaldehyde (HCHO) (Lee et al., 2023). In this study, Level-2 HCHO data (GEMS L2 HCHO) from January 2021 to December 2024 were obtained via the National Institute of Environmental Research (NIER) API, with only forward-calculated (FC) retrievals retained to ensure algorithmic consistency and data reliability. Station-level GEMS HCHO values were derived by averaging pixels within a 10 km radius of each Pandora site. Quality control followed conservative filtering criteria, including FinalAlgorithmFlags = 0, cloud radiance fraction < 0.4, and solar zenith angle SZA < 60° (Lee et al., 2024). The inclusion of GEMS provides enhanced temporal sampling relative to polar-orbiting sensors, allowing improved assessment of sub-daily

variability and reducing temporal representativeness errors in satellite–ground validation over Southeast Asia."

Comment 1.2

Comments

Line 78-79: The direct-sun (DS) retrieval assumes negligible scattering, resulting in a nearly uniform sensitivity to the HCHO column regardless of its vertical distribution. Therefore, the statement that DS measurements have higher sensitivity to near-surface pollution is not accurate. Additionally, the cited reference does not support this claim. The authors should revise this statement and provide an appropriate explanation of the DS sensitivity characteristics. The citation to Herman et al. (2009) is also not appropriate in this context, as sky-scan Pandora observations were not available at that time.

Response 1.2

We thank the reviewer for this important clarification. We agree that the previous statement was not accurate and could be misleading. Direct-sun (DS) retrievals do not inherently have enhanced sensitivity to near-surface HCHO; rather, they provide a total column measurement with relatively uniform vertical sensitivity under clear-sky conditions. In the revised manuscript, we have removed this statement and revised the discussion to more accurately describe the distinction between DS and SS observations in terms of sampling characteristics and spatial representativeness, rather than vertical sensitivity. In particular, the discussion now emphasizes that DS retrievals sample a narrow solar beam and are therefore more susceptible to localized variability, while SS retrievals provide a more spatially integrated measurement. We have also removed the inappropriate citation to Herman et al. (2009) and replaced it with more relevant references where appropriate.

Revised Text

Section 1 (Introduction)

"Direct-sun retrievals sample the atmospheric column along a narrow solar beam, while sky-scan observations integrate scattered radiation across multiple viewing angles. Differences between direct-sun and sky-scan retrievals are primarily associated with sampling characteristics and spatial representativeness."

Comment 1.3

Line 81–82: The statement "Despite these fundamental differences, most previous validation studies have implicitly treated Pandora HCHO as a single product" is unclear. Direct-sun (DS) and sky-scan (SS) Pandora HCHO retrievals have different sensitivities and are generally treated separately in validation studies. However, Rawat et al. (2025) proposed an approach to combine DS and SS observations into a single product by accounting for column biases and differences in integration time.

The authors should clarify this statement and distinguish between studies that treat DS and SS separately and approaches that explicitly combine the two datasets.

Response 1.3

We thank the reviewer for this clarification. We agree that the original statement was oversimplified and did not adequately distinguish between different approaches in the literature. In the revised manuscript, we have clarified that DS and SS retrievals are often treated separately in validation studies, but their implications for satellite validation—particularly in terms of representativeness and sampling differences—are not always explicitly evaluated. We also acknowledge recent work, such as Rawat et al. (2025), which proposes a framework to combine DS and SS observations by accounting for systematic differences. The text has been revised to better reflect these distinctions and to position the present study as focusing on the explicit evaluation of DS and SS behavior, rather than combining them into a single product.

Revised Text

Section 1 (Introduction)

"Direct-sun (DS) and sky-scan (SS) retrievals are often analyzed separately in validation studies due to their differing measurement characteristics. Recent work has proposed approaches to combine DS and SS observations by accounting for systematic differences in bias and sampling (Rawat et al., 2025). However, the extent to which these retrieval geometries influence satellite–ground agreement, particularly in terms of spatial-temporal representativeness, remains insufficiently quantified."

Comment 1.4

I am surprised that the manuscript mentions several satellite missions such as TROPOMI, TEMPO, and Sentinel-4 but does not discuss GEMS, which was a key motivation for deploying Pandora instruments in the Asian domain under the PAN-Asia network. In addition, it is unclear why the analysis focuses only on OMI data when more recent satellite products such as TROPOMI and GEMS are available. The authors should justify the use of OMI alone or consider incorporating these newer datasets, which provide improved spatial and temporal coverage for validation studies.

Response 1.4

We thank the reviewer for this important suggestion and fully agree. In the revised manuscript, the satellite analysis has been substantially expanded beyond OMI. Specifically, we now include TROPOMI and GEMS alongside OMI to provide a more comprehensive evaluation of satellite–Pandora consistency. TROPOMI offers improved spatial resolution, while GEMS provides high-temporal-resolution observations over Southeast Asia, which is particularly relevant for assessing temporal representativeness in tropical environments. OMI is retained as a reference dataset due to its long-term

continuity and stable sampling characteristics, allowing consistent comparison across retrieval geometries. The inclusion of all three sensors enables a more robust and physically meaningful assessment of the roles of spatial resolution and temporal sampling in satellite validation.

Revised Text

Section 1 (Introduction)

"Recent satellite instruments such as the TROPOspheric Monitoring Instrument (TROPOMI) provide substantially higher spatial resolution than OMI and enable improved detection of localized HCHO enhancements under favorable conditions (Lee et al., 2024; Su et al., 2020). However, higher spatial resolution alone does not eliminate representativeness errors when comparing satellite and ground-based observations, particularly in heterogeneous tropical environments (Boersma et al., 2016). TROPOMI retrievals remain sensitive to cloud fraction, aerosol loading, and surface reflectance, and their smaller pixel size can increase sensitivity to localized plumes that may not be representative of broader atmospheric columns (De Smedt et al., 2018). In addition, OMI's coarser spatial footprint provides a stable reference for diagnosing first-order effects related to spatial representativeness. Complementing these polar-orbiting sensors, the Geostationary Environment Monitoring Spectrometer (GEMS) offers hourly observations over East and Southeast Asia, enabling improved characterization of diurnal variability and reducing temporal sampling mismatches in satellite–ground comparisons. The combined use of OMI, TROPOMI, and GEMS therefore provides a comprehensive framework to disentangle the relative roles of spatial resolution, temporal sampling, and retrieval geometry in satellite validation. In this context, differences between Pandora Direct-sun and Sky-scan observations can be evaluated more robustly across multiple observational scales, providing improved insight into the factors governing satellite–ground consistency in tropical environments."

Comment 1.5

It is difficult to use Figures 2 and 3 to intercompare the direct-sun (DS) and sky-scan (SS) observations. A more appropriate approach would be to compare temporally (within 5-10 minutes) matched DS and SS measurements (e.g., nearest observations) in a scatter plot to better assess the consistency and performance of the two observing modes.

Response 1.5

We thank the reviewer for this valuable suggestion and agree that temporally matched comparisons provide a more rigorous assessment of DS–SS consistency. In the revised manuscript, we have implemented this approach. Direct-sun (DS) and sky-scan (SS) observations are first paired within a ±5 min window, and their consistency is evaluated using scatter plot analysis and correlation metrics, including stratification by retrieval quality. This replaces the previous distribution-based comparison

and enables a more direct and physically meaningful assessment of agreement between the two observing modes.

Revised Text

Section 2.1.1 (Quality Control and Data Pairing)

"To improve the robustness of ground-based HCHO observations used for intercomparison and satellite validation, an uncertainty-based quality control (QC) protocol following the methodological framework of Rawat et al. (2025) was applied to contemporaneous Pandora direct-sun (DS) and sky-scan (SS) observations. DS and SS retrievals were first paired within a 5 min tolerance window."

Section 3.2 (DS–SS Comparison)

"The nine-panel correlation analyses (Figs. 3) reveal that DS–SS agreement depends strongly on retrieval quality category, with the highest correlations observed when both measurements fall within the high-quality regime (QF = 0, 10)."

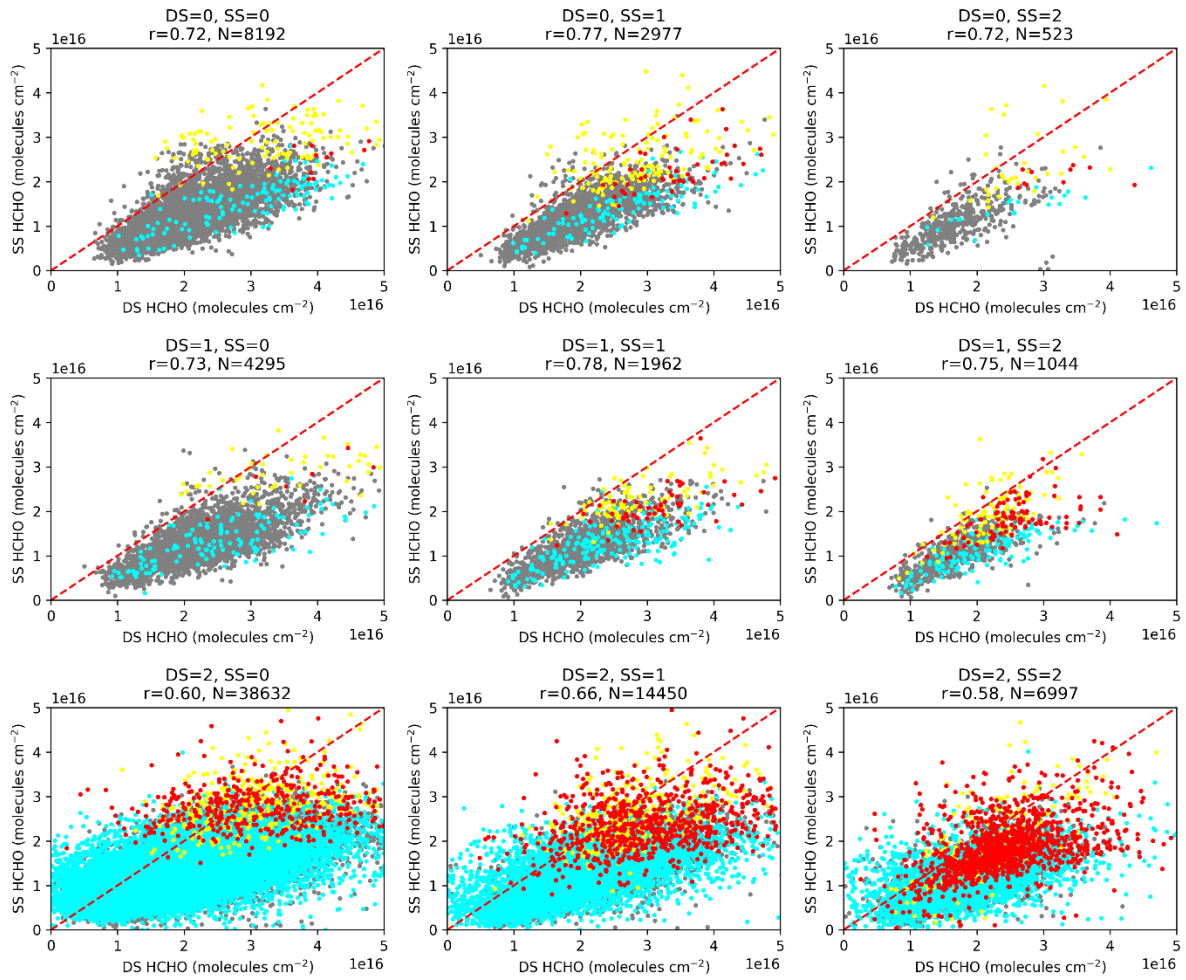


Figure 3. Nine-panel plot of correlation between contemporaneous Pandora HCHO column amounts: direct-sun (DS) vs sky-scan (SS) for each quality category, following the Rawat et al. (2025, AMT) QC

method at Bangkok station. Panels are organized by DS and SS quality categories (0 = high, 1 = medium, 2 = low). Each panel shows the scatter of DS vs SS HCHO (molecules cm^{-2}), with points color-coded by uncertainty thresholds: gray = both below cutoff, cyan = DS above cutoff, yellow = SS above cutoff, red = both above cutoff. The red dashed line represents the 1:1 relationship, and the correlation coefficient (r) and number of matched observations (N) are indicated in each panel.

Comment 1.6

For pandora data quality: it appears that the authors use all flagged data (0, 1, 2, 10, 11, 12). However, medium- and low-quality data can sometimes contain large uncertainties and may require additional filtering (e.g., based on uncertainties, fitting WRMS or other quality criteria). At the same time, strictly removing all data flagged as 12 can sometimes eliminate a large portion of the dataset if only the highest-quality PGN flags are retained. Therefore, applying additional quality screening, similar to the approach proposed in Rawat et al. (2025), would likely strengthen the robustness of the analysis rather than using all flagged data without further evaluation. Additionally, the Pandora quality flag has three broad groups (Assured, Not-Assured, and Unusable). However, Pandora quality flags contain more detailed information, specifically, the units digit (0, 1, 2) indicates high-, medium-, and low-quality retrievals, respectively, while the tens digit indicates the Not-Assured. Thus, data flagged as 0 or 10 are generally considered high-quality and suitable for scientific use, whereas 1 or 11 and 2 or 12 indicate medium and low quality and require additional scrutiny for use (Gebetsberger et al., 2022).

Response 1.6

We thank the reviewer for this important and detailed comment. We fully agree that relying solely on Pandora quality flags without additional screening is insufficient, and that medium- and low-quality retrievals require further evaluation rather than being either fully retained or completely discarded. In response, the revised manuscript adopts an uncertainty-based quality control protocol following Rawat et al. (2025). This approach combines formal Pandora quality flags with independent filtering criteria, including relative uncertainty (<10%), spectral fitting residual (WRMS < 0.01), and spatial representativeness constraints (MHxD < 20 km for sky-scan retrievals). This allows retention of physically meaningful observations while removing retrieval artefacts, without overly restricting the dataset. We have also clarified the interpretation of Pandora quality flags, explicitly distinguishing between high- (0, 10), medium- (1, 11), and low-quality (2, 12) retrievals, and using these categories primarily for diagnostic analysis rather than strict exclusion. These revisions significantly improve the robustness and physical consistency of the dataset used in the analysis.

Revised Text

Section 2.1.1 (Quality Control)

"A high-quality reference subset was then defined using Pandora quality flags QF = 0 or 10 for both DS and SS retrievals, and dynamic absolute uncertainty thresholds were calculated separately for DS and

SS as the mean plus three standard deviations of the uncertainty in this subset. Matched observations were retained when either both DS and SS absolute uncertainties were below these dynamic thresholds or both relative uncertainties were below 10 %. Additional filters required WRMS < 0.01 for both DS and SS retrievals and, for sky-scan observations, maximum horizontal distance (MHxD) < 20 km when available. Pandora quality flags were subsequently used to classify observations into high-quality (QF = 0, 10), medium-quality (QF = 1, 11), low-quality (QF = 2, 12), and unusable (QF ≥ 20) categories for diagnostic analysis. This procedure reduces the influence of retrieval noise, poor spectral fits, and unfavorable viewing geometry prior to satellite collocation."

Comment 1.7

For Figure 4, the time-series analysis is presented using multiple subplots, which makes it difficult to clearly see the overall temporal behavior. I recommend consolidating the information into a single figure showing the hourly and daily variations, and additionally including monthly averages to better illustrate the temporal patterns in the dataset.

Response 1.7

We thank the reviewer for this helpful suggestion regarding figure clarity. We agree that the original time-series presentation using multiple subplots made it difficult to clearly interpret the overall temporal behavior. In the revised manuscript, the time-series figures have been substantially restructured. Rather than focusing on multiple temporal aggregations (hourly/daily) in separate panels, the updated figures are simplified to better highlight the key features relevant to this study, particularly the impact of quality control and the comparison between DS and SS retrievals. This revised presentation improves readability and ensures that the figures are more directly aligned with the scientific objectives of the manuscript.

Revised Text

Figure 4, 7, 10 (Revised)

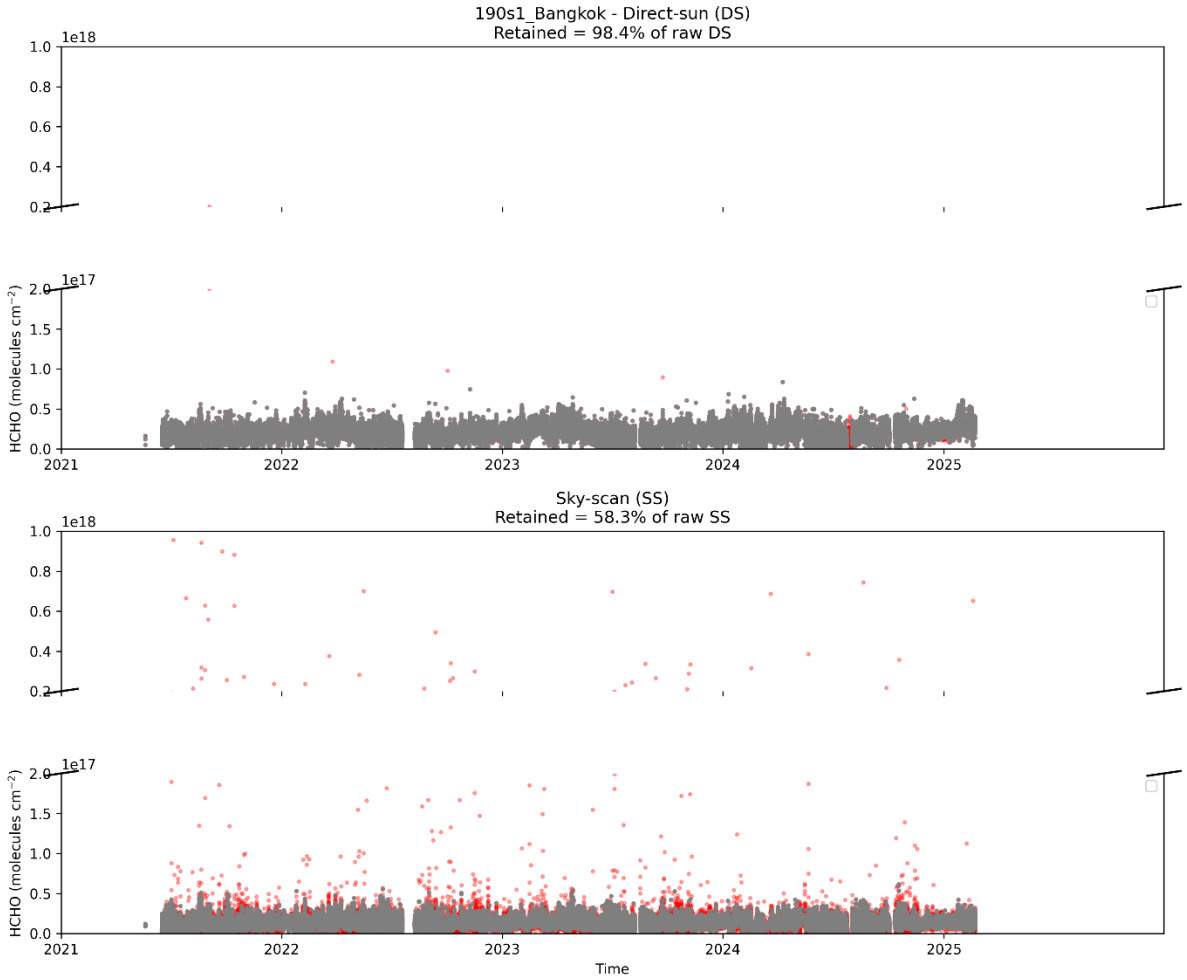


Figure 4. Time series of Pandora HCHO column amounts for direct-sun (DS) and sky-scan (SS) measurements following the Rawat et al. (2025, AMT) QC method at the Pandora Bangkok station. The upper panels show DS HCHO, and the lower panels show SS HCHO. Removed data points failing quality control (QC) are highlighted in red, while retained measurements are shown in gray. Broken y-axes are used to display both low and high concentration ranges. The percentage of removed points due to QC is indicated in the DS panel titles.

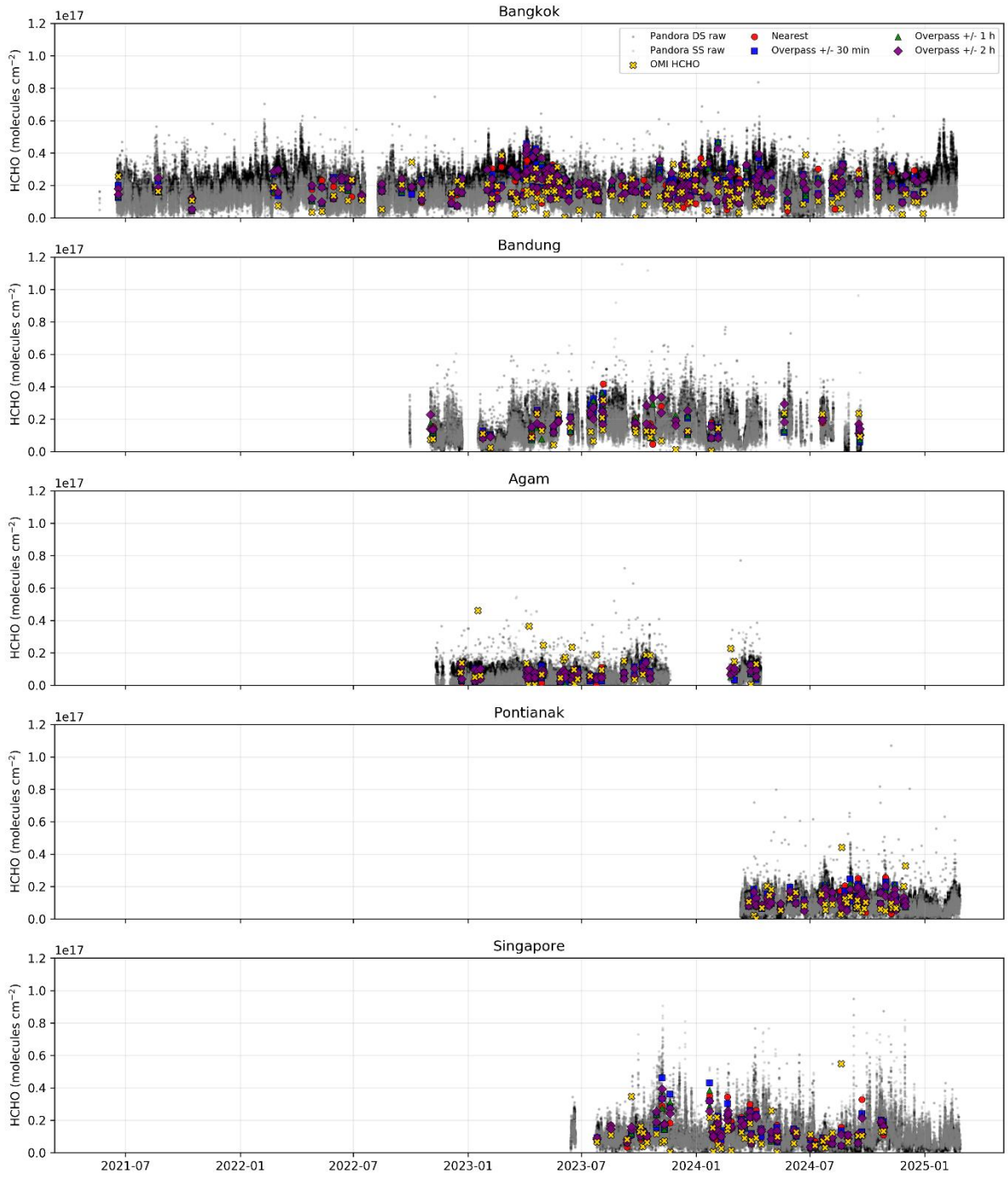


Figure 7. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated OMI observations at five Southeast Asian stations. OMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ±30 min, ±1 h, and ±2 h.

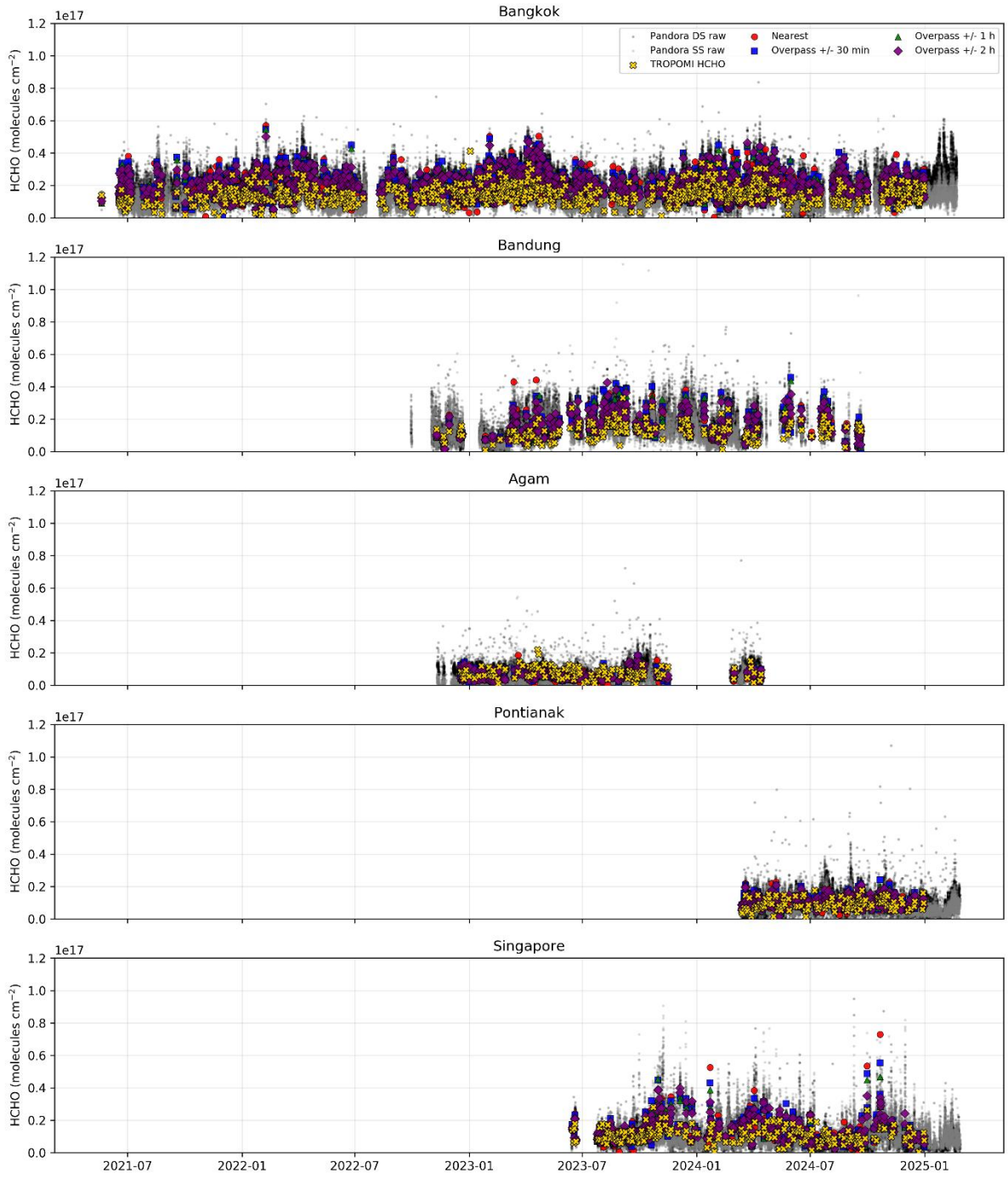


Figure 10. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated TROPOMI observations at five Southeast Asian stations. TROPOMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ±30 min, ±1 h, and ±2 h.

Comment 1.8

Table 4: The temporal averaging for Pandora is described as 00:00–23:00 LT, which is unclear since Pandora instruments only operate during daytime under sunlight conditions (unless moon). Currently, there are no nighttime observations. Please clarify.

Response 1.8

We thank the reviewer for pointing out this ambiguity and fully agree that the original description was not appropriate, as Pandora observations are limited to daytime conditions. In the revised manuscript, this issue has been resolved by removing the previous temporal averaging framework (Table 4) entirely. The satellite–Pandora comparison has been redesigned using a time-based collocation approach, including nearest-time matching and overpass-window averaging centered on the satellite observation time. This ensures physical consistency with the actual temporal sampling of both Pandora and satellite measurements.

Revise text

Section 2.3 (Collocation Strategy)

"Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Comment 1.9

Although I understand the motivation for averaging OMI pixels to reduce noise and increase the number of collocated observations, the motivation for averaging all Pandora measurements over the entire day is not clearly justified. Pandora observations exhibit strong diurnal variability, and averaging over the full day may mask important temporal variability relevant for satellite validation. Recent work has emphasized the importance of carefully accounting for spatial representativeness when using Pandora data for satellite comparisons. Also Park et al., (2026) shows that increasing the collocation radius generally improves the R^2 between Pandora and TROPOMI for HCHO, whereas the opposite behavior is often observed for NO_2 due to its stronger spatial heterogeneity. I recommend that the authors provide a clear justification for this averaging approach. Alternatively, they could restrict the analysis to the afternoon Pandora observation windows (E3, E6, and E9). If robustness is a concern due to limited sampling, using daily averages may provide a more representative comparison. However, using morning averages to compare with OMI observations does not appear justified, given the differences in overpass time and the strong diurnal variability in trace gas columns.

Response 1.9

We thank the reviewer for this important comment and fully agree that averaging Pandora observations over the full day is not appropriate for satellite validation, particularly given the strong diurnal variability of HCHO. In the revised manuscript, this issue has been fully addressed by removing the previous daily and fixed-time averaging framework (E1–E9). The analysis has been redesigned using a time-based collocation approach, including nearest-time matching (±2 h) and overpass-window averaging centered on the satellite observation time (±30 min, ±1 h, ±2 h). This ensures that Pandora observations are temporally consistent with satellite measurements and avoids biases associated with full-day or mismatched temporal averaging. We agree with the reviewer that diurnal variability and representativeness are critical considerations, and these are now explicitly accounted for in the revised collocation framework.

Revise text

Section 2.3 (Collocation Strategy)

"Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Comment 1.10

For the comparison between Pandora and OMI in Figure 5, the reported best performance for E2 and E8 appears questionable in terms of both temporal representativeness and robustness. It is unclear why morning Pandora observations would provide the best agreement with OMI, given the differences in overpass timing and the strong diurnal variability of trace gas columns. In addition, deriving statistical relationships from very limited numbers of collocated data points may not provide robust conclusions. I recommend that the authors consider using E6 or E9. If sampling robustness is a concern, E4 or E7 could also be considered, or the temporal matching windowagain could be slightly relaxed to include additional observations. Similarly, the analysis presented in Figure 7, which relies on a single satellite pixel, may not be sufficient. A spatial averaging approach using multiple nearby pixels would likely provide a more representative comparison. Finally, the analysis shown in Figure 8 also appears to suffer from both robustness issues and potential timing mismatches, which should be carefully reconsidered to ensure meaningful satellite–Pandora comparisons. It is also unclear why the OMI analysis is limited to only one year (2024). This choice is not justified in the manuscript, particularly since several Pandora sites have longer periods of data availability.

Response 1.10

We thank the reviewer for this detailed and important comment. We fully agree that the previous E1–E9 experimental framework, including the use of fixed temporal windows (e.g., morning periods), limited collocation samples, and single-pixel comparisons, was not sufficiently robust for satellite validation. In the revised manuscript, this entire analysis has been removed and replaced with a physically consistent collocation framework. Specifically, Pandora–satellite comparisons are now performed using nearest-time matching (±2 h) and overpass-centered averaging windows (±30 min, ±1 h, ±2 h), ensuring temporal consistency with satellite observations. In addition, satellite HCHO columns are derived using multi-pixel averaging within a defined spatial radius (10 km), rather than relying on single-pixel values, to improve spatial representativeness. Furthermore, the analysis is no longer limited to OMI or a single year. The revised study includes OMI, TROPOMI, and GEMS, and utilizes the full available Pandora dataset (2021–2024), significantly improving the statistical robustness and representativeness of the results.

Revised text

Section 2.3 (Collocation Strategy & Dataset Scope)

"The analysis includes observations from OMI, TROPOMI, and GEMS over the period 2021–2024, allowing a more robust and statistically consistent evaluation of satellite–ground agreement across multiple observational platforms. The overall methodology of the study is illustrated in Figure 2. Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

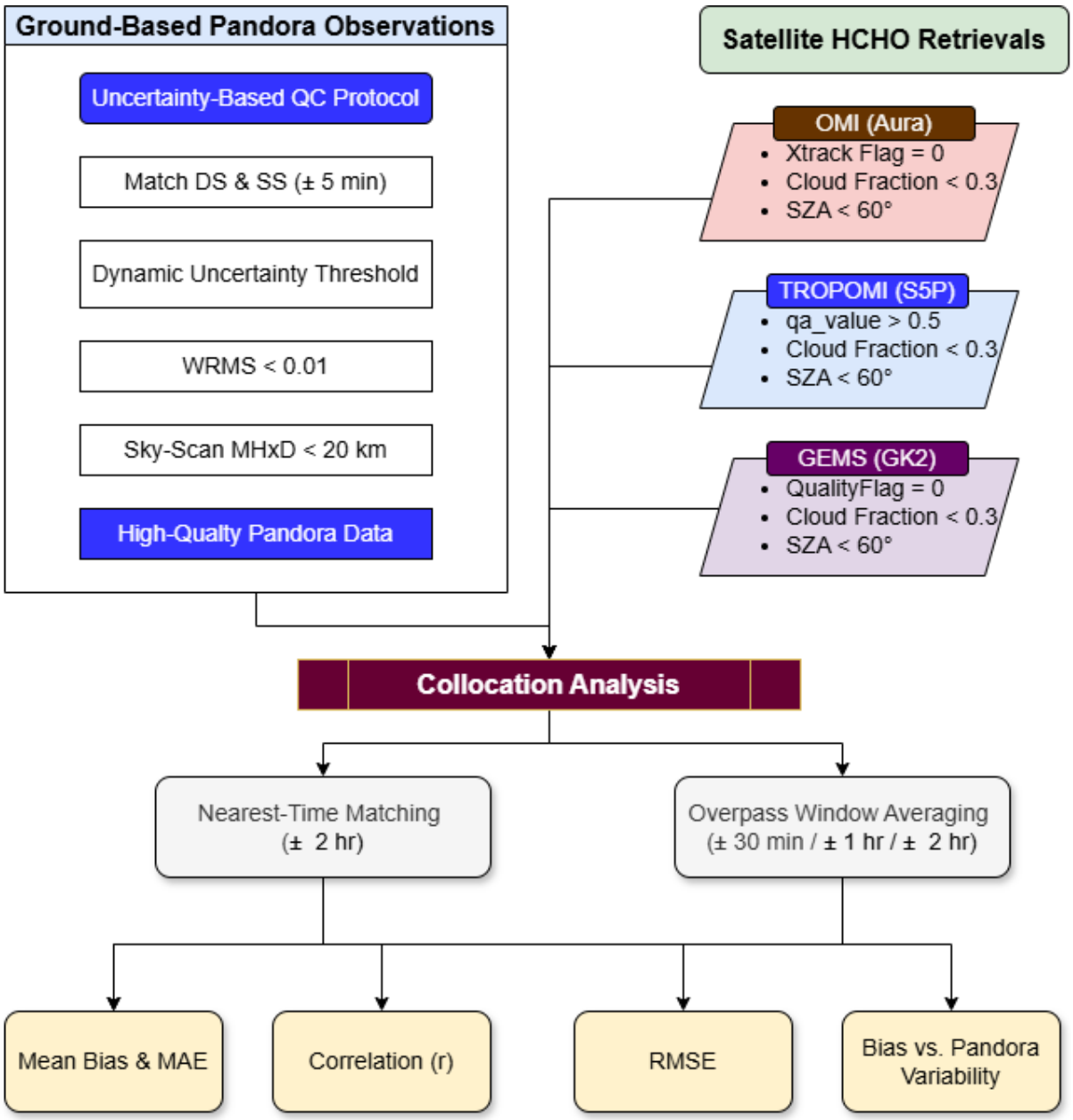


Figure 2. Flowchart illustrating the satellite–Pandora HCHO validation framework applied in this study. The methodology includes uncertainty-based quality control of Pandora observations following **Rawat et al. (2025)**, standard quality screening of OMI, TROPOMI and GEMS retrievals, temporal collocation using multiple overpass windows, and statistical evaluation of bias, error metrics, and representativeness effects in tropical environments.

Comment 1.11

Again, it is difficult to understand why the SZA analysis is relevant for the OMI comparison, since OMI observations occur near early afternoon when the Sun is generally high in the sky, even during winter in tropics. It is also unclear whether the SZA values used in the analysis are derived from OMI or from Pandora observations. If the SZA values are based on Pandora measurements, the authors should clearly

explain how the daily averaged Pandora SZA is used to justify the SZA dependence in the OMI comparison.

Response 1.11

We thank the reviewer for this important clarification and agree that the previous SZA-based analysis was not sufficiently justified in the context of OMI comparisons, given its fixed early afternoon overpass time. In the revised manuscript, the emphasis on SZA-dependent analysis has been removed and de-emphasized. The interpretation of satellite–Pandora differences is now based primarily on temporal collocation and spatial representativeness, rather than solar geometry effects. This revision avoids ambiguity related to the derivation and interpretation of SZA and ensures that the analysis is more directly aligned with the physical factors controlling satellite–ground agreement.

Comment 1.12

Section 4.3, titled "Bias Correction and Retrieval Optimization," appears misleading, as no clear evidence of bias correction or retrieval optimization is presented. The satellite comparison results shown in the previous section contain substantial errors, and no robust or consistent bias statistics are demonstrated to support the claims made in the section.

Response 1.12

We thank the reviewer for this important observation and agree that the previous section title and framing were misleading, as the analysis did not present a formal bias correction or retrieval optimization. In the revised manuscript, this section has been removed, and the discussion has been restructured. The interpretation now focuses on a diagnostic assessment of the factors controlling satellite–ground agreement, including retrieval geometry, temporal sampling, and spatial representativeness, rather than implying bias correction. This revision ensures that the conclusions are consistent with the evidence presented and avoids overstating the scope of the analysis.

References:

Park, J.-U., Lim, S., Hanisco, T. F., Abuhassan, N., Place, B. K., Pandey, A., Cede, A., Tiefengraber, M., Gebetsberger, M., Park, J., Choi, J., Crawford, J. H., Song, C.-K., & Kim, S.-W. (2026). Global analysis of nitrogen dioxide and formaldehyde column densities from the Pandora global network: Variability and implications for satellite validation. Remote Sensing of Environment, 335, 115249. https://doi.org/10.1016/j.rse.2026.115249

Rawat, P., Crawford, J. H., Travis, K. R., Judd, L. M., Demetillo, M. A. G., Valin, L. C., Szykman, J. J., Whitehill, A., Baumann, E., and Hanisco, T. F.: Maximizing the scientific application of Pandora

column observations of HCHO and NO2, Atmos. Meas. Tech., 18, 2899–2917, https://doi.org/10.5194/amt-18-2899-2025, 2025.

DimiRM4AQ_DataQualityFlagging GenericProcedureEvolution_TN_2019008_v7.pdf (last access: 27 September 2023), 2022.

Reviewer 2

Comment 2.1

The authors present a comparison of ground based HCHO retrievals from 5 Pandora sites with retrievals from OMI. The authors compare both direct sun and sky scan retrievals from Pandora to OMI and use those comparisons to suggest the use of sky scans would be preferable for satellite validation in the tropics. This work presents interesting data from an understudied part of the world. Unfortunately, the manuscripts conclusions rely on an incomplete analysis and questionable experimental design. Substantial revisions are needed before this work can be published.

Response 2.1

We thank the reviewer for the constructive assessment and for recognizing the value of the dataset in an understudied tropical region. We agree that the original manuscript required substantial improvements in both experimental design and analytical rigor. In response, the manuscript has been fundamentally revised. The previous OMI-only analysis and multi-scenario experimental design have been replaced with a physically consistent collocation framework based on nearest-time matching and overpass-centered averaging. In addition, an uncertainty-based quality control protocol following Rawat et al. (2025) has been implemented to ensure that all analyses are based on robust and physically meaningful Pandora observations. Furthermore, the study has been expanded to include TROPOMI and GEMS, alongside OMI, providing a more comprehensive multi-sensor evaluation across different spatial and temporal scales. The comparison between direct-sun and sky-scan retrievals is now performed using temporally matched pairs with quality-stratified analysis, enabling a more rigorous assessment of their consistency and behavior.

Revised Text

Section 1 (Introduction)

"In this study, we present a comprehensive evaluation of Pandora HCHO observations across Southeast Asia, explicitly distinguishing between Direct-sun and Sky-scan retrievals and assessing their consistency with multiple satellite products (OMI, TROPOMI, and GEMS). By applying an uncertainty-based quality-control framework and a unified temporal collocation strategy, this work aims to quantify how retrieval geometry, temporal sampling, and spatial representativeness jointly influence satellite–ground agreement in tropical environments."

Section 2.1.1 (Quality Control)

"To improve the robustness of ground-based HCHO observations used for intercomparison and satellite validation, an uncertainty-based quality control (QC) protocol following the methodological framework of Rawat et al. (2025) was applied to contemporaneous Pandora direct-sun (DS) and sky-scan (SS)

observations. DS and SS retrievals were first paired within a 5 min tolerance window. A high-quality reference subset was then defined using Pandora quality flags QF = 0 or 10 for both DS and SS retrievals, and dynamic absolute uncertainty thresholds were calculated separately for DS and SS as the mean plus three standard deviations of the uncertainty in this subset. Matched observations were retained when either both DS and SS absolute uncertainties were below these dynamic thresholds or both relative uncertainties were below 10 %. Additional filters required WRMS < 0.01 for both DS and SS retrievals and, for sky-scan observations, maximum horizontal distance (MHxD) < 20 km when available. Pandora quality flags were subsequently used to classify observations into high-quality (QF = 0, 10), medium-quality (QF = 1, 11), low-quality (QF = 2, 12), and unusable (QF ≥ 20) categories for diagnostic analysis. This procedure reduces the influence of retrieval noise, poor spectral fits, and unfavorable viewing geometry prior to satellite collocation."

Section 2.3 (Collocation Strategy)

"To evaluate the consistency between ground-based and satellite-derived HCHO columns, filtered Pandora observations were collocated with station-level OMI, TROPOMI and GEMS retrievals using a time-based matching framework designed to account for differences in temporal sampling. The overall methodology of the study is illustrated in Figure 2. Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Section 2.2 (Satellite Data – addition of TROPOMI and GEMS)

"TROPOMI, launched in 2017, provides substantially finer spatial sampling than OMI and improved signal-to-noise performance. For the product version used here, the nominal pixel size is approximately 5.5 × 3.5 km² (De Smedt et al., 2021). The TROPOMI HCHO product (S5P OFFL HCHO) is derived using a similar DOAS framework but includes updated air-mass factor calculations and surface reflectance treatment (Su et al., 2020). Station-level TROPOMI HCHO values were extracted from pixels within a 10 km radius of each Pandora site. Quality screening followed recommended criteria, including qa_value ≥ 0.5, cloud fraction cloud_fraction_crb < 0.3, and SZA < 60° (De Smedt et al., 2021; Dimitropoulou et al., 2021). TROPOMI can be regarded as the next-generation continuation of the UV–visible trace-gas observing capability established by OMI, providing improved spatial resolution and signal-to-noise performance while maintaining similar measurement principles and orbital sampling. The temporal overlap between OMI and TROPOMI enables consistent long-term validation of satellite HCHO retrievals and facilitates assessment of algorithm evolution across successive instrument generations. The inclusion of both OMI and TROPOMI allows evaluation of

retrieval consistency across successive satellite generations. While OMI provides a long-term observational baseline beginning in 2004, TROPOMI extends this record with enhanced spatial resolution and improved sensitivity to sub-pixel variability. The overlap period between the two sensors enables assessment of temporal continuity in satellite HCHO products and supports robust validation of long-term atmospheric composition trends.

Satellite observations from the Geostationary Environment Monitoring Spectrometer (GEMS) onboard the GEO-KOMPSAT-2B platform were additionally used to complement polar-orbiting measurements. GEMS provides hourly hyperspectral observations over East and Southeast Asia, enabling improved characterization of diurnal variability in tropospheric formaldehyde (HCHO) (Lee et al., 2023). In this study, Level-2 HCHO data (GEMS L2 HCHO) from January 2021 to December 2024 were obtained via the National Institute of Environmental Research (NIER) API, with only forward-calculated (FC) retrievals retained to ensure algorithmic consistency and data reliability. Station-level GEMS HCHO values were derived by averaging pixels within a 10 km radius of each Pandora site. Quality control followed conservative filtering criteria, including FinalAlgorithmFlags = 0, cloud radiance fraction < 0.4, and solar zenith angle SZA < 60° (Lee et al., 2024). The inclusion of GEMS provides enhanced temporal sampling relative to polar-orbiting sensors, allowing improved assessment of sub-daily variability and reducing temporal representativeness errors in satellite–ground validation over Southeast Asia."

Comment 2.2

Major issues:

This paper doesn't work without a more robust intercomparison of the two ground-based datasets to support the conclusions. Comparing both to OMI and discussing the differences between each and OMI is not sufficient. The authors point out that SZA dependent uncertainties exist in OMI products, so why are the authors conducting their analysis assuming OMI is the more trustworthy observation? We use ground-based measurements to evaluate satellite-based retrievals, not the other way around. How do the retrieved columns compare to each other? What are the conditions where they diverge from each other. What are the conditions when the direct sun and sky scan agree and disagree? Are there potential explanations that might impact the utility of each for satellite validation?

Response 2.2

We thank the reviewer for this important and insightful comment. We fully agree that a robust intercomparison between Direct-sun (DS) and Sky-scan (SS) Pandora retrievals must form the foundation of the analysis, and that satellite observations should not be treated as the reference standard. In the revised manuscript, the analysis has been fundamentally restructured to prioritize DS–SS

intercomparison. A dedicated section (Section 3.2) now presents a comprehensive evaluation of DS–SS consistency using temporally matched pairs (±5 min) and quality-stratified correlation analysis, allowing direct assessment of agreement between the two retrieval geometries independent of satellite data. This analysis explicitly quantifies the conditions under which DS and SS agree and diverge. The results show that DS and SS retrievals exhibit strong agreement under high-quality conditions ($r > 0.7$), indicating that both measurement modes provide consistent representations of the HCHO column when retrieval uncertainties are well constrained. Divergence between DS and SS is primarily observed under lower-quality conditions, where increased retrieval noise, atmospheric heterogeneity, and viewing geometry effects introduce variability. In addition, systematic differences in variability are identified: DS retrievals exhibit a larger dynamic range and stronger short-timescale variability, while SS retrievals provide smoother and more spatially integrated column estimates. Satellite comparisons are therefore used not as a reference standard, but as an independent observational framework to interpret how these differences in sampling characteristics influence satellite–ground consistency. The revised manuscript explicitly avoids treating satellite retrievals as truth and instead uses them to diagnose the role of spatial representativeness and temporal sampling.

Revised Text

Section 3.2 (DS-SS Intercomparison)

"The nine-panel correlation analyses (Figs. 3) reveal that DS–SS agreement depends strongly on retrieval quality category, with the highest correlations observed when both measurements fall within the high-quality regime (QF = 0, 10). Across all stations, high-quality DS–SS pairs exhibit correlation coefficients typically exceeding 0.70, indicating strong agreement between retrieval geometries under well-constrained uncertainty conditions. Bangkok demonstrates the most robust behaviour, with correlations reaching approximately $r \approx 0.78$ in the high-quality category."

Section 3.2 (Differences between DS-SS)

"Differences between DS and SS retrievals arise primarily from sampling characteristics rather than systematic bias, with DS observations exhibiting greater sensitivity to short-timescale variability and SS retrievals providing more spatially integrated column estimates. High-quality observations yield robust agreement across all stations, while lower-quality measurements introduce increased variability and reduced correlation. The application of uncertainty-based QC therefore represents a critical step in ensuring the reliability of Pandora HCHO datasets for atmospheric analysis and satellite validation."

Section 4.0 Discussion

"Satellite observations are not treated as a reference standard in this study; instead, they provide an independent observational framework to evaluate how differences in retrieval geometry and sampling characteristics influence satellite–ground agreement."

Comment 2.3

Sky scan retrievals are not sensitive to the whole column, one would expect that the direct sun retrieval would typically be higher since it is sensitive to the whole column. Sky scans that use a temporally local zenith reference are typically only sensitive to the lowest 2 km of the atmosphere. This doesn't necessarily imply mean direct sun retrievals are "highly sensitive to episodic enhancements" as a general rule, but they are more likely to pick up lofted plumes than a sky scan observation where the plume would impact the reference spectrum and not impact the measured slant columns in the same way as a direct sun observation.

Response 2.3

We thank the reviewer for this important clarification regarding the sensitivity characteristics of sky-scan (SS) and direct-sun (DS) retrievals. We agree that DS retrievals represent the total column along the solar beam, while SS retrievals—depending on retrieval configuration—can have reduced sensitivity to the full column and may be more influenced by the lower troposphere. In the revised manuscript, we have avoided interpreting DS–SS differences in terms of vertical sensitivity, as this would require detailed radiative transfer analysis beyond the scope of this study. Instead, the analysis focuses on observational sampling characteristics and representativeness. Our results show that DS and SS retrievals exhibit strong agreement under high-quality conditions but differ in their variability: DS retrievals display a larger dynamic range and stronger short-timescale fluctuations, while SS retrievals provide smoother and more spatially integrated column estimates. We interpret this behaviour primarily in terms of sampling geometry and sensitivity to localized variability, rather than attributing it directly to vertical sensitivity differences. We have revised the manuscript to clarify this interpretation and to avoid overstating the role of vertical sensitivity in explaining DS–SS differences.

Revised Text

Section 1 (Introduction – refined clarification)

“Differences between direct-sun and sky-scan retrievals are primarily associated with sampling characteristics and spatial representativeness. While the two retrieval modes may differ in their effective sensitivity to atmospheric structure, this study focuses on their observational behaviour and consistency rather than explicit vertical sensitivity differences.”

Section 3.2 (Interpretation refinement)

“The larger variability observed in DS retrievals and the smoother behaviour of SS observations are interpreted as a consequence of differences in sampling geometry and sensitivity to localized atmospheric variability, rather than systematic differences in vertical sensitivity.”

Section 4 (Discussion – strengthening statement)

"The observed differences between DS and SS retrievals reflect the interplay between measurement geometry and atmospheric heterogeneity, with DS capturing localized variability and SS providing a more spatially integrated representation of the atmospheric column."

Comment 2.4

The chosen experiments for comparison don't all have utility for satellite evaluation, so it is unclear why these 9 scenarios were chosen.

1. Given that formaldehyde columns generally have a strong temperature and sunlight dependence and thus vary throughout the day, I'm unclear what the utility of daily averaging is in a satellite evaluation context, where the overpass time is known. Most studies just consider the average around the overpass time .
2. The daytime averaging period is given as 07-09 local, is this a typo or should the label be changed from daytime to early morning? Assuming a typo, for measurements that require sunlight, what is the utility of separate daytime and daily averages? Aren't they pretty much the same aside from some less reliable measurements in low light conditions that would typically be discarded anyway?
3. Similarly, given Rayleigh scattering limits the effective horizontal pathlength of the Pandora measurements to ~20 km under clear sky conditions in this fit window, it makes little sense to average over 2 adjacent OMI pixels (5x5) for comparison as the Pandora may not even be sampling adjacent pixels let alone two over.

Response 2.4

We thank the reviewer for this detailed and insightful comment. We fully agree that the previous nine-scenario (E1–E9) framework, including the use of daily and fixed-time averaging, was not well suited for satellite validation given the strong diurnal variability of HCHO and the known satellite overpass timing. In the revised manuscript, this entire experimental design has been removed and replaced with a physically consistent collocation framework. Pandora observations are now matched to satellite measurements using nearest-time pairing (±2 h) and overpass-centered averaging windows (±30 min, ±1 h, ±2 h), ensuring temporal consistency with satellite sampling and avoiding biases associated with full-day or mismatched temporal averages. In addition, satellite–Pandora comparisons are now performed using multi-pixel averaging within a 10 km radius, rather than fixed adjacent pixel selection, to better account for spatial representativeness and the effective sampling scale of Pandora observations.

Revised Text

Section 2.3 (Collocation Strategy)

"Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Section 2.3 (Spatial representativeness clarification)

"Satellite HCHO columns were calculated by averaging all valid pixels within a 10 km radius of each Pandora site, providing a spatially representative estimate consistent with the effective sampling scale of ground-based observations."

Comment 2.5

There doesn't appear to be sufficient data quality checks done on the Pandora data. While there is a case to be made for not relying solely on Pandora L2 QC flags (e.g. Rawat et al 2025), one should still check fit quality (RMS) and do cloud screening before comparing to satellite based measurements. For example, the statistics presented for your retrievals at Agam show unrealistically large columns with no explanation. Are these actual events or retrieval artifacts?

Response 2.5

We thank the reviewer for this important comment and fully agree that robust quality control is essential for reliable Pandora–satellite comparisons. In the revised manuscript, we have implemented an uncertainty-based quality control protocol following Rawat et al. (2025), which goes beyond the use of Pandora L2 quality flags alone. This includes explicit filtering based on spectral fitting residual (WRMS < 0.01), relative uncertainty (<10%), and dynamic absolute uncertainty thresholds, as well as MHxD constraints for sky-scan retrievals. These criteria effectively remove retrieval artefacts associated with poor spectral fits, cloud contamination, and unfavorable viewing geometry. Regarding the reviewer's concern about unrealistically large HCHO columns at Agam, we confirm that such extreme values in the original manuscript were primarily associated with low-quality retrievals and high-uncertainty conditions. After applying the revised QC protocol, these outliers are substantially reduced, and the retained dataset exhibits physically consistent variability. The impact of QC filtering, including the removal of such artefacts, is now explicitly demonstrated in the revised time-series analysis.

Revised Text

Section 2.1.1 (Quality Control)

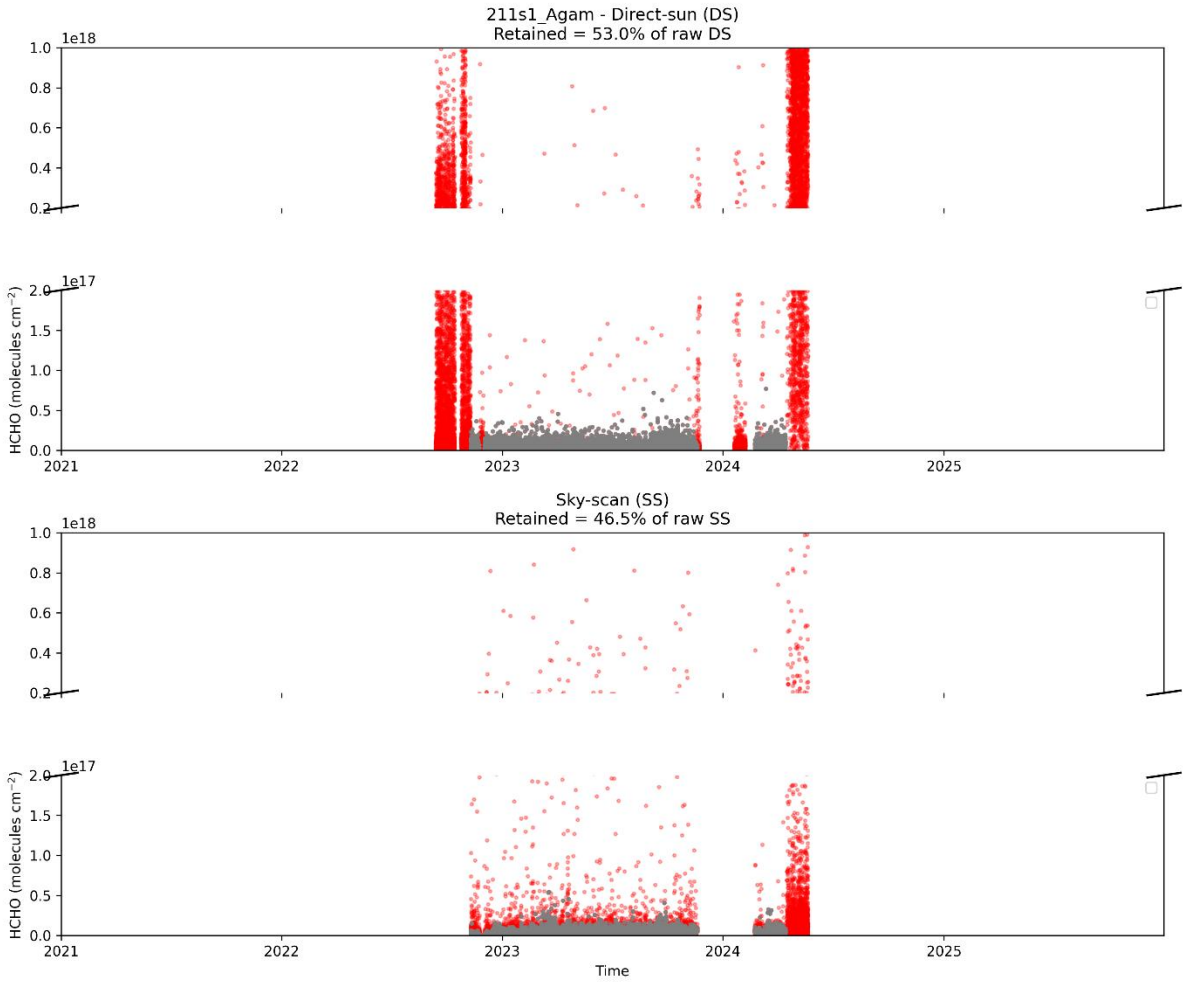
"Additional filters required WRMS < 0.01 for both DS and SS retrievals, and, for sky-scan observations, maximum horizontal distance (MHxD < 20 km) when available. Matched observations were retained only when uncertainty-based criteria were satisfied, ensuring removal of retrieval artefacts associated with poor spectral fitting and unfavorable measurement conditions."

Section 3.1 (QC impact)

"The application of uncertainty-based QC substantially reduces extreme outliers, particularly at sites such as Agam, where high-uncertainty retrievals can otherwise produce unrealistically large column values."

Figure 4 (supporting evidence)

"Removed data points failing quality control (QC) are highlighted in red, illustrating the impact of uncertainty-based filtering in eliminating unrealistic retrievals."



Same as Figure 4 but at the Pandora Agam station.

Comment 2.6

It looks like the Pandora data are filtered when making the Figures 4 and 5, but not when calculating the statistics in Table 3. You need consistent treatment throughout the analysis.

Response 2.6

In the revised manuscript, all comparative analyses (e.g., DS–SS correlation and satellite validation) are based on the QC-filtered dataset, while Table 3 presents both raw and after QC statistics to illustrate the impact of filtering.

Revised Text

Section 2.3 (Methodology Flowchart)

"The analysis includes observations from OMI, TROPOMI, and GEMS over the period 2021–2024, allowing a more robust and statistically consistent evaluation of satellite–ground agreement across multiple observational platforms. The overall methodology of the study is illustrated in Figure 2."

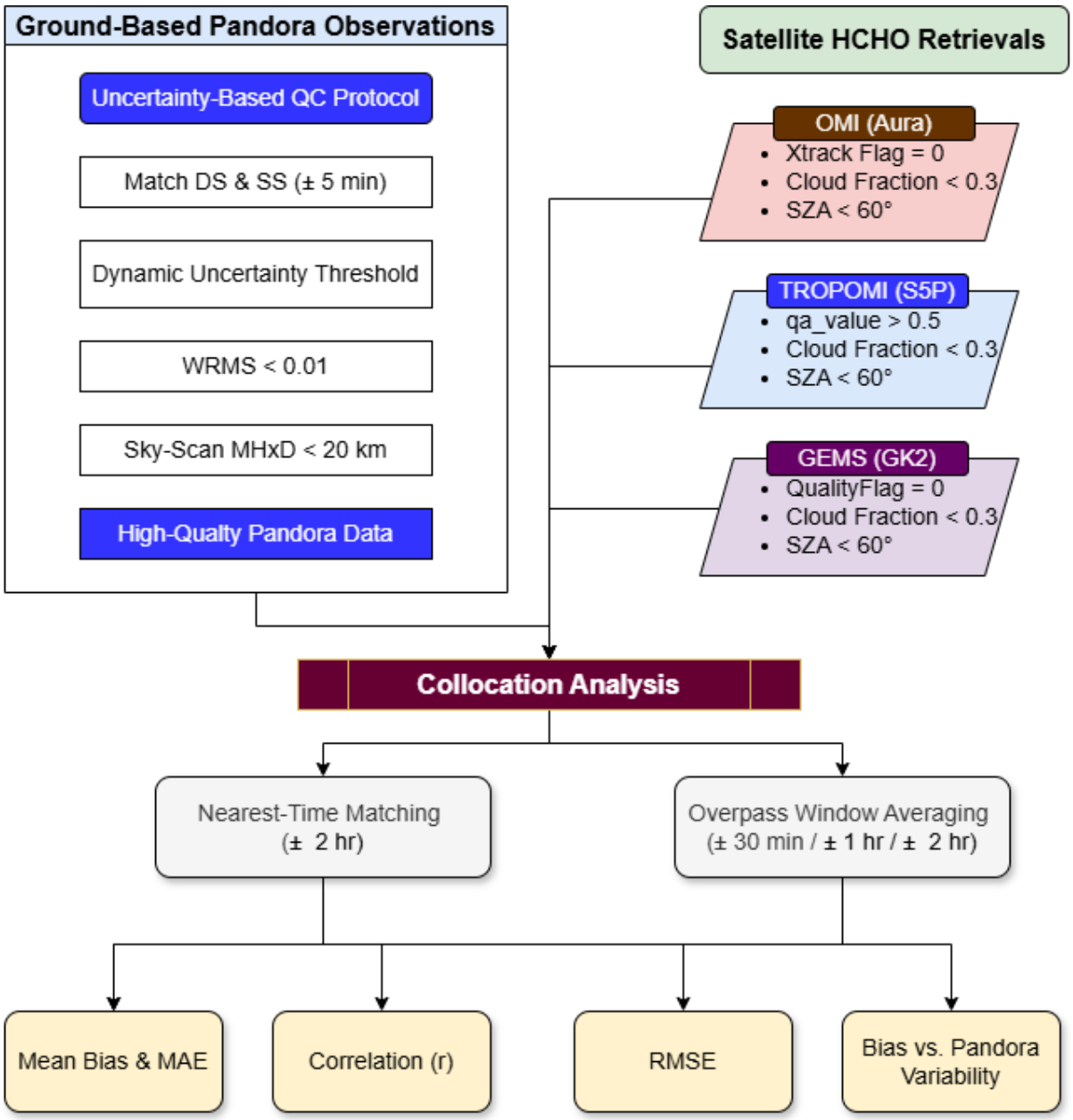


Figure 2. Flowchart illustrating the satellite–Pandora HCHO validation framework applied in this study. The methodology includes uncertainty-based quality control of Pandora observations following **Rawat et al. (2025)**, standard quality screening of OMI, TROPOMI and GEMS retrievals, temporal collocation using multiple overpass windows, and statistical evaluation of bias, error metrics, and representativeness effects in tropical environments.

Section 3.1 (Descriptive Statistics Before QC and After QC)

"The statistical distributions of HCHO columns before and after QC filtering (Table 3) confirm that the protocol primarily removes extreme outliers while preserving the central tendency of the observations. At most stations, mean and median HCHO values remain nearly unchanged following QC application, indicating that the filtering does not introduce systematic bias."

Table 3. Descriptive statistics of Pandora Level-2 formaldehyde (HCHO) retrieved from Direct-sun (DS) and Sky-scan (SS; rfuh5p1-8) observations at selected Southeast Asian stations. Statistics are shown for contemporaneous matched DS–SS pairs before and after applying the Rawat quality control (QC) protocol. Values are reported as mean ± standard deviation (SD), median with interquartile range (IQR; Q1–Q3), and minimum–maximum. All HCHO columns are expressed in units of $\times 10^{15}$ molecules cm^{-2}.

Station	Dataset	Mean ± SD	Median (IQR)	Min–Max	N
(a) Direct-sun (DS) HCHO					
Bangkok	Raw	21.2 ± 7.88	20.6 (16.0–25.7)	0.018–200	80,336
	After QC	21.3 ± 7.83	20.7 (16.1–25.8)	0.018–83.6	79,072
Bandung	Raw	19.6 ± 44.4	15.0 (10.2–21.0)	0.012–988	34,248
	After QC	16.4 ± 8.29	15.3 (10.5–21.1)	0.012–76.9	30,268
Agam	Raw	63.0 ± 166.6	9.93 (6.76–14.0)	0.013–999	35,504
	After QC	9.01 ± 4.07	8.93 (6.40–11.2)	0.013–77.0	18,804
Pontianak	Raw	11.9 ± 7.90	11.7 (8.93–14.4)	0.013–569	25,694
	After QC	11.8 ± 5.17	11.7 (9.00–14.4)	0.013–107	25,095
Singapore-NUS	Raw	10.8 ± 6.89	9.33 (6.53–13.2)	0.012–131	39,791
	After QC	10.9 ± 7.23	9.30 (6.36–13.3)	0.012–131	32,455
(b) Sky-scan (SS) HCHO					
Bangkok	Raw	12.8 ± 10.8	11.9 (8.31–16.1)	0.013–956	135,664
	After QC	13.2 ± 5.57	12.6 (9.43–16.3)	0.025–61.3	79,072
Bandung	Raw	13.0 ± 17.6	11.0 (6.67–16.6)	0.014–923	47,161
	After QC	13.1 ± 7.18	11.8 (7.91–17.1)	0.026–116	30,268
Agam	Raw	7.40 ± 26.9	4.39 (2.52–6.78)	0.010–992	40,440
	After QC	4.93 ± 3.07	4.51 (2.92–6.40)	0.010–54.5	18,804
Pontianak	Raw	8.10 ± 15.0	6.75 (4.46–9.77)	0.010–919	36,279
	After QC	8.22 ± 3.82	7.62 (5.53–10.2)	0.027–45.2	25,095
Singapore-NUS	Raw	10.9 ± 11.1	9.02 (6.05–13.4)	0.013–883	61,473
	After QC	11.7 ± 7.34	9.90 (7.11–14.0)	0.024–90.6	32,455

Comment 2.7

Minor Points

Figures 2 and 3: I think your analysis would be better served by correlation plots of these data rather than frequency distributions. If the authors want to present frequency distributions, all three retrievals should be present on the same axis for each site, so the reader can more easily compare the distributions.

Response 2.7

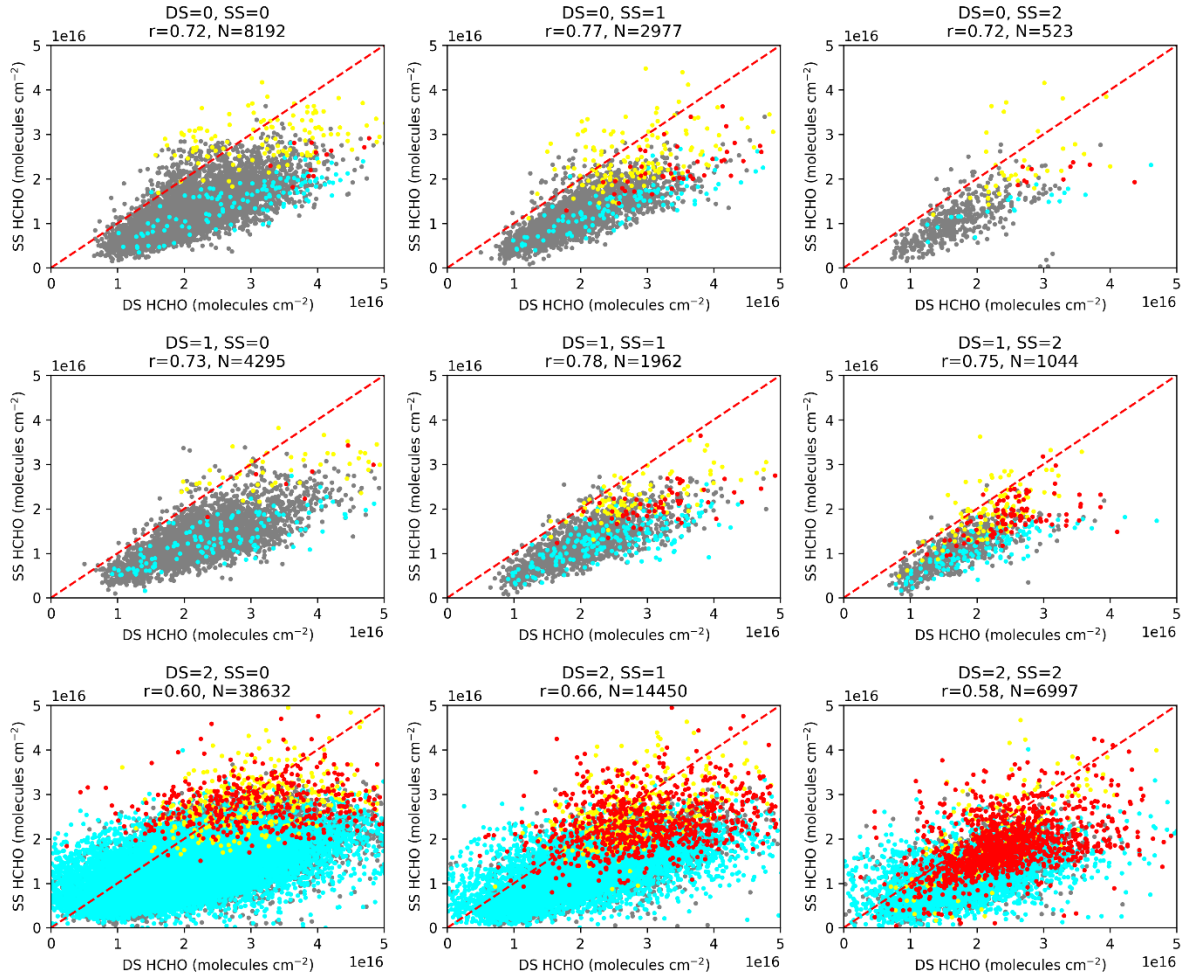
We thank the reviewer for this helpful suggestion and agree that correlation-based analysis provides a more direct assessment of consistency between retrievals. In the revised manuscript, the DS–SS comparison has been updated to use scatter plot (correlation) analysis of temporally matched pairs,

replacing the previous reliance on frequency distributions. This allows a more quantitative evaluation of agreement between retrieval modes across different quality categories. Frequency distributions have been retained only where appropriate to illustrate overall variability, but the primary analysis now emphasizes correlation-based diagnostics.

Revised Text

Section 3.2 (DS–SS Comparison)

"The nine-panel correlation analyses (Figs. 3) reveal that DS–SS agreement depends strongly on retrieval quality category, with the highest correlations observed when both measurements fall within the high-quality regime (QF = 0, 10)."



referemc

Figure 3. Nine-panel plot of correlation between contemporaneous Pandora HCHO column amounts: direct-sun (DS) vs sky-scan (SS) for each quality category, following the Rawat et al. (2025, AMT) QC method at Bangkok station. Panels are organized by DS and SS quality categories (0 = high, 1 = medium, 2 = low). Each panel shows the scatter of DS vs SS HCHO (molecules cm^{-2}), with points color-coded by uncertainty thresholds: gray = both below cutoff, cyan = DS above cutoff, yellow = SS above cutoff, red = both above cutoff. The red dashed line represents the 1:1 relationship, and the correlation coefficient (r) and number of matched observations (N) are indicated in each panel.

Comment 2.8

Line 241: Elevated has an ambiguous meaning when discussing atmospheric measurements, do you mean aloft or enhanced relative to background.

Response 2.8

In the revised manuscript, this line has been removed from the text due to the substantial revisions.

Comment 2.9

OMI while providing a long timeseries is not really the most widely used HCHO product used these days, the community would likely find more benefit from comparisons with TROPOMI and GEMS. Spatial averaging can be utilized to deal with the Pandora path crossing multiple pixels.

Response 2.9

We thank the reviewer for this valuable comment and fully agree that comparisons with more recent satellite products are important for current applications. In the revised manuscript, the analysis has been expanded to include both TROPOMI and GEMS, in addition to OMI. TROPOMI provides higher spatial resolution, while GEMS offers high-temporal-resolution observations over Southeast Asia, making them highly relevant for satellite validation in this region. As a result, the study is no longer centered on OMI alone but instead adopts a multi-sensor framework to evaluate satellite–Pandora consistency across different spatial and temporal scales. OMI is retained to provide continuity with previous validation studies and to serve as a stable reference for examining first-order representativeness effects. In addition, satellite–Pandora comparisons are now performed using multi-pixel averaging within a 10 km radius, ensuring that the satellite data are spatially representative of the effective Pandora sampling scale.

Revised Text

Section 2.3 (Satellite Data)

"To evaluate the consistency between ground-based and satellite-derived HCHO columns, filtered Pandora observations were collocated with station-level OMI, TROPOMI and GEMS retrievals using a time-based matching framework designed to account for differences in temporal sampling. The analysis includes observations from OMI, TROPOMI, and GEMS over the period 2021–2024, allowing a more robust and statistically consistent evaluation of satellite–ground agreement across multiple observational platforms."

Section 2.3 (Spatial Representativeness)

"Satellite HCHO columns were derived by averaging all valid pixels within a 10 km radius of each Pandora site, ensuring consistency with the effective spatial sampling of ground-based observations."

References:

Dimitropoulou, E., Hendrick, F., Friedrich, M. M., Tack, F., Pinardi, G., Merlaud, A., et al. (2022). Horizontal distribution of tropospheric NO2 and aerosols derived by dual-scan multi-wavelength multi-axis differential optical absorption spectroscopy (MAX-DOAS) measurements in Uccle, Belgium. Atmospheric Measurement Techniques, 15(15), 4503–4529. https://doi.org/10.5194/amt-15-4503-2022

Rawat, P., Crawford, J. H., Travis, K. R., Judd, L. M., Demetillo, M. A. G., Valin, L. C., et al. (2025). Maximizing the scientific application of Pandora column observations of HCHO and NO_2. Atmospheric Measurement Techniques, 18(13), 2899–2917. https://doi.org/10.5194/amt-18-2899-2025

Reviewer 3

Comment 3.1

This manuscript presents a much-needed investigation of Pandora HCHO products, including both direct-sun and sky-scan retrievals. Given there are limited studies evaluating these data products, particularly in Southeast Asia, the work addresses an important need. The authors analyze observations from five Pandora instruments across this region over several years, comparing variability between direct-sun and sky-scan measurements, as well as against satellite observations from OMI. While the study has the potential to make a valuable contribution, substantial revisions are necessary to address concerns related to the methodology and the interpretation of the results.

Response 3.1

We thank the reviewer for the positive assessment and for recognizing the importance of this study in an understudied region. We also agree that substantial improvements were required in the original manuscript. In response, the manuscript has been extensively revised, including the implementation of an uncertainty-based quality control protocol (Rawat et al., 2025), a rigorous DS–SS intercomparison using temporally matched pairs, and a redesigned satellite–ground collocation framework based on overpass-centered temporal matching. In addition, the analysis has been expanded to include TROPOMI and GEMS, alongside OMI, providing a more comprehensive multi-sensor evaluation.

Revised Text

Section 1 (Introduction – final paragraph)

"In this study, we present a comprehensive evaluation of Pandora HCHO observations across Southeast Asia, explicitly distinguishing between Direct-sun and Sky-scan retrievals and assessing their consistency with multiple satellite products (OMI, TROPOMI, and GEMS). By applying an uncertainty-based quality-control framework and a unified temporal collocation strategy, this work aims to quantify how retrieval geometry, temporal sampling, and spatial representativeness jointly influence satellite–ground agreement in tropical environments."

Comment 3.2

line 24-26: Sky-scan does not report total vertical columns. Only the tropospheric columns.

Response 3.2

We thank the reviewer's comment and agree sky-scan does not essentially report total vertical columns. To avoid confusion, we revised to Pandora Level-2 HCHO columns for clarity.

Revised Text

Abstract

"This study evaluates Pandora Level-2 HCHO columns from five Southeast Asian stations (2021–2024), distinguishing between direct-sun (DS) and sky-scan (SS) observations…"

Comment 3.3

line 81-87: I'm not sure this is true that previous studies don't differentiate between sky-scan and Direct sun modes for HCHO. I'd say it is more that there have been very few studies that look into DS HCHO at all because of known issues. Please reword this section. Same thing on line 121.

Response 3.3

We thank the reviewer for this clarification and agree that the original statement was not appropriately framed. We acknowledge that DS and SS retrievals are generally distinguished in the literature. In the revised manuscript, we have reworded this section to reflect that the key gap is not the lack of differentiation, but rather the limited number of studies that systematically evaluate DS-SS retrievals and their implications for satellite validation, particularly in tropical environments.

Revised Test

Section 1 (Introduction)

"Differences between direct-sun and sky-scan retrievals are primarily associated with sampling characteristics and spatial representativeness. While the two retrieval modes may differ in their effective sensitivity to atmospheric structure, this study focuses on their observational behaviour and consistency rather than explicit vertical sensitivity differences. Direct-sun (DS) and sky-scan (SS) retrievals are often analyzed separately in validation studies due to their differing measurement characteristics. Recent work (e.g., Rawat et al., 2025) has proposed approaches to combine DS and SS observations by accounting for systematic differences in bias and sampling. However, the extent to which these retrieval geometries influence satellite–ground agreement, particularly in terms of spatial-temporal representativeness, remains insufficiently quantified."

Comment 3.4

line 113: Weird that you mention TEMPO and Sentinel-4 but not GEMS for a southeast Asia study?

Response 3.4

We thank the reviewer for this comment and fully agree. In the revised manuscript, GEMS has been explicitly included and discussed alongside TROPOMI and OMI, given its strong relevance for Southeast Asia and its role in the PAN-Asia Pandora network. The text has been revised to reflect the importance of GEMS as a geostationary sensor providing high-temporal-resolution observations over the study region.

Revised Text

Section 1.0 (Introduction)

"Complementing these polar-orbiting sensors, the Geostationary Environment Monitoring Spectrometer (GEMS) offers hourly observations over East and Southeast Asia, enabling improved characterization of diurnal variability and reducing temporal sampling mismatches in satellite–ground comparisons (Bak et al., 2019a, b). The combined use of OMI, TROPOMI, and GEMS therefore provides a comprehensive framework to disentangle the relative roles of spatial resolution, temporal sampling, and retrieval geometry in satellite validation."

Comment 3.5

Table 1: State Altitude above sea level?

Response 3.5

We thank the reviewer for this suggestion. The altitude values in Table 1 have been clarified to explicitly indicate that they represent altitude above mean sea level (m a.s.l.).

Revised Text

Table 1. Summary of Pandora monitoring stations used in this study, including location, altitude, product status, and data availability. Data description: Formaldehyde (HCHO) Level 2, Version: rfus5p1-8 and rfuh5p1-8 (Last accessed: 27 Feb 2025).

Station ID	Station Name	Lat	Lon	Altitude (m a.s.l.)	rfus5p1-8 and rfuh5p1-8		
					Product Status	Data Start	Last Updated
190s1	Bangkok	13.7847	100.5400	60	Official	20210520	20250221
210s1	Bandung	-6.8948	107.5865	752	Official	20230611	20240920
211s1	Agam	-0.2046	100.3195	865	Official	20220913	20240521
212s1	Pontianak	0.0415	109.3366	1	Official	20240309	20250226
77s1	Singapore	1.2990	103.7710	77	Official	20230621	20250226

Comment 3.6

line 142: sky-scan does not report the same product as DS. Sky-scan reports 'Tropospheric' column (usually 3-4 km) while DS reports total column.

Response 3.6

We thank the reviewer for this important clarification. We agree that sky-scan (SS) and direct-sun (DS) retrievals do not represent identical quantities. DS retrievals provide total column HCHO along the direct solar beam, whereas SS retrievals are generally more sensitive to the tropospheric column and may not fully capture the total column, depending on the retrieval configuration and atmospheric conditions. In the revised manuscript, we have clarified this distinction to avoid ambiguity and to ensure that the differences between DS and SS retrievals are interpreted appropriately.

Revised Text

Section 2.1 (Pandora Data Description)

"We use Level 2 HCHO products from the rfus5p1-8 and rfuh5p1-8 processing version (last accessed: 27 February 2025), which provides HCHO columns derived from direct-sun and diffuse-sky measurements. Direct-sun (DS) retrievals provide total column HCHO along the solar beam, whereas sky-scan (SS) retrievals represent a tropospheric column derived from multi-angle scattered radiation measurements, with sensitivity that depends on retrieval configuration and atmospheric conditions."

Comment 3.7

line 155: I suggest you reword this sentence because Pandora data quality flags are already confusing. All pandoras belong to the official pandora global network, but this does nothing to guarantee the quality of the data.

Response 3.7

We thank the reviewer for this important clarification. We agree that affiliation with the Pandora Global Network does not inherently guarantee data quality, and that the original wording could be misleading. In the revised manuscript, this statement has been reworded to avoid any implication of automatic data reliability. Instead, we emphasize that data quality is ensured through the application of uncertainty-based quality control criteria within this study.

Revised Text

Section 2.1 (Pandora Data Description)

"All Pandora instruments used in this study are part of the Pandora Global Network; however, data quality is not assumed a priori and is evaluated using uncertainty-based quality control criteria applied in this work."

Comment 3.8

line173-175: You don't need to include the file versions in the results. The methods is sufficient.

Response 3.8

We thank the reviewer for this suggestion and agree that including file version details in the Results section is unnecessary. In the revised manuscript, all dataset version information has been removed from the Results section and retained only in the Methods section, where it is more appropriate.

Comment 3.9

Table 2: move to section 2.1

Response 3.9

We thank the reviewer for this suggestion. We agree that tables describing data characteristics are typically placed in the Methods section. However, in the revised manuscript, Table 2 has been substantially updated to present comparative statistics before and after the application of uncertainty-based quality control (QC). As such, Table 2 now reflects the impact of QC on the dataset, including changes in data distribution and retained observations, and is therefore an integral part of the results analysis rather than a data description. For this reason, we have retained Table 2 in the Results section.

Revised Text

Section 3.1 (Results – reference to Table 2)

"As summarised in Table 2, the QC protocol affects direct-sun (DS) and sky-scan (SS) retrievals differently. DS observations generally exhibit higher intrinsic stability, with only modest reductions in retained measurements at most stations (≤3 % at Bangkok and Pontianak)."

Table 2. Summary of Pandora Level-2 formaldehyde (HCHO) observations from Direct-sun (DS) and Sky-scan (SS) retrievals at selected Southeast Asian stations. Data are categorized by quality flags into High (0,10), Medium (1,11), Low (2,12), and Unusable (≥20). The Rawat quality control (QC) protocol was applied to filter observations based on independent uncertainty thresholds, relative uncertainty (<10%), WRMS (<0.01), and maximum horizontal distance (MHxD < 20 km for SS). Totals represent the number of valid matched DS–SS observation pairs used in the analysis.

Station	Dataset	High	Medium	Low	Unusable	Total
Bangkok	DS Raw	11,693	7,339	61,304	0	80,336
	SS Raw	65,921	35,305	34,438	0	135,664
	DS After QC	11,692	7,301	60,079	0	79,072
	SS After QC	51,119	19,389	8,564	0	79,072
Bandung	DS Raw	8,995	2,113	23,140	0	34,248
	SS Raw	32,766	6,223	8,172	0	47,161
	DS After QC	8,321	1,896	20,051	0	30,268
	SS After QC	26,447	2,448	1,373	0	30,268
Agam	DS Raw	5,438	849	20,375	8,842	35,504
	SS Raw	2,450	445	37,545	0	40,440
	DS After QC	4,582	627	13,413	182	18,804
	SS After QC	1,623	165	17,016	0	18,804
Pontianak	DS Raw	6,779	1,015	17,900	0	25,694
	SS Raw	27,300	1,681	7,298	0	36,279
	DS After QC	6,696	996	17,403	0	25,095
	SS After QC	22,204	443	2,448	0	25,095
Singapore-NUS	DS Raw	5,942	1,771	32,078	0	39,791
	SS Raw	18,633	18,000	24,840	0	61,473
	DS After QC	5,209	1,496	25,750	0	32,455
	SS After QC	13,623	11,279	7,553	0	32,455

Comment 3.10

Table 2: While explaining the data quality is necessary for Pandora discussions, I think there is more relevant information that would explain the Pandoras better. Because basically all data is unassured it doesn't do us much good to focus on that. Instead, I would rather see the 'high' 'medium' and 'low' flags that are also included in the L2 files.

Response 3.10

We thank the reviewer for this helpful suggestion and agree that the high-, medium-, and low-quality classifications provide more meaningful insight than the broader assured/not-assured grouping. In the revised manuscript, Table 2 has been updated to explicitly present the distribution of high- (QF = 0, 10), medium- (QF = 1, 11), and low-quality (QF = 2, 12) retrievals, both before and after the application of uncertainty-based QC. This provides a clearer representation of data quality characteristics and the impact of QC filtering across stations. Please see Response 3.9 for the updated Table 2.

Comment 3.11

The DS HCHO is all unassured because the PGN does not have an official method of assuring that product (I'm not sure what is going on with Singapore, but if you have not I suggest reaching out to the operator to make sure the assured values are real).

Response 3.11

We thank the reviewer for this important clarification regarding the assured status of Pandora HCHO retrievals. We acknowledge that DS HCHO products are often flagged as “unassured” within the PGN framework, reflecting the absence of an official assurance procedure rather than necessarily indicating poor data quality. In the revised manuscript, we do not rely on the assured/not-assured classification to determine data quality. Instead, we apply an uncertainty-based quality control framework (Rawat et al., 2025), which evaluates each observation based on independent criteria such as relative uncertainty, spectral fitting residual (WRMS), and spatial representativeness constraints. This approach ensures that data quality is assessed consistently across all sites, regardless of PGN assurance status. To avoid confusion, the manuscript has been revised to clarify the interpretation of Pandora quality flags and to emphasize that data selection is based on quantitative QC metrics rather than PGN assurance categories.

Revised Text

Section 2.1.1 (Quality Flag Clarification)

“In this way, the assured/not-assured classification within the Pandora Global Network does not directly determine data usability for this study. Instead, data quality is evaluated using uncertainty-based criteria,

including relative uncertainty, spectral fitting residual (WRMS), and additional screening parameters, ensuring consistent selection of physically reliable observations."

Comment 3.12

Once the backlog of manually assuring the data is complete, much of the Sky-scan data should be fine, however this table does not show that. I am surprised that there is no assured data at all for any of these sites.

Response 3.12

We acknowledge that the absence of assured data in Table 2 may appear unexpected, particularly given that sky-scan retrievals are generally considered reliable once formally reviewed within the PGN framework. However, in this study, the analysis does not rely on the assured/not-assured classification, as this status reflects the administrative validation process within the PGN, which may not yet be completed for all datasets. Instead, data quality is evaluated using an uncertainty-based quality control framework (Rawat et al., 2025), which applies consistent and quantitative criteria across all observations. The lack of 'assured' classification for DS retrievals reflects current PGN processing status rather than data invalidity and does not preclude their use when uncertainty-based criteria are satisfied. As a result, Table 2 reflects the raw PGN quality flag distribution, while the actual data selection for analysis is based on independent QC metrics. This approach ensures that high-quality observations are retained regardless of their formal assurance status.

Revised Text

Section 2.1.1 (QC Protocol)

To improve the robustness of ground-based HCHO observations used for intercomparison and satellite validation, an uncertainty-based quality control (QC) protocol following the methodological framework of Rawat et al. (2025) was applied to contemporaneous Pandora direct-sun (DS) and sky-scan (SS) observations. DS and SS retrievals were first paired within a 5 min tolerance window. A high-quality reference subset was then defined using Pandora quality flags QF = 0 or 10 for both DS and SS retrievals, and dynamic absolute uncertainty thresholds were calculated separately for DS and SS as the mean plus three standard deviations of the uncertainty in this subset. Matched observations were retained when either both DS and SS absolute uncertainties were below these dynamic thresholds or both relative uncertainties were below 10 %. Additional filters required WRMS < 0.01 for both DS and SS retrievals and, for sky-scan observations, maximum horizontal distance (MHxD) < 20 km when available. Pandora quality flags were subsequently used to classify observations into high-quality (QF = 0, 10), medium-quality (QF = 1, 11), low-quality (QF = 2, 12), and unusable (QF ≥ 20) categories for diagnostic analysis. This procedure reduces the influence of retrieval noise, poor spectral fits, and unfavorable viewing geometry prior to satellite collocation.

Comment 3.13

Are these data only for the OMI overpass time?

Response 3.13

We thank the reviewer for this question. In the revised manuscript, the data are not limited to the exact OMI overpass time. Pandora observations are collocated with satellite measurements using two approaches: nearest-time matching (±2 h) and overpass-centered averaging windows (±30 min, ±1 h, ±2 h). This allows evaluation of temporal representativeness while ensuring consistency with satellite sampling. All analyses are based on temporally collocated observations rather than fixed overpass-only sampling.

Revised Text

Section 2.3 (Collocation Strategy)

"Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing."

Comment 3.14

Line 188: Reword

Response 3.14

This line has been removed due to the substantial revision in the revised manuscript.

Comment 3.15

Line 189-197: No need to type out everything that is already in Table 2.

Response 3.15

These lines are now removed, partly due to the substantial revision in the revised manuscript.

Comment 3.16

Line 190: I think you should rethink your exclusion criteria. There is no official recommendation for excluding bad quality data, however several studies suggest removing data based on uncertainty (Rawat et al 2024).

Response 3.16

We thank the reviewer for this important suggestion and fully agree that uncertainty-based filtering provides a more robust approach than relying solely on Pandora quality flags. In the revised manuscript, we have implemented an uncertainty-based quality control protocol following Rawat et al. (2025). This includes filtering based on relative uncertainty (<10%), spectral fitting residual (WRMS < 0.01), and additional criteria such as MHxD for sky-scan retrievals, ensuring that only physically reliable observations are retained.

Revised Text

Section 3.1 (Implications of Rawat QC Protocol)

"The application of the uncertainty-based quality-control (QC) protocol following Rawat et al. (2025) substantially improves the statistical robustness and physical representativeness of Pandora Level-2 HCHO retrievals across the Southeast Asian network. The QC procedure integrates formal quality flags with independent uncertainty metrics, including relative uncertainty (<10 %), spectral fitting residual (WRMS < 0.01), and spatial representativeness constraints for sky-scan (SS) observations. The filtering primarily targets retrieval artefacts associated with viewing-geometry sensitivity and low signal-to-noise conditions, while preserving the underlying atmospheric variability."

Comment 3.17

Table3: Don't need version numbers in description.

Response 3.17

Revised. The version number is now removed in Table 3.

Table 3. Descriptive statistics of Pandora Level-2 formaldehyde (HCHO) retrieved from Direct-sun (DS) and Sky-scan (SS) observations at selected Southeast Asian stations. Statistics are shown for contemporaneous matched DS–SS pairs before and after applying the Rawat quality control (QC) protocol. Values are reported as mean ± standard deviation (SD), median with interquartile range (IQR; Q1–Q3), and minimum–maximum. All HCHO columns are expressed in units of $\times 10^{15}$ molecules cm^{-2}.

Station	Dataset	Mean ± SD	Median (IQR)	Min–Max	N
(a) Direct-sun (DS) HCHO					
Bangkok	Raw	21.2 ± 7.88	20.6 (16.0–25.7)	0.018–200	80,336
	After QC	21.3 ± 7.83	20.7 (16.1–25.8)	0.018–83.6	79,072
Bandung	Raw	19.6 ± 44.4	15.0 (10.2–21.0)	0.012–988	34,248
	After QC	16.4 ± 8.29	15.3 (10.5–21.1)	0.012–76.9	30,268
Agam	Raw	63.0 ± 166.6	9.93 (6.76–14.0)	0.013–999	35,504
	After QC	9.01 ± 4.07	8.93 (6.40–11.2)	0.013–77.0	18,804
Pontianak	Raw	11.9 ± 7.90	11.7 (8.93–14.4)	0.013–569	25,694
	After QC	11.8 ± 5.17	11.7 (9.00–14.4)	0.013–107	25,095

Station	Dataset	Mean ± SD	Median (IQR)	Min–Max	N
Singapore-NUS	Raw	10.8 ± 6.89	9.33 (6.53–13.2)	0.012–131	39,791
	After QC	10.9 ± 7.23	9.30 (6.36–13.3)	0.012–131	32,455
(b) Sky-scan (SS) HCHO					
Bangkok	Raw	12.8 ± 10.8	11.9 (8.31–16.1)	0.013–956	135,664
	After QC	13.2 ± 5.57	12.6 (9.43–16.3)	0.025–61.3	79,072
Bandung	Raw	13.0 ± 17.6	11.0 (6.67–16.6)	0.014–923	47,161
	After QC	13.1 ± 7.18	11.8 (7.91–17.1)	0.026–116	30,268
Agam	Raw	7.40 ± 26.9	4.39 (2.52–6.78)	0.010–992	40,440
	After QC	4.93 ± 3.07	4.51 (2.92–6.40)	0.010–54.5	18,804
Pontianak	Raw	8.10 ± 15.0	6.75 (4.46–9.77)	0.010–919	36,279
	After QC	8.22 ± 3.82	7.62 (5.53–10.2)	0.027–45.2	25,095
Singapore-NUS	Raw	10.9 ± 11.1	9.02 (6.05–13.4)	0.013–883	61,473
	After QC	11.7 ± 7.34	9.90 (7.11–14.0)	0.024–90.6	32,455

Comment 3.18

Figure 2: What are the colors of the bars? If nothing, make the bars the same color.

Response 3.18

In the revised manuscript, Figure 2 is removed and replaced with Figure 6.

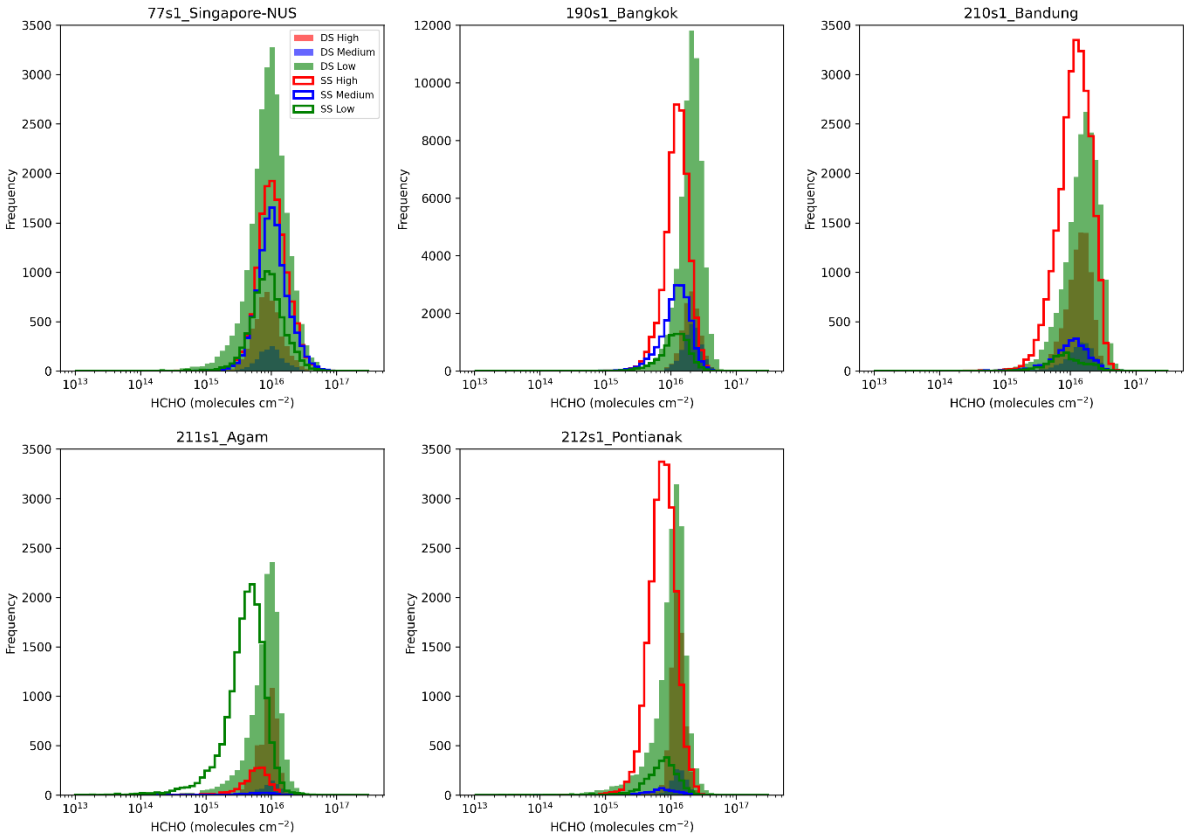


Figure 6. Frequency distributions of filtered HCHO vertical column densities (molecules cm^{-2}) for direct-sun (DS; shaded histograms) and sky-scan (SS; solid line histograms) observations across five stations. Each panel corresponds to a station, with HCHO data grouped by quality flag: high quality (red; QF = 0, 10), medium quality (blue; QF = 1, 11), and low quality (green; QF = 2, 12). The x-axis is shown on a logarithmic scale to capture the wide dynamic range of HCHO values. The y-axis represents the frequency of observations, with consistent limits applied across stations except for Bangkok, which has a higher observation count.

Comment 3.19

Figures 2 and 3: I suggest changing these to a normalized distribution plot. Because some Pandoras have more data than others we would be able to see the differences better.

Comment 3.20

If the message is to compare DS to Sky-scan I also suggest including both results on the same subplot. For example a histogram of the one monitor's DS values in one color and the Sky-scan values in another color on top of that histogram. As it is now, it is hard to compare.

Response 3.19 & 3.20

We thank the reviewer for this helpful suggestion and agree that direct comparison between DS and sky-scan retrievals should be clearly presented. In the revised manuscript, the primary analysis of DS–SS consistency has been updated to use scatter plot (correlation) analysis of temporally matched pairs, which provides a more direct and quantitative assessment of agreement between retrieval modes. As the focus of this study is on evaluating consistency and variability between DS and SS observations, correlation-based diagnostics are considered more appropriate.

Revised Text

Section 3.2 (DS–SS Comparison)

The nine-panel correlation analyses (Figs. 3) reveal that DS–SS agreement depends strongly on retrieval quality category, with the highest correlations observed when both measurements fall within the high-quality regime (QF = 0, 10). Across all stations, high-quality DS–SS pairs exhibit correlation coefficients typically exceeding 0.70, indicating strong agreement between retrieval geometries under well-constrained uncertainty conditions.

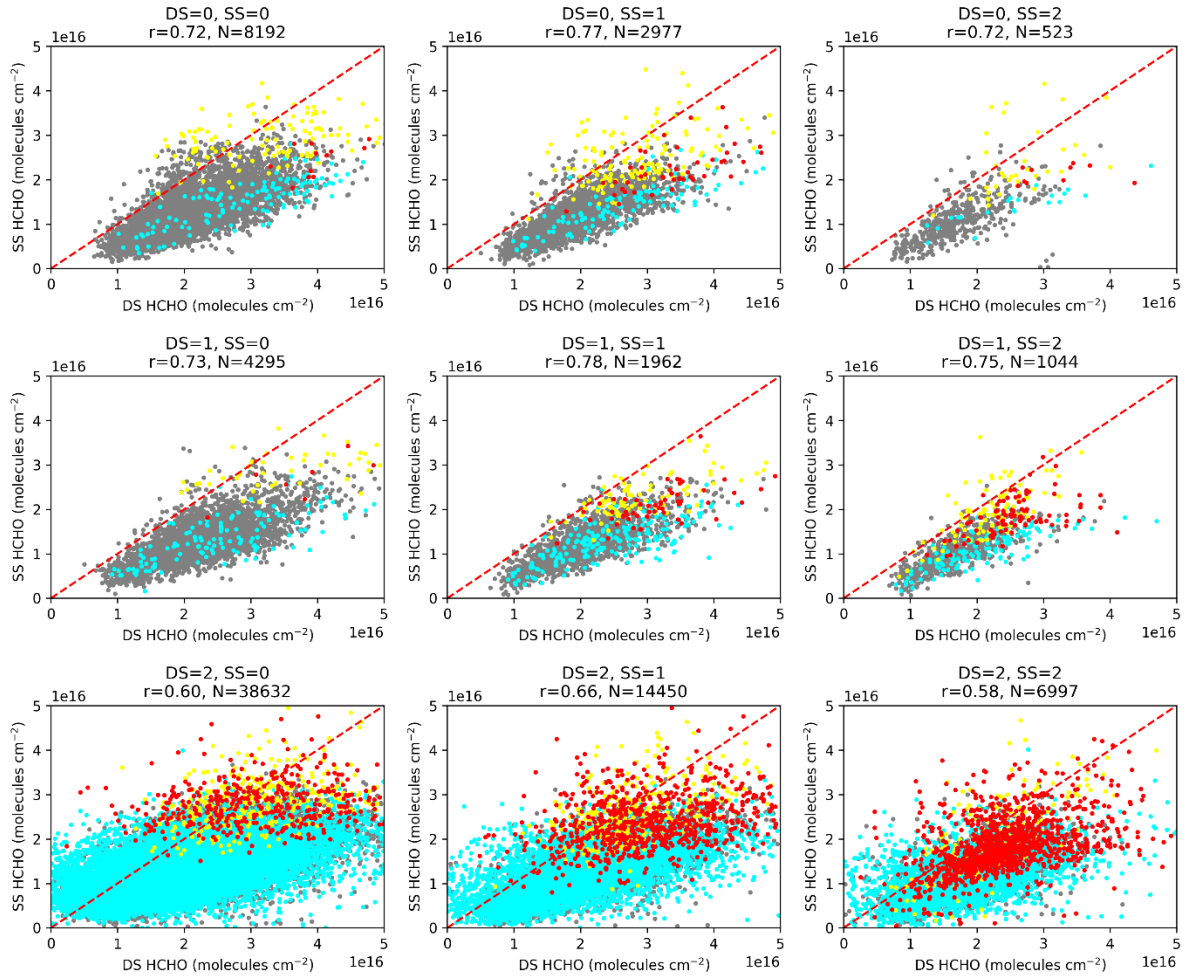


Figure 3. Nine-panel plot of correlation between contemporaneous Pandora HCHO column amounts: direct-sun (DS) vs sky-scan (SS) for each quality category, following the Rawat et al. (2025, AMT) QC method at Bangkok station. Panels are organized by DS and SS quality categories (0 = high, 1 = medium, 2 = low). Each panel shows the scatter of DS vs SS HCHO (molecules cm^{-2}), with points color-coded by uncertainty thresholds: gray = both below cutoff, cyan = DS above cutoff, yellow = SS above cutoff, red = both above cutoff. The red dashed line represents the 1:1 relationship, and the correlation coefficient (r) and number of matched observations (N) are indicated in each panel.

Comment 3.21

Figure 4: In this figure description you change your filtering methods to also remove data above 50e15. This needs to be consistent throughout the entire results and stated in methodology. What is the reasoning behind this number? Instead I suggest filtering based on uncertainty and that would most likely give a similar result.

Response 3.21

We thank the reviewer for this important comment and fully agree that applying an arbitrary fixed threshold (e.g., 50×10^{15} molecules cm^{-2}) is not appropriate without consistent justification. In the revised manuscript, this threshold-based filtering has been removed entirely. All data screening is now performed using an uncertainty-based quality control framework following Rawat et al. (2025),

including criteria based on relative uncertainty, spectral fitting residual (WRMS), and additional quality constraints. This ensures a physically consistent and objective selection of valid observations. The QC approach is now applied consistently across all analyses, and no additional ad hoc thresholds are used.

Revised Text

Section 2.1.1 (QC Protocol)

"To improve the robustness of ground-based HCHO observations used for intercomparison and satellite validation, an uncertainty-based quality control (QC) protocol following the methodological framework of Rawat et al. (2025) was applied to contemporaneous Pandora direct-sun (DS) and sky-scan (SS) observations. DS and SS retrievals were first paired within a 5 min tolerance window. A high-quality reference subset was then defined using Pandora quality flags QF = 0 or 10 for both DS and SS retrievals, and dynamic absolute uncertainty thresholds were calculated separately for DS and SS as the mean plus three standard deviations of the uncertainty in this subset. Matched observations were retained when either both DS and SS absolute uncertainties were below these dynamic thresholds or both relative uncertainties were below 10 %. Additional filters required WRMS < 0.01 for both DS and SS retrievals and, for sky-scan observations, maximum horizontal distance (MHxD) < 20 km when available."

Comment 3.22

You reuse letters in labeling the panels in this figure (and Figs 2-3). Each panel needs a unique letter. I also suggest combining the hourly and daily figures together so we can easily see the differences.

Response 3.22

The panel labeling has been revised so that each subplot is assigned a unique identifier, ensuring clarity and consistency across all figures in the revised manuscript. Regarding the suggestion to combine hourly and daily plots, these figures have been removed, and the result visualization have been reorganized in the revised manuscript to improve readability and to better align with the key analysis objectives. Rather than combining multiple temporal aggregations, the revised figures focus on different collocation strategies, which present the most relevant information clearly.

Revised Text

Section 3.4 (Impacts of Collocation Strategy)

"The impact of temporal collocation strategy differs markedly between OMI and TROPOMI. For OMI (Fig. 8), expanding the collocation window from nearest to ±2 h results in only marginal changes in correlation and error metrics, indicating limited sensitivity to temporal averaging. For instance, at Bangkok, correlation remains nearly unchanged (r = 0.11–0.12), while RMSE (~1.13–1.16 × 10^{16} molecules cm^{-2}) and MAE (~9.0–9.9 × 10^{15} molecules cm^{-2}) show minimal improvement. In contrast,

TROPOMI (Fig. 11) demonstrates clearer benefits from temporal averaging. At Bangkok, the correlation increases from r = 0.23 (nearest) to r = 0.32 (±2 h), accompanied by a reduction in RMSE from 9.35×10^{15} to 8.80×10^{15} molecules cm^{-2} and a decrease in MAE from 7.47×10^{15} to 7.07×10^{15} molecules cm^{-2}. A similar but more subtle improvement is observed at Singapore, where the correlation remains consistently high ($r \approx 0.46–0.47$), while RMSE decreases from 8.34×10^{15} to 6.47×10^{15} molecules cm^{-2} and MAE from 5.42×10^{15} to 4.81×10^{15} molecules cm^{-2} when applying a ±2 h window. These results indicate that TROPOMI retrievals benefit from temporal averaging while maintaining strong correlation, reflecting improved representation of short-timescale variability compared to OMI."

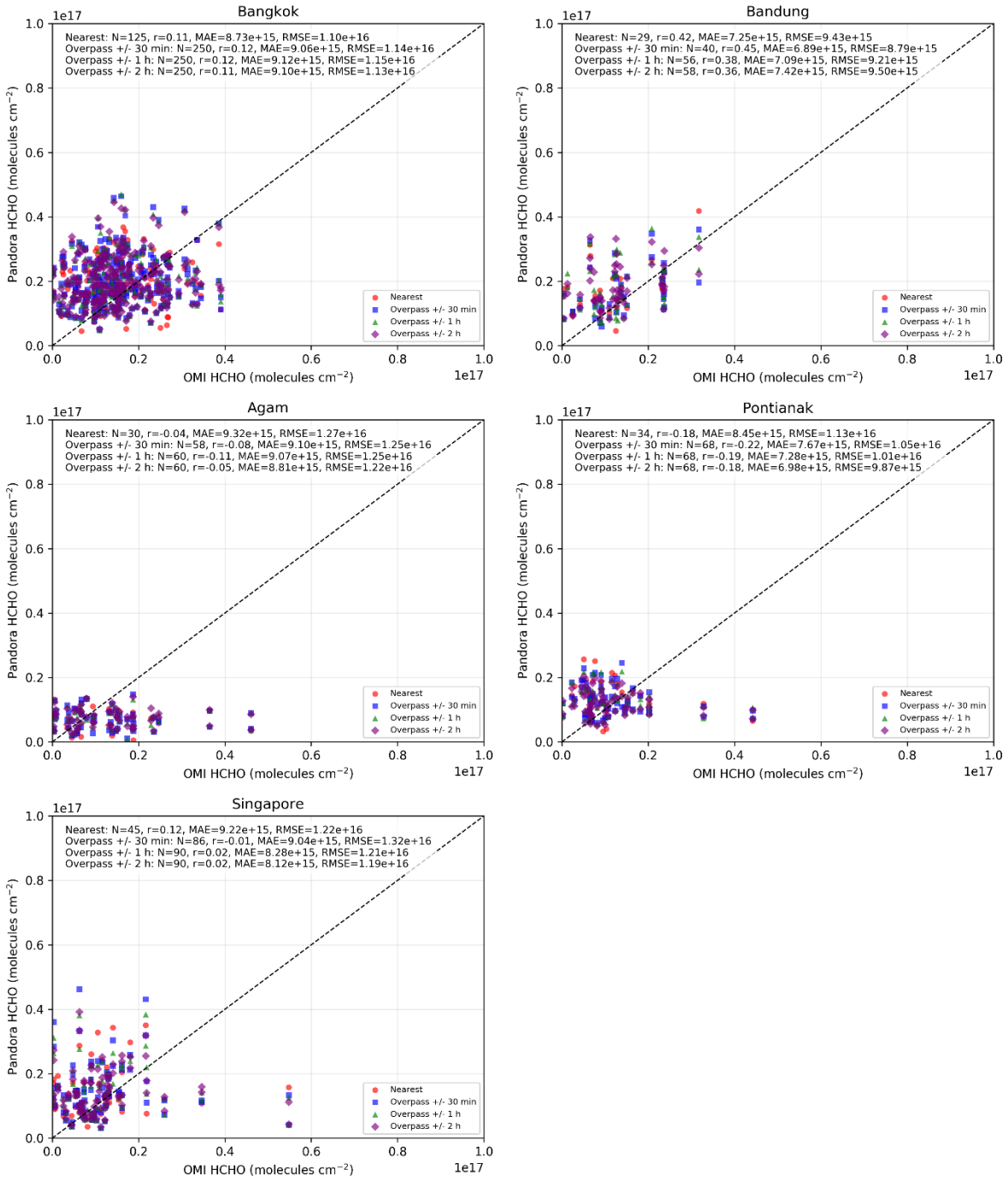


Figure 8. Scatter plots comparing Pandora and OMI HCHO column retrievals for different temporal collocation strategies. Each panel corresponds to a measurement station and includes the 1:1 reference line. Reported statistics include sample size (N), mean absolute error (MAE), root-mean-square error (RMSE), and Pearson correlation coefficient (r).

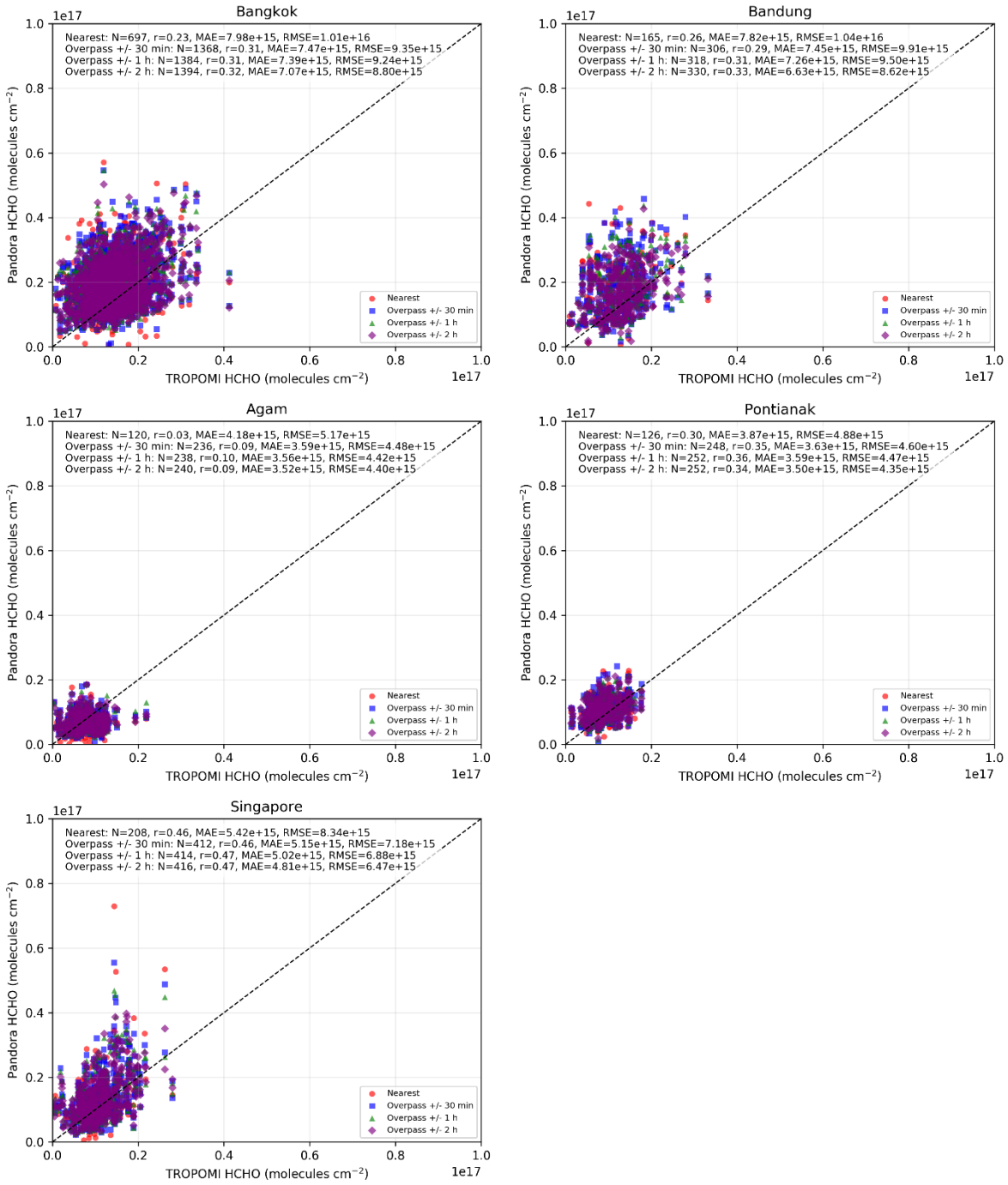


Figure 11. Scatter plots comparing Pandora and TROPOMI HCHO column retrievals for different temporal collocation strategies. Each panel corresponds to a measurement station and includes the 1:1 reference line. Reported statistics include sample size (N), mean absolute error (MAE), root-mean-square error (RMSE), and Pearson correlation coefficient (r).

Comment 3.23

Table 4: Under "remarks" are the distances needed? I don't understand what that is referring to other than the OMI spatial averaging column.

Response 3.23

In the revised manuscript, Table 4 and the associated experimental framework have been removed as part of the redesign of the collocation methodology. The analysis now uses a unified time-based collocation approach, and the ambiguity associated with the previous table has been eliminated.

Comment 3.24

Why noontime if the OMI overpass is closer to 1pm? What are the time windows for "daytime". Pandora does not report data at nighttime.

Response 3.24

We thank the reviewer for this comment and fully agree that fixed time windows such as “noontime” or “daytime” are not appropriate for satellite validation, particularly given the well-defined OMI overpass time and the strong diurnal variability of HCHO. In the revised manuscript, this framework has been removed entirely. Pandora observations are now collocated with satellite data using overpass-centered temporal matching, including nearest-time pairing (±2 h) and symmetric averaging windows (±30 min, ±1 h, ±2 h) around the satellite overpass time. This approach ensures temporal consistency and avoids ambiguity associated with fixed time definitions.

Revised Text

Section 2.3 (Collocation Straetgy)

“Two complementary approaches were applied. First, a nearest-time matching method paired each satellite observation with the closest Pandora measurement within a ±2 h tolerance window. Second, an overpass-window averaging method was used, in which all Pandora observations within symmetric windows centered on the satellite overpass time were averaged to form representative ground-based column estimates. Three temporal windows were tested (±30 min, ±1 h, and ±2 h) to assess sensitivity to temporal smoothing.”

Comment 3.25

Figure 5: x tick times need to be more clear. State month.

Response 3.25

The x-axis labels in Figure 7 and Figure 10 (after revised) have been updated to explicitly include month information, improving clarity and readability of the time series.

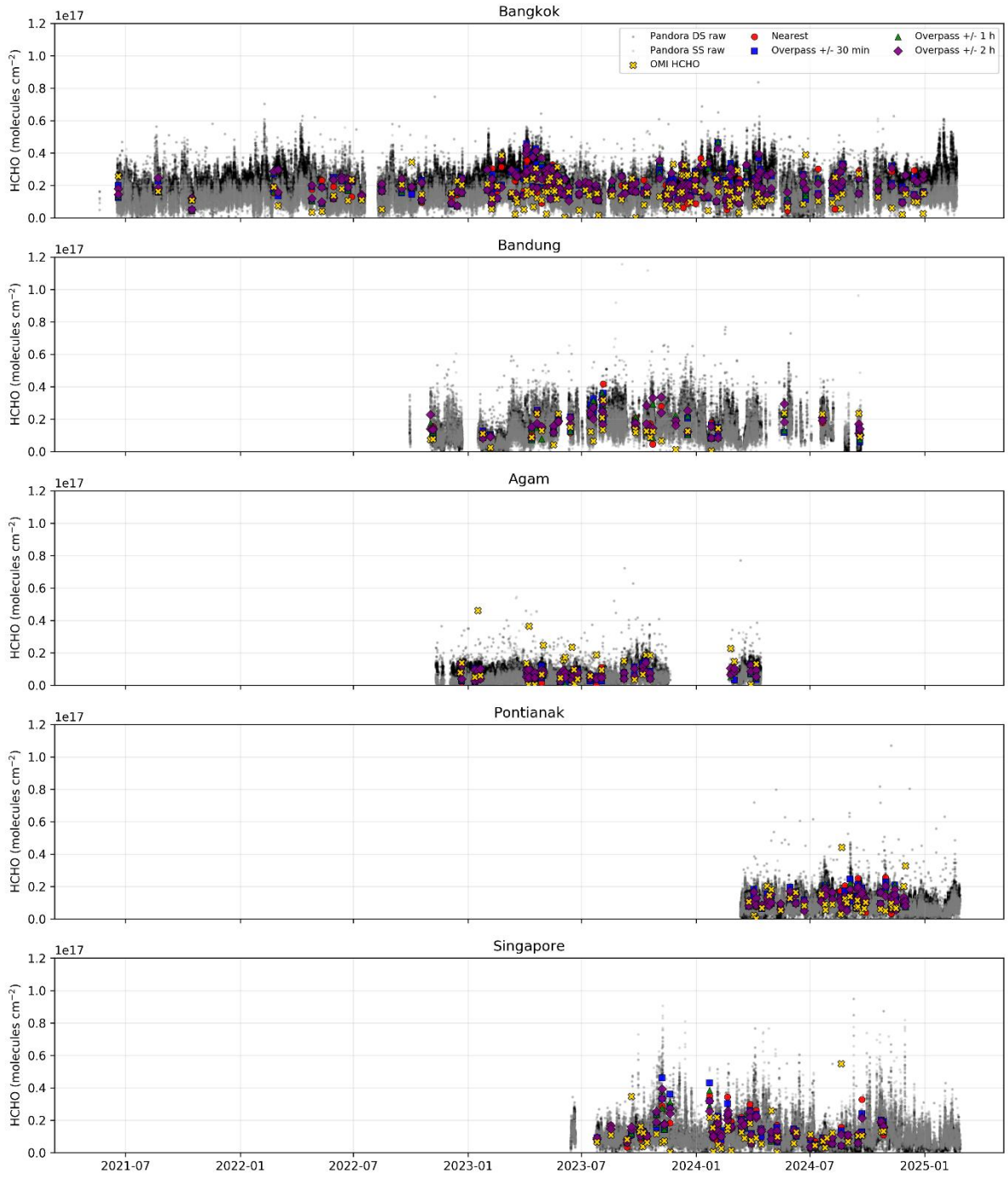


Figure 7. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated OMI observations at five Southeast Asian stations. OMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ±30 min, ±1 h, and ±2 h.

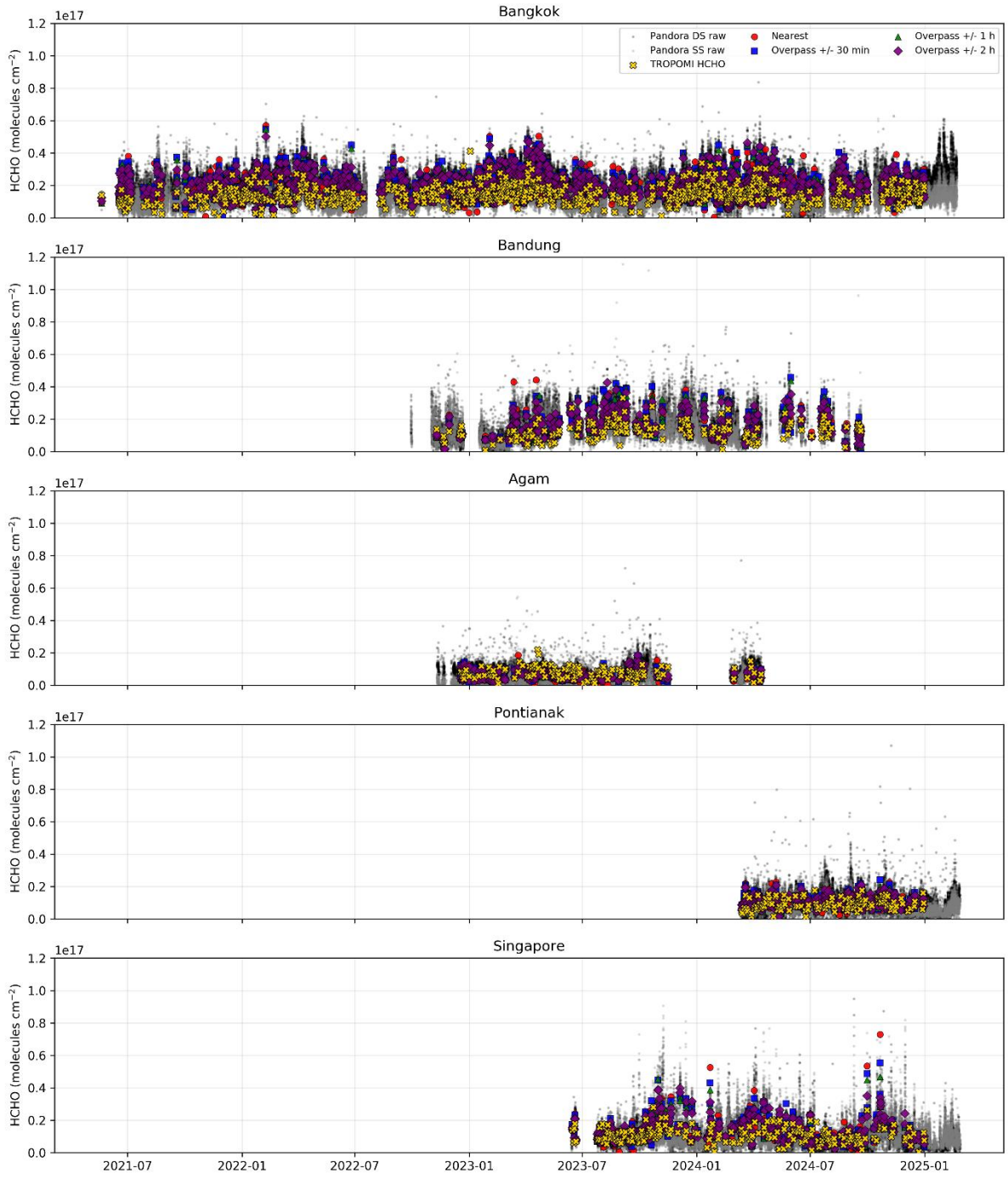


Figure 10. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated TROPOMI observations at five Southeast Asian stations. TROPOMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ±30 min, ±1 h, and ±2 h.

Comment 3.26

Reword the description. The different experiments E1-E9 are referring to OMI not the Pandoras, right?

Response 3.26

We thank the reviewer for this comment and agree that the description of the E1–E9 experiments was unclear. In the revised manuscript, the entire E1–E9 experimental framework has been removed and replaced with a unified, physically consistent collocation methodology. As a result, this ambiguity no longer arises in the revised version.

Comment 3.27

I think you are being a little hasty in determining which experiment is performing best. This needs more discussion in the text. In Table 5 there does not seem to be a clear winner, rather each Pandora monitor works best with a different comparison method.

Response 3.27

We thank the reviewer for this important comment and agree that the identification of a single "best" experiment in the previous analysis was not sufficiently robust. In the revised manuscript, the E1–E9 experimental framework and Table 5 have been removed and replaced with a unified collocation approach based on overpass-centered temporal matching. As a result, the comparison is no longer framed in terms of selecting a single optimal method. Instead, the revised discussion emphasizes how satellite–ground agreement depends on temporal window selection and site-specific conditions, without implying a universal best configuration. This provides a more physically consistent and generalizable interpretation.

Revised Text

Section 3.4 (Impact of Collocation Strategy)

"The impact of temporal collocation strategy differs markedly between OMI and TROPOMI. For OMI (Fig. 8), expanding the collocation window from nearest to ±2 h results in only marginal changes in correlation and error metrics, indicating limited sensitivity to temporal averaging. For instance, at Bangkok, correlation remains nearly unchanged ($r = 0.11–0.12$), while RMSE (~$1.13–1.16 \times 10^{16}$ molecules cm^{-2}) and MAE (~$9.0–9.9 \times 10^{15}$ molecules cm^{-2}) show minimal improvement. In contrast, TROPOMI (Fig. 11) demonstrates clearer benefits from temporal averaging. At Bangkok, the correlation increases from $r = 0.23$ (nearest) to $r = 0.32$ (±2 h), accompanied by a reduction in RMSE from 9.35×10^{15} to 8.80×10^{15} molecules cm^{-2} and a decrease in MAE from 7.47×10^{15} to 7.07×10^{15} molecules cm^{-2}. A similar but more subtle improvement is observed at Singapore, where the correlation remains consistently high ($r \approx 0.46–0.47$), while RMSE decreases from 8.34×10^{15} to 6.47×10^{15} molecules cm^{-2} and MAE from 5.42×10^{15} to 4.81×10^{15} molecules cm^{-2} when applying a ±2 h window.

These results indicate that TROPOMI retrievals benefit from temporal averaging while maintaining strong correlation, reflecting improved representation of short-timescale variability compared to OMI."

Comment 3.28

Figure 6: No need for different colors in this figure.

Response 3.28

This figure is now removed in the revised manuscript.

Comment 3.29

Line 315-319: This figure needs to be discussed in greater detail. I think this could be an important figure to show the comparison between OMI and Pandora HCHO distributions.

Response 3.29

We thank the reviewer for this helpful suggestion and agree that this figure provides important insight into the comparison between Pandora and satellite HCHO observations. In the revised manuscript, the discussion has been substantially expanded. For the DS–SS comparison, we provide a more detailed interpretation of the distributional characteristics, highlighting how differences in distribution shape and spread reflect retrieval quality and sampling-related variability.

For the satellite comparison, the analysis has been fully redesigned to provide a more comprehensive and physically grounded evaluation. Specifically, we now assess Pandora–satellite consistency using multiple complementary diagnostics, including:
(i) time-series comparison,
(ii) correlation analysis (DS and SS combined),
(iii) bias as a function of Pandora short-timescale variability,
(iv) separate correlation analysis for DS and SS retrievals.
This multi-dimensional framework provides a more complete interpretation of satellite–ground agreement and allows clearer attribution of discrepancies to temporal sampling and spatial representativeness effects.

Revised Text

Section 3.3 (Distributional characteristics DS-SS retrievals)

"The frequency distributions highlight important station-dependent differences in DS–SS consistency. At the Pandora Singapore-NUS station, DS and SS distributions largely overlap across all QF categories, indicating strong consistency between retrieval geometries irrespective of quality classification. However, this behaviour is not observed at the other stations. At Bangkok, Bandung,

Pontianak, and Agam, SS retrievals are systematically skewed towards lower HCHO values compared to DS, particularly in the medium- and low-quality regimes. This systematic shift suggests a geometry-dependent bias, where SS measurements tend to underestimate column HCHO relative to DS under less favourable retrieval conditions. Such discrepancies are reduced in the high-quality category but remain evident overall, highlighting the importance of quality filtering when combining DS and SS observations for quantitative analysis."

Section 3.4 Impact of temporal collocation and sub-pixel variability on OMI and TROPOMI validation

Section 3.5 Role of High-Temporal-Resolution Observations: Insights from GEMS HCHO Retrievals

Comment 3.30

Figures 7and 8: combine into one showing E1 in one color, and E2/E8 in another for easier comparison. Are these the daily comparisons? Noontime?

Response 3.30

We thank the reviewer for this suggestion and agree that the previous figure design and labeling were not sufficiently clear. In the revised manuscript, Figures 7 and 8 and the associated E1–E9 experimental framework have been removed as part of the redesign of the analysis. The satellite–Pandora comparison is now presented using a unified and physically consistent approach, including time-series comparison, correlation analysis, and overpass-centered temporal matching, which avoids the ambiguity associated with the previous experiment-based figures. As a result, the issues related to figure combination and unclear temporal definitions (e.g., daily vs. noontime) no longer arise in the revised version.

Comment 3.31

lines 320-325: More discussion on this figure as well. This seems to be the main point of the paper yet only a few sentences.

Response 3.31

We thank the reviewer for this important comment and agree that this figure represents a central result of the study and requires more detailed discussion. In the revised manuscript, the interpretation of this figure has been substantially expanded to provide a clearer and more comprehensive explanation of the differences between DS and SS retrievals in the context of satellite comparison. The discussion now explicitly addresses how differences in variability, correlation strength, and bias reflect the roles of temporal sampling and spatial representativeness, and how these factors influence the relative

performance of DS and SS observations in satellite validation. After revision, this figure is replaced by separate correlation analysis for DS and SS retrievals in Figure 13 and Figure 14.

Revised Text

Section 3.4 (Comparison of DS-SS Pandora and Satellite HCHO)

"A refined comparison between direct-sun (DS) and sky-scan (SS) retrieval geometries (Figs. 13–14) indicates that their relative performance depends strongly on the statistical metric considered. DS retrievals consistently exhibit higher correlation with TROPOMI, particularly at urban-influenced sites (e.g. Singapore: DS r ≈ 0.50–0.52 vs SS r ≈ 0.43–0.44; Bangkok: DS r up to ≈ 0.51 vs SS r up to ≈ 0.32), reflecting their stronger sensitivity to short-term variability and localized variability. However, SS retrievals can simultaneously achieve lower error magnitudes, as demonstrated at Bangkok where SS exhibits reduced RMSE and MAE compared to DS despite slightly lower correlation. Quantitatively, this improvement is substantial, with RMSE reduced by ~10–30 % and MAE by ~5–20 % in SS relative to DS depending on the collocation window, indicating a more consistent agreement in absolute column magnitude. This apparent inconsistency arises from differences in spatial representativeness: DS measurements sample a narrow atmospheric column and therefore capture fine-scale variability that enhances correlation but increases mismatch with the spatially averaged satellite pixel, whereas SS retrievals integrate multiple viewing directions and better approximate the satellite footprint, leading to reduced RMSE and MAE. This behaviour is most pronounced in urban environments with strong spatial gradients, while in low-HCHO regions such as Agam and Pontianak, SS retrievals show comparable or slightly improved agreement across both correlation and error metrics (e.g. Pontianak: SS r ≈ 0.40 vs DS r ≈ 0.39, with RMSE ~3.4 × 10^{15} vs ~5.1 × 10^{15} molecules cm^{-2}). Overall, these results demonstrate that DS retrievals are not universally superior; rather, DS and SS provide complementary strengths, with DS better capturing temporal variability and SS offering improved spatial representativeness for satellite validation in heterogeneous tropical environments."

Comment 3.32

Section 4.1: This discussion and figure 9 should still be under results. You are presenting new information.

Response 3.32

We thank the reviewer for this suggestion. We agree that the distinction between results and discussion should be clearly maintained. In the revised manuscript, the content associated with this analysis has been reorganized to ensure that the presentation of results and their interpretation are appropriately separated. The quantitative results are now presented within the Results section, while the Discussion section focuses on the interpretation and broader implications of these findings. After revision, this

figure is now replaced by Figure 9 and Figure 12, which show the bias as a function of Pandora short-timescale variability.

Revised Text

Section 3.4 (Bias-variability Relationship)

"The bias–variability relationships (Figs. 9 and 12) further highlight fundamental differences in retrieval behaviour. For OMI, correlations between Pandora sub-daily variability and absolute bias are generally weak or inconsistent (e.g. r = −0.07 to 0.08 at Bangkok, r = −0.34 to −0.19 at Agam), indicating that OMI errors are not strongly linked to local temporal heterogeneity. In contrast, TROPOMI exhibits clearer and more physically consistent relationships, particularly at Singapore (DS: r = 0.33; SS: r = 0.63) and Pontianak (DS: r = 0.35; SS: r = 0.27), where increased short-timescale variability leads to larger satellite–ground discrepancies. Moreover, TROPOMI maintains lower overall error magnitudes compared to OMI, with RMSE typically below ~9×10^{15} molecules cm^{-2} and MAE below ~5×10^{15} molecules cm^{-2} at most sites. These results indicate that while TROPOMI remains sensitive to sub-pixel variability, its errors are more physically interpretable and systematically linked to atmospheric heterogeneity, whereas OMI discrepancies are dominated by coarse spatial resolution and representativeness limitations."

Comment 3.33

Figure 10: I don't think a and b are necessary. the colors and markers on this figure are difficult to read. Instead of SZA versus HCHO column, try SZA versus uncertainty.

Response 3.33

We thank the reviewer for this helpful suggestion. We agree that the original SZA-based analysis and figure presentation were not optimal for addressing the main objectives of the study. In the revised manuscript, the SZA-dependent analysis has been removed and de-emphasized, and the associated figure has been replaced by diagnostics that more directly reflect retrieval uncertainty and representativeness effects. As a result, Figure 10 is now removed and the issues related to panel labeling, marker readability, and the choice of SZA-based visualization no longer arise.

Comment 3.34

General/Conclusions:

Why not include GEMS in this analysis? The time range should be perfect and there are several figures showing the hourly Pandora columns.

Response 3.34

We thank the reviewer for this important comment and fully agree on the relevance of GEMS for this study, particularly given its high-temporal-resolution observations over Southeast Asia. In the revised manuscript, GEMS has been explicitly included in the analysis, alongside OMI and TROPOMI. The comparison framework has been expanded to incorporate GEMS in the time-series analysis, correlation assessment, and variability diagnostics. The inclusion of GEMS allows us to better evaluate the role of temporal sampling and diurnal variability, which is particularly important in tropical environments.

Revised Text

Section 3.5 Role of High-Temporal-Resolution Observations: Insights from GEMS HCHO Retrievals

Comment 3.35

Further development into understanding the data quality (and the data products) is necessary. For example, it is not made clear in the manuscript that DS is total column, while sky-scan is only the lower portion of the troposphere. The Pandoras are also pointing in different directions throughout the day in the DS mode.

Response 3.35

We thank the reviewer for this important comment and agree that clearer description of the Pandora retrieval characteristics is necessary. In the revised manuscript, we have clarified that direct-sun (DS) retrievals represent total column HCHO along the solar beam, while sky-scan (SS) retrievals represent a tropospheric column derived from multi-angle scattered radiation measurements, with sensitivity dependent on retrieval configuration and atmospheric conditions. We have also added clarification that DS observations are obtained by tracking the Sun throughout the day, resulting in changing viewing geometry and sampling direction, which can introduce additional variability in the observed column due to spatial heterogeneity. In contrast, SS observations provide a more spatially integrated measurement over a broader field of view.

Revised Text

Section 2.1 (Pandora Data Description)

"Direct-sun (DS) retrievals provide total column HCHO along the solar beam, whereas sky-scan (SS) retrievals represent a tropospheric column derived from multi-angle scattered radiation measurements, with sensitivity that depends on retrieval configuration and atmospheric conditions."

Section 4 (Discussion – strengthened interpretation)

"The application of uncertainty-based quality control improves the robustness of Pandora observations, while the separation of DS and SS retrievals reveals complementary strengths in capturing variability

and spatially representative column structure. The observed differences between DS and SS retrievals reflect the interplay between measurement geometry and atmospheric heterogeneity, with DS capturing localized variability and SS providing a more spatially integrated representation of the atmospheric column."

Comment 3.36

Several figures could be removed/combined.

Response 3.36

In the revised manuscript, the figures have been carefully reviewed and reorganized. Redundant figures have been removed or consolidated. The main text now focuses on the key figures that directly support the primary conclusions.

Comment 3.37

This paper is presented as a comparison between Pandora and OMI, however the methods used for the comparison require further justification.

Response 3.37

We thank the reviewer for this important comment and agree that the comparison methodology required clearer justification in the original manuscript. In the revised manuscript, the comparison framework has been fully redesigned and explicitly justified. The analysis is now based on:

(i) an uncertainty-based quality control protocol (Rawat et al., 2025),
(ii) a rigorous DS–SS intercomparison using temporally matched pairs (instead of hourly or daily averages), and
(iii) a physically consistent collocation approach based on nearest-time matching and overpass-centered averaging.

In addition, the study has been expanded to include TROPOMI and GEMS, providing a multi-sensor context that strengthens the interpretation of satellite–ground differences. The revised methodology is now clearly described and consistently applied throughout the manuscript.