International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(Research Article)

Check for updates

Satellite-based tropospheric pollutant monitoring using sentinel-5P: An end-to-end processing and analysis framework

Sayali Shripad Mahulkar * and Abhay Kamlakar Khamborkar

Department of Statistics, Institute of Science, R.T.Road, Civil Lines, Nagpur, Maharashtra, India 440001.

International Journal of Science and Research Archive, 2026, 19(02), 055-066

Publication history: Received on 25 March 2026; revised on 30 April 2026; accepted on 02 May 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.2.0954

Abstract

Air quality monitoring has traditionally relied on sparse ground-based measurement networks, limiting spatial coverage and accessibility in data-poor regions. Satellite remote sensing, particularly the Sentinel-5 Precursor (S5P) mission's Tropospheric Monitoring Instrument (TROPOMI), offers unprecedented opportunities for global atmospheric composition monitoring at high spatial and temporal resolution. This paper presents a comprehensive methodological framework for acquiring, processing, and integrating Sentinel-5P satellite data for air quality estimation. We detail the complete workflow from data discovery and download through the Copernicus Data Space Ecosystem (CDSE) API, extraction and preprocessing of Level-2 geophysical products, integration with reanalysis meteorological data, physical conversion of satellite column measurements to surface-level concentrations, and rigorous quality assurance procedures. Specific attention is given to the critical role of Planetary Boundary Layer Height (PBLH) from ERA5 reanalysis in converting satellite-observed atmospheric column densities to ground-level pollutant concentrations. This framework demonstrates how open-access satellite data and modern cloud-based processing infrastructure enable researchers in all regions to access high-quality air quality data without dependence on costly proprietary systems. The methodology is globally scalable and applicable to multiple pollutant species, including nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3), making it an essential resource for air quality researchers, environmental scientists, and policy makers seeking to understand and address air pollution on a global scale [5][13].

Keywords: Sentinel-5P; TROPOMI; Air Quality; Satellite Remote Sensing; Data Acquisition; Atmospheric Column; Boundary Layer Height; Air Pollution Monitoring

1. Introduction

1.1. The Global Data Gap in Air Quality Monitoring

Air pollution represents one of the world's most pressing environmental health challenges. The World Health Organization estimates that approximately 7 million premature deaths annually are attributable to outdoor air pollution exposure, with particulate matter and trace gases contributing substantially to this burden. However, despite this enormous public health impact, air quality monitoring infrastructure remains severely inequitable globally[12]. While developed nations maintain dense ground-based monitoring networks, approximately 36% of countries lack any systematic air pollution monitoring capability. Nine out of ten individuals living in the world's nearly one billion least developed regions have no access to air quality information.

This global monitoring gap persists because traditional ground-based monitoring requires substantial capital investment, ongoing operational costs, technical expertise, and institutional infrastructure that many countries, particularly in low- and middle-income regions, cannot sustain [15]. As a result, knowledge about air quality patterns,

* Corresponding author: Sayali Shripad Mahulkar

pollution sources, and health exposure remains limited precisely in regions experiencing the most severe air quality challenges.

1.2. The Sentinel-5P Mission and TROPOMI Instrument

The European Union's Copernicus Programme represents a transformative shift toward open-access Earth observation data. Launched on October 13, 2017, the Sentinel-5 Precursor (S5P) satellite carries the Tropospheric Monitoring Instrument (TROPOMI), a cutting-edge multispectral spectrometer designed specifically for atmospheric composition monitoring. TROPOMI provides daily global observations of key atmospheric constituents, including nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), methane (CH_4), and formaldehyde (HCHO), along with aerosol properties [1], [7].

The TROPOMI instrument represents a significant advancement over previous atmospheric monitoring satellites through its combination of high spatial resolution (5.5 km × 3.5 km for most products, with capabilities approaching 1 km^2 resolution for some applications), global daily coverage (2,600 km swath width), and comprehensive spectral coverage spanning ultraviolet, visible, near-infrared, and shortwave-infrared wavelengths[5]. These characteristics make TROPOMI particularly valuable for identifying air pollution sources, tracking pollutant dispersion, and quantifying atmospheric composition at scales relevant to urban and regional air quality decisions.

1.3. Data Accessibility and the Copernicus Ecosystem

A critical advantage of Sentinel-5P data is its complete open-access distribution through the Copernicus Data Space Ecosystem (CDSE), a modern cloud-based infrastructure operated by the European Commission. Unlike proprietary satellite systems requiring expensive licenses, Sentinel-5P data are freely available to any researcher globally through standardized Application Programming Interfaces (APIs) that integrate seamlessly with contemporary data processing workflows. This democratisation of access represents a fundamental shift, enabling researchers in resource-constrained settings to access the same high-quality data as institutions in wealthy nations [3].

1.4. Objective and Scope

This paper presents a detailed methodological framework for acquiring and processing Sentinel-5P TROPOMI data for air quality applications. The framework addresses a critical challenge in satellite-based air quality monitoring: converting satellite measurements of total atmospheric column density (molecules/m^2) into ground-level pollutant concentrations ($\mu g/m^3$) that are directly relevant to environmental health and regulatory standards. The complete workflow encompasses data discovery, acquisition, extraction, temporal and spatial alignment with meteorological data, physical conversion to surface concentrations, and comprehensive quality assurance procedures [12],[7].

By documenting this end-to-end methodology in detail, the paper aims to: (1) reduce barriers to entry for researchers new to satellite-based air quality monitoring, (2) standardize processing procedures to enable data comparability across studies and regions, (3) highlight critical data integration steps that connect satellite observations to ground-level reality, and (4) demonstrate how open-access data infrastructure can democratize access to air quality information globally [11],[16].

2. Sentinel-5P TROPOMI System Specifications and Data Products

2.1. Mission Architecture and Orbital Characteristics

The Sentinel-5 Precursor satellite operates in a polar, sun-synchronous orbit at an altitude of approximately 824 km, completing orbits with a mean local solar time equator-crossing time of 13:30 (±15 minutes). This orbital geometry ensures consistent observation geometry and stable overpass timing, critical for producing comparable time series data. The sun-synchronous nature of the orbit maintains consistent solar illumination angles across seasons, reducing systematic variations in measurement geometry [13].

The TROPOMI instrument is a push-broom grating spectrometer operated in nadir-viewing geometry, observing Earth's atmosphere directly below the satellite. This configuration provides a constant cross-track swath width of approximately 2,600 km, achieving global coverage with daily frequency at the equator. At higher latitudes, orbital overlap provides additional revisit opportunities, with polar regions observable multiple times per day.

2.2. Spectral Characteristics and Instrument Design

TROPOMI's spectral coverage spans seven discrete wavelength bands strategically selected to maximize measurement capability across target atmospheric constituents:

- **UV-1 Band** (270–300 nm): Ultraviolet measurements for ozone retrieval
- **UV-2 Band** (300–320 nm): Extended ultraviolet for trace gas detection
- **Visible Band** (350–500 nm): Primary wavelength for NO_2 and HCHO retrieval
- **NIR-1 Band** (675–775 nm): Near-infrared for water vapor and surface properties
- **NIR-2 Band** (745–773 nm): Additional near-infrared for molecular oxygen
- **SWIR-1 Band** (2305–2385 nm): Shortwave-infrared for CO and CH_4
- **SWIR-2 Band** (2340–2390 nm): Extended shortwave-infrared measurements

This diverse spectral range enables simultaneous retrieval of multiple atmospheric constituents from single observations. Each spectral band is characterized by its unique response function, continuously monitored through on-board calibration sources maintained since launch to track instrument degradation and maintain long-term radiometric consistency [17].

2.3. Spatial Resolution and Data Product Levels

TROPOMI Level-1B (L1B) data products provide georeferenced raw radiance measurements at the instrument's native pixel resolution. Standard ground pixel dimensions are 5.5 km (along-track) × 3.5 km (cross-track), though processing improvements implemented in August 2019 gradually enhanced along-track resolution to approach near-native TROPOMI sampling. Specialized product configurations and processing enhancements have enabled scientific research products at resolutions approaching 1 km^2resolution in select applications. Level-2 (L2) data products represent the primary deliverables for air quality research. These products provide geophysical quantities directly derived from spectral analysis, including:

- **Tropospheric Column Densities:** Total atmospheric burden (molecules/m^2) of target pollutants integrated from ground level to the top of the troposphere
- **Stratospheric Column Densities:** Upper atmospheric contributions, essential for calculating tropospheric-only values
- **Quality Assurance Flags:** Per-pixel confidence metrics (0–1 scale) indicating retrieval reliability.
- **Ancillary Parameters:** Supporting information, including air mass factors, cloud properties, and surface reflectance characteristics
- **Spatial Coordinates:** Latitude and longitude for each pixel center with associated uncertainty estimates

Two processing streams produce L2 products: Near-Real-Time (NRT) products available within 2–3 hours of observation, and Offline (OFFL) products available within 5–7 days with complete quality control and retrieval refinements applied. For research requiring the highest data quality and consistency, OFFL products are strongly preferred, as they incorporate comprehensive air mass factor corrections, advanced cloud screening, and full quality flagging procedures.

2.4. Pollutant-Specific Data Products

TROPOMI delivers four core offline Level-2 products essential for air quality analysis: OFFL_L2__NO2 provides tropospheric nitrogen dioxide columns (nitrogen_dioxide_tropospheric_column, mol/m^2), serving as a sensitive tracer for urban traffic emissions and photochemical smog formation; OFFL_L2__SO2 captures total vertical sulfur dioxide columns (sulfur_dioxide_total_vertical_column, mol/m^2), crucial for identifying industrial stacks and volcanic activity; OFFL_L2__CO monitors total atmospheric carbon monoxide columns (carbonmonoxide_total_column, mol/m^2), indicating combustion processes from vehicles and biomass burning; and OFFL_L2__O3 measures total ozone columns (ozone_total_vertical_column, mol/m^2), distinguishing stratospheric depletion from tropospheric pollution enhancement.

Each product is delivered as a NetCDF-4 hierarchical file containing measurements from a single satellite overpass (approximately 10-minute observation window). The hierarchical structure separates measurement data, quality metrics, and supporting information into organized groups, enabling efficient data access and processing [4] [10]

3. Data Access Infrastructure: Copernicus Data Space Ecosystem

3.1. Platform Components

The Copernicus Data Space Ecosystem (CDSE) transitions from legacy FTP systems to a modern cloud-native architecture using REST APIs and OAuth 2.0 authentication. It integrates three essential components:

- Catalogue API: Employs OData query syntax for product discovery, supporting spatial filtering (bounding boxes/polygons), temporal ranges, and product metadata across all Sentinel missions. Users can verify data availability before bulk downloads.
- Download API: Provides secure file access with intelligent load balancing and automatic retry mechanisms. Typical download speeds reach 100–500 MB/s on standard broadband, enabling efficient processing of multi-year datasets.
- Visualization API: Supports on-demand data previewing, subset extraction, and quality validation—ideal for confirming spatial/temporal coverage before committing to full file downloads.

3.2. Authentication Workflow

Registration at the Copernicus Data Space Ecosystem portal provides free Client ID/Secret credentials for OAuth 2.0 token generation (24-hour validity). The standard workflow follows:

- Client Credentials Flow: Exchange Client ID/Secret for bearer token via CDSE authentication service
- Token Usage: Apply bearer token to Catalogue/Download API calls
- Token Refresh: Automatically renew expired tokens using refresh tokens

Supported Libraries streamline implementation:

- Python: requests-oauthlib + sentinelsat
- R: httr2 package
- JavaScript: node-fetch with OAuth middleware

This infrastructure eliminates proprietary satellite data costs, making global TROPOMI observations accessible to researchers worldwide without institutional budgets.

3.3. Query Construction and Data Discovery

The Catalogue API employs OData (Open Data Protocol) syntax, a standardized query language for web services. Geographic filtering employs Well-Known Text (WKT) POLYGON format, where coordinates are specified as longitude-latitude pairs in the WGS84 (EPSG:4326) reference system. The platform returns paginated results with up to 1,000 products per page, requiring pagination handling for large queries. A single city-scale region over one calendar year typically yields 300–400 satellite overpasses per pollutant, collectively requiring 300–400 individual L2 product files for download.

3.4. File Organization and Download Management

Efficient large-scale data acquisition requires systematic organization of download workflows. Recommended practice involves:

- **Temporal Chunking:** Break annual queries into quarterly segments to manage API request complexity and individual file sizes. Each quarter's query typically returns 80–120 product files per pollutant.

- **Parallel Downloads:** Implement concurrent download threads (4–8 parallel connections) to maximize bandwidth utilization while respecting platform resource constraints. The CDSE platform implements adaptive rate limiting to prevent abuse while supporting legitimate bulk requests.

- **Integrity Verification:** Implement SHA-256 checksum verification for all downloaded files. CDSE provides checksums in API responses, enabling automated integrity confirmation and detection of incomplete transfers.

- **Metadata Logging:** Record all download attempts, including timestamp, file size, checksum, and success/failure status. This logging enables progress tracking, resume capability for interrupted downloads, and auditing of data acquisition.

4. NetCDF Data Extraction and Preprocessing

4.1. NetCDF File Format and Hierarchical Structure

Sentinel-5P TROPOMI Level-2 products are distributed as NetCDF-4 (Network Common Data Form version 4) files, a self-documenting binary format designed for scientific data. NetCDF-4 employs hierarchical organization, enabling logical grouping of related data variables.

A typical TROPOMI L2 file structure comprises:

File: S5P_OFFL_L2__NO2__20240115T062145_20240115T080314_NNNN_TT_XXXXXX_YYYYMMDDTHHMMSS.nc

4.2. Data Extraction Using Python

The Python ecosystem provides excellent tools for NetCDF data extraction. The recommended approach uses the xarray library, which presents NetCDF hierarchies as intuitive Pandas-like objects:

```
import xarray as xr
import numpy as np
from datetime import datetime
# Open NetCDF file
file_path = "file_name.nc"
ds = xr.open_dataset(file_path, group='PRODUCT')
# Extract core variables
latitude = ds['latitude'].values  # Shape: (450, 4096)
longitude = ds['longitude'].values  # Shape: (450, 4096)
no2_column = ds['nitrogen_dioxide_tropospheric_column'].values  # mol/m²
qa_value = ds['qa_value'].values  # 0-1 scale
time_array = ds['time'].values  # Seconds since 2010-01-01
```

4.3. Quality Assurance Filtering

TROPOMI quality assurance flags represent the crucial first step in data preprocessing. Each satellite pixel receives a quality score (qa_value) ranging from 0 to 1, representing confidence in the retrieval algorithm's performance under that specific atmospheric condition:

- **qa_value ≥ 0.75:** Excellent quality, high confidence retrievals suitable for all applications
- **0.5 ≤ qa_value < 0.75:** Good quality, moderate confidence, suitable for most research
- **0.3 ≤ qa_value < 0.5:** Acceptable quality, reduced confidence for sensitive applications
- **qa_value < 0.3:** Poor quality, unreliable retrievals prone to large errors

For research requiring high data reliability, filtering to qa_value ≥ 0.75 is recommended. This typically retains 40–60% of satellite pixels per observation, balancing data retention against quality assurance. Additional filtering steps include:

- **Non-Physical Values:** Reject negative column densities (physically impossible) and zero values (failed retrieval algorithms)
- **Cloud Contamination:** Reject measurements with cloud fraction > 0.5 (excessive cloud cover compromises retrieval accuracy)
- **Extreme Outliers:** Flag and review values exceeding 5 standard deviations from regional climatology (potential data corruption)

4.4. Spatial Reference and Coordinate Systems

TROPOMI measurements are provided in WGS84 (EPSG:4326) geographic coordinates, where latitude ranges from -90° (South Pole) to +90° (North Pole), and longitude ranges from -180° to +180° (or 0° to 360°, depending on data provider convention). Each pixel is referenced to its center coordinates, with dimensions of approximately 5.5 km × 3.5 km.

For subsequent spatial operations—including matching with ground-based monitoring stations or reanalysis meteorological data—it is critical to consistently maintain the same coordinate reference system. Coordinate transformations between different projections (e.g., WGS84 to UTM or Web Mercator) introduce small systematic errors that accumulate over large datasets. Best practice involves maintaining all coordinates in WGS84 throughout processing, with coordinate transformation deferred to final visualization steps.

5. Planetary Boundary Layer Height: Critical Meteorological Context

5.1. Planetary Boundary Layer Physics and Air Quality Implications

The planetary boundary layer (PBL) represents the lowermost atmosphere where Earth's surface directly influences air motion through friction, heating, and moisture exchange. This typically extends from ground level to heights ranging from 300 m to 3,000 m, depending on time of day, atmospheric stability, and surface characteristics.

The PBL height (PBLH) is a critical determinant of surface air quality because it directly controls the volume of air available to dilute and disperse ground-level pollution. Under conditions of shallow PBLH (weak vertical mixing), pollutants emitted near the surface remain confined to a smaller atmospheric volume, resulting in high surface concentrations. Conversely, deep PBLH values (strong vertical mixing during convectively active periods) enable pollutants to disperse vertically into a larger volume, resulting in lower surface concentrations even under constant emissions.

The relationship between PBLH and surface pollution is well-established empirically. Studies across China demonstrated strong negative correlations between PBLH and particulate matter ($PM_{2.5}$) concentrations, with correlations remaining robust across seasonal variations and geographic regions. This makes PBLH a critical input parameter for converting satellite column measurements (which integrate the entire atmospheric column) to surface concentrations relevant for human health and regulatory compliance.

5.2. ERA5 Reanalysis and PBLH Retrieval

The European Centre for Medium-Range Weather Forecasts (ECMWF) produces the ERA5 atmospheric reanalysis, representing the state-of-the-art global reanalysis product. ERA5 integrates observations from multiple sources (ground stations, satellites, radiosondes, aircraft) using advanced data assimilation techniques to produce a physically consistent estimate of the complete 3D atmospheric state.

Key ERA5 Characteristics:

- Temporal Coverage: 1950–present, with continuous updates (5-day latency to real-time)
- Spatial Resolution: 0.25° × 0.25° (approximately 31 km grid spacing globally)
- Temporal Resolution: Hourly (0:00 to 23:00 UTC daily)
- Vertical Resolution: 137 atmospheric levels from surface to 1 Pa (approximately 80 km altitude)
- Data Variables: 200+ atmospheric, land, and oceanic variables, including PBLH

Crucially, ERA5 PBLH is not directly measured but rather computed from atmospheric profile data using a standardized algorithm. The bulk Richardson number criterion defines PBLH as the height where the Richardson number first exceeds a critical threshold (~0.25), indicating the transition from convectively mixed air to stable, non-turbulent air:

$$Ri_b(z)=\frac{g \cdot \Delta z \cdot [\theta_v(z)-\theta_v(z_s))}{\theta_v(z_s) \cdot [u(z)^2+v(z)^2+(b \cdot u^f)^2]}$$

Where:

- **g** = gravitational acceleration (9.81 m/s^2)
- **Δz** = height difference ($z - z_s$)
- **θ_v** = virtual potential temperature, accounting for moisture effects
- **z_s** = surface elevation
- **u, v** = horizontal wind components
- **u^f** = friction velocity (surface turbulence intensity)
- **b** = empirical tuning parameter (typically 0.27)

This physical formulation ensures that PBLH estimates reflect actual atmospheric thermodynamic and dynamical conditions, providing a scientifically grounded basis for column-to-surface conversion.

5.3. Downloading ERA5 PBLH Data

ERA5 data are distributed through the Copernicus Climate Data Store (CDS) using the cdsapi Python interface. The download workflow differs from CDSE because CDS serves gridded climate data rather than individual satellite observations.

Recommended download configuration specifies:

- **Dataset:** reanalysis-era5-single-levels
- **Variable:** boundary_layer_height
- **Time:** 08:00 UTC daily (corresponding to 13:30 IST for the Indian subcontinent; UTC + 5:30)
- **Grid Resolution:** 0.25° × 0.25° (native ERA5 resolution)
- **Spatial Domain:** Bounding box encompassing target region(s)
- **Temporal Domain:** Complete calendar year or targeted seasons

The resulting NetCDF file contains hourly PBLH values on a regular geographic grid. File sizes are modest (30–100 MB per quarter) compared to satellite products, and download times typically measure minutes rather than hours. This makes PBLH data acquisition a negligible computational burden compared to satellite data processing.

5.4. Temporal and Spatial Alignment of Satellite and PBLH Data

A fundamental data integration challenge arises from different temporal and spatial resolutions of satellite and meteorological data:

- **Sentinel-5P Observations:** Single overpass ~13:30 local solar time, varying ±15 minutes daily; 5.5 km × 3.5 km pixel footprint
- **ERA5 PBLH Data:** 08:00 UTC daily; 0.25° × 0.25° (~31 km) grid resolution

5.4.1. Temporal Alignment Strategy:

For satellite observations within ±30 minutes of the ERA5 reference time (08:00 UTC), the PBLH value is assigned directly with negligible temporal difference. For satellite observations exceeding ±30 minutes from the reference time, linear interpolation between adjacent hourly ERA5 values provides physically reasonable estimates:

$$PBLH(t_{sat})=PBLH(t_1)+\frac{(t_{sat}\text{-}t_1)}{(t_2\text{-}t_1)}\cdot[PBLH(t_2)\text{-}PBLH(t_1))$$

Where t_1 and t_2 represent adjacent hourly ERA5 times bracketing the satellite observation time. In practice, TROPOMI observations cluster tightly around 13:30 local time (±1 hour globally), minimizing interpolation requirements for most datasets.

5.4.2. Spatial Alignment Strategy:

For each satellite pixel, identify the nearest ERA5 grid point using efficient spatial indexing (KDTree nearest-neighbor search). Assign the ERA5 PBLH value for that grid point to the satellite pixel. This approach recognizes that ERA5 represents average conditions over ~31 km areas, while TROPOMI provides point measurements at ~5.5 km resolution. The nearest-neighbor approach preserves spatial structure while acknowledging the fundamental resolution difference.

6. Physical Conversion: From Column Density to Surface Concentration

6.1. Fundamental Measurement Challenge

The satellite measures atmospheric column density (molecules/m^2), representing the total number of pollutant molecules integrated vertically from ground to the top of the troposphere. This column density is NOT directly comparable to surface concentration measurements (μg/m^3) from ground stations because the column represents an atmospheric burden distributed across a height range of 10–15 km, not a point measurement at ground level.

This measurement difference is not a data quality issue but rather a fundamental distinction in what the two observing systems measure. Satellite column density provides valuable global information about atmospheric pollution burden and long-range transport, but must be converted to surface concentration to:

- Compare with ground-based air quality monitoring networks
- Assess compliance with surface-level air quality standards (e.g., WHO, EPA, India CPCB guidelines)
- Estimate human exposure and health impacts
- Validate chemical transport models

6.2. Column-to-Surface Conversion Framework

The conversion from satellite column density to surface concentration relies on a simplified vertical mixing model. Under the assumption that pollutants are relatively uniformly distributed within the planetary boundary layer and negligible pollutant concentration exists above the boundary layer (appropriate for most near-surface pollutants), the surface concentration can be approximated as:

$$n_{surface}\left(molecule\frac{s}{c}m^3\right) = \frac{N_{column}\left(molecule\frac{s}{c}m^2\right)}{PBLH(cm)}$$

Where the column density is integrated only over the planetary boundary layer height rather than the entire troposphere. This represents a simplification because:

- Actual pollutant profiles are rarely uniform vertically
- Pollutants are occasionally transported above the PBLH
- Vertical gradients change with atmospheric stability and mixing strength

Despite these limitations, this vertical averaging model has demonstrated reasonable performance across diverse geographic regions and seasons, provided that ground-level data are used for model training and validation.

Converting from number density to mass concentration requires multiplying by molecular weight:

$$C_{surface}\left(\mu\frac{g}{m^3}\right) = \frac{N_{column}\left(mo\frac{l}{m^2}\right) \times MW\left(\frac{g}{m}ol\right) \times 10^6\left(\mu\frac{g}{g}\right)}{PBLH(m)}$$

Where molecular weight constants are:

- **NO_2:** 46 g/mol
- **SO_2:** 64 g/mol
- **CO:** 28 g/mol
- **O_3:** 48 g/mol

6.3. Implementation and Error Considerations

6.3.1. Error Propagation and Uncertainty

The conversion introduces uncertainties from multiple sources:

- **Column Density Uncertainty:** TROPOMI retrievals have typical uncertainties of 20–50%, depending on atmospheric conditions
- **PBLH Uncertainty:** ERA5 PBLH has an estimated uncertainty of 10–30%, depending on the atmospheric regime
- **Model Assumptions:** The uniform vertical mixing assumption may introduce systematic errors of 20–50% in specific geographic regions

Combined uncertainty in surface concentration estimates typically ranges from 30–70% of the calculated value. This uncertainty is substantially higher than direct ground-based measurements (±5–10%) but remains acceptable for many research applications, including:

- Identifying pollution hotspots and source regions
- Assessing long-term air quality trends
- Filling spatial gaps in sparse monitoring networks
- Supporting regulatory decision-making in data-poor regions

For applications requiring higher accuracy, satellite estimates should be calibrated against available ground-truth measurements using statistical correction methods (e.g., linear regression, random forest models, etc.).

7. Quality Assurance and Validation Procedures

7.1. Temporal Coverage Assessment

Comprehensive quality assurance begins with the evaluation of data temporal coverage. Calculate the following metrics:

- **Calendar Days with Observations:** Count unique calendar days (target: 360–366 for annual datasets)
- **Missing Dates:** Identify gaps and classify as:
    - Cloud/corruption (satellite temporary unavailability)
    - Orbital mechanics (normal seasonal gaps at certain latitudes)
    - Data processing failures (investigate and resolve)
- **Observations per Day:** Calculate mean and standard deviation (typical: 1 observation/day at equator, multiple at higher latitudes)

7.2. Spatial Coverage and Geographic Representation

Verify spatial coverage for research regions:

- **Latitude/Longitude Bounds:** Confirm pixel coordinates match the expected geographic domain
- **Pixel Count by Region:** Count observations in subregions to identify geographic bias
- **Spatial Completeness:** Calculate percentage of grid cells with at least one observation (target: >90%)

```
import numpy as np
# Verify geographic bounds
print(f"Latitude Range: {df['latitude'].min():.2f}° to {df['latitude'].max():.2f}°")
print(f"Longitude Range: {df['longitude'].min():.2f}° to {df['longitude'].max():.2f}°")
# Calculate spatial completeness
grid_size = 0.1  # 0.1 degree resolution; lat_bins = np.arange(-90, 90.1, grid_size)
lon_bins = np.arange(-180, 180.1, grid_size)
grid_count = np.histogram2d(df['latitude'], df['longitude']; bins=[lat_bins, lon_bins])[0]
spatial_completeness = 100 * (grid_count > 0).sum() / grid_count.size
print(f"Spatial Completeness: {spatial_completeness:.1f}%")
```

7.3. Statistical Validation

Comprehensive statistical characterization identifies data quality issues:

```
df['surface_no2'].describe()
# Output summary statistics (mean, std, min, 25%, 50%, 75%, max)
# Check for normality (typically violated for concentration data)
from scipy import stats
skewness = stats.skew(df['surface_no2'].dropna())
kurtosis = stats.kurtosis(df['surface_no2'].dropna())
# Identify outliers
mean = df['surface_no2'].mean(); std = df['surface_no2'].std()
outliers = (df['surface_no2'] > mean + 5*std) | (df['surface_no2'] < mean - 5*std)
```

7.4. Quality Flag Performance Evaluation

Assess the effectiveness of quality assurance filters:

```
# Retention by QA threshold
qa_thresholds = [0.3, 0.5, 0.75, 0.9]
for qa_thresh in qa_thresholds:

    retained = (df['qa_value'] >= qa_thresh).sum()
```

```
retention_pct = 100 * retained / len(df)

 mean_quality = df[df['qa_value'] >= qa_thresh]['qa_value'].mean()

print(f"QA ≥ {qa_thresh}: {retained} observations ({retention_pct:.1f}%), mean QA = {mean_quality:.3f}")
```

7.5. Validation Against Ground-Truth Measurements

Where ground-based monitoring data are available, perform direct validation:

```
from sklearn.metrics import r2_score, mean_squared_error
# Collocate satellite and ground observations
# (Implementation requires specific ground data source)
r_squared = r2_score(ground_measurements, satellite_estimates)
rmse = np.sqrt(mean_squared_error(ground_measurements, satellite_estimates))
bias = (satellite_estimates - ground_measurements).mean()
```

8. Applications and Impact

8.1. Research Applications

The processed and verified Sentinel-5P datasets enable diverse research applications:

- **Air Quality Monitoring:** High-spatial-resolution maps of NO_2, SO_2, CO, and O_3 distribution identify pollution hotspots, industrial emission sources, and urban-rural pollution gradients inaccessible through sparse ground monitoring networks. This information supports environmental justice research identifying disproportionate pollution exposure in economically disadvantaged communities.
- **Health Impact Assessment:** Satellite-derived air quality estimates combined with population exposure models and epidemiological dose-response relationships enable estimation of air pollution-attributable health burdens. Studies using TROPOMI NO_2 data have demonstrated significant associations with cardiovascular and respiratory outcomes across global populations.
- **Source Attribution and Emission Modeling:** High-resolution satellite observations of pollution plumes enable visual identification and mapping of emission sources, validating bottom-up emission inventories. Satellite data constrain chemistry transport models by providing independent observations of atmospheric composition, improving model fidelity for future air quality forecasting.
- **Climate-Air Quality Interactions:** Satellite observations capture how meteorological variability influences air quality, enabling quantification of seasonal and interannual variations. Understanding these interactions is essential for predicting how climate change will alter future air quality in different regions.

8.2. Regulatory and Policy Applications

- **Environmental Standards and Compliance:** Satellite data provide objective, spatially comprehensive information about whether geographic regions meet ambient air quality standards. This is particularly valuable in countries with limited monitoring infrastructure, where satellite data may represent the primary available air quality information.
- **Transboundary Pollution Assessment:** Long-range pollutant transport crosses political boundaries, creating transboundary air quality issues. Satellite observations provide unbiased, continuous spatial coverage, enabling quantification of pollution contribution from different geographic sources, supporting policy negotiations between nations.
- **Real-Time Air Quality Information:** Integration of satellite data with numerical weather prediction models enables production of short-term air quality forecasts, supporting public health warnings and advisory systems.

8.3. Global Development and Capacity Building

- **Democratization of Air Quality Data:** By providing open-access satellite data and standardized processing methodologies, this framework enables researchers in low-resource settings to access the same air quality information as well-funded institutions. This addresses fundamental equity issues in environmental monitoring and environmental health research.

9. Limitations, Uncertainties, and Future Improvements

9.1. Acknowledged Limitations

- **Spatial Resolution:** TROPOMI's 5.5 km × 3.5 km footprint, while unprecedented for atmospheric composition monitoring, remains coarser than ground monitoring networks (~100–1000 m typical spacing). This limits the ability to characterize fine-scale urban pollution variability.
- **Temporal Sampling:** A single daily overpass at ~13:30 local time provides limited temporal resolution. Diurnal cycles of pollution and rapid pollution episodes cannot be captured. Geostationary satellites offer potential future improvement through hourly observations.
- **Column-to-Surface Conversion Uncertainty:** The vertical averaging assumption may introduce 30–70% uncertainty in surface concentration estimates, particularly in regions with complex topography or extreme atmospheric conditions.
- **Data Gaps:** Cloud cover, over snow/ice, and technical issues occasionally prevent observations over target regions. High-latitude coverage varies seasonally due to orbital geometry.

9.2. Emerging Improvements and Future Directions

- **Geostationary Satellites:** The Geostationary Environment Monitoring Spectrometer (GEMS) aboard South Korean satellites provides hourly observations over East Asia, and similar instruments (TEMPO over North America, Sentinel-4 over Europe) are entering operation. These will revolutionize air quality monitoring through near-continuous temporal coverage, enabling hourly forecasting and rapid pollution event detection.
- **Machine Learning Integration:** Deep learning models trained on satellite data combined with meteorological variables and chemical transport model outputs increasingly enable more accurate surface concentration estimates, with R^2 values exceeding 0.8 in some applications. These approaches reduce reliance on PBLH data and capture nonlinear relationships missed by simplified physical models.
- **Multi-Source Data Fusion:** Combining TROPOMI with complementary satellite observations (aerosol optical depth from MODIS, high-resolution urban morphology from optical satellites, land-use data) in unified machine learning frameworks improves estimation accuracy and spatial resolution.

10. Conclusion

Sentinel-5P TROPOMI satellite data represents a transformative opportunity for air quality research and monitoring globally. By providing open-access observations of atmospheric composition at high spatial and temporal resolution, the mission addresses fundamental equity issues in environmental monitoring that have historically limited air quality knowledge in data-poor regions.

This paper has presented a comprehensive methodological framework for acquiring, processing, and analyzing Sentinel-5P data to generate actionable air quality information. The framework demonstrates how open-access satellite data infrastructure (Copernicus Data Space Ecosystem), coupled with standardized reanalysis meteorological data (ERA5) and modern computational tools, enables researchers worldwide to generate analysis-ready air quality datasets without dependence on proprietary systems or expensive licenses.

The step-by-step methodology from API-based data discovery through quality-assured surface concentration estimation provides a complete template that researchers can adapt to specific geographic regions and research questions. By reducing technical barriers and providing transparent, reproducible procedures, the framework aims to democratize access to high-quality air quality information.

As geostationary satellites enter operation and machine learning methods increasingly advance surface concentration retrieval accuracy, satellite-based air quality monitoring will become increasingly central to global environmental health surveillance. The methodologies and insights presented here provide a foundation for leveraging these emerging capabilities to advance environmental justice, support evidence-based policy, and ultimately improve air quality and human health globally.

Compliance with ethical standards

Acknowledgments

The authors sincerely thank the Institute of Science, Nagpur, and its director for providing the necessary resources and support to conduct this research. Appreciation is also extended to dataspace.copernicus.eu for their timely assistance.

Disclosure of conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Boersma, K. F., Eskes, H. J., & Brinksma, E. J. (2018). Error analysis for tropospheric NO_2 retrieval from space. Journal of Geophysical Research, 109, D04311.

[2] Copernicus Climate Change Service (C3S). (2025). ERA5: Fifth-generation ECMWF atmospheric reanalysis. Retrieved from https://cds.climate.copernicus.eu/

[3] European Commission. (2025). Copernicus Data Space Ecosystem: User Guide and API Documentation. Retrieved from https://dataspace.copernicus.eu/

[4] Goldberg, D. L., Saide, P. E., Lamsal, L. N., de Foy, B., Lu, Z., Woo, J. H., ... & Streets, D. G. (2019). A top-down assessment of nitrogen oxide emissions from major point and area sources in the Southern United States using TROPOMI observations. Environmental Research Letters, 14(10), 104008.

[5] Griffin, D., Zhao, X., McLinden, C. A., Boersma, F., Bourassa, A., Dammers, E., ... & Dickson, N. (2019). High-resolution mapping of nitrogen dioxide with TROPOMI: First results and validation over the Canadian oil sands. Geophysical Research Letters, 46(1), 195-202.

[6] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999-2049.

[7] Liu, X., Ding, S., Lin, X., Pan, Y., Han, J., Wang, Y., ... & Khosrawi, F. (2021). An improved tropospheric NO_2 retrieval algorithm for the Sentinel-5P TROPOMI instrument. Atmospheric Measurement Techniques, 14(11), 7297-7326.

[8] S. Mahulkar and A. Khamborkar, "Advances in Satellite Remote Sensing for Air Quality Assessment: A Comprehensive Review" doi: https://doi.org/10.30574/wjarr.2025.28.2.3905

[9] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., ... & Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), 4349-4383.

[10] Shen, L., Zhai, S., Zhu, L., Xie, M., & Liu, X. (2023). Global retrieval of TROPOMI tropospheric HCHO and NO_2 vertical columns. Atmospheric Measurement Techniques, 18(3), 1561-1581.

[11] Su, T., Li, Z., & Kahn, R. (2018). Relationships between the planetary boundary layer height and surface pollutants derived from the Chinese environmental monitoring network. Atmospheric Chemistry and Physics, 18(21), 15921-15935.

[12] The Health Effects Institute. (2025). Satellite-derived air quality estimation for environmental health research: Research needs and recommendations. Boston, MA: Health Effects Institute.

[13] Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., ... & Levelt, P. F. (2012). TROPOMI on the Sentinel-5 Precursor: A GMES mission for atmospheric composition monitoring. Remote Sensing of Environment, 120, 70-83.

[14] Wang, Z., Maeda, T., Hayashi, M., Kubo, M., & Komori, D. (2020). Hourly scale spatiotemporal rainfall prediction using a convolutional neural network. Water Resources Research, 56(10), e2019WR027santiago.

[15] Yang, Q., Yuan, Q., Li, T., Shen, H., & Zhang, L. (2023). A synchronized estimation of hourly surface concentrations of six criteria air pollutants using a multi-output random forest model. Nature Communications, 14(1), 4825.

[16] Zhang, Y., Liu, X., Ma, Y., & Li, Z. (2021). Aerosol physical-optical properties and PBLH under different pollution accumulation pathways. Geophysical Research Letters, 48(17), e2021JD035236.