

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 ($QF \geq 20$) 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 $qa_value \geq 0.5$, cloud fraction $cloud_fraction_crb < 0.3$, and $SZA < 60^\circ$ (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 $SZA < 60^\circ$ (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).”

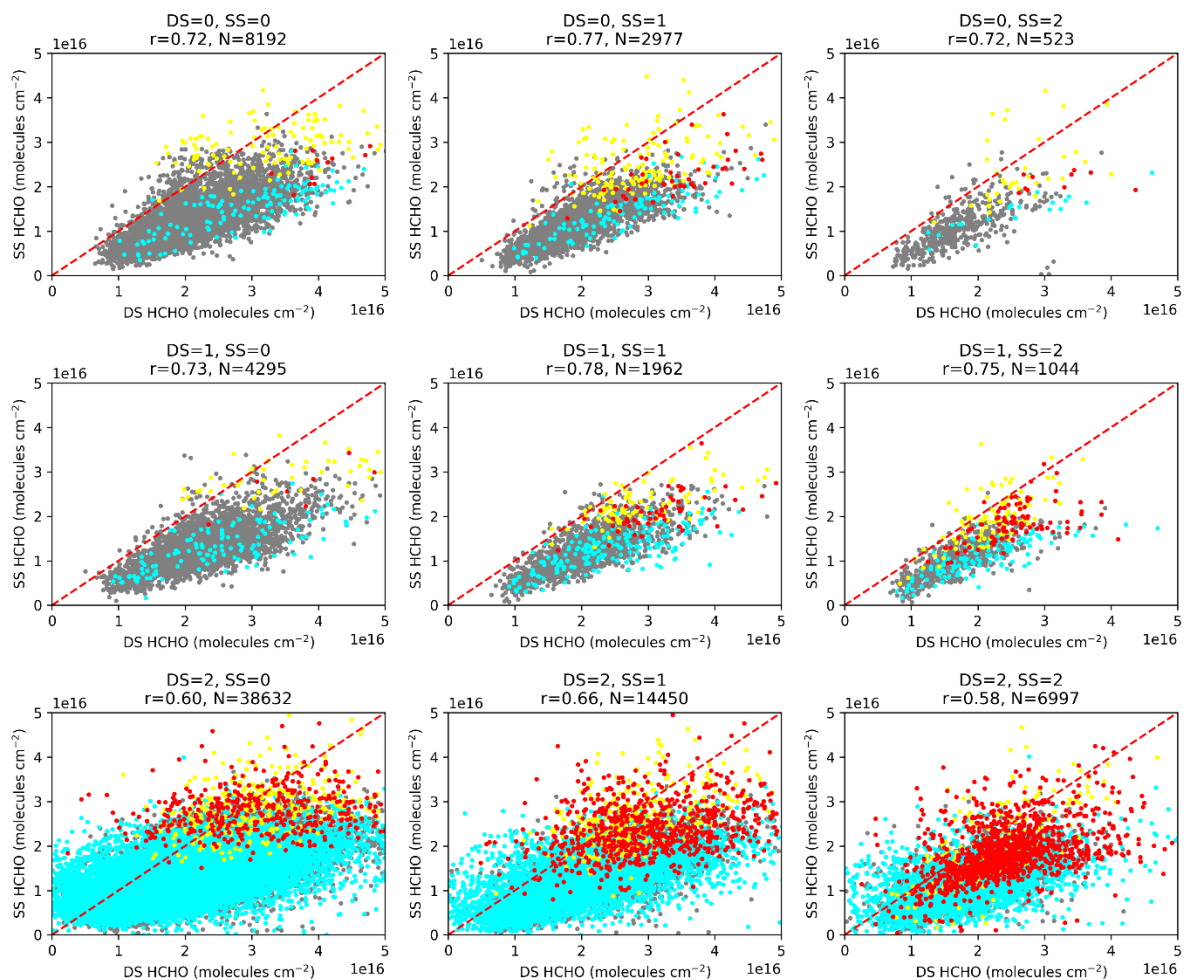


Figure 3. Nine-panel plot of correlation between contemporaneous Pandora HCHO column amounts: direct-sun (DS) vs sky-scan (SS) for each quality category, following the Rawat et al. (2025, AMT) QC

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)

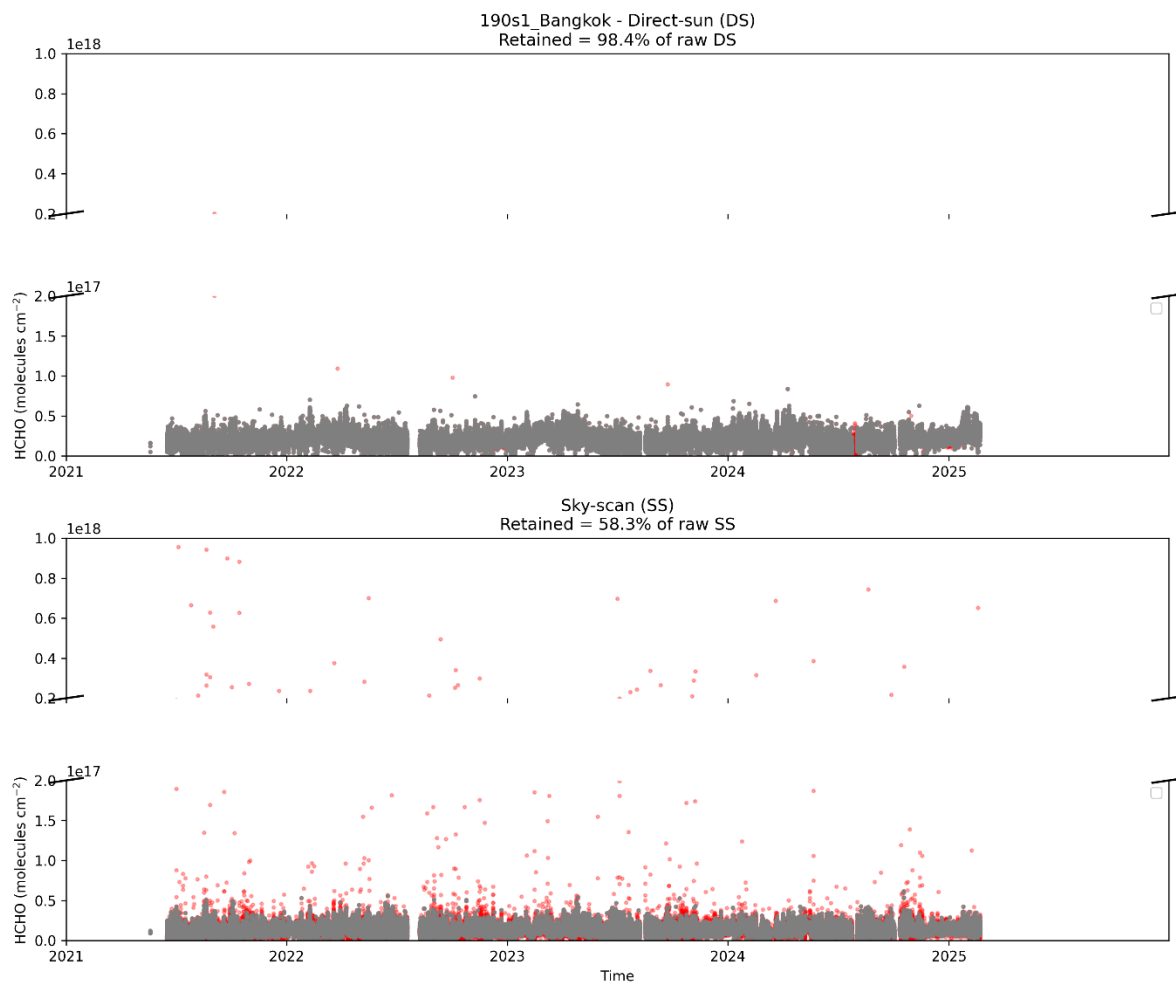


Figure 4. Time series of Pandora HCHO column amounts for direct-sun (DS) and sky-scan (SS) measurements following the Rawat et al. (2025, AMT) QC method at the Pandora Bangkok station. The upper panels show DS HCHO, and the lower panels show SS HCHO. Removed data points failing quality control (QC) are 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.

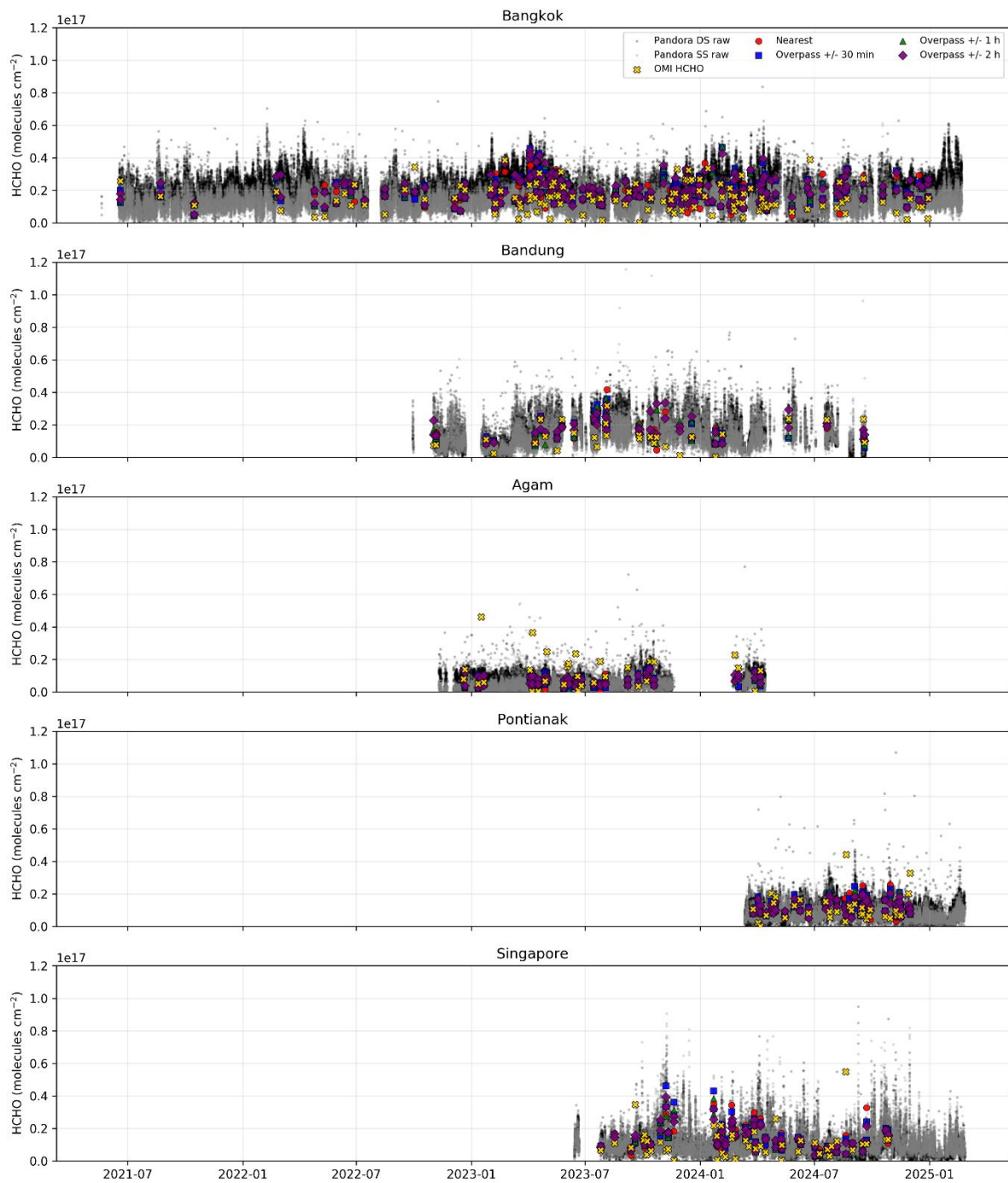


Figure 7. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated OMI observations at five Southeast Asian stations. OMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ± 30 min, ± 1 h, and ± 2 h.

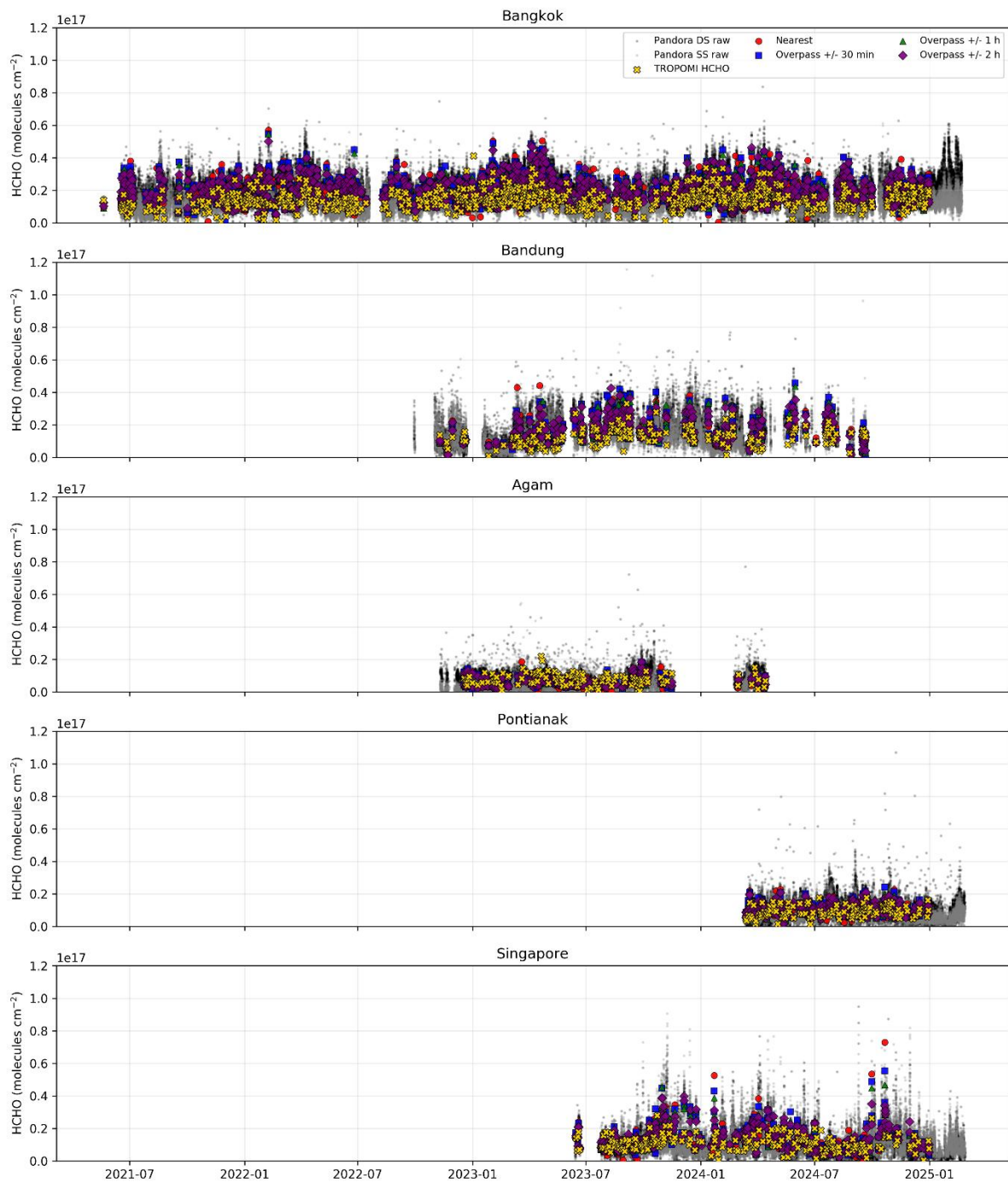


Figure 10. Time series of Pandora HCHO column measurements (DS and SS) and temporally collocated TROPOMI observations at five Southeast Asian stations. TROPOMI-Pandora data are shown for four collocation approaches: nearest-time matching and overpass-centred averaging windows of ± 30 min, ± 1 h, and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

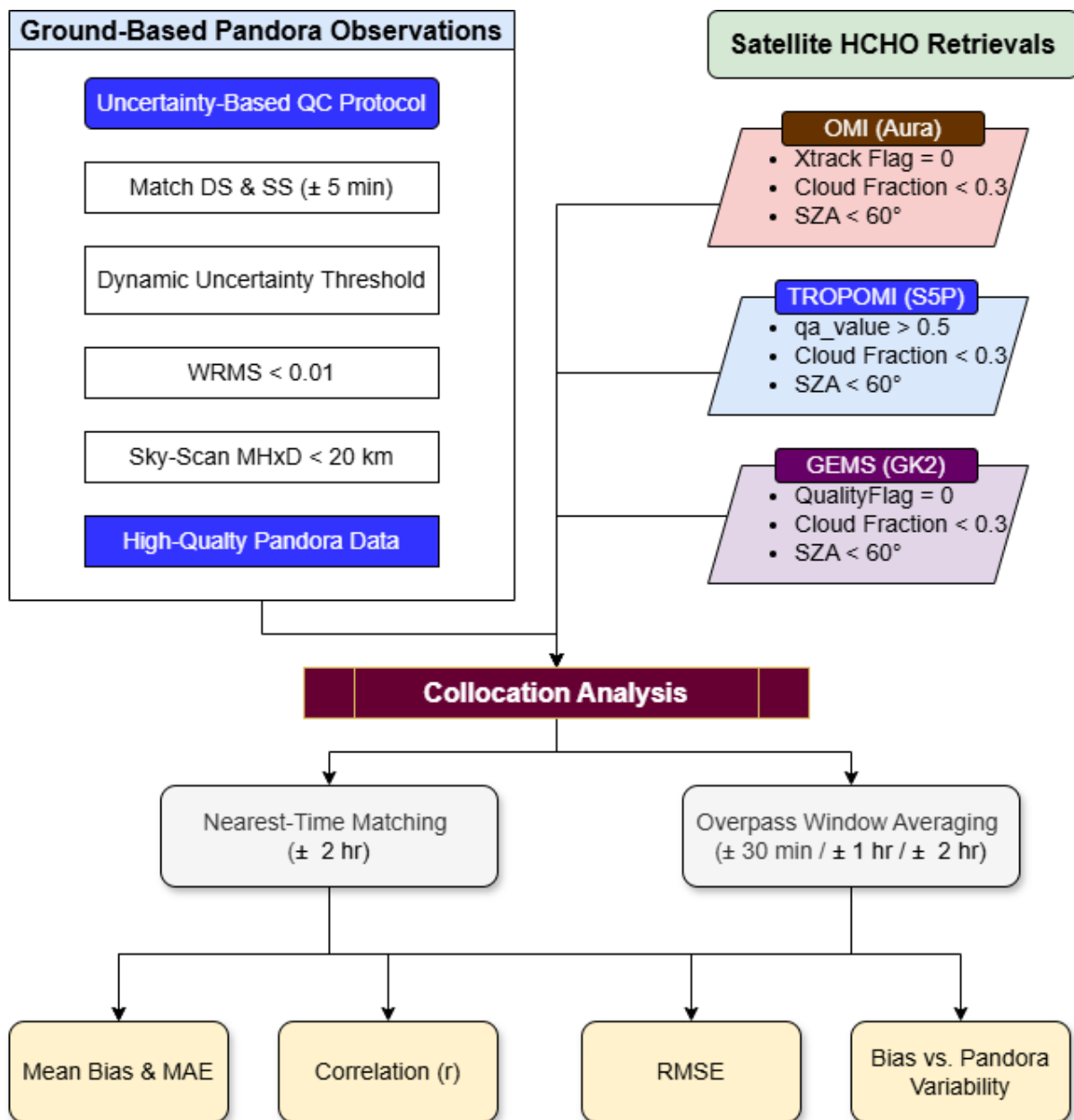


Figure 2. Flowchart illustrating the satellite–Pandora HCHO validation framework applied in this study. The methodology includes uncertainty-based quality control of Pandora observations following [Rawat et al. \(2025\)](#), standard quality screening of OMI, TROPOMI and GEMS retrievals, temporal collocation 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.