

Aanshik Ramakant Dubey¹, Anvi Jain², Veday Deshmukh³, Sahil Pharate⁴

*Department of Computer Science & Engineering
MIT Art, Design & Technology University, Pune, India*

Guided by: Prof. Yuvraj Nikam
*Department of Computer Science & Engineering
MIT Art, Design & Technology University, Pune, India*

Abstract

Ambient air pollution remains a major environmental health concern in India, and the Central Pollution Control Board (CPCB) National Air Quality Index (AQI) is the primary instrument used to communicate multi-pollutant exposure risk to the public. Because ground-based monitoring stations are spatially sparse, large parts of the country lack a locally representative AQI value. A separate atmospheric-chemistry problem concerns tropospheric formaldehyde (HCHO), a short-lived trace gas that serves as a proxy for volatile organic compound emissions and is closely associated with biomass-burning activity; HCHO is not one of the eight pollutants used in the CPCB AQI formulation and must therefore be treated as a distinct scientific variable rather than as an AQI input. This paper proposes a two-objective satellite-to-surface framework for India that (i) estimates a CPCB-consistent surface AQI by learning a mapping from INSAT-3D Aerosol Optical Depth, Sentinel-5P TROPOMI trace-gas columns, and ERA5/IMDAA reanalysis meteorology to CPCB ground-station pollutant concentrations using a CNN-LSTM architecture, and (ii) characterizes spatio-temporal HCHO hotspots during biomass-burning periods by computing a climatological standardized-anomaly score from TROPOMI HCHO columns and relating flagged hotspots to MODIS/VIIRS active-fire detections and reanalysis wind fields. The two objectives share a common data-acquisition and spatio-temporal alignment pipeline but remain scientifically and evaluatively separate. This paper presents the system architecture, the CPCB sub-index and AQI aggregation formulation, the CNN-LSTM mathematical formulation, the HCHO anomaly-detection formulation, and a reproducible experimental-evaluation protocol based on spatially and temporally held-out validation. No experimental results are reported, since model training and validation have not yet been performed; parameters such as dataset size, spatial/temporal resolution of the final fused product, and performance metrics will be determined during experimental validation. The intended contribution is a reproducible, dual-objective pipeline that keeps CPCB-regulated AQI estimation and trace-gas hotspot characterization scientifically distinct while sharing a common multi-source data infrastructure.

Keywords: Air Quality Index, Biomass Burning, Formaldehyde, Satellite Remote Sensing

1. INTRODUCTION

Ambient air pollution is a leading environmental health risk, and the Central Pollution Control Board (CPCB) introduced the National Air Quality Index (AQI) in 2014 to translate ground-monitored concentrations of eight regulated pollutants -- PM_{2.5}, PM₁₀, NO₂, SO₂, CO, O₃, NH₃, and Pb -- into a single, publicly interpretable value [1], [2]. The composite AQI is obtained by computing a pollutant-specific sub-index from CPCB breakpoint tables and taking the maximum sub-index across at least three measured pollutants, one of which must be a particulate matter species [1]. However, CPCB continuous ambient air-quality monitoring stations are concentrated in urban centers, so most of the geographic area of India does not have a nearby station from which a representative AQI can be computed.

Satellite Aerosol Optical Depth (AOD) has been used as a proxy for surface particulate loading because it integrates the column-total aerosol content, which is statistically related to ground-level PM_{2.5} and PM₁₀ under suitable atmospheric conditions [3], [4]. India additionally benefits from the geostationary INSAT-3D Imager AOD product, which provides AOD retrievals over the Indian subcontinent at approximately 30-minute temporal resolution -- a revisit frequency not available from polar-orbiting sensors such as MODIS [3].

A scientifically distinct problem concerns tropospheric formaldehyde (HCHO). HCHO is a short-lived intermediate product of volatile organic compound (VOC) oxidation and is strongly enhanced during agricultural residue burning and forest-fire events, making its column density a recognized tracer of biomass-burning emissions [5]. The Tropospheric Monitoring Instrument (TROPOMI) aboard Sentinel-5P retrieves HCHO tropospheric columns daily at an along-track resolution of approximately 5.5 km x 3.5 km using a Differential Optical Absorption Spectroscopy (DOAS) algorithm [6]. Because HCHO is not among the eight pollutants defined in the CPCB AQI methodology, it cannot be incorporated into the composite AQI computation and must instead be analyzed as an independent trace-gas / atmospheric-chemistry variable.

This work is motivated by India's National Clean Air Programme (NCAP), which targets a reduction in particulate pollution in non-attainment cities, and by the general policy interest in spatially continuous air-quality information to complement the CPCB ground network. The remainder of this paper is organized as follows. Section II reviews related literature on satellite-based surface pollutant estimation, HCHO retrieval, and fire detection. Section III states the problem and the research gap. Section IV presents the proposed system architecture. Section V describes the methodology, and Section VI presents the mathematical formulation. Section VII defines the experimental setup and evaluation protocol. Section VIII discusses the design, Section IX states the limitations, and Section X concludes with future work.

II. LITERATURE REVIEW

A. Satellite-Based Surface Pollutant / AQI Estimation

Mishra et al. [3] developed and validated the operational INSAT-3D Imager AOD retrieval algorithm over the Indian subcontinent and adjoining ocean, reporting correlation coefficients of approximately 0.70-0.91 against AERONET sun-photometer measurements over land and 0.79-0.91 over ocean, and 0.74-0.89 (land) and 0.65-0.97 (ocean) against co-located MODIS-Aqua AOD. This work establishes INSAT-3D AOD as a validated, high-temporal-frequency input for the Indian domain but is limited to the AOD product itself and does not extend to surface pollutant concentration or AQI estimation. Van Donkelaar et al. [4] combined satellite AOD (MODIS, MISR, SeaWiFS) with a chemical transport model (GEOS-Chem) and ground-based calibration to produce global geophysical-statistical estimates of surface PM_{2.5}, demonstrating that combining satellite columnar data with a physically informed model can substantially improve surface PM_{2.5} estimation relative to AOD alone; the approach is global in scope and is not tailored to the CPCB multi-pollutant AQI framework or to India-specific meteorological regimes such as the Indo-Gangetic Plain boundary layer.

B. Satellite-Derived HCHO Monitoring

De Smedt et al. [5] examined long-term trend detection in satellite-retrieved tropospheric HCHO columns using GOME, SCIAMACHY, and OMI, showing that HCHO column retrievals can reveal both long-term VOC emission trends and short-term enhancements associated with biomass burning, while also identifying instrument-to-instrument consistency and retrieval-noise as key limitations for trend and anomaly detection. Veefkind et al. [6] describe the TROPOMI instrument and its Level-2 trace-gas retrieval products, including HCHO, noting the improved spatial resolution and daily global revisit relative to its predecessors, which is the basis for the HCHO input used in this work.

C. Biomass-Burning and Fire Detection

Giglio et al. [7] describe the MODIS Collection 6 active-fire detection algorithm, which provides global fire-pixel detections at approximately 1 km nominal resolution from the Terra and Aqua satellites. Schroeder et al. [8] extended this contextual detection approach to the VIIRS instrument at 375 m resolution, improving the detection of smaller and lower-intensity fires relevant to agricultural residue burning. Both products are widely used to co-locate active-fire activity with trace-gas column enhancements, which is the basis of the fire-HCHO relationship analyzed in this work.

D. Deep Learning for Spatio-Temporal Prediction

Convolutional Neural Networks (CNNs) are established for extracting spatial patterns from gridded imagery, and Long Short-Term Memory (LSTM) recurrent networks are established for modelling temporal dependencies in sequential data [9]. Combined CNN-LSTM architectures, which apply a CNN to each time step of a gridded input sequence before passing the resulting spatial feature sequence to an LSTM, are used in this work as a means to jointly capture the spatial structure of columnar satellite inputs and the temporal persistence of pollutant concentrations, consistent with the choice of methodology indicated in the original problem statement for this project.

E. Research Gap

The reviewed literature indicates that INSAT-3D AOD retrieval [3] and satellite-to-surface PM_{2.5} modelling [4] have each been separately validated, that HCHO trend and column-retrieval methods [5], [6] exist independently of surface AQI estimation, and that MODIS/VIIRS fire products [7], [8] are

established tools for locating biomass-burning activity. However, the reviewed sources do not combine an India-specific, CPCB-consistent multi-pollutant AQI estimation pipeline with an independently scoped HCHO hotspot-characterization pipeline that shares the same multi-source acquisition and alignment infrastructure. This is a comparatively narrow and conservative gap: rather than claiming that no related work exists, this paper identifies that the specific combination -- CPCB sub-index-consistent AQI estimation using INSAT-3D and TROPOMI inputs, alongside a scientifically separate HCHO anomaly/fire-correlation pipeline for the Indian domain -- has not been demonstrated as an integrated system in the reviewed sources, which motivates the framework proposed in Sections IV-VI.

III. PROBLEM STATEMENT AND RESEARCH GAP

Given INSAT-3D AOD, Sentinel-5P TROPOMI trace-gas columns (NO₂, SO₂, CO, O₃, HCHO), CPCB ground-station pollutant concentrations, ERA5/IMDAA reanalysis meteorology, and MODIS/VIIRS active-fire products, the problem addressed in this paper is twofold and intentionally kept scientifically separate: Objective 1 is to estimate a CPCB-consistent surface AQI over India by predicting ground-level concentrations of the CPCB-regulated pollutants from columnar and meteorological inputs; Objective 2 is to characterize the spatio-temporal distribution of HCHO hotspots during biomass-burning periods and to relate flagged hotspots to fire activity and wind fields. The two objectives share a common data-acquisition and spatio-temporal alignment stage (Section V-A) but use distinct models, distinct scientific outputs, and distinct evaluation protocols, because HCHO is not a CPCB AQI pollutant and must not be aggregated into the composite AQI.

IV. PROPOSED SYSTEM ARCHITECTURE

Fig. 1 shows the overall proposed architecture. Multi-source data are acquired, passed through quality control (cloud/QA filtering), spatially and temporally aligned to a common analysis basis, and fused into a multi-source feature stack. From this shared stack, the pipeline branches into two scientifically independent paths: the AQI estimation branch, in which a CNN-LSTM model predicts surface pollutant concentrations that are converted to a CPCB composite AQI; and the HCHO hotspot branch, in which a climatological anomaly score is computed from TROPOMI HCHO columns and combined with MODIS/VIIRS fire co-location and reanalysis wind fields to produce a characterized hotspot map. The two branches are combined only at the shared pre-processing stage and are not merged at the output stage, consistent with the requirement in Section III that HCHO not be treated as an AQI pollutant.

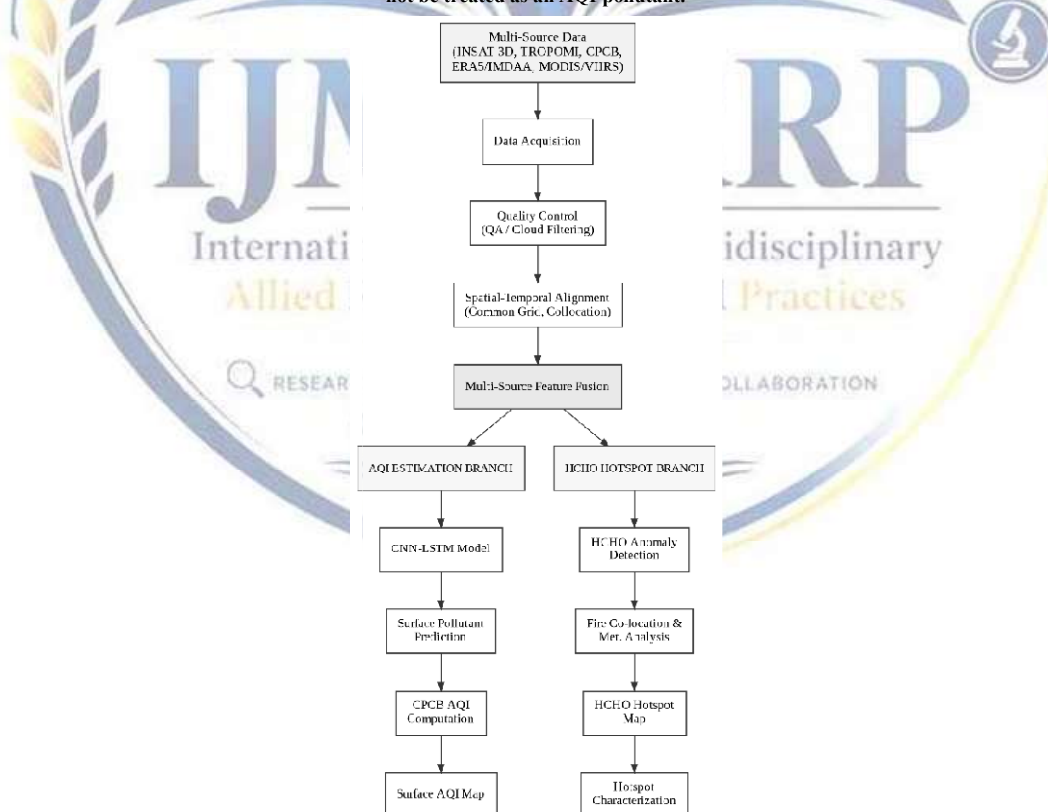


Fig. 1. Proposed overall system architecture, showing the shared acquisition/alignment stage and the two scientifically separated branches.

V. METHODOLOGY

A. Data Acquisition and Spatio-Temporal Alignment

Fig. 2 shows the shared pre-processing pipeline. INSAT-3D AOD is delivered on a 0.1 deg x 0.1 deg grid at approximately 30-minute temporal resolution [3], while TROPOMI trace-gas columns are delivered per-orbit at approximately 5.5 km x 3.5 km resolution with daily global coverage [6]. Because these products differ in native spatial resolution, both are reprojected onto a single common analysis grid; MODIS (~1 km) and VIIRS (375 m) fire-pixel detections [7], [8] are aggregated onto the same grid as fire-count fields, and ERA5/IMDAA reanalysis fields are interpolated onto the same grid. CPCB ground stations, which report point measurements rather than gridded fields, are matched to their nearest enclosing grid cell (station-to-grid collocation) so that a predicted grid-cell concentration can be supervised against the corresponding station record. All sources are then temporally aligned to the CPCB reporting convention -- a 24-hour averaging period for particulate matter, NO₂, SO₂, NH₃, and Pb, and an 8-hour averaging period for O₃ and CO [1] -- after which missing values arising from cloud cover, orbital gaps, or station downtime are handled by exclusion of the corresponding time-step rather than by imputed substitution, and continuous input features are scaled prior to model input.

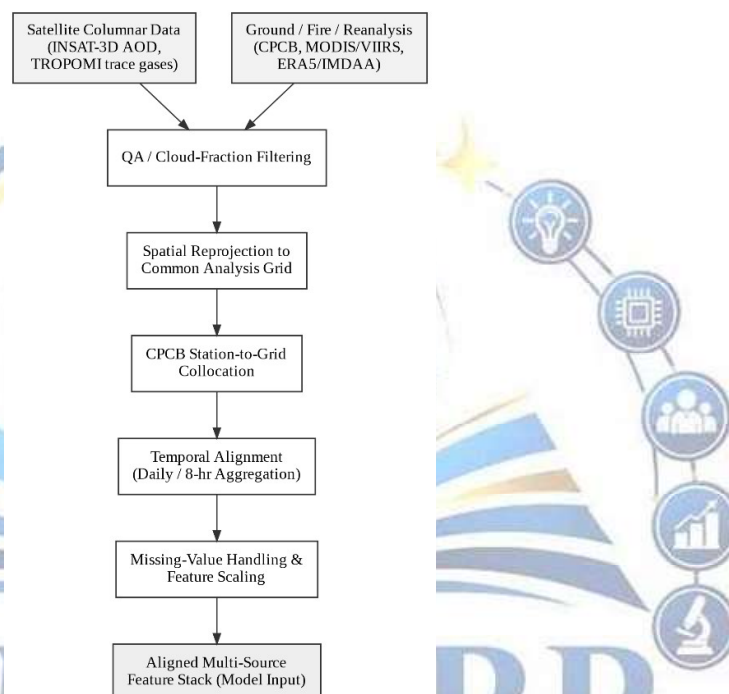


Fig. 2. Proposed multi-source data preprocessing and spatio-temporal alignment workflow.

B. Objective 1 -- CNN-LSTM Surface AQI Estimation

The proposed methodology for Objective 1 uses a CNN-LSTM architecture, selected because it is explicitly identified as a suitable hybrid deep-learning approach for this task in the original problem statement, and because the problem combines a spatial estimation task (predicting a pollutant field over a two-dimensional grid) with a temporal estimation task (capturing diurnal and seasonal persistence in pollutant concentrations). Fig. 3 shows the model architecture. A convolutional sub-network is applied independently to the gridded multi-channel input at each time step to extract local spatial patterns -- for example, spatial gradients in AOD associated with regional haze or dust transport -- without assuming a fixed relationship between a single pixel's AOD and its pollutant concentration. The resulting sequence of spatial feature maps is passed to an LSTM layer, which is used specifically because pollutant concentrations exhibit temporal persistence and diurnal cycles (for example, boundary-layer collapse in the evening) that a purely spatial model cannot represent; the LSTM's gating mechanism allows the network to retain relevant temporal context across multiple time steps while discarding transient noise. The LSTM output is passed through a fully connected layer to produce the predicted concentration for each CPCB-regulated pollutant, from which the CPCB sub-index and composite AQI are computed deterministically (Section VI-C). This paper does not claim that the CNN-LSTM architecture is superior to alternative architectures; the choice is a design decision justified by the spatio-temporal nature of the problem and by the original problem statement, and comparative evaluation against baseline models is defined as part of the experimental protocol in Section VII.

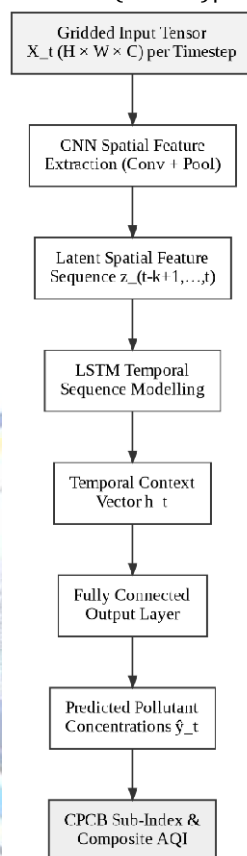


Fig. 3. Proposed CNN-LSTM surface pollutant estimation architecture.

C. Objective 2 -- HCHO Hotspot Detection and Fire Correlation

The proposed methodology for Objective 2 (Fig. 4) uses a single, clearly defined hotspot-detection approach: climatological standardized-anomaly thresholding, rather than an unspecified choice between multiple candidate techniques. TROPOMI HCHO columns are acquired and passed through the QA and cloud-fraction filtering described in Section V-A, then aggregated to daily (or, where daily coverage is insufficient, multi-day) composites on the common analysis grid. For each grid cell, a climatological mean and standard deviation are constructed from a multi-year reference period outside the biomass-burning season, and the standardized anomaly of the current HCHO column relative to this baseline is computed (Section VI-D). A grid cell/time-step pair is flagged as a hotspot when this anomaly exceeds a threshold. Flagged hotspots are then co-located with MODIS/VIIRS active-fire counts on the same grid and time step, and the Pearson correlation between fire counts and HCHO anomaly is computed over the flagged burning-period cells to quantify the strength of the fire-HCHO relationship (Section VI-E). Finally, ERA5/IMDAA wind-direction and wind-speed fields at the time of each flagged hotspot are used to perform a qualitative upwind/downwind comparison between the hotspot location and nearby fire detections, in order to assess whether the spatial offset between a fire detection and its associated HCHO enhancement is consistent with the reanalysis wind field. This qualitative wind check is descriptive rather than a full trajectory reconstruction; back-trajectory modelling (e.g., a HYSPLIT-type analysis) is not part of the current proposed methodology and is identified as future work in Section X, since it is not presently implemented in this pipeline.

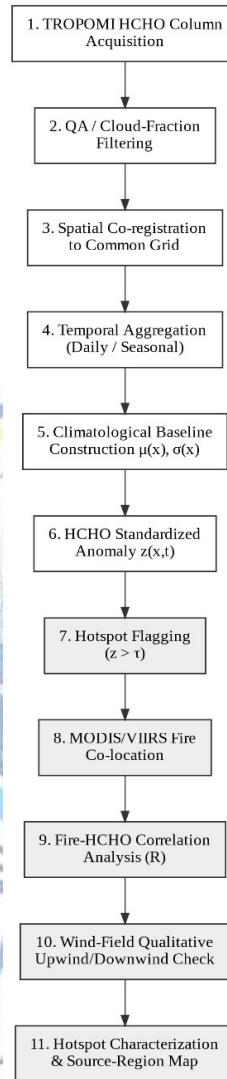


Fig. 4. Proposed HCHO hotspot-detection and fire-correlation workflow.

VI. MATHEMATICAL FORMULATION

A. CNN Spatial Feature Extraction

Let X_t in $\mathbb{R}^{H \times W \times C}$ denote the gridded, aligned input tensor at time t , comprising C channels (AOD, TROPOMI trace-gas columns, and reanalysis meteorological variables) over a spatial domain of height H and width W . A convolutional feature extractor f_{cnn} maps each time slice to a latent spatial feature representation:

$$z_t = f_{cnn}(X_t; \theta_{cnn}) \quad (1)$$

where θ_{cnn} denotes the learnable convolutional parameters.

B. LSTM Temporal Modelling

The sequence of latent spatial features $z_{t-k+1}, \dots, z_t$ over a temporal window of length k is passed to an LSTM layer to obtain a temporally contextualized representation h_t :

$$h_t = f_{lstm}(z_{t-k+1}, \dots, z_t; \theta_{lstm}) \quad (2)$$

The predicted concentration vector $\hat{y}_t$, with one component per CPCB-regulated pollutant p in $\{PM_{2.5}, PM_{10}, NO_2, SO_2, CO, O_3\}$, is obtained through a fully connected output layer:

$$\hat{y}_t = W_o h_t + b_o \quad (3)$$

The model parameters (θ_{cnn} , θ_{lstm} , W_o , b_o) are trained by minimizing the mean squared error between predicted and CPCB-measured concentrations over a training set of size N :

$$L = (1/N) * \sum_{i=1}^N \| \hat{y}_i - y_i \|^2$$

C. CPCB Sub-Index and Composite AQI

For a predicted concentration C_p of pollutant p , the CPCB pollutant-specific sub-index I_p is computed by piecewise-linear interpolation between the CPCB breakpoint concentrations BP_{LO} and BP_{HI} (Table I) and their corresponding index values I_{LO} and I_{HI} [1], [2]:

$$I_p = [(I_{HI} - I_{LO}) / (BP_{HI} - BP_{LO})] * (C_p - BP_{LO}) + I_{LO} \quad (5)$$

The composite AQI is the maximum sub-index across all pollutants for which a valid concentration is available, subject to the CPCB requirement that at least three pollutants be available, one of which must be PM_{2.5} or PM₁₀ [1]:

$$AQI = \max(I_{PM2.5}, I_{PM10}, I_{NO2}, I_{SO2}, I_{CO}, I_{O3}) \quad (6)$$

Sub-index values are not computed for HCHO, NH₃, or Pb in this study; NH₃ and Pb are outside the scope of the selected satellite/reanalysis input channels, and HCHO is deliberately excluded from (6) because it is not a CPCB AQI pollutant (Section III).

D. HCHO Standardized Anomaly

Let $\phi(x, t)$ denote the aggregated TROPOMI HCHO tropospheric column at grid location x and time t , and let $\mu(x)$ and $\sigma(x)$ denote the climatological mean and standard deviation of ϕ at that location over a reference period outside the biomass-burning season. The standardized HCHO anomaly is:

$$z(x, t) = (\phi(x, t) - \mu(x)) / \sigma(x) \quad (7)$$

A location-time pair is flagged as a hotspot when $z(x, t)$ exceeds a threshold τ , i.e., $z(x, t) > \tau$, where τ is set empirically during experimental validation (Section VII).

E. Fire-HCHO Correlation

Let $F(x, t)$ denote the co-located MODIS/VIIRS active-fire count at grid cell x and time t . The Pearson correlation coefficient R between fire counts and HCHO anomaly, computed over the set of flagged burning-period grid cells, is:

$$R = \text{cov}(F, z) / (\sigma_F * \sigma_z) \quad (8)$$

where σ_F and σ_z are the standard deviations of F and z , respectively, over the same set of grid cells and time steps. R quantifies the strength of the fire-HCHO relationship and does not by itself constitute a transport analysis; the qualitative wind-field comparison described in Section V-C is used to assess spatial consistency with the reanalysis wind field.

VII. EXPERIMENTAL SETUP AND EVALUATION

A. Dataset and Data Sources

Table I lists the multi-source datasets used in the proposed framework, together with the verified spatial and temporal resolutions reported in the corresponding technical documentation and validation literature.

TABLE I
MULTI-SOURCE DATASETS USED IN THE PROPOSED FRAMEWORK

Dataset	Variable	Source	Spatial Res.	Temporal Res.	Role
INSAT-3D	AOD	ISRO [3]	0.1° grid	~30 min	AQI input
TROPOMI	NO ₂ , SO ₂ , CO, O ₃ , HCHO	ESA [6]	~5.5×3.5 km	Daily	Both branches
CPCB CAAQMS	PM _{2.5} , PM ₁₀ , NO ₂ , SO ₂ , CO, O ₃	CPCB [1]	Point	24-hr/8-hr [1]	AQI ground truth
MODIS	Active fire	NASA [7]	~1 km	2x/day	HCHO input
VIIRS	Active fire	NASA [8]	375 m	2x/day	HCHO input
ERA5/IMDAA	Wind, BLH, temp., humidity	ECMWF/NCMRWF	TBD	TBD	Both branches

As shown in Table I, the exact spatial and temporal resolution to be used for ERA5/IMDAA reanalysis fields is marked TBD (to be determined) rather than an assumed value; this will be fixed during implementation once the specific reanalysis product and processing level are selected.

B. Dataset Split and Leakage Avoidance

For Objective 1, a naive random split of individual station-day observations would allow near-identical observations from the same station and adjacent days to appear in both the training and test sets, artificially inflating apparent performance. To avoid this, the evaluation protocol will use a combined spatial and temporal split: a subset of CPCB stations will be held out entirely (spatial split, testing generalization to unseen locations) and a contiguous, non-overlapping time period will be held out separately (temporal split, testing generalization to unseen periods), following the general leakage-avoidance principle applicable to spatio-temporal satellite-to-ground prediction tasks. The exact station subset, time period, and split ratios will be determined once the dataset described in Table I has been assembled.

C. Objective 1 Evaluation Protocol

The surface pollutant regression model will be evaluated on the held-out split using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), the Pearson correlation coefficient (R), and the coefficient of determination (R^2) between predicted and CPCB-measured pollutant concentrations, computed per pollutant. A deterministic AOD-to-pollutant baseline (linear regression, without the CNN-LSTM temporal component) is defined as the comparison baseline against which the proposed model will be assessed; this comparison will indicate the incremental contribution of the temporal modelling component but is not assumed in advance to favour the proposed model.

D. Objective 2 Evaluation Protocol

Because no independently validated ground-truth hotspot dataset is available for this project, the HCHO hotspot-characterization component will not be evaluated using classification accuracy. Instead, the following criteria, which do not require a labelled ground truth, are used: (i) the fire-HCHO Pearson correlation coefficient R defined in (8), computed over the biomass-burning study period; (ii) the spatial overlap between flagged HCHO hotspots and MODIS/VIIRS fire-detection clusters; (iii) the temporal correspondence between the onset/decline of flagged hotspots and known biomass-burning-season windows (e.g., the post-monsoon paddy-residue burning period in the Indo-Gangetic Plain); (iv) the persistence of a flagged hotspot across consecutive available overpasses; and (v) a qualitative visual assessment of the resulting spatial and time-series maps. Table II summarizes this evaluation protocol.

TABLE II

PLANNED EVALUATION FRAMEWORK

Objective	Metric / Criterion	Basis
1: Surface AQI	RMSE, MAE, R, R^2	CPCB held-out data
2: HCHO Hotspot	Fire-HCHO correlation R, Eq.(8)	MODIS/VIIRS co-location
2: HCHO Hotspot	Spatial overlap, temporal correspondence, persistence	Qualitative / descriptive

No numerical values are reported in Table II because the corresponding experiments have not yet been executed; the table defines the evaluation protocol only, consistent with the current stage of the project.

VIII. DISCUSSION

The proposed framework is designed to keep CPCB-regulated AQI estimation and HCHO hotspot characterization scientifically separate while sharing a common data-acquisition and alignment infrastructure. The CNN-LSTM component is designed to jointly capture spatial structure in columnar satellite inputs and temporal persistence in pollutant concentrations; whether this design choice yields a measurable benefit over the AOD-to-pollutant regression baseline defined in Section VII-C is a question for experimental validation and is not assumed here. Similarly, the HCHO anomaly-thresholding approach is designed to avoid the ambiguity of choosing between multiple unrelated candidate techniques

(e.g., clustering versus thresholding) by committing to a single, climatologically grounded formulation; the sensitivity of hotspot detection to the choice of threshold τ and to the length of the climatological reference period is an open methodological question that will be examined once TROPOMI data are processed. The qualitative wind-field comparison described in Section V-C is explicitly a descriptive check rather than a quantitative transport model, and its findings should be interpreted accordingly.

IX. LIMITATIONS

Several limitations are anticipated at the design stage. First, the relationship between satellite-column observations (AOD, trace-gas columns) and surface-level concentrations is influenced by boundary-layer height, vertical aerosol/gas distribution, and meteorological conditions that are not fully captured by the input feature set, which may limit prediction accuracy under certain synoptic conditions. Second, the differing native spatial resolutions of INSAT-3D (0.1 deg grid), TROPOMI (~5.5 km x 3.5 km), MODIS (~1 km), and VIIRS (375 m) require a common-grid reprojection that necessarily discards some sub-grid detail. Third, cloud contamination and TROPOMI quality-assurance filtering can create data gaps, particularly during the monsoon season, that reduce the number of valid observations available for both objectives. Fourth, CPCB ground stations are unevenly distributed and concentrated in urban areas, which constrains the spatial diversity of the training and validation data for Objective 1. Fifth, HCHO retrieval and MODIS/VIIRS fire-detection products carry their own retrieval uncertainty and omission/commission error, which propagates into the fire-HCHO correlation computed in (8). Finally, the wind-field comparison in Section V-C is qualitative and does not constitute a validated atmospheric-transport model, and the generalization of both branches across seasons and geographic regions has not been established and will require dedicated evaluation.

X. CONCLUSION AND FUTURE SCOPE

This paper presented a two-objective satellite-to-surface framework for India that (i) is designed to estimate a CPCB-consistent surface AQI using a CNN-LSTM model driven by INSAT-3D AOD, Sentinel-5P TROPOMI columns, and reanalysis meteorology, and (ii) is designed to characterize HCHO hotspots during biomass-burning periods through a climatological standardized-anomaly formulation combined with MODIS/VIIRS fire co-location and a qualitative wind-field check. The two objectives were deliberately kept scientifically and evaluatively separate because HCHO is not a CPCB AQI pollutant. The mathematical formulation of both branches was presented, along with a reproducible experimental protocol that avoids spatio-temporal data leakage and that does not assume classification accuracy for hotspot detection in the absence of a labelled ground truth. No experimental results are claimed at this stage; parameters such as the anomaly threshold τ , the climatological reference period, the CNN-LSTM hyperparameters, and the final dataset size will be determined during experimental validation. Future work includes: assembling and finalizing the ERA5/IMDAA reanalysis product and resolution; conducting the spatially and temporally held-out experiments defined in Section VII; comparing the CNN-LSTM model against the defined regression baseline; and, if subsequently justified by the results, extending the qualitative wind-field check into a quantitative back-trajectory analysis (e.g., a HYSPLIT-type model) for the transport component of Objective 2.

ACKNOWLEDGMENT

The authors thank Prof. Yuvraj Nikam, Department of Computer Science & Engineering, MIT Art, Design & Technology University, Pune, for guidance during the formulation of this work.

REFERENCES

- [1] Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India, "National Air Quality Index," Technical Report, New Delhi, 2014.
- [2] Central Pollution Control Board (CPCB), "About National Air Quality Index," CPCB, New Delhi. [Online resource].
- [3] M. K. Mishra, "Retrieval of Aerosol Optical Depth From INSAT-3D Imager Over Asian Landmass and Adjoining Ocean: Retrieval Uncertainty and Validation," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 10, pp. 5484-5508, 2018.

- [4] A. van Donkelaar, R. V. Martin, M. Brauer, et al., "Global Estimates of Fine Particulate Matter Using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors," *Environmental Science & Technology*, vol. 50, no. 7, pp. 3762-3772, 2016.
- [5] I. De Smedt, T. Stavrou, J.-F. Muller, R. J. van der A, and M. Van Roozendaal, "Trend Detection in Satellite Observations of Formaldehyde Tropospheric Columns," *Geophysical Research Letters*, vol. 37, no. 18, L18808, 2010.
- [6] J. P. Veefkind, I. Aben, K. McMullan, et al., "TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications," *Remote Sensing of Environment*, vol. 120, pp. 70-83, 2012.
- [7] L. Giglio, W. Schroeder, and C. O. Justice, "The Collection 6 MODIS Active Fire Detection Algorithm and Fire Products," *Remote Sensing of Environment*, vol. 178, pp. 31-41, 2016.
- [8] W. Schroeder, P. Oliva, L. Giglio, and I. A. Csizsar, "The New VIIRS 375 m Active Fire Detection Data Product: Algorithm Description and Initial Assessment," *Remote Sensing of Environment*, vol. 143, pp. 85-96, 2014.
- [9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

