POLITECNICO
MILANO 1863

SCHOOL OF CIVIL
ENVIRONMENTAL
AND LAND MANAGEMENT
ENGINEERING

Sentinel-5P Spatio-Temporal Gap Filling For NO_2 and SO_2 Data

Master of Science Degree in
Geoinformatics Engineering

Author: **Zhanbin Wu**
Student ID: 240246
Advisor: Prof. Maria Antonia Brovelli
Co-advisors: Vasil Yordanov, Jesus Rodrigo Cedeno Jimenez

Academic Year: 2024-25

Abstract

Air quality satellite products, such as Sentinel-5P NO_2 and SO_2, are often hindered by substantial spatio-temporal gaps caused by cloud cover, surface reflectance, and sensor-related constraints. This thesis investigates the missing value patterns of the NO_2 and SO_2 data in the Po Valley region in Northern Italy during 2019–2023, and proposes two models to reconstruct data gaps(missingness or missing observations). One model is a LightGBM baseline and the other is a 3D convolutional neural network (3D CNN); both are trained on the same dataset, with model-specific parameters tuned to maximize performance.

Statistical analysis reveals a 5-year average missing rate of 45.4% for NO_2 and 77.4% for SO_2, with pronounced seasonality, particularly in autumn and winter. To reconstruct these gaps, we train two models that learns joint spatial–temporal dependencies. Both models ingest auxiliary variables, including historical NO_2/SO_2 lags, meteorological drivers (e.g., temperature, wind, pressure), and static factors (e.g., land cover, population density).

Training is carried out using synthetically masking pixels to simulate realistic gap scenarios. This aim to enhance the continuity and usability of Sentinel-5P observations, supporting downstream applications in urban air pollution monitoring, environmental modeling, and policy-making in data-sparse conditions.

In masked validation on 2023, both models reconstruct large gaps, with the 3D CNN yielding lower errors than LightGBM—while LightGBM is competitive and substantially faster.

Keywords: Satellite Gap-filling, Sentinel-5P, NO_2, SO_2, LightGBM, 3D CNN, Air Quality

Abstract in lingua italiana

I prodotti satellitari per la qualità dell'aria, come NO_2 e SO_2 di Sentinel-5P, sono spesso ostacolati da notevoli lacune spazio-temporali dovute alla copertura nuvolosa, alla riflettanza della superficie e a vincoli legati al sensore. Questa tesi analizza i pattern di valori mancanti dei dati di NO_2 e SO_2 nella regione della Pianura Padana, nel Nord Italia, nel periodo 2019–2023, e propone due modelli per ricostruire i gap (missingness o osservazioni mancanti). Un modello è una baseline LightGBM e l'altro è una rete neurale convoluzionale tridimensionale (3D CNN); entrambi sono addestrati sullo stesso dataset, con parametri specifici del modello ottimizzati per massimizzare le prestazioni.

L'analisi statistica evidenzia un tasso medio di mancanze su 5 anni pari al 45.4% per NO_2 e al 77.4% per SO_2, con una stagionalità marcata, in particolare in autunno e inverno. Per ricostruire tali gap, alleniamo due modelli che apprendono dipendenze spazio–temporali congiunte. Entrambi i modelli utilizzano variabili ausiliarie, incluse le serie storiche con ritardi di NO_2/SO_2, forzanti meteorologiche (ad es., temperatura, vento, pressione) e fattori statici (ad es., copertura del suolo, densità di popolazione).

L'addestramento viene effettuato utilizzando pixel di mascheramento sintetico per simulare scenari di gap realistici. L'obiettivo è migliorare la continuità e l'usabilità delle osservazioni di Sentinel-5P, supportando applicazioni a valle nel monitoraggio dell'inquinamento atmosferico urbano, nella modellazione ambientale e nell'elaborazione di politiche in condizioni di dati sparsi.

Nella validazione con mascheramento sul 2023, entrambi i modelli ricostruiscono ampie lacune, con la 3D CNN che ottiene errori inferiori rispetto a LightGBM—mentre LightGBM rimane competitivo e sostanzialmente più rapido.

Parole chiave: Gap-filling satellitare, Sentinel-5P, NO_2, SO_2, LightGBM, CNN 3D, qualità dell'aria

Contents

Introduction

Air quality monitoring has become increasingly critical as urbanization accelerates [1]. Satellite-based observations provide wide-area coverage that complements sparse ground-based monitoring networks [2]. Among current missions, the European Space Agency's Sentinel-5 Precursor (TROPOMI) delivers daily measurements of trace gases such as nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), enabling detailed regional to urban-scale analyses [3].

However, the operational use of these data is limited by spatio-temporal data gaps caused by cloud cover, sensor-specific filtering, and other atmospheric conditions [4]. These gaps break the continuity of daily concentration fields, thereby limiting the reliability of the data for downstream applications such as emission inversion and health exposure assessment [3]. The Po Valley in Northern Italy—one of Europe's most industrialized and densely populated basins, bordered by the Alps and the Apennines—offers a compelling case where stagnation and wintertime conditions exacerbate data gaps [5, 6].

This research develops two models for Sentinel-5P atmospheric imagery. One model is a LightGBM baseline [7] and the other is a 3D convolutional neural network (3D CNN) [8]. The approachs reconstructs missing pixels using meteorology, land cover, topography, temporal patterns and local spatial context, and yields gap-filled datasets that improve continuity and usability for urban air-quality analysis and policy support [9].

The thesis quantifies data gaps from 2019 to 2023 over the Po Valley, details data processing and models design, and evaluates reconstructions with ground observations and uncertainty estimates. The resulting products aim to facilitate reliable use of Sentinel-5P for environmental monitoring in data-sparse conditions.

1 | Background and Research Objectives

1.1. Research Background

Satellite-based atmospheric monitoring has greatly enhanced regional-to-global air-quality observation. Unlike ground-based systems, satellites provide extensive spatial coverage and consistent measurements, enabling the tracking of pollution sources and transport patterns. The Sentinel-5P mission, with its Tropospheric Monitoring Instrument (TROPOMI), delivers daily global high-resolution data, since 2018, on key pollutants such as nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), which have been available on platforms such as Google Earth Engine.

Despite these advantages, Sentinel-5P data are often affected by gaps due to cloud cover, high solar zenith angles, and strict quality control—especially during winter in the Northern Hemisphere. These limitations hinder applications such as emission monitoring, policy evaluation, and public-health guidance, particularly in critical regions like the Po Valley where continuous and reliable air-quality data are essential.

1.2. Problem Formulation

1.2.1. Motivation

Accurate monitoring of NO_2 and SO_2 is vital for assessing air quality and health impacts. While Sentinel-5P provides unprecedented coverage, data gaps reduce its practical utility. Current interpolation methods often fail to capture complex atmospheric processes, lacking temporal and meteorological integration [10]. Advances in deep learning, especially 3D CNNs, offer new ways to effectively reconstruct missing data by combining satellite imagery with auxiliary inputs [8].

1.2.2. Problem Statement

Daily Sentinel-5P Level-3 NO_2 and SO_2 imagery over the Po Valley exhibits substantial, spatially fragmented gaps caused by clouds, viewing geometry, and stringent L2–L3 quality assurance (QA) filtering. These gaps limit time-series analysis and urban exposure assessment. The aim is to reconstruct the missing pixels and to produce spatially and temporally consistent daily NO_2 and SO_2 maps for the period 2019–2023.

Inputs

The following multi-source datasets are integrated and aligned on a common spatio-temporal grid to serve as inputs for the models:

1. Primary data: Sentinel-5P Level-3 NO_2 and SO_2 daily grids with QA masks (2019–2023, Northern Italy).
2. Dynamic predictors (ERA5): planetary boundary layer height, 2-m air temperature, 10-m wind speed and direction (from u, v), total precipitation, clear-sky surface net radiation, surface net thermal radiation, surface pressure.
3. Static geography: elevation and slope, land cover, population density.
4. Temporal features: day-of-year encoded with sine and cosine, and day-of-week or weekday flag.
5. Spatio-temporal priors: lagged columns (for example $t-1$) and neighborhood statistics such as 3×3 means, computed separately for NO_2 and SO_2.

Outputs

The workflow produces two kinds of outputs:

1. Gap-filled, daily NO_2 and SO_2 datasets.
2. Two predictive models (LightGBM and 3D-CNN) that provides per-pixel reconstructions for both pollutants.

Objectives

Building upon the research background and challenges outlined above, this study aims to develop a robust, integrated framework for reconstructing the extensive gaps in Sentinel-5P NO_2 and SO_2 data over the Po Valley. The specific objectives are as follows:

1. Quantify the extent, frequency, and seasonal distribution of missing data in Sentinel-

5P NO_2 and SO_2 over the Po Valley during 2019–2023, establishing a benchmark for the severity of data gaps.

2. Develop two models: a LightGBM model and a 3D CNN model for spatio–temporal patterns—trained on the same dataset with model-specific hyperparameter tuning.
3. Evaluate reconstructions against withheld valid observations and in synthetic-gap tests.
4. Demonstrate practical utility by generating continuous time-series and improved air-pollution maps from the gap-filled datasets, illustrating value for environmental monitoring and decision-making.
5. Provide an open, reproducible pipeline for Sentinel-5P data reconstruction that can be adapted to other regions and satellite products.

Challenges

Based on preliminary inspection of the Sentinel–5P data and prior studies, we identify several challenges that motivate our design choices. The following items summarize the issues to be tackled:

1. High and seasonally variable coverage gaps: daily level-3 NO_2 and SO_2 fields over the Po Valley frequently exhibit extensive missing areas, with winter/autumn conditions particularly affected by clouds, solar-zenith geometry and QA filtering.
2. Strong spatio-temporal dependence: both pollutants depend on nearby pixels and recent days, so models must capture local spatial structure and short-term dynamics.
3. Multi-source harmonization: aligning meteorology, static layers, and satellite data on a common daily grid while preventing data leakage.
4. Operational scale: five-year daily stacks with multi-layer inputs require efficient training and inference, and robust handling of extreme missing patterns, especially for SO_2.

1.3. Study Area: The Po Valley, Italy

The Po Valley in Italy represents an ideal area for studying atmospheric pollution and validating satellite-based data reconstruction methods. Bordered by the Alps to the north and west and the Apennines to the south, the region features predominantly flat terrain that forms a characteristic basin topography. This basin effect, combined with

frequent temperature inversions during winter, inhibits pollutant dispersion and leads to the accumulation of airborne contaminants [11].

As one of Italy's most densely populated and economically active regions, the Po Valley suffers from severe air pollution, characterized by high concentrations of secondary aerosols[11] . These pollutants form in the atmosphere from reactions between agricultural ammonia and industrial as well as vehicular emissions [11, 12] . This distinct seasonality provides a realistic and challenging setting for evaluating the performance and robustness of spatio-temporal data reconstruction methods under varying pollution scenarios.

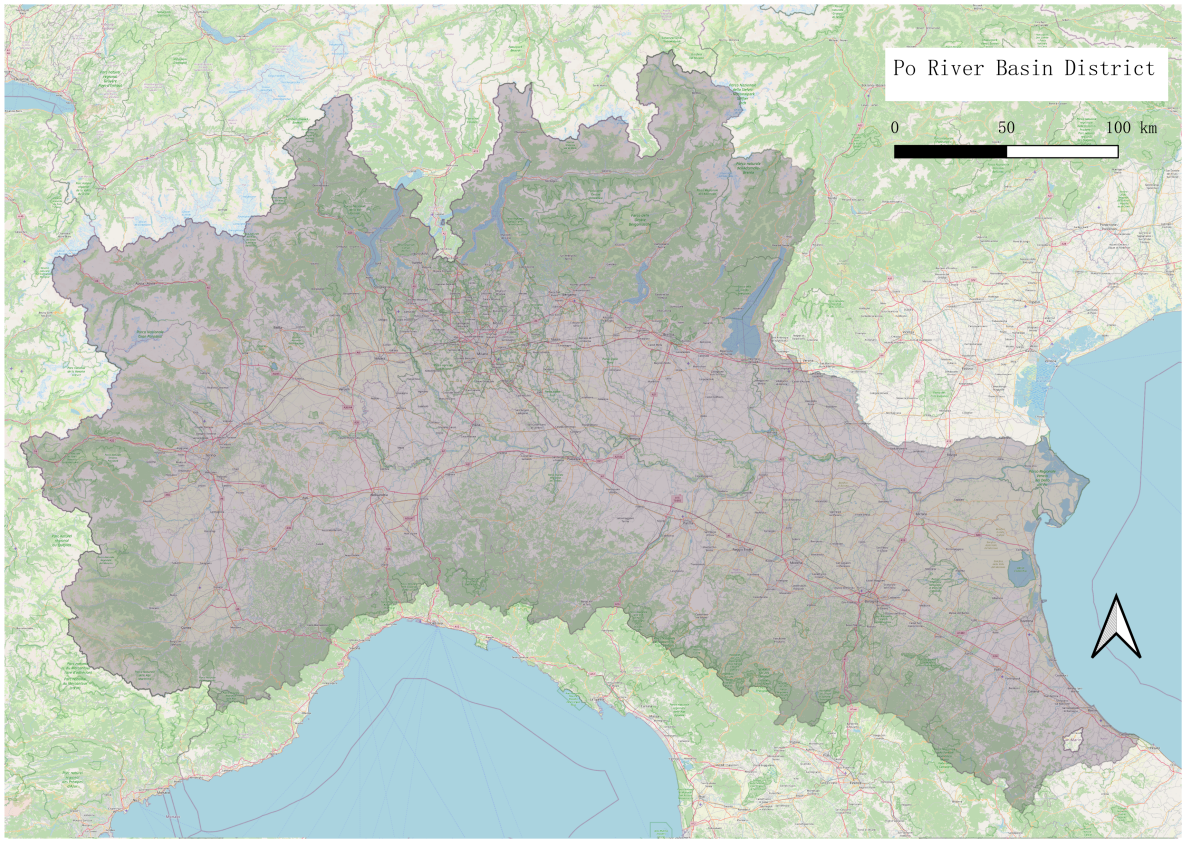


Figure 1.1: Study area: Po River Plain; outline = Po River Basin District. Basemap: © OpenStreetMap contributors (ODbL). Map created by the author using QGIS.

1.4. Overview of Sentinel-5P and Target Pollutants

The Sentinel-5P satellite, launched in 2017, carries the TROPOMI instrument which provides daily global monitoring of atmospheric composition at high spatial resolution. This represents a substantial improvement over previous satellite missions, enabling enhanced characterization of pollution sources and improved urban-scale monitoring capabilities.

Nitrogen dioxide[13] serves as a critical indicator of anthropogenic activity, primarily originating from fossil fuel combustion in transportation and industrial sectors. Its relatively short atmospheric lifetime makes it an excellent tracer for local and regional emission sources. Sulfur dioxide[14] emissions predominantly arise from industrial processes, power generation facilities, and volcanic activity, with significant implications for both air quality and climate through aerosol formation processes.

Both NO_2 and SO_2 data products incorporate comprehensive QA frameworks that flag conditions affecting measurement reliability (e.g., cloud contamination, surface albedo effects, and suboptimal viewing geometries) [15, 16]. While these quality controls ensure scientific data integrity, they simultaneously introduce substantial gaps in the observational record. The resulting incomplete datasets present significant challenges for continuous air quality monitoring and trend analysis, particularly in regions with frequent cloud cover or challenging retrieval conditions.

2 | Literature Review

2.1. Satellite-based Atmospheric Monitoring

Satellite remote sensing has become an indispensable tool for global atmospheric monitoring, providing extensive spatial coverage that powerfully complements sparse ground-based station networks. Advanced imaging spectrometers on polar-orbiting platforms, such as the TROPOMI instrument aboard Sentinel-5P, deliver daily (or near-daily) observations of trace-gas column densities (e.g., NO_2, SO_2) at kilometer-scale resolution, alongside essential quality assurance (QA) metadata that determine data usability [17].

Unlike relatively stable land-surface variables (e.g., LST or soil moisture), atmospheric pollutants exhibit pronounced spatio-temporal dynamics driven by intermittent emissions, meteorologically influenced advection and diffusion, and diurnal variations in planetary boundary layer height [18–20]. This strong non-stationarity means that gaps in satellite products are seldom random. Instead, missing values arise predominantly from cloud and aerosol obscuration, unfavorable viewing geometries (e.g., high solar zenith angles), and retrieval screening based on QA flags—factors that are often intrinsically correlated with the underlying atmospheric state being measured [21].

Operational processing pipelines (e.g., using HARP/harpconvert) apply strict filtering to ensure scientific-grade quality, which further accentuates the MNAR problem. Standard steps include enforcing thresholds for cloud fraction (e.g., `cloud_fraction_crb < 0.3`), solar geometry (e.g., `solar_zenith_angle < 60°`), and QA values (e.g., `QA > 0.5`), often combined with alignment to overpass-local-time windows to minimize representativeness errors [13]. While these screenings yield high-quality, analysis-ready datasets, they can introduce systematically non-random gaps (MNAR), which bias downstream analyses if not handled properly [15, 16, 22, 23]. MNAR (Missing Not At Random) means the probability that an observation is missing depends on its unobserved true value (even after conditioning on observed covariates), in contrast to MCAR/MAR [22, 23]. This underscores the critical need for robust gap-filling schemes that respect atmospheric dynamics, preserve critical features like pollution peaks, and quantify reconstruction uncertainty.

2.2. Spatio-temporal Gaps in Remote Sensing

Gaps in remote sensing time series are a fundamental characteristic rather than merely a data-availability issue, and they must be explicitly characterized to guide reconstruction method selection and evaluation. We categorize gaps into three primary types [24, 25]:

1. Random and sparse gaps: from sensor noise or isolated retrieval failures.
2. Structured and systematic gaps: from clouds, orbital swaths, or sensor failure (e.g., Landsat 7 SLC-off), creating coherent missing regions.
3. Missing-Not-At-Random (Missing Not At Random) gaps: where missingness correlates with the target variable (e.g., high pollution causing algorithm failure and systematic undersampling).

In this study the dominant mechanisms are (1)–(2): cloud screening and orbital/terrain effects that create coherent gaps in S5P NO_2/SO_2. MNAR may occur locally, but it is not identifiable without extra assumptions [22].

The prevalence of MNAR gaps necessitates evaluation metrics beyond pixel-wise accuracy (e.g., RMSE). A comprehensive assessment should consider [26, 27]:

- Temporal consistency (preserving cycles and trends).
- Spatial coherence (maintaining gradients and plume structure).
- Peak preservation (avoiding attenuation of extremes).
- Physical plausibility (e.g., non-negative concentrations).
- Uncertainty quantification (providing prediction intervals).

Gap-filling methodologies generally follow three paradigms [28]:

1. Temporal extrapolation (using a pixel's own history).
2. Spatial propagation (using nearby valid pixels).
3. Cross-variable learning (using correlated auxiliary data).

Hybrid approaches that combine these paradigms are often essential for addressing complex atmospheric MNAR patterns.

2.3. Gap-filling Methods

Remote sensing gap-filling methods are broadly categorized into three paradigms: deterministic interpolation, machine learning (ML), and deep learning (DL). Their applicability hinges on how they leverage information—from the data's own spatio-temporal neighborhood, from auxiliary predictor variables, or by learning complex patterns directly from data patches.

The following subsections critically analyze a representative method from each category, evaluating its strengths and limitations for reconstructing atmospheric pollutants. This review motivates our hybrid approach, which integrates concepts across these paradigms to address the specific challenges of NO_2 and SO_2 gap-filling.

2.3.1. Deterministic Spatio-temporal Interpolation

For variables exhibiting strong seasonal regularity and smooth spatio-temporal dynamics, such as land surface temperature (LST) and evapotranspiration (ET), deterministic methods based on spatial and temporal neighbors offer a simple and effective solution. Siabi et al. [21] proposed the Differential Dynamic Search Distance Algorithm (DDSDA), a spatio-temporal gap-filling workflow that operates without requiring model training. The method proceeds by: (1) identifying valid pixels in a spatial neighborhood around the gap; (2) selecting auxiliary images from the same seasonal period that contain valid data at the gap location; (3) matching spatial patterns between the target (gapped) and auxiliary images; (4) estimating the missing value using a scaled difference derived from the pattern match; and (5) combining multiple estimates from different auxiliary images and neighbors using distance-based weighting.

The authors demonstrated that DDSDA achieves high accuracy ($R^2 > 0.9$) on MODIS LST and ET products, even for very large gaps (up to 96% missing data), while maintaining computational efficiency and linear time complexity. Its key advantages are simplicity, minimal parameterization, and complete parallelizability, as each gap pixel is processed independently.

However, the method's core assumptions—that spatial patterns are temporally transferable and that sufficient valid auxiliary images exist—limit its applicability for atmospheric pollutants like NO_2 and SO_2. These gases exhibit irregular, quality-driven data gaps and rapid, nonlinear dynamics driven by emissions and meteorology, which violate the method's prerequisites. Furthermore, DDSDA does not integrate auxiliary predictive variables (e.g., meteorological fields) and provides no inherent uncertainty quantification.

In our work on Sentinel-5P NO_2/SO_2 reconstruction, we draw inspiration from the concept of leveraging spatio-temporal neighborhoods. However, to address the above limitations, we employ a learning-based framework where a LightGBM model integrates meteorological and static geographical predictors, and a subsequent 3D CNN refines the spatio-temporal structure. This approach is more robust for handling the sporadic, large gaps and complex physicochemical processes characteristic of trace gas time series.

2.3.2. Machine Learning Approaches

Liu et al. [9] present a robust gap-filling approach for satellite-based soil moisture time series by integrating satellite observations, model-driven knowledge, and spatio-temporal machine learning. Their study addresses the challenge that large, irregular gaps limit the usability of daily products and proposes a workflow that couples spatio-temporal learning with spatial statistical calibration.

Their framework comprises three main components: (1) explanatory variables from multi-source data are screened and bias-corrected to ensure consistency; (2) a machine-learning regressor (Random Forest) is trained within an adaptive spatio-temporal window, leveraging pixels from nearby locations and adjacent days to predict missing values; (3) model residuals are spatially calibrated using geographically weighted regression and smoothed with a Gaussian filter to reduce systematic biases and enhance spatial coherence.

Methodologically, the contribution lies in combining rigorous variable processing, neighborhood-aware learning that respects local spatio-temporal structure, and an explicit residual-calibration step. The approach was validated against in situ measurements and showed superior performance compared to baseline methods.

The relevance of this work to our study is twofold. Conceptually, we adopt the same principles of multi-source integration and spatio-temporal context learning. Practically, we adapt and extend this framework for atmospheric trace gases by employing a two-stage model (LightGBM prior followed by 3D CNN refinement) and tailoring the predictor set to air-quality processes—including boundary-layer dynamics, wind fields, precipitation, radiation, land cover, population, and road proximity. We also retain the evaluation protocol using space-time holdout and seasonal stratification, with special attention paid to winter days with near-zero native coverage for SO_2.

2.3.3. Partial-Convolution–based Spatio-temporal Approaches

Appel et al. [29] proposed a deep learning approach to address the challenge of extensive gaps in satellite image time series. Their method employs a U-Net architecture integrated with three-dimensional partial convolutional layers. Each layer takes both the data and a corresponding binary mask as input, which allows the model to ignore missing values during convolution operations while progressively updating the mask to reflect filled regions.

The model processes spatio-temporal patches (e.g., of size $128 \times 128 \times 16$). The encoder utilizes strided partial convolutions for downsampling, while the decoder uses upsampling layers and skip connections to reconstruct the original resolution. During training, artificial gaps are generated using Gaussian random fields to simulate missing data patterns. The model is optimized using the Adam algorithm, with a loss function based on the Mean Absolute Error (MAE) computed solely on the artificially masked pixels.

In comprehensive benchmarks, the proposed method was compared against naive baselines (e.g., block-wise mean, temporal interpolation), statistical approaches (e.g., gapfill based on local quantile regression, and stmra as a multi-resolution approximation of Gaussian processes), and a standard 3D convolutional neural network (Conv3D). The large partial convolution model (STpconv_L) achieved the best performance in terms of RMSE, correlation coefficient (CC), and R^2, while yielding competitive MAE. Notably, its inference speed was orders of magnitude faster than the statistical methods.

The authors also note limitations, including a tendency to produce overly smooth predictions due to the pixel-wise loss, the lack of inherent uncertainty quantification, and potential performance degradation under extrapolation to out-of-distribution data.

For our application involving Sentinel-5P NO_2 and SO_2 time series, we adopt the core concepts of mask-aware partial convolutional blocks and synthetic gap training. We extend this foundation by incorporating a LightGBM prior that leverages meteorological and static predictors to enhance initial estimates. Furthermore, our evaluation is specifically tailored to address the challenges of winter periods with prevalent near-zero concentration values.

2.4. AI for Atmospheric Reconstruction

The evolution from deterministic interpolation to machine and deep learning represents a paradigm shift in reconstructing atmospheric data. Current state-of-the-art approaches increasingly favor hybrid models that leverage both spatio-temporal information and aux-

iliary predictive variables (e.g., meteorology, topography) to address the critical challenge of Missing-Not-At-Random (MNAR) gaps. This shift emphasizes moving beyond mere gap-filling towards generating physically plausible and uncertainty-aware reconstructions that preserve critical features like pollution plumes and extreme events. Our work builds upon this foundation by implementing and evaluating such an integrated approach.

3 | Data and Preprocessing

This chapter documents the data acquisition and preprocessing workflow used to build a unified, analysis-ready panel for Sentinel-5P NO_2 and SO_2 over the Po Valley (2019–2023). We describe how each dataset is sourced and formatted, and how raw products are transformed into harmonized inputs for subsequent modeling and analysis. Specifically, we detail: (1) the Sentinel-5P observations and gridding scheme; (2) the auxiliary datasets used to inform reconstruction (ERA5 meteorology, static geographical layers, and anthropogenic indicators).

3.1. Data Sources and Features Description

Table 3.1: Data and Features Description

Category	Data	Description
Basic Data	AOI	Northern Italy, Po River Plain.
	NO_2	band = `tropospheric_NO2_column_number_density`.
	SO_2	band = `SO2_column_number_density`.
Meteorological Variables	boundary_layer height (BLH)	Low PBLH traps NO_2/SO_2 near the surface; high PBLH allows dispersion.
	wind_speed	Higher wind speeds lead to dilution and lower NO_2/SO_2. wind_speed $= \sqrt{u\_\text{wind}^2 + v\_\text{wind}^2}$
	wind_direction	Identifies upwind source direction (traffic/industrial). wind_direction $= \text{atan2}(u\_\text{wind}, v\_\text{wind}) \times 180/\pi$
	total precipitation	Rain removes NO_2/SO_2 via wet scavenging, reducing concentration.
	2m_temperature	Warmer temperatures accelerate $NO_2 \rightarrow O_3$ conversion and alter stability.
	surface_net _solar_radiation (clear sky)	Amount of solar radiation reaching the surface; enhances photochemistry, affects NO_2/SO_2.
	surface_net _thermal _radiation	Net longwave radiation; influences energy balance, stability, and dispersion.
	surface_pressure	Lower pressure in valleys may trap pollutants; higher pressure often links to stable, stagnant air.
Static Geographical Variables	elevation	High-altitude or valley areas may accumulate pollutants differently.
	slope	Terrain complexity affects accumulation and ventilation (valleys vs. ridges).
	lulc (Land Use / Cover)	Urban/industrial emit more; forests/water emit less (used as one-hot categories).
Anthropogenic Variables	population _density	Higher population means more traffic and energy use. Perfer to use only 2020.
Temporal Variables	day_of_week	Captures weekday–weekend differences in emissions ("weekend effect").
	day_of_year	Seasonal cycle (e.g., winter heating, summer photochemistry).
Spatial Variables	no2_lag_1day / so2_lag_1day	Yesterday's levels predict today's; It reflects persistence (shifted by 1 day).
	mean_no2 _neighbor/mean _so2_neighbor	Neighborhood mean within a local window; useful for spatial interpolation.

Data	Dataset / Derivation and Format	Source
AOI	Po River Basin District boundary (SHP)	adbpo.it
NO2	Sentinel-5P TROPOMI OFFL L3 NO_2, per-year multiband GeoTIFF (2019–2023)	Google Earth Engine
SO2	Sentinel-5P TROPOMI OFFL L3 SO_2, per-year multiband GeoTIFF (2019–2023)	Google Earth Engine
boundary_layer _height (BLH)	ERA5 single levels — boundary layer height	Climate Data Store
wind_speed	ERA5 single levels — 10 m wind components (u10, v10)	Climate Data Store
wind_direction	ERA5 single levels — 10 m wind components (u10, v10)	Climate Data Store
total precipitation	ERA5 single levels — total precipitation	Climate Data Store
2m_temperature	ERA5 single levels — 2m temperature	Climate Data Store
surface_net_solar _radiation, clear sky	ERA5 single levels — surface net solar radiation (clear-sky)	Climate Data Store
surface_net_thermal _radiation	ERA5 single levels — surface net thermal radiation	Climate Data Store
surface_pressure	ERA5 single levels — surface pressure	Climate Data Store
elevation	SRTM 30m (SRTMGL1_003)	Google Earth Engine
slope	Derived from SRTM DEM	Google Earth Engine
day_of_week	Weekday weights 2019–2023 (from S5P L3 NO_2/SO_2 daily global means; weekday=1, weekend=weekend/weekday)	Python + pandas
day_of_year	DOY (sin/cos), NO_2-2019 grid; bands=days (365/366); per-band constant.	Python + pandas
lulc (Land Use/Cover)	ESA WorldCover 2020 v100 (10m)	Google Earth Engine
population_density	WorldPop UN-Adjusted 2020 (1km)	worldpop.org
no2_lag_1day / so2_lag_1day	Computed from S5P L3 NO_2/SO_2: for day d = 0.5×d + 0.5×(d1); day1 blends previous year's last day if available; NaN→same-day global mean; 0→1e6.	Python + CAMS Data
mean_no2_neighbor / mean_so2_neighbor	Computed from Sentinel-5P TROPOMI OFFL L3 NO_2/SO_2 (2019–2023): 3×3 neighborhood mean (excluding center), NaN handling as described.	Python + NO2 / SO2 Data

Table 3.2: Data sources and formats.

Table 3.1 summarizes the variables used in this study, grouped into Basic, Meteorological, Static Geographicall, Anthropogenic, Temporal and Spatial features, and explains their physical rationale for NO_2/SO_2 variability. In turn, Table 3.2 enumerates the concrete datasets, file formats, and providers that instantiate those variables (e.g., Sentinel-5P products, ERA5 fields, DEM/land cover, and derived spatio-temporal context features). Together, these tables function as a concise data dictionary: the first clarifies *what* each feature represents and why it matters, while the second specifies *where* it comes from and *how* it is stored. All listed sources are harmonized to a common grid and temporal cadence before feature derivation (lags and neighborhood statistics), ensuring consistency and full reproducibility of the modeling pipeline.

3.2. Sentinel-5P NO_2 and SO_2 Data

3.2.1. Data Download and Aggregation

We generated year-wise daily composites of Sentinel-5P TROPOMI NO_2 using Google Earth Engine (GEE) [30–33]. For each calendar year (2019–2023), we filtered the Level-3 Near-Real-Time product `COPERNICUS/S5P/NRTI/L3_NO2` by AOI and date, and selected the tropospheric column density band (`tropospheric_NO2_column_number_density`). For every day within the target year, all available orbits were aggregated to a daily mean image; if a day had no observations inside the AOI, a masked (empty) image was inserted. Crucially, we encode each calendar day as one band; therefore, each annual GeoTIFF contains 365/366 bands (one per day) stacked into a single multiband raster, preserving a complete and ordered daily time axis for subsequent alignment and gap diagnostics.

The daily bands were renamed using a human-readable key (`DD-MM-YYYY_NO2`), clipped to the AOI, and exported at ∼1.1 km nominal scale (`scale = 1113.2` m) in `EPSG:4326` with the pattern `NO2_Daily_Multiband_{YYYY}.tif`. In parallel, we recorded "missing dates" (calendar days with no valid granules over the AOI), computed as the set difference between the full date sequence and the collection's time stamps, to quantify annual/seasonal completeness and to drive the gap-filling experiments. The same procedure was applied to SO_2 using `COPERNICUS/S5P/NRTI/L3_SO2` and the band `SO2_column_number_density`. All rasters are stored as `Float32` for consistency across days and years.

After export, we performed an interactive sanity check in Google Colab. Annual multiband GeoTIFFs were read with rasterio, daily bands were rendered with matplotlib using a common color scale set by the 99th percentile across all years, and ipywidgets sliders enabled browsing by year and day. This interactive visualization verified band–date align-

ment and AOI clipping, and helped flag days with missing or anomalous coverage. The same procedure was applied to SO_2 Data.

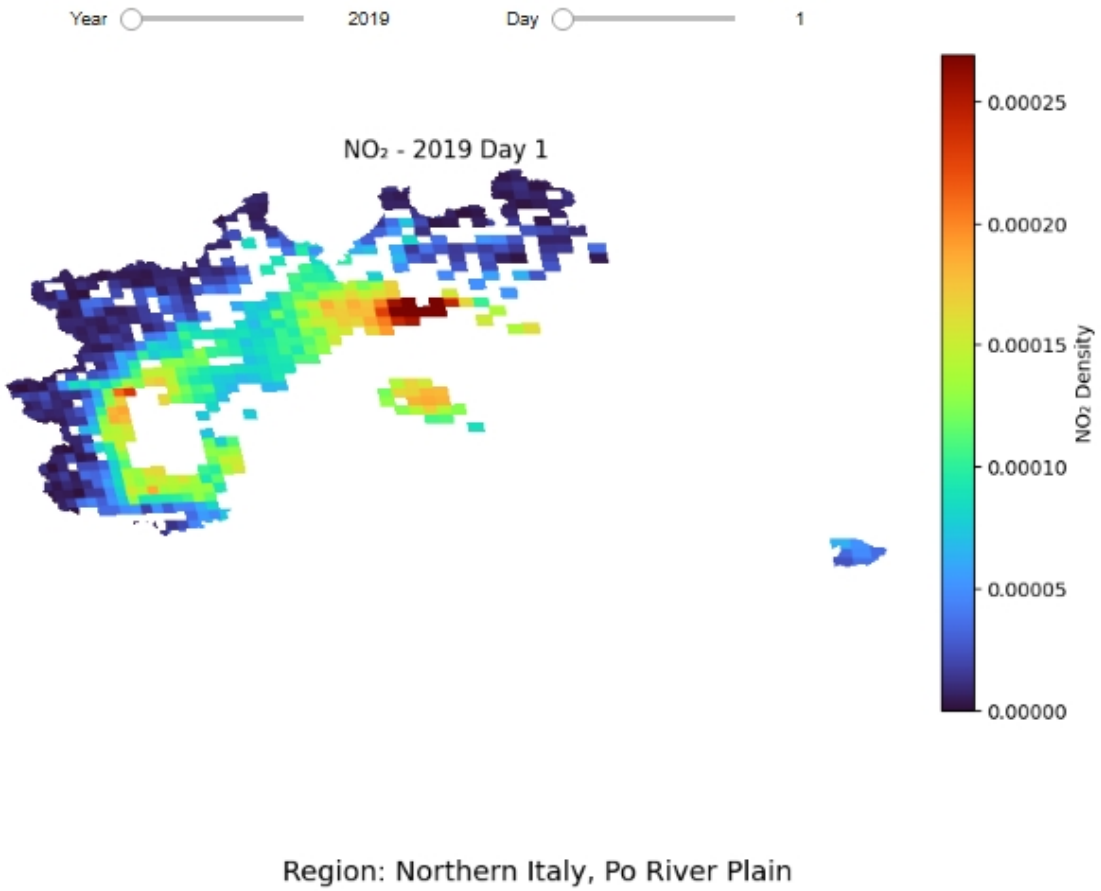


Figure 3.1: Sanity check view of daily NO_2 fields (example: 2019 Day 1). The color bar shows tropospheric column density in mol m^{-2}.

3.3. Auxiliary Datasets

To support both the descriptive gap analysis and the gap-filling models, we assembled a set of auxiliary predictors that capture atmospheric dynamics, underlying geography, human activity, temporal cycles, and local spatio-temporal context. Variables were selected based on established links with pollutant formation and dispersion (e.g., boundary-layer height, wind, precipitation, radiation, temperature), and harmonized to the Sentinel-5P grid and daily cadence (hourly ERA5 aggregated to daily means/sums; spatial resampling).

These variables are grouped into five families: (1) meteorological drivers, (2) static geographical features, (3) anthropogenic indicators, (4) temporal features, and (5) spatial context features—and are summarized in Tables 3.1–3.2. The following subsections detail

the rationale, preprocessing, and formats for each group.

3.3.1. Meteorological Drivers (ERA5)

We use ERA5 single-level reanalysis fields from the Copernicus Climate Data Store (CDS)[34, 35]. All meteorological variables are retrieved with the *same* CDS request over Northern Italy (5–15°E, 41–48°N) for 2019–2023, stored as NetCDF on a regular longitude and latitude grid (EPSG:4326) at 0.25° resolution, with four hourly stamps per day (12:00–15:00 UTC). This yields identical native dimensions `(time, latitude, longitude)` across variables (domain size 29×41) and a total of 7,304 time steps. The ERA5 domain fully encloses the Sentinel-5P AOI used in this study.

The following eight single-level fields are used: 2 m temperature (`t2m`), 10 m wind components (`u10`, `v10`), boundary-layer height (`blh`), total precipitation (`tp`), surface pressure (`sp`), surface net solar radiation under clear-sky conditions, and surface net thermal radiation. All share the same spatio-temporal resolution and coverage; only physical meaning and native units differ.

3.3.2. Static Geographical Features

To capture terrain and surface characteristics that modulate pollutant accumulation and dispersion, we use three time-invariant layers from Google Earth Engine(GEE): (1) elevation from the SRTM GL1 DEM (30 m) [36], (2) slope derived from the same DEM, and (3) land use/land cover (LULC) from ESA WorldCover 2020 v100 (10 m) [37]. These static features summarise topographic confinement/ventilation (elevation, slope) and emission/uptake differences across land covers (LULC).

SRTM GL1 provides a global 1-arcsecond DEM referenced to WGS84/EPSG:4326 with metres as elevation units.

Slope is derived from the DEM with a 3×3 Horn finite-difference operator: local east–west and north–south gradients are estimated from the surrounding 8 cells, then combined to obtain the slope angle in degrees. This is equivalent to the GEE "Horn" implementation [38].

ESA WorldCover 2020 is a 10 m categorical map in EPSG:4326 with 11 classes (e.g., tree cover, cropland, built-up, water).

All layers were exported for the study domain (Po Valley AOI) and inspected against the Sentinel-5P reference raster (EPSG:4326, 0.01°; 300×621 cells, bounds 6.62–12.83°E and

43.64–46.64°N).

3.3.3. Anthropogenic Indicators

We proxy human activity and emission potential with gridded population density from WorldPop UN-Adjusted 2020 (1 km) [39–41]. The native product is provided in EPSG:4326 as people km^{-2} on a regular grid of $\sim$0.008333° ($\approx$1 km) for Italy.

For the national layer (bounds covering Italy), we checked that the coordinates are monotonic with a ~ 1 km spacing, and confirmed complete coverage of the Po Valley AOI. Before masking, summary statistics are biased by the NoData code (e.g., negative mean); after masking NoData and harmonising to the Sentinel-5P reference grid (EPSG:4326, 0.01°, 300×621 cells), the aligned population raster shows a plausible distribution with: range 0–13,963 people km^{-2}, mean 148.4 people km^{-2}, and std 489.3 people km^{-2} (valid pixels: 132,273; non-zero: 119,968). Zero-valued pixels correspond to sparsely populated or uninhabited cells and are retained.

3.3.4. Temporal Features

To encode systematic temporal cycles that affect atmospheric composition we use two date-level descriptors: a weekly cycle (day-of-week plus a pollutant-specific weekend-to-weekday mean ratio) and an annual cycle (day-of-year encoded as a sine–cosine pair). All descriptors are computed on the training period only, then merged into the S5P panel by date and standardised.

The first descriptor captures weekly modulation of human activity by combining day-of-week indicators with a pollutant-specific weekend-to-weekday mean ratio estimated from historical S5P daily means (values < 1 indicate typical weekend reductions; values > 1 indicate increases). The calculation method is as follows.

$$\text{weekend\_weight} = \frac{\text{weekend\_mean}}{\text{weekday\_mean}}$$

The second places each date on a cyclic calendar representation so the model reads its position within the year—i.e., season timing and strength—without a discontinuity between December 31 and January 1.

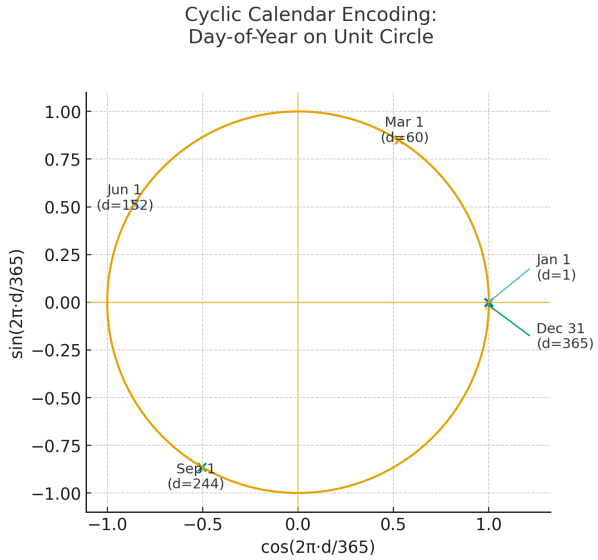


Figure 3.2: Cyclic calendar encoding of the day-of-year. Each date d is mapped to a point on the unit circle via $(\cos(2\pi d/365), \sin(2\pi d/365))$. This preserves yearly continuity (e.g., Dec 31 and Jan 1 are adjacent) and helps the model read a date's position within the year without a discontinuity.

3.3.5. Spatial Context Features

We construct two context features for each pollutant that act as simple priors when satellite retrievals are sparse or masked.

(1) One-day CAMS lag prior. `no2_lag_1day` and `so2_lag_1day` are the CAMS concentrations at the same grid cell on day $t-1$. They represent short-term atmospheric persistence driven by emissions and slowly varying meteorology. We use them because CAMS has full spatio-temporal coverage.

(2) Local 3×3 neighbourhood mean on the S5P grid. `mean_no2_neighbor` and `mean_so2_neighbor` are the mean of valid neighbouring pixels in a 3×3 window on the same day t. They represent the local background around the target pixel, exploiting spatial correlation to smooth pixel-level noise and remaining robust when the centre pixel is cloudy or masked.

All context features are keyed by (date, cell), merged into the S5P panel, and standardised on the training split. Together, the lag prior supplies temporal memory and the neighbourhood mean supplies spatial background, improving stability while allowing the model to learn deviations such as hotspots and events.

4 | Methodology and Implementation

This chapter turns the data assets of Chapter 3 into an analysis-ready panel and the final gap-filling system. We first outline the overall workflow and quantify the observation gaps in Sentinel-5P NO_2 and SO_2, then verify and align auxiliary datasets to the S5P grid.

Building on the gap diagnostics, we construct features spanning meteorology, static geography, temporal cycles, and spatio-temporal context, and train two complementary models: a LightGBM and a 3D CNN.

We detail the training setup, and evaluation metrics, followed by the inference pipeline for producing gridded reconstructions and the procedures for output generation and archiving.

4.1. Overall workflow

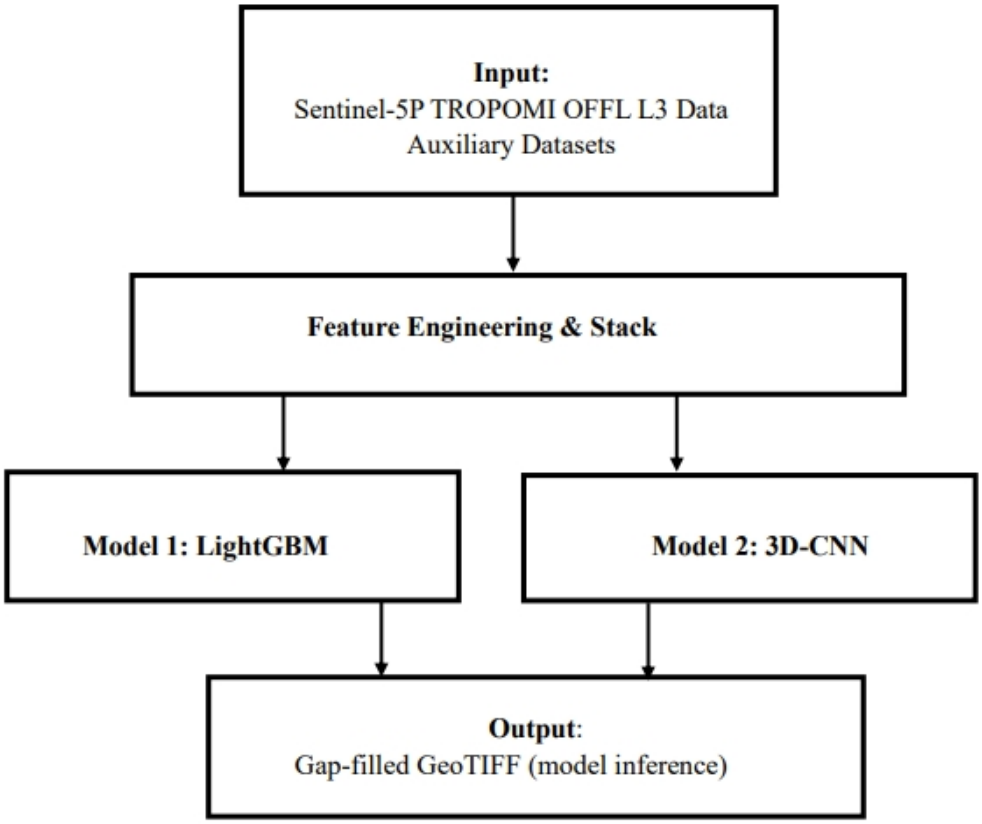


Figure 4.1: Overall workflow from inputs (S5P NO_2/SO_2 and auxiliary datasets) through feature engineering and two complementary models (LightGBM and 3D-CNN) to model-inferred, gap-filled GeoTIFF outputs.

We turn raw observations into daily gap-filled maps in four steps. The Figure 4.1 summarises the pipeline: we start from S5P TROPOMI OFFL L3 NO_2/SO_2 (2019–2023) together with auxiliary sources, harmonise all inputs to the S5P grid (CRS, resolution, AOI clip, pixel alignment), then build a feature stack combining meteorology, static geography, population, temporal signals (day-of-week weights, day-of-year sine/cosine) and spatio-temporal context (CAMS lag-1d, 3×3 neighbour mean). A LightGBM model and a 3D-CNN model are trained and hyperparameter tuning, after which model inference produces gap-filled GeoTIFF outputs that are assessed by masked validation check.

4.2. Data Gap Exploratory Analysis

To characterise missing observations, we define a gap as any S5P pixel-day that is missing or flagged invalid by the product QA/cloud screening. Starting from the daily multi-band images (one band per day), we build a binary gap mask $G(t, \mathbf{x})$ and summarise it across time and space to answer: how often, when, and where the gaps occur, and how they differ between NO_2 and SO_2.

These diagnostics inform downstream choices (daily cadence, feature set, validation splits) and provide a baseline for assessing reconstruction quality.

4.2.1. NO_2 and SO_2 Annual Trend

Annual coverage was quantified by reading the multiband GeoTIFFs (2019–2023) in Google Colab and treating pixels with value ≤ 0 or NaN as missing. For each pixel we summed the number of missing days per year, accumulated over the five years, and divided by the total number of days to obtain a pixel-wise 5-year data gaps ratio. The ratio was then masked to the AOI and summarized using robust statistics (mean, median, standard deviation, and 90th percentile). To facilitate interpretation we also grouped pixels into four categories by ($<$20%, 20–50%, 50–80%, $\geq$80%). The resulting map highlights coherent spatial patterns of coverage loss, while the summary statistics indicate that most pixels fall within the 20–50% range, with a small but non-negligible share in the 50–80% range. The same procedure was applied to SO using its annual multiband stacks.

4.2.2. Pre-2021 Winter Square Analysis

A notable artifact appears in the annual maps: in winters before February 2021 (e.g., 2019–2020 and February 2021), a square, tile-shaped region in the southwestern AOI shows systematically higher gap fractions than surrounding areas. To verify and quantify this, we produced monthly gap-fraction maps for 2019–2023 and inspected them month by month. .

4.2.3. NO_2 and SO_2 Seasonal Data Gaps Pattern

Seasonal patterns were derived by aggregating daily coverage into four fixed seasons: DJF (December–February), MAM (March–May), JJA (June–August), and SON (September–November). For each species and season, the pixel-wise seasonal data gaps ratio was computed as the fraction of days flagged as missing (value ≤ 0 or NaN) within that

season, and then summarized the missing rate statistics over the AOI.

4.2.4. Comparative Analysis of NO_2/SO_2

For comparability across species, AOI-wide pixel statistics were used to contrast NO_2 and SO_2 gap levels and seasonality over 2019–2023.

In this part, yearly daily multi-band GeoTIFFs for NO_2 and SO_2 were read, an AOI mask was built from the shapefile, and pixels with values ≤ 0 or NaN were treated as missing. Missing days were summed and divided by total days to obtain AOI-mean missing rates for 2019–2023; seasonal indices were constructed with `calendar.monthrange` to compute seasonal rates. Results were organized into comparison tables and simple text bars to display NO_2 vs. SO_2 (overall and by season).

4.2.5. Interannual Trends in NO_2/SO_2 Data Gaps

We applied the same pixel-wise trend workflow to NO_2 and SO_2. For each year (2019–2023), we converted the multiband GeoTIFF (one band per day) into a per-pixel data gaps map by counting days with non-positive or NaN values and dividing by the number of days in that year. Using the AOI mask, we stacked the five yearly maps to form a short time series at each pixel. We then tested whether data gaps steadily increased or decreased (Mann–Kendall test) and estimated the rate of change (Sen's slope). Positive slopes indicate increasing data gaps; negative slopes indicate decreasing data gaps. We report (1) a slope map, (2) a p-value map, and (3) a binary significance mask ($p < 0.05$).

4.2.6. Correlation Between Elevation and NO_2/SO_2 Data Gaps

Motivated by the persistent hotspot of high data gaps along the western edge of the AOI in the annual maps, we examined whether topography and cloudiness help explain this pattern via elevation and cloud data gaps correlations.

We firstly examined whether terrain height helps explain the spatial pattern of data gaps. For each species, we first aggregated the daily products (2019–2023) to a per-pixel data gaps ratio (fraction of days with value ≤ 0 or NaN), then resampled a DEM to the same grid and applied the common AOI mask. Using all valid pixels ($n = 94{,}666$), we computed both Pearson (linear) and Spearman (rank) correlations between elevation and data gaps.

4.2.7. Correlation Between Cloudiness and NO_2/SO_2 Data Gaps

In the part, we retrieved daily cloud fraction from Sentinel-5P (TROPOMI) in Google Earth Engine for 2019–2023, using the same AOI and grid as the NO_2 and SO_2 products. Cloud fraction is in $[0, 1]$ and denotes the share of a pixel covered by cloud (0 cloud-free, 1 fully cloud-covered).

And then, we co-registered the pollutant and cloud stacks and, for each pixel, computed (1) the five-year data gaps ratio of the pollutant (count of days with value ≤ 0 or NaN divided by total days in 2019–2023) and (2) the five-year mean cloud fraction, averaging over valid cloud observations at that pixel (ignoring NaNs). We then evaluated pixel-wise Pearson and Spearman correlations between data gaps and cloudiness.

Diagnostics include a map of pollutant data gaps, a map of mean cloud fraction, and a hexbin plot of cloud fraction versus data gaps with a fitted linear trend. Because SO_2 cloud fractions in the AOI are very small, the SO_2 hexbin uses an x-axis of 0 to 0.1 for readability, while NO_2 uses 0 to 1. This workflow yields AOI-wide correlation summaries together with spatial and distributional context.

4.3. Auxiliary Datasets Preprocess

4.3.1. Meteorological Drivers

We load the ERA5 single–level 2 m temperature NetCDF and the AOI boundary, auto–detect the time coordinate and the `t2m` variable, and verify native metadata (CF-compliant, units in K) and grid spacing (0.25° lat/lon) fully covering the S5P AOI. For visualization only, values are converted to °C; all subsequent processing is performed on the native field. Temporal completeness is checked on control dates (four hourly records per day around 12:00–15:00 UTC, consistent with Section 3.3.1). We then aggregate `t2m` to daily means and reproject each day to the S5P template grid (EPSG:4326; identical width/height/transform) using bilinear resampling so that the output stack matches S5P in (time, y, x). The result is one multi–band GeoTIFF per year (2019–2023; one band per day) plus a 5-year mean map, with NaN used as NoData. As for checking, we report per–year ranges and summary statistics of the aligned stacks and confirm exact agreement of CRS, transform and dimensions with the S5P reference. The same protocol (daily aggregation, reprojection to the S5P grid, AOI masking, and check) is applied to all other ERA5 drivers (`blh`, `u10`/`v10`, `tp`, `sp`, clear-sky surface net solar/thermal radiation), without unit conversion.

Three ERA5 drivers(wind speed, wind direction and total precipitation) require a small amount of extra processing before being aligned to the S5P grid.

Wind requires a small difference before alignment. The 10m components `u10` and `v10` are time-aligned by inner join on date and averaged to daily means, then reprojected. We derive speed and meteorological "from" direction, $|\mathbf{V}| = \sqrt{u^2 + v^2}$ $(\mathrm{m\,s^{-1}})$ and $\theta = \big(\mathrm{atan2}(-u, -v)\cdot 180/\pi\big) \bmod 360$ (deg), and observe day-by-day mean absolute differences $< 10^{-2}\mathrm{ms}^{-1}$ between the speed from (u, v) and the QA rasters. In modelling we use the aligned components `u10`/`v10`; if a directional feature is needed, the angle is encoded as $(\sin\theta, \cos\theta)$ to avoid circularity.

Total precipitation (`tp`) differs because it is provided as accumulations in metres. We sum sub-daily steps to obtain daily totals, convert to $\mathrm{mm\,day^{-1}}$ ($\times 1000$), clip tiny negatives to zero, and then align to the S5P grid as above.

All remaining ERA5 drivers (`t2m`, `blh`, `sp`, clear-sky surface net solar/thermal radiation) follow the generic route: verify coverage and metadata, aggregate to daily *means*, reproject to the S5P template (EPSG:4326; identical width/height/transform), apply the AOI mask, and store as multi-band GeoTIFFs with NaN as NoData.

4.3.2. Static Geographical Features

Static covariates provide terrain and land-use context for gap filling. We harmonised (1) elevation (SRTM DEM), (2) terrain slope (derived from SRTM), and (3) ESA WorldCover 2020 land-use/land-cover (LULC) to the Sentinel-5P template grid (EPSG:4326, identical width/height/transform).

As for DEM and slope, the source rasters were read as single bands, NoData was mapped to `NaN`, and values were reprojected to the S5P grid with *bilinear* resampling (continuous variables). We preserved the template's geotransform and CRS and wrote GeoTIFFs with `float32` and `NaN` nodata. As a quick check, we reported valid-pixel counts and basic statistics (min, max, mean, std) after alignment to ensure ranges were physically reasonable (metres for DEM, degrees for slope).

As for land-use/land-cover data, two ESA WorldCover tiles were mosaicked, then reprojected to the S5P grid using nearest-neighbour resampling[1]. After alignment we enumerated present classes and produced a one-hot stack: for each class c (excluding background), a binary layer $\mathbb{1}\{\mathrm{LULC} = c\}$ was saved (`uint8`, 0/1 with 0 as nodata), along with per-class pixel counts and percentages. All outputs inherit the S5P grid definition

[1] Categorical data require nearest-neighbour to avoid class mixing.

to guarantee pixel-wise joining with the dynamic features.

4.3.3. Anthropogenic Indicators

Human activity is proxied with the WorldPop UN–Adjusted 2020 gridded population (1 km, people km^{-2}). The source raster is read and the declared NoData value is mapped to NaN (zeros are retained as genuine low values); data are kept in `float32` with native units. We then reproject the layer to the S5P template grid (identical CRS, geotransform, and width/height) using bilinear resampling and write a single-band, grid-aligned GeoTIFF. A simple check reports valid pixel count, range, and mean/std after reprojection. The population layer is a static feature: at feature-stack time it is broadcast across days (no temporal smoothing), and only standardized during model fitting.

4.3.4. Temporal Features

For each pollutant (NO_2, SO_2) we compute a daily, domain-wide mean from the S5P L3 multi-band rasters (one band per day, 2019–2023), then estimate a data-driven "weekend effect". Let $\overline{y}_{\mathrm{wk}}$ and $\overline{y}_{\mathrm{we}}$ denote the mean concentrations over weekdays and weekends, respectively; we define a scalar weekend weight $w = \overline{y}_{\mathrm{we}}/\overline{y}_{\mathrm{wk}}$. A date-level series is then created with `weekday_weight` $= 1$ on weekdays and $= w$ on weekends, together with integer `day_of_week` (0–6) and a Boolean `is_weekend`. The resulting CSV files (2019–2023) contain columns {`date`, `concentration`, `day_of_week`, `is_weekend`, `weekday_weight`} and are used later to provide a simple prior for weekly modulation in learning models.

To represent seasonality without discontinuity at year boundaries, we encode the day-of-year (DOY) using a two-dimensional cyclic embedding. For each year (2019–2023), we create daily arrays of $\sin(2\pi\,\mathrm{DOY}/L)$ and $\cos(2\pi\,\mathrm{DOY}/L)$, where L is the actual length of the year (365/366). These are written as multi-band rasters (one band per day) on the S5P reference grid (dimensions and georeferencing taken from the 2019 NO_2 product), yielding two date-aligned features `sin_doy` and `cos_doy`. The pair jointly captures smooth seasonal progression and is robust to model choices that assume linearity in inputs.

Using 2019–2023 S5P daily means, the estimated weekend weights are $w_{\mathrm{NO_2}} \approx 0.854$ and $w_{\mathrm{SO_2}} \approx 0.951$ (defined as weekend/weekday mean ratio). We set the date-level modifier to 1 on weekdays and to w on weekends.

4.3.5. Spatial Context Features

First, for NO_2 and SO_2 we derive a 1-day lag prior from CAMS surface fields (kg/kg). Times are converted from UTC to local time (Europe/Rome), then two local windows are averaged per calendar day (00:00–13:00 and 15:00–23:59). The daily value is a weighted mean based on the number of available time steps [17].

Units are preserved (kg/kg) to serve as a prior rather than a directly comparable observation. The daily stacks are reprojected to the Sentinel-5P template (EPSG:4326, 0.01°, 300×621), bilinear-resampled, clipped to the AOI without changing shape, and exported as one multi-band file per gas (2019–2023).

Secondly, from the S5P daily composites we compute a 3×3 neighbourhood mean (excluding the centre pixel), producing a same-day "context" value for each cell. Pixels with explicit NoData sentinels are masked; for SO_2 we retain zeros (valid very low values) and treat only negatives as missing. Remaining gaps are filled with the same-day global mean for that band; if a band has no valid pixels, a small positive fallback is used to keep the stack consistent. Outputs are multi-band rasters aligned to the S5P grid.

All context features pass a strict readiness check against the S5P template: identical dimensions (300×621), CRS equality (EPSG:4326), affine transform match (within numerical tolerance), complete 1-day cadence, reasonable value ranges, and coverage statistics recorded per day. Configuration files (JSON) capturing CRS, grid transform, time span, units, compression, and AOI metadata are saved for reproducibility.

4.4. Feature Engineering and Stack

4.4.1. Feature Stack Construction

To let the model see all relevant signals at the same pixel and time, all layers are combined. A single layer (e.g., NO_2) is not enough, because gap patterns depend on weather, land cover, and past values. Stacking these layers gives joint context.

A feature stack is a set of co-registered rasters (same grid, CRS, and resolution) treated as channels. For each date t we form a multi-channel array $X_t \in \mathbb{R}^{H \times W \times C}$, where C includes dynamic layers (NO_2/SO_2 lags, temperature, wind, pressure) and static layers (elevation, slope, land cover, population). All layers are aligned to the Sentinel-5P grid pixel by pixel.

The stack is the input to the models. LightGBM uses per-pixel feature vectors sampled

from X_t, while the 3D CNN ingests spatio-temporal blocks $\mathcal{X} \in \mathbb{R}^{T\times H\times W\times C}$, so it can learn patterns across space and time. This design also makes preprocessing reproducible and ablation straightforward (add or remove a channel).

Feature engineering follows a daily, pixel-aligned design on the S5P grid (EPSG:4326; 300×621) for NO_2/SO_2. All layers are reprojected to the template (bilinear for continuous, nearest neighbour for categorical) and clipped to the AOI. Product NoData and out-of-range values are mapped to `NaN`; for NO_2 values ≤ 0 are treated as missing, while for SO_2 zeros are kept and negatives are masked. Each day's inputs are saved under `/Feature_Stacks/NO2_{year}/NO2_stack_{YYYYMMDD}.npz` and `/Feature_Stacks/SO2_{year}/SO2_stack_{YYYYMMDD}.npz`.

For NO_2 (first implementation), daily stacks were produced in a *dictionary* layout: each feature was stored as a named array in the `.npz`. This design is reproducible and matches the scaler's channel list, but reading requires rebuilding the $H \times W \times C$ matrix from keys, which adds I/O and risks name–channel mismatches.

For SO_2 (improved implementation), the builder writes a matrix-first file: a ready-to-use tensor $X \in \mathbb{R}^{H\times W\times C}$ plus minimal metadata (fixed `feature_order`, indices for continuous/one-hot/no-scale). Channels mirror NO_2 (terrain, population, WorldCover one-hot, ERA5 meteorology, sin / cos of day-of-year, weekday weight, wind diagnostics, 3× 3 neighbour mean, lag at t–1) and add SO_2-specific signals: a monthly `so2_climate_prior` to stabilise winter. This climatological prior was introduced as an additional feature to filter out wintertime observations, where retrievals are frequently affected by low solar irradiance and high noise, in order to test whether such seasonally informed information could improve gap-filling performance. In subsequent ablation tests, the monthly `so2_climate_prior` is treated as optional.

Across both pollutants, all files use the same grid, masks, and a fixed channel order aligned with the scaler. Typical daily files are ∼5 MB. This harmonised representation (common grid and ordering) lets the 3D-CNN take $T \times H \times W \times C$ blocks directly, and allows LightGBM to sample per-pixel feature vectors without extra conversion. The SO_2 matrix format removes the need to rebuild tensors at load time, which was a limitation of the NO_2 dictionary format.

We summarise the feature stacks with four tables: (1) overall comparison (Table 4.1); (2) channel counts by family (Table 4.2); (3) naming and storage/technical specs (Table 4.3); and (4) modelling implications (Table 4.4).

Table 4.1: Overall comparison of NO_2 and SO_2 feature stacks. Columns "NO_2" and "SO_2" report the values for each pollutant; "Notes" clarifies the metric. "Data coverage (% of y)" is the fraction of valid pixels in the target raster y (per day, averaged over 2019–2023).

Aspect	NO_2	SO_2	Notes
Total files (2019–2023)	1,826	1,826	Same temporal coverage
Typical file size	∼5.01 MB	∼5.54 MB	SO_2 slightly larger
Storage format	`dictionary`	`matrix`	Different internal layout
Predictor channels (C)	29	30	SO_2 adds a climatology prior
Data coverage (% of y)	34.70	26.93	Pixel-level valid ratio

Table 4.2: Feature categories and counts

Category	NO_2	SO_2	Notes
Terrain (DEM, slope)	2	2	Same
Population	1	1	Same (name may differ)
LULC one-hot	10	10	Same number; encoding labels may differ
Temporal (doy sine, cosine, weekday weight)	3	3	Same
Meteorological (ERA5)	8	8	Same
Wind (u10, v10, optional encodings)	3	3	For modelling we use u_{10}, v_{10}
Spatial context	2	2	CAMS lag $(t-1)$ and 3×3 neighbour mean
Climatology	0	1	SO_2 monthly climatology prior
TOTAL	**29**	**30**	SO_2 has +1 feature

Table 4.3: Naming differences and storage/tech specs

Type	NO_2 name	SO_2 name	Impact / note
Population	`pop`	`population`	Minor naming difference
Solar radiation	`ssr_clr`	`ssr_clear`	Minor naming difference
Lag feature	`no2_lag_1day`	`so2_lag1`	CAMS lag at $t-1$, pollutant-specific name
Neighbour feature	`no2_neighbor`	`so2_neighbor`	Pollutant-specific name
LULC one-hot	`lulc_class_*`	`lulc_class_*`	Same count, labels may differ
File extension	`.npz`		NumPy compressed
Data type	`float32`		
Spatial grid	300 × 621		EPSG:4326 template
Feature dims	29×300×621	30×300×621	channel×H×W
Total dataset size	∼9.15 GB	∼10.11 GB	2019–2023

Table 4.4: Implications for modelling and recommendations

Aspect	NO_2	SO_2	Recommendation
Training priority	higher	medium	Start with NO_2 (better coverage)
Pre-processing	minimal	moderate	Handle missing data carefully
Model channels	29	30	Separate models per pollutant
Feature engineering	standard	enhanced	Leverage SO_2 climatology prior
Format	dictionary	matrix	Consider standardising to a common matrix layout
Data split	2019–2021 train, 2022 val, 2023 test		Time-based split

4.4.2. Feature Standardization

Having defined the daily feature stacks, we standardize the continuous channels using a single global scaler estimated on the 2019–2021 training data. We compute global per-channel means and standard deviations on the 2019–2021 training stacks and reuse them at both training and inference to keep a fixed scale across regions and seasons. Per channel (μ_c, σ_c) are estimated online with Welford's algorithm over valid pixels (mask = 1), excluding non-finite values; land-use one-hot channels are marked no-scale ($\mu = 0,\ \sigma = 1$). Normalization uses $z = (x - \mu_c)/\sigma_c$. For NO_2 stacks, channels are read by `source_key`; for SO_2 stacks, arrays are taken from `X` $(C \times H \times W)$ using indices resolved from `feature_names`. We save `mean`, `std`, and the aligned vectors `mean_vec`/`std_vec` together with metadata (version, channel list and signature, units, seed) to `artifacts/scalers/{NO2,SO2}/meanstd_global_2019_2021.npz`, and append a record to `metadata.jsonl`.

4.5. Machine Learning Model: LightGBM

LightGBM is a gradient-boosted decision tree (GBDT) model that builds an ensemble of shallow trees to capture non-linear effects and feature interactions. It suits our tabular feature stack (continuous, binary, one-hot) and handles missing values natively by learning default split directions, so heavy imputation of auxiliary inputs is unnecessary. It trains fast at scale, is robust to feature scaling, and provides simple importance measures, making it a strong, low-cost baseline for gap filling using temporal (lags, DOY), spatial (neighbour mean), static and meteorological predictors. We therefore use LightGBM to validate the feature stack and masking protocol before moving to the 3D-CNN; models are trained on 2019–2021, tuned on 2022, and evaluated on 2023 with masked validation [7].

We adopt LightGBM as a baseline gap-filling model to verify the correctness of our feature

stack and data pipeline. Its efficiency on large tabular data, ability to capture non-linear interactions, and interpretable feature importance make it suitable for this validation. The model is trained on 2019–2021 data, uses 2022 for early stopping, and is evaluated on 2023, respecting temporal dependencies. No artificial gaps are injected—the model learns directly from TROPOMI's natural missingness patterns.

For each day, the feature stack is read and flattened to (N_{pixels}, C) so that all pixels' features serve as inputs. Labels are taken only from observed pixels: for NO_2 the target y is used as is, while for SO_2 an additional filter $y > 0$ is applied. Missing features receive light imputation (column mean or zero), and channel names are mapped to the scaler's `channel_list` to ensure alignment.

The model is a LightGBM regressor (GBDT) with typical settings such as `num_leaves=127` and `learning_rate=0.03`; early stopping uses `stopping_rounds=300`.

All data are on the unified S5P grid (EPSG:4326; 300×621) with a fixed channel order consistent with the scaler. NO_2 stacks were initially stored as a dictionary-based `.npz` that requires rebuilding the tensor at load time; SO_2 adopts a matrix-first, channel-first layout ($C \times H \times W$) for direct loading.

Table 4.5: LightGBM hyperparameters used for the baseline gap-filling model (both pollutants).

Parameter	Value
objective, metric	regression; {L1, L2}
boosting_type	gbdt
num_leaves	127
learning_rate	0.03
min_data_in_leaf	300
feature_fraction	0.8
bagging_fraction / bagging_freq	0.8 / 1
max_bin	255
n_estimators	1000 (early stopping 300)
random_state	42; threads = CPU cores

4.6. Deep Learning Model: 3D CNN

A 3D CNN applies convolution over *time*×*height*×*width* so each kernel jointly captures temporal evolution and spatial context. This is well-suited to S5P gap filling, where

missingness and plumes are spatio-temporally correlated across consecutive days. Compared with 2D CNNs, 3D CNNs model short-range temporal dynamics directly, offering a simple, stable baseline for week-long windows without extra sequence modules [29] .

We split the data by time (2019–2021 train, 2022 validation, 2023 test) on the S5P grid (EPSG:4326; 300×621), built sliding windows of $T = 7$ days, and trained on 64×64 patches. Training ran on CUDA with optional Automatic Mixed Precision, fixed learning rate/batch size/epochs, global-norm gradient clipping, a ReduceLROnPlateau scheduler, and early stopping (patience-based). Targets were z-scored with (μ_y, σ_y).

For each day, the NO_2 stack is loaded, re-keyed to the scaler's `channel_list`, and assembled as channel-first $X \in \mathbb{R}^{C\times H\times W}$ with (y, mask). Sliding windows build (T, C, H, W) and (T, H, W). Artificial gaps are injected during training: time-drop (hide some earlier frames) and space-drop (hide random pixels), never touching the last frame. Patches are sampled at random for training (retry if the last-frame valid ratio $<$ `min_valid_ratio`, up to 20 attempts) and centre-cropped for validation/test. Features are standardised by NaN→0 then per-channel z-score via broadcast; invalid targets are set to 0 for tensoring, and a label map with -999 marks ignored pixels.

The model is a 3D CNN with three Conv3D→GroupNorm→ReLU blocks (channels 32→64→128) and a 1×1×1 head to one output channel (Figure 4.2). Inputs are passed as (B, T, C, H, W) (internally permuted to (B, C, T, H, W)); no sigmoid is applied so the network learns in the original scale.

Table 4.6: 3D-CNN training configuration.

Temporal window	$T = 7$ days
Patch size	64×64
Artificial gaps	time: 0.2 (frames 0-5), space: 0.15 (frames 0-5)
Optimizer	Adam (1×10^{-4})
Loss	masked MSE (last frame only)
Data workers	NO_2: 4, SO_2: 8

The training pipelines are identical except for pollutant-specific target normalization, feature name mappings, and the aforementioned computational settings.

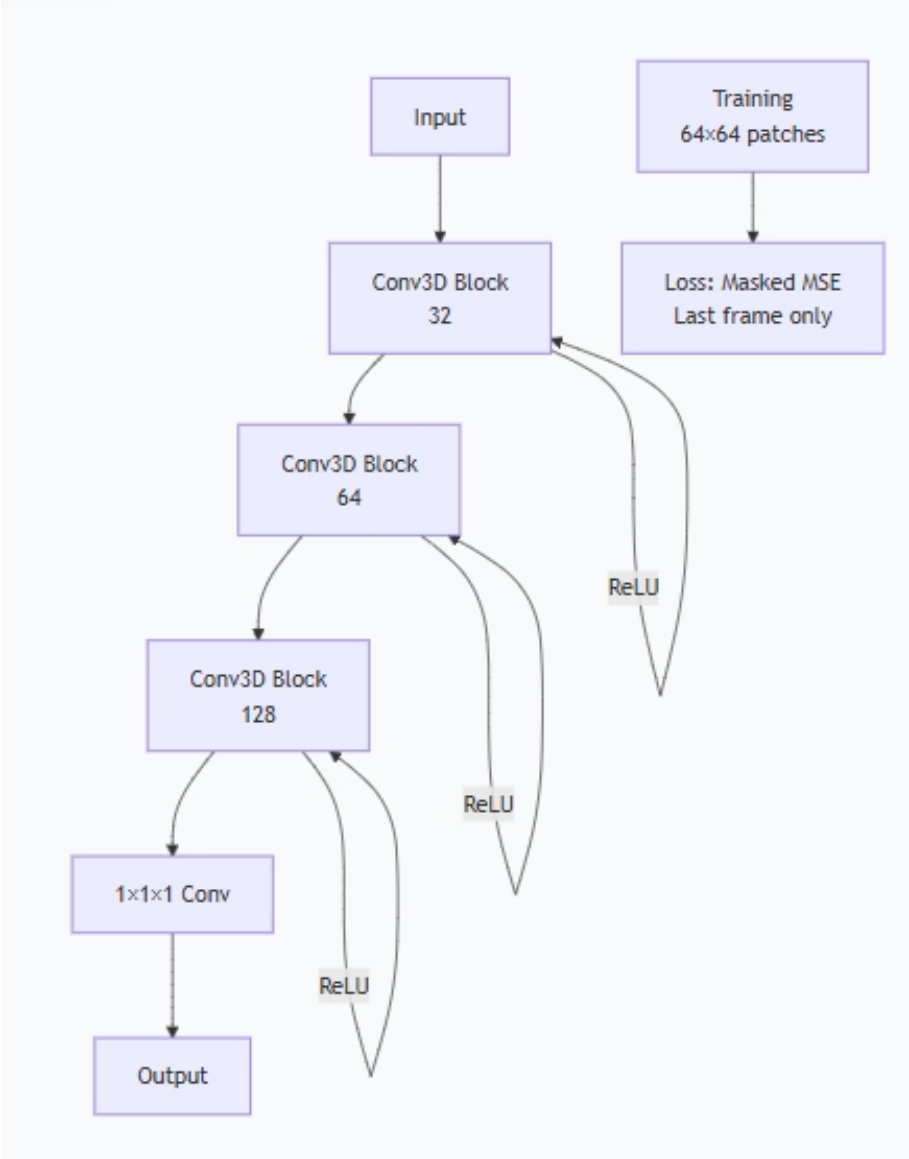


Figure 4.2: 3D CNN with three Conv3D→GroupNorm→ReLU blocks

4.7. Evaluation: Masked Validation

4.7.1. LightGBM

We evaluate gap filling on the 2023 test split with a masked (counterfactual) validation that injects missingness only at inference time. For each test day, we sample a fraction $r \in \{0.1, 0.2, 0.3\}$ of *valid* pixels on the last frame and hide them while preserving their true targets. At masked locations all feature channels are hidden, so the model must reconstruct using unmasked context. To stress-test temporal/spatial dependence and avoid leakage, we optionally enable a *sensitive-features* mode that explicitly sets the lag and

neighbour predictors (`lag1`, `neighbor`) to NaN at masked pixels, relying on LightGBM's native missing-value handling.

The inference pipeline matches training: we load the trained LightGBM model and the global scaler, align feature names to the canonical order (with mapping where needed), apply per-channel z-scoring to continuous variables, and keep designated non-scaled features. Naturally missing inputs in the stack are imputed with in-file means; only the artificial validation masks remain as NaN. Targets follow pollutant-specific rules: for NO_2 we evaluate on all finite targets; for SO_2 we retain only $y > 0$. Files (days) with fewer than 10 valid target pixels are excluded.

Metrics are computed exclusively on masked pixels using finite prediction–target pairs: RMSE, MAE, R^2, Pearson correlation, bias, and normalised errors $\mathrm{NRMSE}_{\mathrm{std}}$ (by target standard deviation) and $\mathrm{NRMSE}_{\mathrm{range}}$ (by target range). For each configuration (r, sensitive mode) we repeat the evaluation with different random seeds and report mean $\pm$ standard deviation; we also track the total count of held-out valid pixels across the processed days.

4.7.2. 3D CNN

We evaluate the trained NO_2 3D-CNN with a masked (holdout) validation that mimics missing pixels. For each target day in 2023 we build a $T = 7$ window (backfilling from 2022 if needed), assemble features in the scaler's channel order (with name mapping), pad/crop to the S5P grid, replace NaNs by 0, and standardise by the per-channel mean/std. We also load the target y and its validity mask m.

Masking is applied only on valid pixels of the last frame. We set: (1) Random Pixel masking—uniformly sample a given fraction of valid pixels over the full year; (2) Block32 masking—tile the last frame with fixed-size squares (e.g., 32×32) over the full year until the target ratio is reached; (3) optional Sensitive-channel masking like Masked Block32—at held-out locations, zero all time steps and all channels, and additionally zero channels like `neighbor` to prevent leakage, which apply on Block over the full year; (4) Monthly Block32—restrict evaluation to selected months to probe seasonality.

The network predicts the last frame; predictions are de-normalised and compared with ground truth only on held-out pixels. We report RMSE, MAE, R^2, Pearson correlation, $\mathrm{NRMSE}_{\mathrm{std}}$ (by target std). The sample size is the total count of held-out valid pixels across processed days. Each configuration is repeated `REPEATS` times with different seeds and summarised by mean±std; days with too few valid or held-out pixels (below `MIN_HOLD`) are skipped, and `MAX_WINDOWS` limits runtime if needed.

4.8. Gap-filling Inference and Output Generation

We load the trained model and the global per-channel scaler; the scaler's `channel_list` fixes the feature order for inference (with the same key mapping as in training). For each 2023 day, sentinels/out-of-range targets are set to NaN, features are assembled in that order, continuous channels are z-scored (LULC one-hot left unscaled), and missing inputs remain NaN (LightGBM handles NaNs). Pixels are flattened, predicted, reshaped to the native raster, and written as one band per day; bands share the S5P grid/CRS/transform copied from a reference file. Outputs are tiled, compressed Float32 GeoTIFFs with NaN as NoData, plus logged per-band valid counts and ranges. The procedure is model-agnostic; for CNNs, predictions are de-standardized to physical units before stacking.

4.9. Computing Resources and Environment

4.9.1. Hardware Configuration

All experiments were executed on Google Colab Pro+ GPU runtimes, utilizing a single NVIDIA A100 GPU (40 GB VRAM) with a Linux x86_64 environment. Google Drive storage was mounted for persistent storage of datasets, model checkpoints, and output artifacts.

4.9.2. Software Stack

The software environment was based on Python 3.10, with key dependencies including: NumPy and Pandas for array operations and I/O; PyTorch (CUDA-enabled) for CNN model development; LightGBM for gradient boosting baselines; built-in NumPy functions for evaluation metrics; and Rasterio/GDAL for geospatial data processing.

4.9.3. Training Configuration

Both NO_2 and SO_2 3D-CNN models shared the same architectural design, consisting of three 3D convolutional layers with group normalization and ReLU activations. The models were trained using the Adam optimizer with weight decay (1e-6), a batch size of 8, patch size of 64×64 pixels, and a 7-day temporal context window.

A key aspect of 3D-CNN training was the injection of artificial gaps to teach genuine inpainting: random earlier time steps (ratio=0.2) and spatial pixels in those frames (ratio=0.15) were masked during training, while the final frame remained intact for super-

vision. This encouraged the network to recover signals from spatiotemporal context.

Key differences in training configuration included:

- Input Channels: SO_2 models had 30 input channels (including a monthly climatology prior), compared to 29 channels for NO_2.
- Data Loading: SO_2 used 8 data loading workers for improved I/O performance, while NO_2 used 4 workers.
- Normalization: Separate target value normalization parameters were used for each pollutant, reflecting their different concentration ranges.

This configuration remained within the 40 GB VRAM capacity of the A100 GPU, with `pin_memory` enabled to optimize data transfer between CPU and GPU.

5 | Results

5.1. Descriptive Results of Data Gaps Analysis

5.1.1. Results of NO_2 and SO_2 Annual Trend

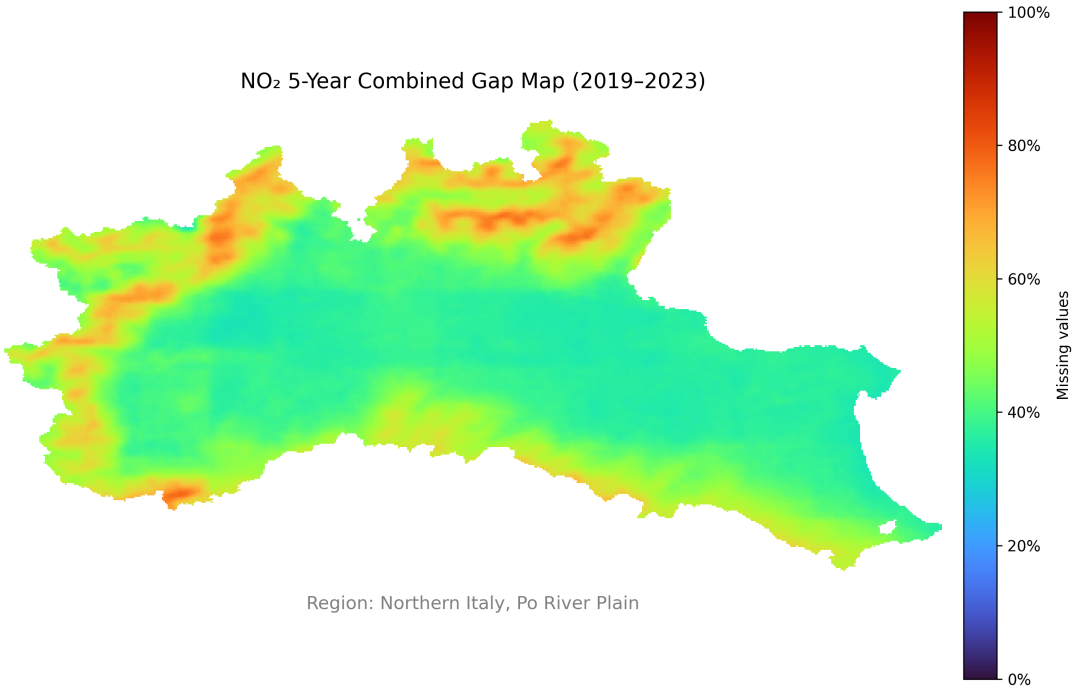


Figure 5.1: NO_2 5-Year Combined Gap Map (2019–2023). Color bar shows ratio (0–1).

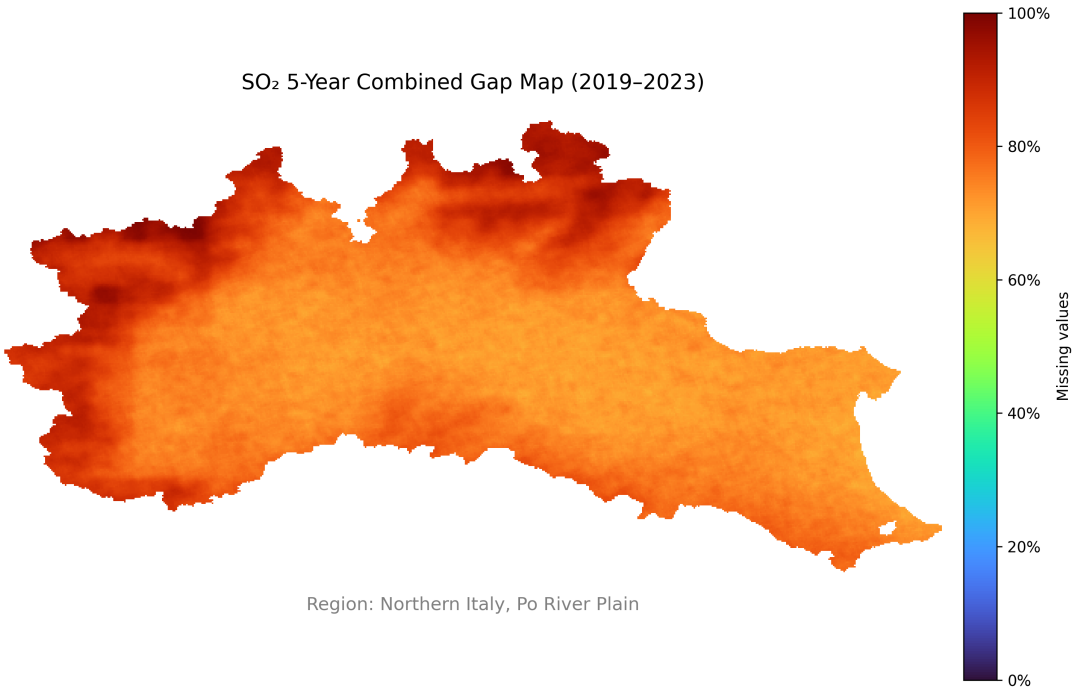


Figure 5.2: SO_2 5-Year Combined Gap Map (2019–2023). Color bar shows ratio (0–1).

Species	Mean	Median	Std. dev.	90th pct.
NO_2 (2019–2023)	0.454 (45.4%)	0.412 (41.2%)	0.099	0.607 (60.7%)
SO_2 (2019–2023)	0.774 (77.4%)	0.749 (74.9%)	0.068	0.889 (88.9%)

Species	$<20\%$	20–50%	50–80%	$\geq 80\%$
NO_2	0.0%	70.4%	29.6%	0.0%
SO_2	0.0%	0.0%	73.1%	26.9%

Table 5.1: Five-year summary by species over the AOI (2019–2023). Ratios are computed per pixel from daily coverage, treating values ≤ 0 or `NaN` as missing. Std. dev. = standard deviation of pixel-wise five-year data gaps ratios; 90th pct. = value below which 90% of pixels fall.

The gap maps in Figures 5.1 and 5.2 align with the summary in Table 5.1. Over the AOI, NO_2 shows moderate data loss: the five-year mean (median) data gaps is 45.4% (41.2%), with a limited spread (std. 0.099) and a 90th percentile of 60.7% (i.e., 90% of pixels are at or below 60.7%). Most pixels fall in the 20–50% category. By contrast, SO_2 exhibits markedly poorer coverage, with a mean (median) of 77.4% (74.9%), std. 0.068, and a 90th percentile of 88.9% (i.e., 90% of pixels are at or below 88.9%); the majority

of pixels lie in the 50–80% range and the remainder at ≥80%. The close mean–median pairs suggest skewed but not extreme distributions, while the level shift for SO_2 indicates substantially sparser valid observations. Practically, these patterns motivate stronger reliance on temporal and neighborhood/context features and more cautious validation for SO_2, whereas NO_2 benefits from comparatively denser sampling.

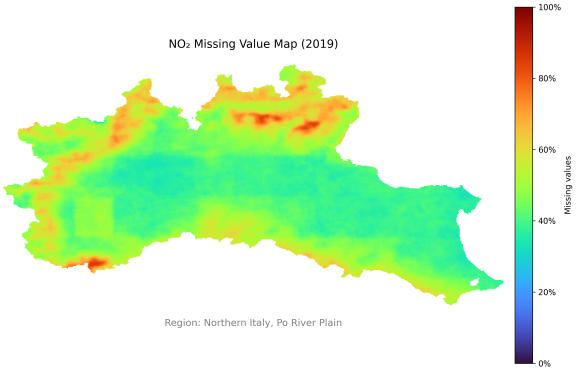
Compared to NO_2, SO_2 shows systematically higher gap fractions because the SO_2 retrieval applies stricter QA screening under large solar-zenith angles and in cloudy/snow/ice scenes, conditions common in winter mid-latitudes; moreover, background SO_2 columns are typically near zero, so many retrievals fail QA or become non-positive and are masked by our ≤ 0 rule [15, 16].

We also summarized annual data gaps by year (2019–2023). For each year, we computed the pixel-wise fraction of missing days within the AOI and reported the missing rate statistics.

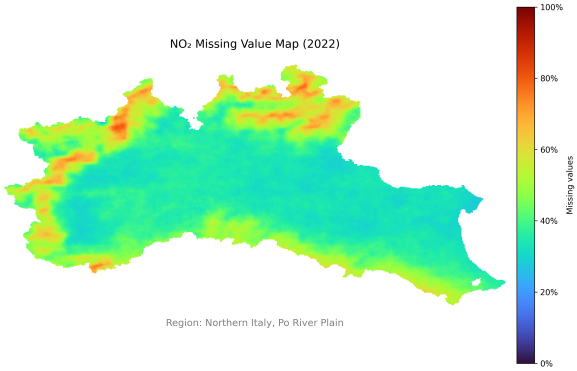
The table 5.2 summarizes annual statistics for NO_2 and SO_2 across the full study period (2019–2023), whereas the figures 5.3 that four maps provide illustrative snapshots of the spatial patterns in two representative years. The examples (2019 and 2022) are not intended to single out specific years but to visualize typical contrasts: NO_2 generally exhibits moderate gaps with basin-wide coherence, while SO_2 remains highly incomplete and spatially pervasive. These spatial structures are consistent with the distributional summaries in the table and with the multi-year maps shown earlier; similar patterns are observed in other years and months. All maps use a fixed color scale (fraction of missing days, 0–1) for comparability.

Table 5.2: Annual data gaps statistics within the AOI by species (2019–2023). Ratios are pixel-wise fractions of missing days; the 90th percentile is the value below which 90% of pixels fall.

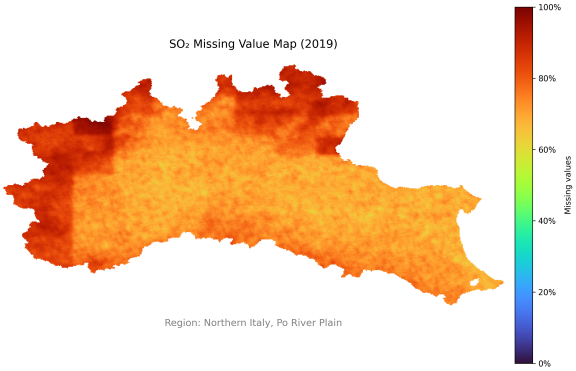
Species	Year	Mean	Median	Std. dev.	90th pct.
NO_2	2019	0.469 (46.9%)	0.447 (44.7%)	0.092	0.608 (60.8%)
	2020	0.462 (46.2%)	0.434 (43.4%)	0.088	0.593 (59.3%)
	2021	0.485 (48.5%)	0.444 (44.4%)	0.116	0.666 (66.6%)
	2022	0.413 (41.3%)	0.370 (37.0%)	0.101	0.573 (57.3%)
	2023	0.441 (44.1%)	0.395 (39.5%)	0.113	0.611 (61.1%)
SO_2	2019	0.751 (75.1%)	0.723 (72.3%)	0.075	0.874 (87.4%)
	2020	0.772 (77.2%)	0.746 (74.6%)	0.074	0.896 (89.6%)
	2021	0.791 (79.1%)	0.767 (76.7%)	0.075	0.918 (91.8%)
	2022	0.770 (77.0%)	0.751 (75.1%)	0.066	0.877 (87.7%)
	2023	0.786 (78.6%)	0.764 (76.4%)	0.067	0.893 (89.3%)



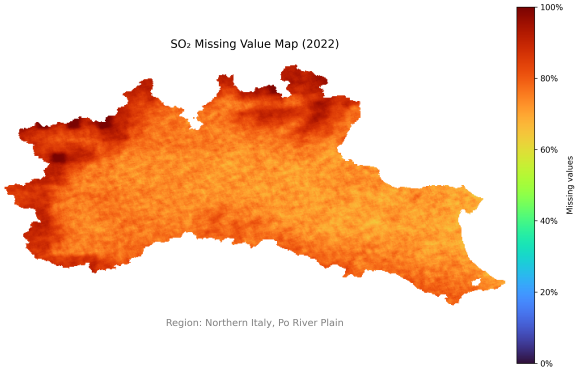
(a) NO_2 2019



(b) NO_2 2022



(c) SO_2 2019



(d) SO_2 2022

Figure 5.3: Examples of annual maps over the AOI. Panels show NO_2 (top) and SO_2 (bottom) for 2019 and 2022. All images use the same color scale (fraction of missing days, 0–1) to enable direct visual comparison.

5.1.2. Results of Pre-2021 Winter Square Analysis

Month-by-month inspection confirms that this pattern is confined to winter months before 2021-02 and disappears thereafter (no reoccurrence in winters of 2021–2022 and 2022–2023). While the exact cause is unclear, the square footprint and the sharp temporal cutoff suggest a nonphysical origin (e.g., orbit tiling/coverage geometry, quality-flag screening, or product processing changes).

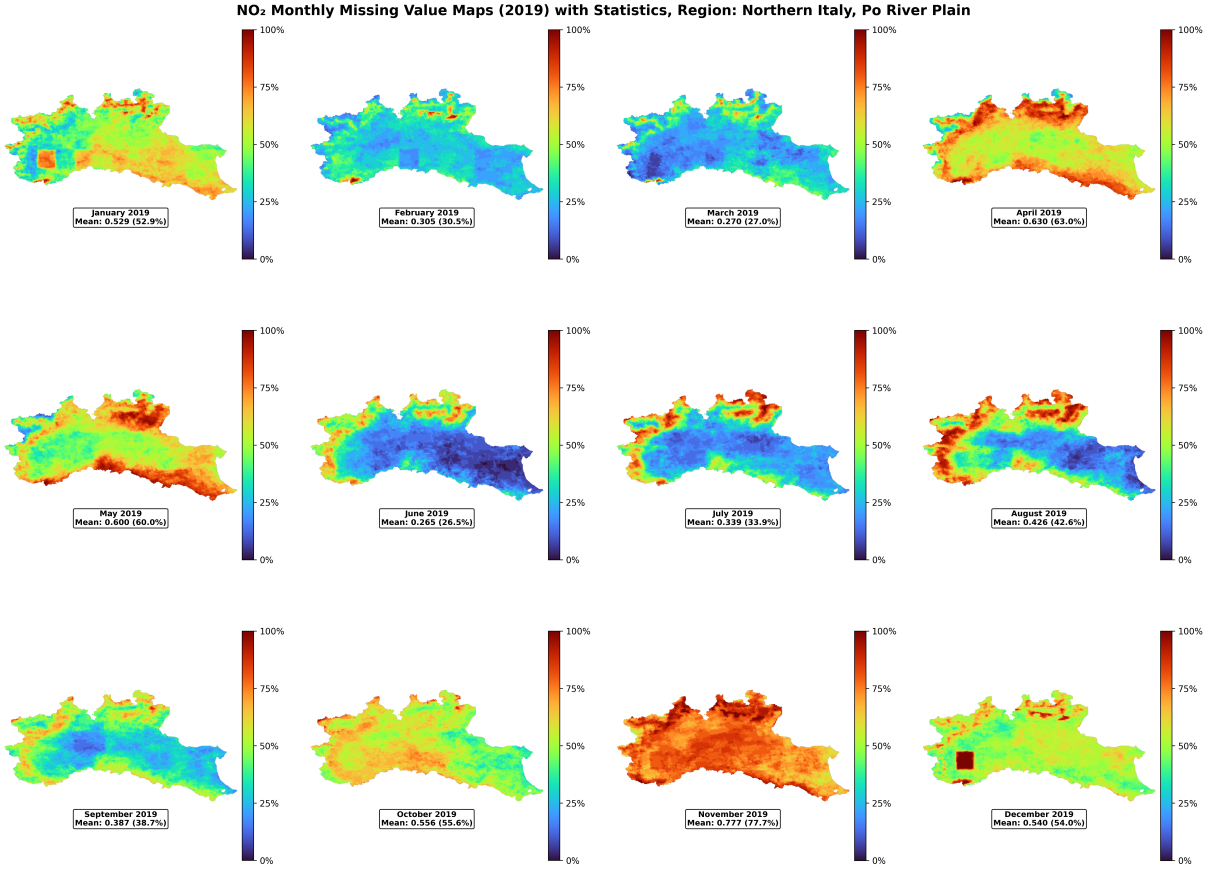


Figure 5.4: Monthly maps for NO_2 in 2019 (Po Valley). A square, tile-shaped high-data gaps patch is clearly visible in the southwestern AOI during winter months (e.g., Jan–Feb and Dec), consistent with the pre-2021 winter artefact. Color bar shows the fraction of missing days (0–1) with a fixed scale across months.

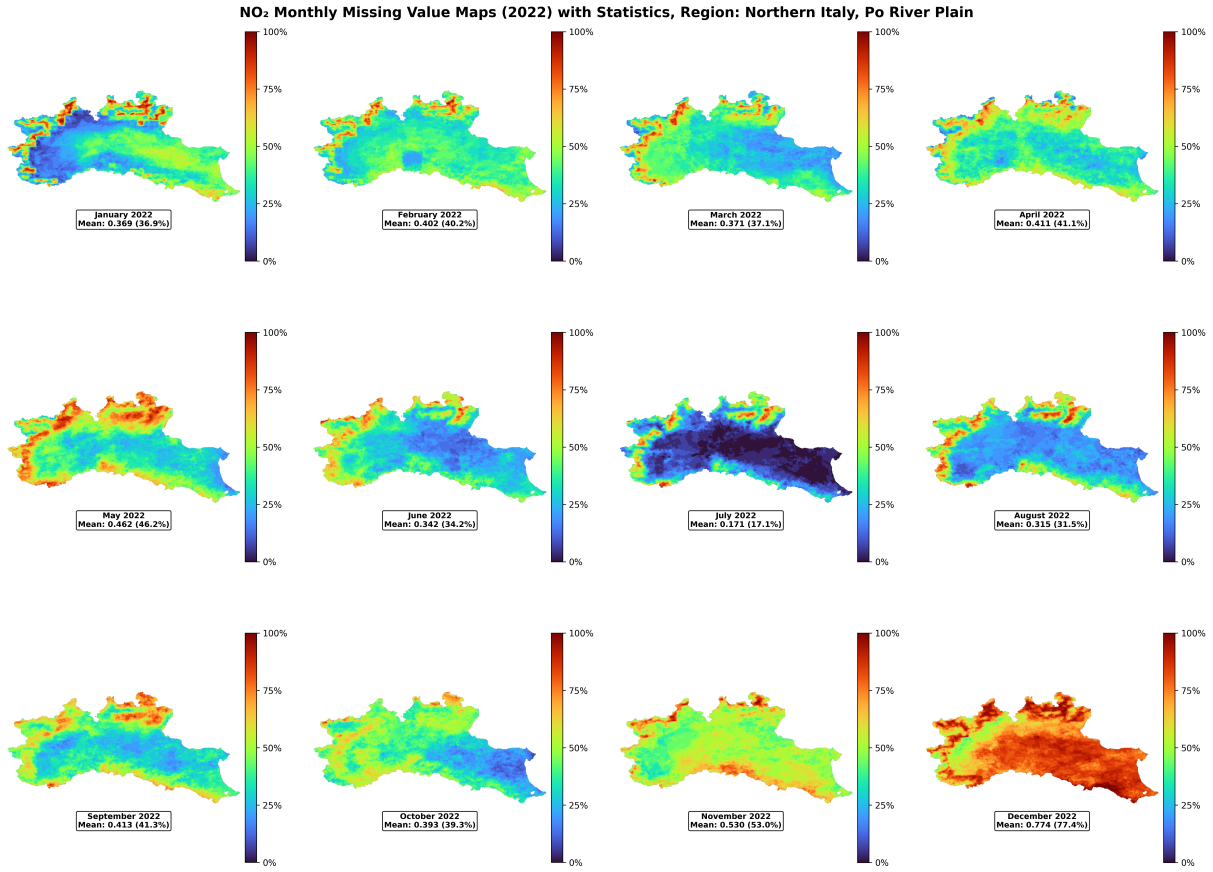
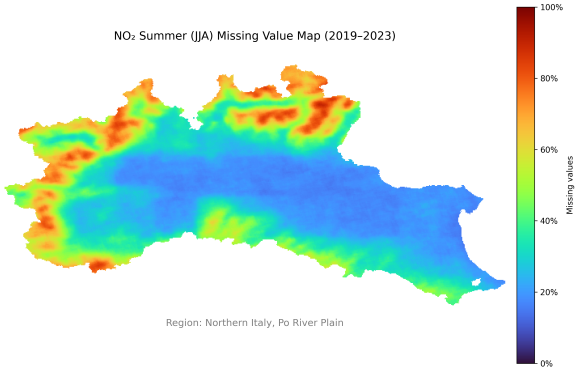


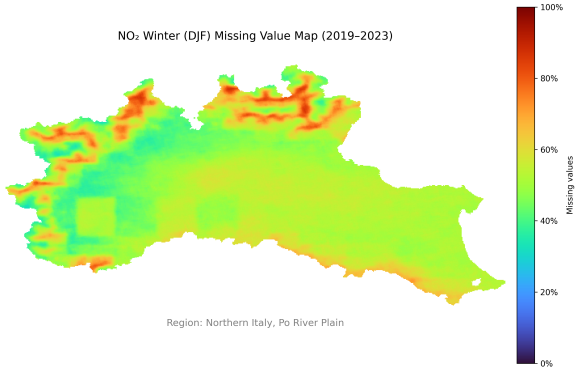
Figure 5.5: Monthly data gaps maps for NO_2 in 2022 (Po Valley). The square winter artefact observed in 2019 is absent; spatial patterns are smoother and broadly consistent across months. Color bar shows the fraction of missing days (0–1) with a fixed scale across months.

5.1.3. Results of NO_2 and SO_2 Seasonal Data Gaps Pattern

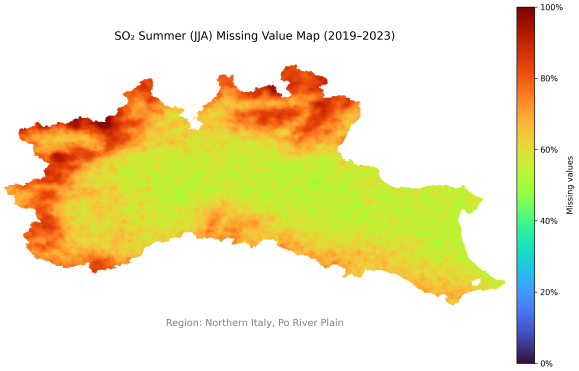
Seasonal patterns are pronounced and consistent across years: NO_2 shows moderate gaps that are higher in December–January–February (DJF) and lower in June–July–August (JJA), while SO_2 exhibits persistently high data gaps, with the strongest deficits in DJF.



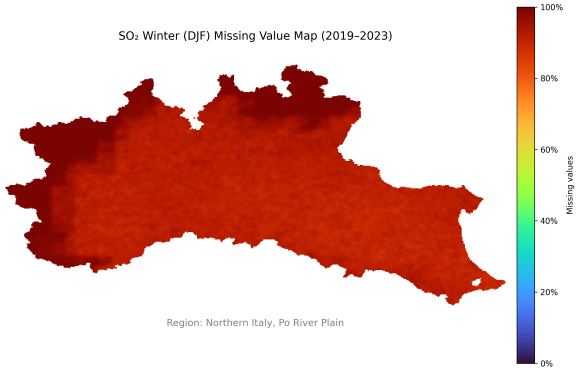
(a) NO_2 Summer (JJA)



(b) NO_2 Winter (DJF)



(c) SO_2 Summer (JJA)



(d) SO_2 Winter (DJF)

Figure 5.6: Seasonal data gaps maps pooled over 2019–2023. Panels compare summer (JJA) and winter (DJF) for NO_2 (top) and SO_2 (bottom). All maps use a fixed color scale (fraction of missing days, 0–1) to enable direct visual comparison.

Species	Season	Mean	Median	Std. dev.	90th pct.
NO_2	Spring (MAM)	0.460 (46.0%)	0.422 (42.2%)	0.119	0.637 (63.7%)
NO_2	Summer (JJA)	0.356 (35.6%)	0.298 (29.8%)	0.180	0.654 (65.4%)
NO_2	Autumn (SON)	0.475 (47.5%)	0.455 (45.5%)	0.078	0.596 (59.6%)
NO_2	Winter (DJF)	0.527 (52.7%)	0.523 (52.3%)	0.079	0.627 (62.7%)
SO_2	Spring (MAM)	0.725 (72.5%)	0.676 (67.6%)	0.121	0.957 (95.7%)
SO_2	Summer (JJA)	0.643 (64.3%)	0.617 (61.7%)	0.095	0.793 (79.3%)
SO_2	Autumn (SON)	0.800 (80.0%)	0.793 (79.3%)	0.039	0.846 (84.6%)
SO_2	Winter (DJF)	0.931 (93.1%)	0.918 (91.8%)	0.035	1.000 (100.0%)

Table 5.3: Seasonal data gaps summary over the AOI (pooled 2019–2023). Ratios are pixel-wise fractions of missing days within each season.

Species	Season	<20%	20–50%	50–80%	≥80%
NO_2	Spring (MAM)	0.0%	65.9%	33.9%	0.2%
NO_2	Summer (JJA)	25.4%	52.0%	21.4%	1.1%
NO_2	Autumn (SON)	0.0%	67.2%	32.8%	0.0%
NO_2	Winter (DJF)	0.0%	32.2%	67.3%	0.5%
SO_2	Spring (MAM)	0.0%	0.0%	77.1%	22.9%
SO_2	Summer (JJA)	0.0%	0.1%	90.6%	9.3%
SO_2	Autumn (SON)	0.0%	0.0%	56.0%	44.0%
SO_2	Winter (DJF)	0.0%	0.0%	0.0%	100.0%

Table 5.4: Share of AOI pixels by seasonal data gaps category (2019–2023).

Seasonal patterns are consistent across species (see Fig. 5.6 and Tabs. 5.3–5.4). For NO_2, summer (JJA) shows the lowest gaps (mean 35.6%) and winter (DJF) the highest (52.7%); for SO_2, data gaps is high in all seasons and peaks in winter (DJF mean 93.1%, 90th pct. 100%). Category shares reinforce this contrast: in JJA, 25.4% of NO_2 pixels fall below 20% data gaps, whereas SO_2 has essentially no pixels below 50% in any season.

5.1.4. Results of Comparative Analysis of NO_2/SO_2

Overall, SO_2 exhibits substantially higher data gaps than NO_2 (average 77.4% vs. 45.4%). Seasonally, both species are worst in winter (DJF) and best in summer (JJA), but SO_2 remains high in every season. The seasonal contrast between species (SO_2 minus NO_2) is smallest in spring and largest in winter.

	NO_2 mean	SO_2 mean	Δ (SO_2–NO_2)
Overall (2019–2023)	45.4%	77.4%	+32.0 pp
Spring (MAM)	46.0%	72.5%	+26.5 pp
Summer (JJA)	35.6%	64.3%	+28.7 pp
Autumn (SON)	47.5%	80.0%	+32.5 pp
Winter (DJF)	52.7%	93.1%	**+40.4 pp**

Table 5.5: Average data gaps (AOI means) for NO_2 vs. SO_2 (2019–2023) and seasonal differences. Percentages denote the pixel-wise fraction of missing days; Δ reports percentage-point differences.

The observed differences in data gaps patterns indicate that SO_2 and NO_2 may require distinct modeling considerations. For SO_2, it is plausible to emphasize temporal/context features and to adopt validation that reflects DJF scarcity; for NO_2, leveraging spatial

structure may be relatively more effective. These are hypotheses to be tested rather than definitive prescriptions.

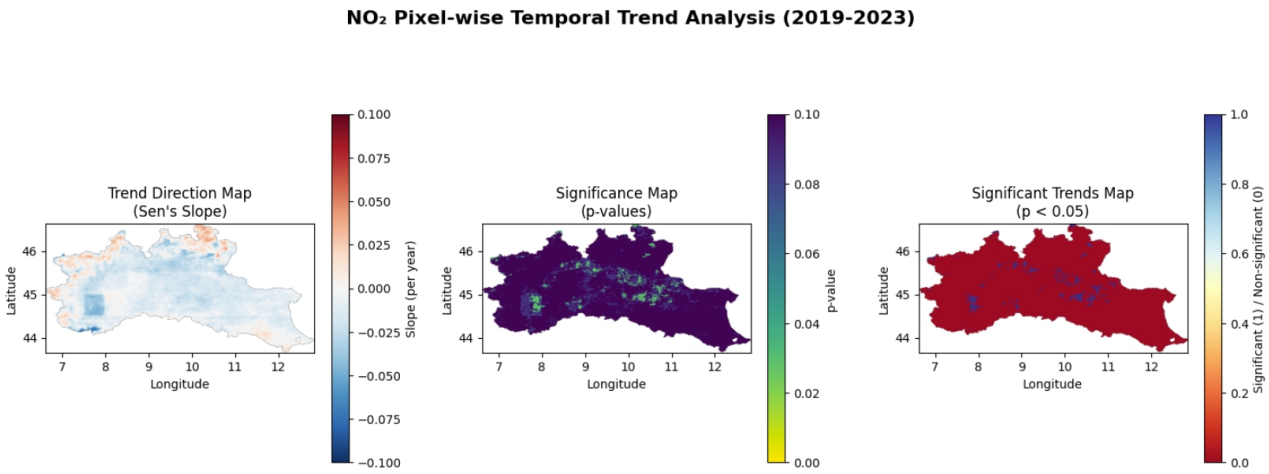
5.1.5. Results of NO_2/SO_2 Interannual Trends

Regarding the interannual Trends, we report (1) a slope map, (2) a p-value map, and (3) a binary significance mask ($p < 0.05$).

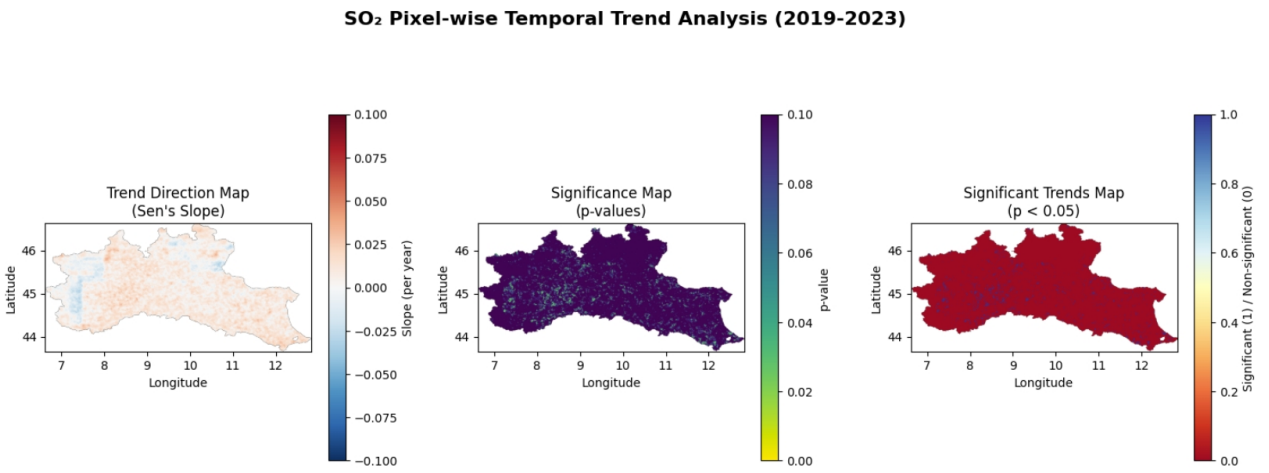
For NO_2, only a small share of pixels shows statistically significant trends ($p < 0.05$; 3.2%), with 159 increasing and 2,896 decreasing; magnitudes are modest (median Sen's slope $\approx -0.011\,\mathrm{yr}^{-1}$). SO_2 likewise has a low fraction of significant pixels (3.2%), but these are mostly increases (2,669 increasing vs. 334 decreasing) with small, spatially patchy slopes; relative to NO_2, coherent trend patches are rarer. In both products, significant pixels cluster in limited subregions.

Metric	NO_2	SO_2
Years analyzed	2019–2023	2019–2023
Total pixels	94,666	94,666
Significant trends ($p < 0.05$)	3,055 (3.2%)	3,003 (3.2%)
Increasing trends	159 (0.2%)	2,669 (2.8%)
Decreasing trends	2,896 (3.1%)	334 (0.4%)
Mean Sen's slope (yr^{-1})	−0.0101	+0.0069
Median Sen's slope (yr^{-1})	−0.0112	+0.0077
Std. dev. of slope	0.0133	0.0091
Mean p-value	0.4657	0.4896

Table 5.6: Pixel-wise temporal trend summary of data gaps for NO_2 and SO_2 over 2019–2023. Sen's slope is the median pairwise slope of annual data gaps per pixel (units: change in data gaps per year).



(a) NO_2 (2019–2023). Only **3.2%** of pixels are significant (**159** increasing, **2,896** decreasing); mean (median) Sen's slope -0.0101 $(-0.0112)\,\mathrm{yr}^{-1}$.



(b) SO_2 (2019–2023). **3.2%** of pixels are significant, dominated by increases (**2,669** increasing, **334** decreasing); mean (median) Sen's slope $+0.0069$ $(+0.0077)\,\mathrm{yr}^{-1}$.

Figure 5.7: Pixel-wise MK trend diagnostics of data gaps (2019–2023). In each row, the three panels show: (left) Sen's slope (yr^{-1}); (middle) MK p-values (0–0.10); (right) binary significance mask with threshold $p < 0.05$ (colorbar: 1 = significant, 0 = non-significant).

Overall, the maps confirm that pixel-wise trends are sparse and of small magnitude, with opposite signs prevailing for NO_2 (slight decreases) and SO_2 (slight increases)

5.1.6. Correlation Between Elevation and NO_2/SO_2 Data Gaps

Data gaps increases markedly with elevation. For NO_2, the correlations are Pearson $r = 0.769$ and Spearman $r = 0.896$; for SO_2, they are even stronger (Pearson $r = 0.920$,

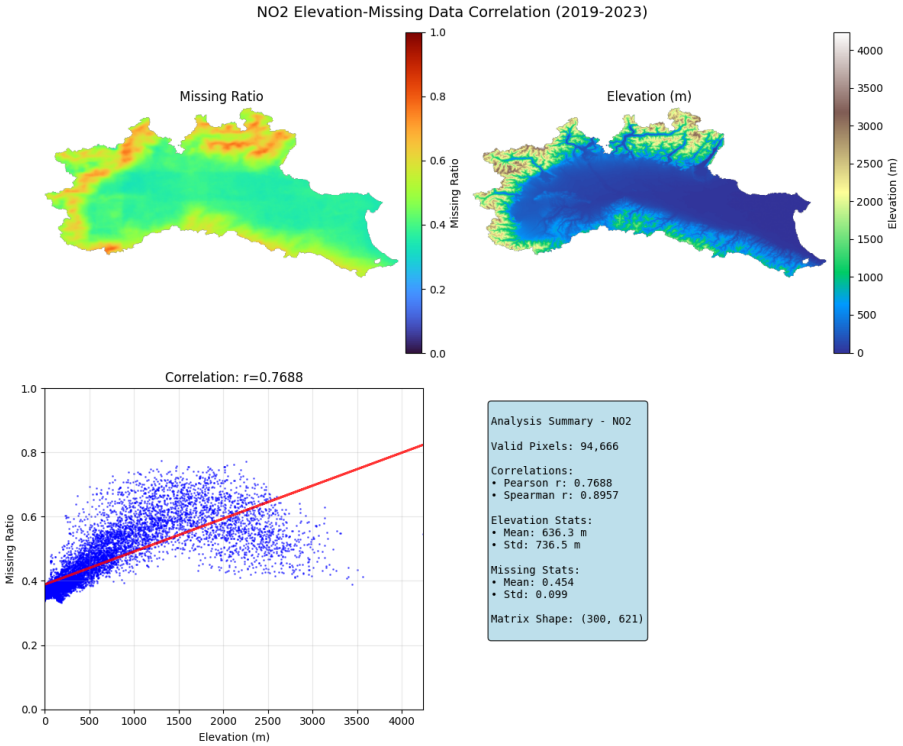
Spearman $r = 0.919$). The maps show higher gap ratios along the western/northwestern, more mountainous flank of the AOI; the scatter plots display a clear upward trend of data gaps with terrain height. In absolute terms, the mean data gaps over 2019–2023 is ≈ 0.454 for NO_2 and ≈ 0.774 for SO_2, consistent with the stronger overall sparsity in SO_2.

The positive elevation–data gaps relationship likely reflects orographic cloud formation and viewing-geometry effects in complex terrain; it does not imply direct causality by topography alone. Because SO_2 is already much sparser, the same terrain-linked cloudiness or quality screening yields a steeper apparent gradient.

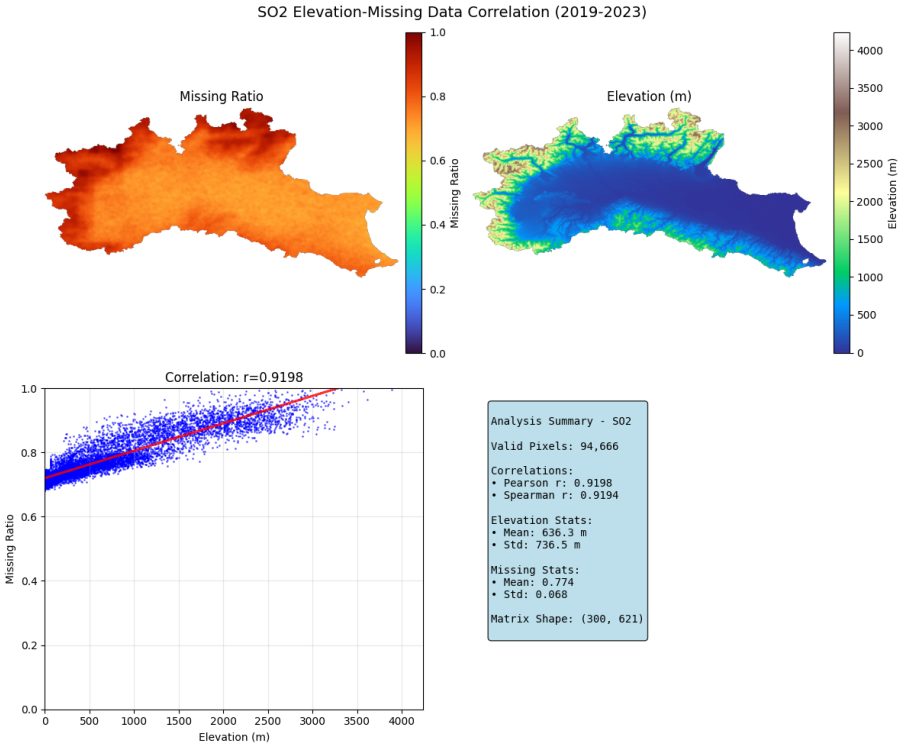
The statistical information tables and visual graphs are below.

Metric	NO_2	SO_2
Valid pixels	94,666	94,666
Pearson r (elev vs. missing)	0.7688	0.9198
Spearman r (rank)	0.8957	0.9194
Mean elevation (m)	636.3	636.3
Mean data gaps	0.454	0.774

Table 5.7: Elevation–data gaps correlation summary (2019–2023).



(a) NO_2 (2019–2023): Pearson $r = 0.769$, Spearman $r = 0.896$.



(b) SO_2 (2019–2023): Pearson $r = 0.920$, Spearman $r = 0.919$.

Figure 5.8: Elevation vs. data gaps (2019–2023). (a)–(b) show NO_2/SO_2.

5.1.7. Correlation Between Cloudiness and NO_2/SO_2 Data Gaps

After quantified how cloudiness relates to pixel-wise data gaps. For NO_2, data gaps increases with cloudiness: Pearson $r = 0.708$ and Spearman $r = 0.790$ (both $p < 10^{-300}$). The AOI-wide means over 2019–2023 are $\bar{m}_{\text{miss}} = 0.454$ (std 0.099) and $\bar{m}_{\text{cloud}} = 0.206$ (std 0.064). For SO_2, cloud fractions are much smaller and tightly clustered ($\bar{m}_{\text{cloud}} = 0.032$, std 0.007), and the observed association is strongly negative (Pearson $r = -0.905$, Spearman $r = -0.743$; both $p < 10^{-300}$). This sign difference reflects that SO_2 gaps remain high ($\bar{m}_{\text{miss}} = 0.774$, std 0.068) even under generally low reported cloudiness, consistent with additional screening and low-SNR filtering dominating SO_2 data loss rather than cloud cover alone.

Overall, cloudiness explains a meaningful share of NO_2 gaps, whereas SO_2 gaps are largely driven by other factors; the cloud–gap relationship for SO_2 should be interpreted with caution given the very narrow cloud dynamic range (mostly < 0.1). The statistical information tables and visual graphs are below.

Metric	NO_2	SO_2
Valid pixels (AOI)	94,666	94,666
Pearson r (cloud vs. missing)	+0.7075	−0.9048
Spearman r (rank)	+0.7899	−0.7427
Mean cloud fraction	0.206 (std 0.064)	0.032 (std 0.007)
Mean data gaps	0.454 (std 0.099)	0.774 (std 0.068)

Table 5.8: Cloud and data gaps correlation summary (2019–2023). Cloud fraction is the five-year mean per pixel; data gaps is the five-year fraction of days flagged missing.

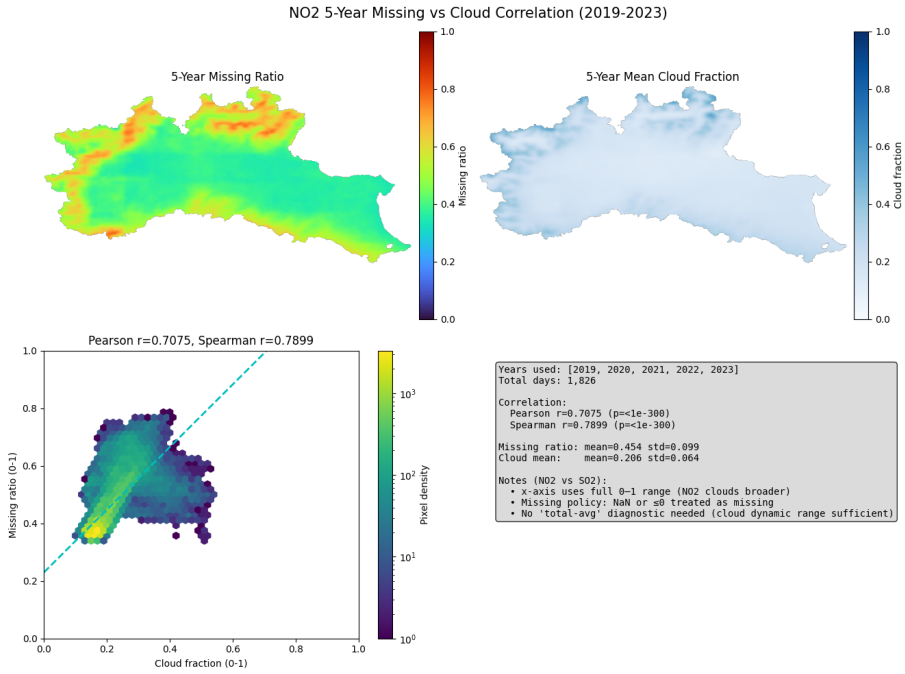


Figure 5.9: NO_2 (2019–2023): five-year data gaps (top left), five-year mean cloud fraction (top right), and pixel-wise cloud and data gaps hexbin with fitted line (bottom left). Pearson $r = 0.708$, Spearman $r = 0.790$.

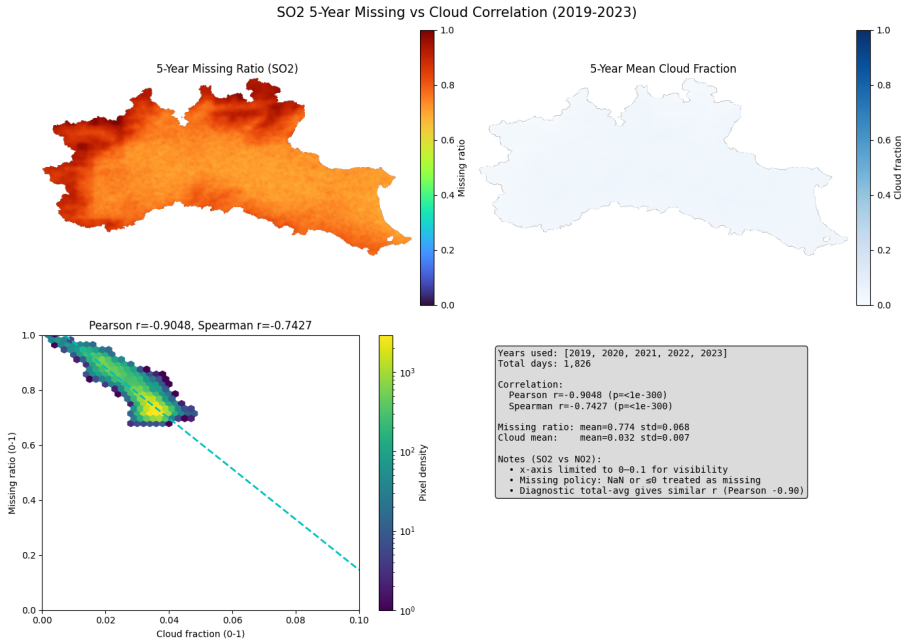


Figure 5.10: SO_2 (2019–2023): same computation as Fig. 5.9. Cloud fractions are narrowly distributed (∼0–0.1), and the observed association is negative (Pearson $r = -0.905$, Spearman $r = -0.743$), indicating SO_2 gaps are largely driven by non-cloud screening.

5.2. Result of Auxiliary Datasets Preprocess

Auxiliary predictors were reprojected to the target CRS, resampled to the modeling grid, clipped to the study boundary, and quality-controlled to remove sentinel values and non-finite entries. Masks were harmonized across years and variables. To verify cross-variable comparability—spatial continuity, interannual consistency, and potential artifacts from clipping and checking—we visualize three representative layers: 2-m air temperature (ERA5 t2m), total precipitation (ERA5 tp, accumulated depth), and surface pressure (ERA5 sp), using annual panels (2019–2023) and a five-year aggregate. All maps share the same grid, extent, and boundary; other predictors underwent the same preprocessing and checks.

Figure 5.11 shows 2-m air temperature annual means in °C. Spatial gradients are smooth with stable interannual variability and no discontinuities, indicating successful reprojection and resampling. Summary statistics: time-series mean 15.64°C, time-series standard deviation 6.89°C; spatial mean 15.64°C and spatial standard deviation 3.91°C, consistent with regional climatology.

Figure 5.12 summarizes total precipitation as annual accumulations and a 2019–2023 multi-year total in meters (ERA5 `tp`, accumulated depth). Spatial patterns and interannual changes are coherent with no tiling or striping artifacts. The multi-year total reports: minimum 0.30 m, maximum 3.71 m, mean 1.09 m, and standard deviation 0.59 m.

Figure 5.13 presents surface pressure annual means and the five-year mean in Pa (ERA5 `sp`). Large-scale gradients are preserved and stable across years. Time-series statistics: minimum 92,699.62 Pa, maximum 97,646.09 Pa, mean 95,609.58 Pa, and standard deviation 685.25 Pa; the five-year spatial mean is 95,609.59 Pa with standard deviation 6,602.14 Pa. These checks verify that auxiliary datasets are spatially aligned, temporally coherent, and numerically well-behaved for subsequent modeling.

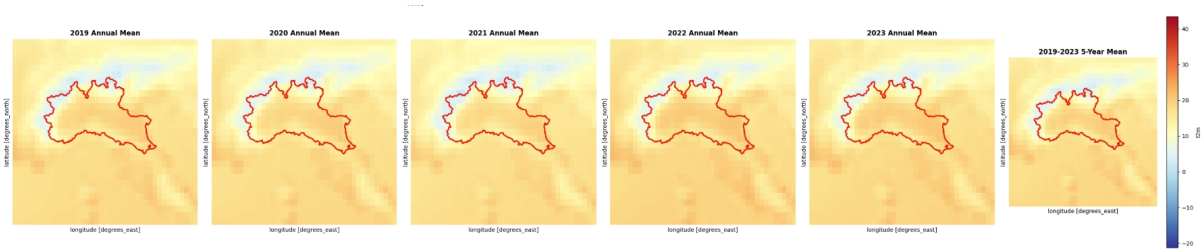


Figure 5.11: Auxiliary example 1: 2-m air temperature annual means (2019–2023) and five-year mean, units in °C.

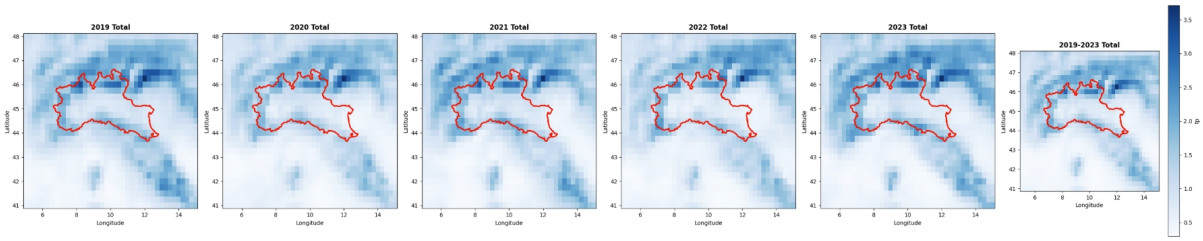


Figure 5.12: Auxiliary example 2: total precipitation—annual totals (2019–2023) and 2019–2023 multi-year total, units in meters (ERA5 `tp`).

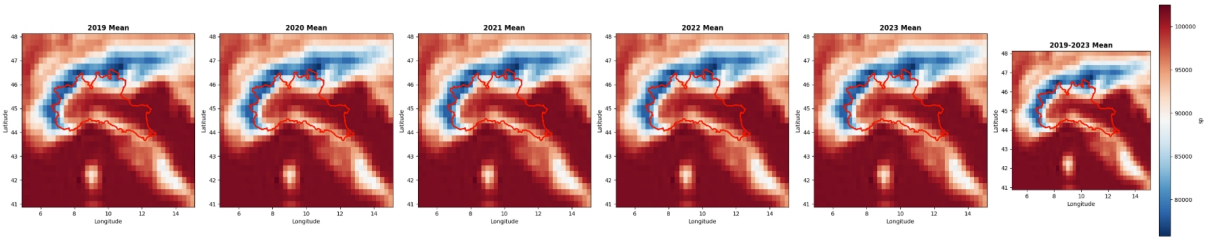


Figure 5.13: Auxiliary example 3: surface pressure annual means (2019–2023) and five-year mean, units in Pa (ERA5 `sp`).

5.3. Model Performance:LightGBM and 3D CNN

5.3.1. LightGBM

In masked validation, a random $r\%$ of the observed test pixels in 2023 (pooled over all days) are hidden and predicted. A valid sample is any masked pixel with an observed target; other inputs are available because non-sensitive missing features are mean-imputed, while the sensitive channels (`lag1`, `neighbor`) are intentionally set to NaN (which LightGBM handles natively). With the same r, NO_2 yields 531,849 / 1,063,699 / 1,595,548 valid masked pixels (10/20/30%), whereas SO_2 yields far fewer, reflecting its lower observational coverage in 2023.

Both pollutants show consistent performance across mask ratios, with SO_2 achieving slightly higher R^2 (0.1327–0.1351) compared to NO_2 (0.0879–0.0888), while NO_2 exhibits better correlation (0.4198–0.4208 vs. 0.3698–0.3724). $\mathrm{NRMSE}_{\mathrm{std}}$ values remain near 1.0 for both species, and bias is negligible across all conditions.

The modest R^2 values indicate limited predictive power when temporal and spatial context features are unavailable, establishing a performance baseline for subsequent 3D-CNN evaluation.

Table 5.9: SO_2 vs NO_2 LightGBM gap-filling masked validation on 2023 (sensitive features hidden).

	10% mask		20% mask		30% mask	
Metric	SO_2	NO_2	SO_2	NO_2	SO_2	NO_2
R^2	0.1344	0.0888	0.1351	0.0880	0.1327	0.0879
$NRMSE_{std}$	0.9304	0.9546	0.9300	0.9550	0.9313	0.9551
Corr	0.3717	0.4208	0.3724	0.4203	0.3698	0.4198
Bias	0.000042	-0.000007	0.000041	-0.000007	0.000042	-0.000007
Samples n	176,969	531,849	353,939	1,063,699	530,908	1,595,548

5.3.2. 3D CNN

This study comprehensively evaluates 3D CNN models for predicting NO_2 and SO_2 concentrations using four sampling strategies, with three independent runs for statistical reliability. Table 5.10 summarizes the comparative performance metrics.

Table 5.10: Validation results of 3D CNN models for NO_2 and SO_2 prediction

Pollutant	**Strategy**	**Sample Size**	**RMSE**	**NRMSE_std**	**MAE**	**R^2**	**Corr**
NO_2	Random Pixel	1,063,659	0.000013	0.2305	0.000010	0.9469	0.9748
	Block32	3,075,753	0.000054	0.9502	0.000037	0.0971	0.4368
	Masked Block	3,075,753	0.000054	0.9502	0.000037	0.0971	0.4368
	March Block	1,013,024	0.000039	0.9349	0.000028	0.1259	0.4351
SO_2	Random Pixel	1,488,433	0.000316	0.5111	0.000220	0.7387	0.8755
	Block32	6,630,254	0.000603	0.9745	0.000353	0.0503	0.3893
	Masked Block	6,630,254	0.000603	0.9745	0.000353	0.0503	0.3893
	March Block	811,056	0.000725	1.0497	0.000474	-0.1018	0.3618

Random pixel sampling demonstrated superior performance for both pollutants. NO_2 achieved exceptional results with $R^2 = 0.9469$, Corr = 0.9748, and minimal errors (RMSE = 0.000013, MAE = 0.000010), while SO_2 showed strong performance with $R^2 = 0.7387$ and Corr = 0.8755. Block sampling strategies performed moderately for NO_2 ($R^2 = 0.0971$) but poorly for SO_2 ($R^2 = 0.0503$), with masked neighborhood processing showing negligible impact. March block sampling revealed significant differences, achieving the best block strategy performance for NO_2 ($R^2 = 0.1259$) but worst for SO_2 ($R^2 = -0.1018$). NO_2 models consistently outperformed SO_2 across all strategies, with the performance gap most pronounced in random pixel sampling.

5.4. Gap-filling Inference and Output Generation

This section presents the gap-filling of both LightGBM and 3D-CNN models for NO_2 and SO_2 concentration mapping across selected dates in 2023. The visual comparisons demonstrate the distinct characteristics and complementary strengths of each modeling approach in reconstructing complete pollution fields from partial satellite observations.

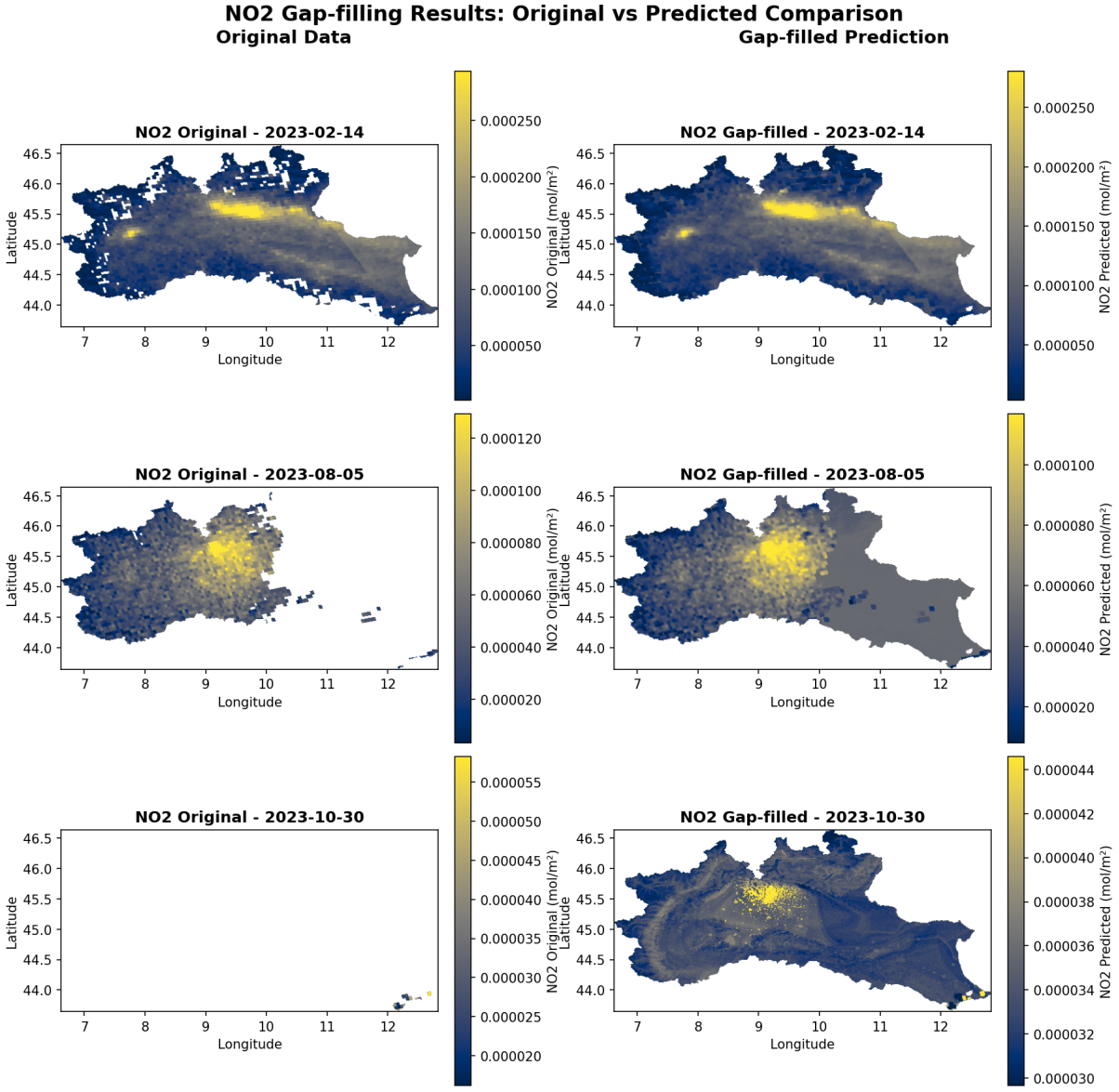


Figure 5.14: NO_2 LightGBM: three 2023 dates. Left: originals; right: gap-filled. Predictions extend urban/transport plumes and fill fragmented swaths.

The model can reconstruct partially observed scenes and even fully missing scenes; however, the quantitative accuracy of the reconstructions still needs improvement.

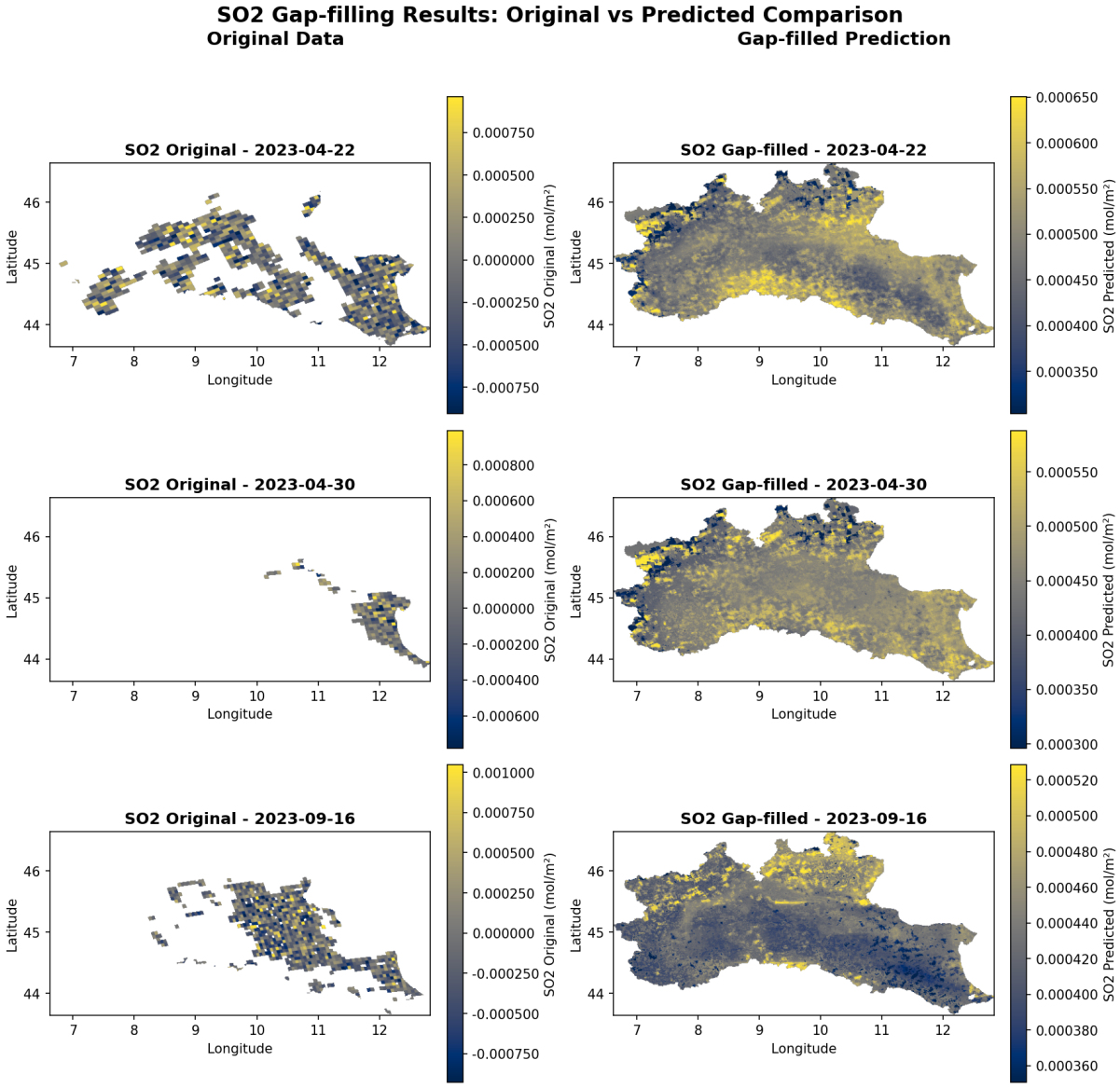


Figure 5.15: SO_2 LightGBM: predictions provide continuous backgrounds where observations are sparse; local enhancements remain visible.

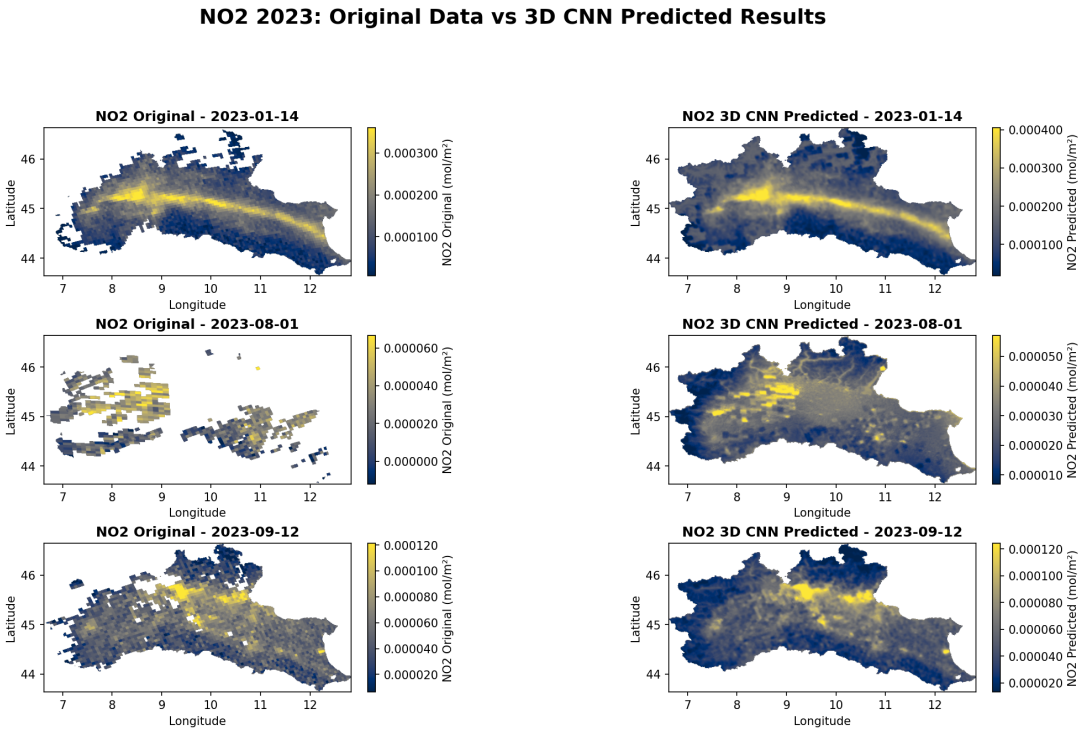


Figure 5.16: NO_2 3D-CNN: smoother, more spatially coherent gap-filling while preserving major hotspots and plume axes.

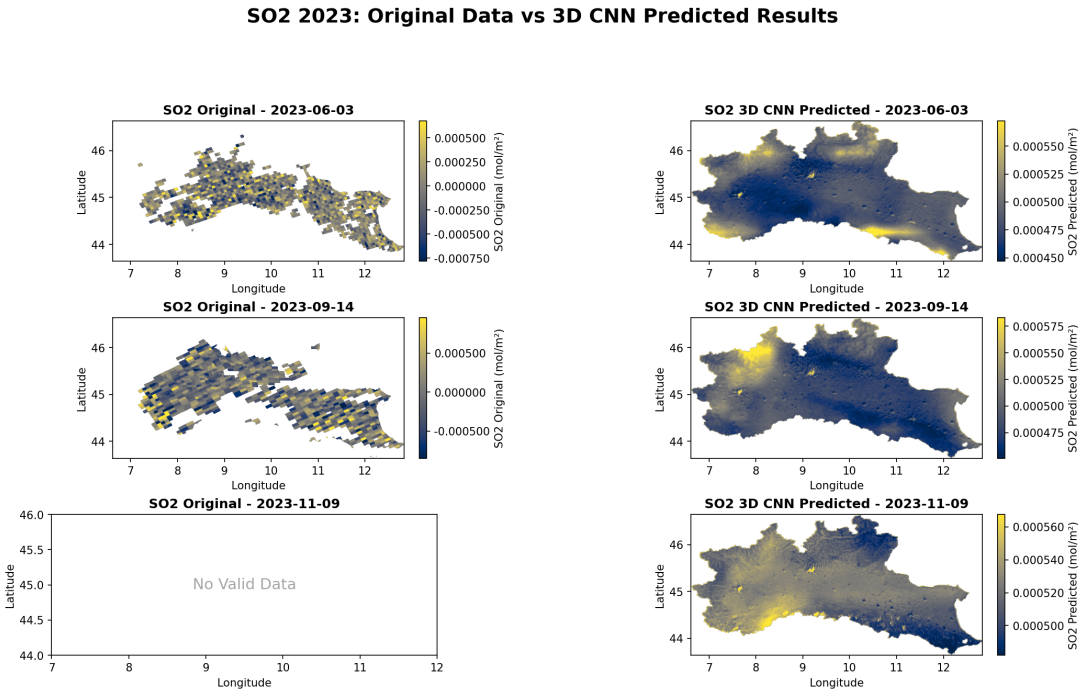


Figure 5.17: SO_2 3D-CNN: complete, plausible fields even on near-empty days; fine-scale contrast is damped relative to LightGBM.

6 | Conclusions and Future Work

This work characterised multi-year missingness in Sentinel-5P NO_2/SO_2 over the Po Valley and proposed a reproducible gap-filling pipeline. Using a harmonised multi-layer feature stack and two models (LightGBM, 3D-CNN), we combined meteorology, geography, and temporal/spatial priors. Masked validation indicates that both models recover large gaps effectively, with a clear trade-off between accuracy and computational cost.

6.1. Limitations

ERA5 drivers at 0.25° and static covariates introduce scale mismatches with S5P; global standardisation and fixed 7-day windows may underfit seasonal and synoptic variability. The neighbour and lag features, while effective, can implicitly propagate local biases; sensitive-channel masking mitigates but does not eliminate this risk. Validation is internal (masked held-out pixels) and lacks dense ground-based co-validation; uncertainty is not explicitly quantified. Finally, models were tuned for a single region and period, so generalisation outside the Po Valley and post-2023 is untested.

6.2. Future Work

Extend to new regions (Alpine arc/Europe) and species (O_3, CO, HCHO), and release gap-filled maps with uncertainty. Explore hybrid/foundation models (spatio-temporal Transformers, ConvLSTM/ViT hybrids, GNNs on dynamic grids) to capture long-range transport. Add physics-aware constraints (non-negativity, seasonal climatology, simple advection priors) and optionally assimilate CAMS to stabilise reconstructions. Establish a small benchmarking suite (shared masks, kriging/interp/partial-conv U-Net baselines, unified metrics) for transparent comparison. Finally, package an automated open pipeline (Docker/Conda + CI) with an API/CLI for on-demand inference to maximise reuse in operational and policy workflows.

Bibliography

[1] World Health Organization. Who global air quality guidelines: Particulate matter (pm2.5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, 2021. URL https://www.who.int/publications/i/item/9789240034228.

[2] David G. Streets, Timothy Canty, Gregory R. Carmichael, Benjamin de Foy, Russell R. Dickerson, Bryan N. Duncan, David P. Edwards, John A. Haynes, Daven K. Henze, Marc R. Houyoux, Daniel J. Jacob, Nickolay A. Krotkov, Lok N. Lamsal, Yang Liu, Zifeng Lu, Randall V. Martin, Gabriele G. Pfister, Robert W. Pinder, Ross J. Salawitch, and Kevin J. Wecht. Emissions estimation from satellite retrievals: A review of current capability. *Atmospheric Environment*, 77:1011–1042, 2013. doi: 10.1016/j.atmosenv.2013.05.051. URL https://doi.org/10.1016/j.atmosenv.2013.05.051.

[3] J. P. Veefkind, I. Aben, K. McMullan, H. Forster, J. de Vries, G. Otter, J. Claas, H. J. Eskes, J. F. de Haan, Q. Kleipool, M. van Weele, O. Hasekamp, R. Hoogeveen, J. Landgraf, R. Snel, P. Tol, P. Ingmann, R. Voors, B. Kruizinga, R. Vink, and P. F. Levelt. 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*, 120:70–83, 2012. doi: 10.1016/j.rse.2011.09.027. URL https://doi.org/10.1016/j.rse.2011.09.027.

[4] ESA Sentinel-5P Mission Performance Centre. Sentinel-5p level-2 product user manual: Nitrogen dioxide (no_2). Technical report, European Space Agency (ESA), 2022.

[5] European Environment Agency. Nitrogen dioxide — no_2, 2025. URL https://www.eea.europa.eu/en/analysis/publications/air-quality-status-report-2025/nitrogen-dioxide. Published 09 Apr 2025.

[6] Carmine Serio, Guido Masiello, and Angela Cersosimo. No_2 pollution over selected cities in the po valley in 2018–2021 and its possible effects on boosting covid-19 deaths. *Heliyon*, 8(8):e09978, 2022. doi: 10.1016/j.heliyon.2022.e09978. URL https://doi.org/10.1016/j.heliyon.2022.e09978.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 3146–3154, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

[8] Huyan Fu, Zhenfeng Shao, Peng Fu, Xiao Huang, Tao Cheng, and Yewen Fan. Combining atc and 3d-cnn for reconstructing spatially and temporally continuous land surface temperature. *International Journal of Applied Earth Observation and Geoinformation*, 108:102733, 2022. doi: 10.1016/j.jag.2022.102733. URL https://doi.org/10.1016/j.jag.2022.102733.

[9] Kai Liu, Xueke Li, Shudong Wang, and Hongyan Zhang. A robust gap-filling approach for european space agency climate change initiative (esa cci) soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning. *Hydrology and Earth System Sciences*, 27:577–598, 2023. doi: 10.5194/hess-27-577-2023.

[10] David J. Lary, Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016. doi: 10.1016/j.gsf.2015.07.003. URL https://doi.org/10.1016/j.gsf.2015.07.003.

[11] G. Pirovano, C. Colombi, A. Balzarini, G. M. Riva, V. Gianelle, and G. Lonati. Pm2.5 source apportionment in lombardy (italy): Comparison of receptor and chemistry-transport modelling results. *Atmospheric Environment*, 2015. doi: 10.1016/j.atmosenv.2015.01.073.

[12] European Environment Agency. Impacts of air pollution on ecosystems, 2022. URL https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2022/impacts-of-air-pollution-on-ecosystems. Published 24 Nov 2022.

[13] Sentinel-5p tropomi no_2 — offline level 3 (copernicus/s5p/offl/l3_no2), 2018. URL https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2#description.

[14] Sentinel-5p tropomi so_2 — offline level 3 (copernicus/s5p/offl/l3_so2), 2018. URL https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_SO2#description.

[15] KNMI and partners. *Sentinel-5P Level 2 Product User Manual: Nitrogen Dioxide*, 2022. URL https://sentinels.copernicus.eu/documents/247904/2474726/Sen

tinel-5P-Level-2-Product-User-Manual-Nitrogen-Dioxide.pdf. Includes QA value definition and conditions (clouds, albedo, geometry).

[16] DLR and partners. *Sentinel-5P Level 2 Product User Manual: Sulphur Dioxide (SO_2)*, 2024. URL https://sentiwiki.copernicus.eu/__attachments/1673595/S5P-L2-DLR-PUM-400E%20-%20Sentinel-5P%20Level%202%20Product%20User%20Manual%20Sulphur%20Dioxide%20SO2%202024%20-%202.8.0.pdf. Lists QA parameters, processing quality flags and surface classification.

[17] Jesus Rodrigo Cedeño Jimenez and Maria Antonia Brovelli. Estimating ground-level no_2 concentrations using machine learning exclusively with remote sensing and era5 data: The mexico city case study. *Remote Sensing*, 16:3320, 2024. doi: 10.3390/rs16173320.

[18] John H. Seinfeld and Spyros N. Pandis. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, Hoboken, NJ, 3 edition, 2016. ISBN 978-1-118-94740-1.

[19] Roland B. Stull. *An Introduction to Boundary Layer Meteorology*. Springer, Dordrecht, 1988. ISBN 978-90-277-2768-3. doi: 10.1007/978-94-009-3027-8.

[20] V. Comegna and A. Basile. Temporal stability of spatial patterns of soil water storage in a cultivated vesuvian soil. *Geoderma*, 62(1–3):299–310, 1994. doi: 10.1016/0016-7061(94)90042-6.

[21] Negar Siabi, Seyed Hossein Sanaeinejad, and Bijan Ghahraman. Effective method for filling gaps in time series of environmental remote sensing data: An example on evapotranspiration and land surface temperature images. *Computers and Electronics in Agriculture*, 193:106619, 2022. doi: 10.1016/j.compag.2021.106619.

[22] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2 edition, 2002.

[23] Joseph L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.

[24] Daniel J. Weiss, Peter M. Atkinson, Samir Bhatt, Bonnie Mappin, Simon I. Hay, and Peter W. Gething. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:106–118, 2014. doi: 10.1016/j.isprsjprs.2014.10.001.

[25] Florian Gerber, Rogier de Jong, Michael Schaepman, and Reinhard Furrer. Predicting missing values in spatio-temporal remote sensing data. *IEEE Transactions on*

Geoscience and Remote Sensing, pages 1–13, 2018. doi: 10.1109/TGRS.2017.2785240. Early Access.

[26] Yves Julien and Jose Sobrino. Optimizing and comparing gap-filling techniques using simulated ndvi time series from remotely sensed global data. *International Journal of Applied Earth Observation and Geoinformation*, 76:93–111, 2019. doi: 10.1016/j.jag.2018.11.008.

[27] Ronggao Liu, Rong Shang, Yang Liu, and Xiaoliang Lu. Global evaluation of gap-filling approaches for seasonal ndvi with considering vegetation growth trajectory, protection of key point, noise resistance and curve stability. *Remote Sensing of Environment*, 189:164–179, 2017. doi: 10.1016/j.rse.2016.11.023.

[28] Chao Zeng, Huanfeng Shen, and Liangpei Zhang. Recovering missing pixels for landsat etm+ slc-off imagery using multi-temporal regression analysis and a regularization method. *Remote Sensing of Environment*, 131:182–194, 2013. doi: 10.1016/j.rse.2012.12.012.

[29] Marius Appel. Efficient data-driven gap filling of satellite image time series using deep neural networks with partial convolutions. *Artificial Intelligence for the Earth Systems*, 2024. doi: 10.1175/AIES-D-22-0055.1. Early Online Release.

[30] Google earth engine code editor, 2025. URL `https://code.earthengine.google.com/`. Accessed: 2025-09-24.

[31] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. doi: 10.1016/j.rse.2017.06.031.

[32] Google Earth Engine Catalog. Copernicus/s5p/offl/l3_no2 — tropomi offline no_2 level-3, 2024. URL `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2`.

[33] Google Earth Engine Catalog. Copernicus/s5p/offl/l3_so2 — tropomi offline so_2 level-3, 2024. URL `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_SO2`.

[34] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. Hogan, E. Hólm,

M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

[35] Era5 single levels reanalysis, 2025. URL `https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels`. Accessed: 2025-09-24.

[36] NASA JPL. Nasa shuttle radar topography mission (srtm) global 1 arc second v003, 2013. Accessed via Google Earth Engine: USGS/SRTMGL1_003.

[37] D. Zanaga, R. Van De Kerchove, D. Daems, and ... Esa worldcover 10 m 2020 v100, 2021.

[38] Berthold K. P. Horn. Hill shading and the reflectance map. *Proceedings of the IEEE*, 69(1):14–47, 1981. doi: 10.1109/PROC.1981.11918.

[39] Andrew J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4: 170004, 2017. doi: 10.1038/sdata.2017.4.

[40] WorldPop, University of Southampton. Worldpop un-adjusted population density, 2020, 2020. URL `https://www.worldpop.org/`. 1 km gridded population density (people/km^2), EPSG:4326.

[41] Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, 10(2):e0107042, 2015. doi: 10.1371/journal.pone.0107042.

A | Appendix A

Code repository: `https://github.com/zhanbinw/s5p-gapfill-no2-so2`
Note: This is process code from the research journey. It contains redundant, exploratory, and experimental scripts/notebooks kept for transparency and reproducibility.

Dataset: 10.5281/zenodo.17226552

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to everyone who supported me throughout my studies and research.

First, I am deeply grateful to my supervisor, **Prof. Maria Antonia Brovelli**. Her rigorous scholarship and broad vision have benefited me immensely, and her guidance at every stage of this thesis has been invaluable. I also greatly admire her discerning leadership in building a strong research team and her generosity in providing me with substantial resources.

I am especially indebted to my co-supervisor, **Vasil Yordanov**. He is meticulous and highly responsible, consistently responding with remarkable speed and executing with great efficiency. His explanations were always clear, and from the very beginning of this thesis to its completion he offered support that I find hard to put into words. His prompt follow-ups and commitment to closing the loop on every issue have earned my deepest respect. Under his patient supervision, I progressed from initial confusion about the topic to systematically completing each part of the work. He encouraged me when things went well and, when mistakes occurred, he pointed them out with rigor and taught me how to improve. I am profoundly grateful for this journey. I also wish to thank my second co-supervisor, **Jesus Rodrigo Cedeno Jimenez**, for his patience and generosity. Whenever I was uncertain, he provided detailed guidance that greatly helped my understanding. Working with Vasil and Rodrigo made this thesis process both productive and rewarding, and I learned a great deal from them.

I would also like to thank **Prof. Carlo Iapige De Gaetani** for kindly facilitating the connection with my thesis supervisors at a crucial moment. Although we have not met in person, his trust and support from afar were instrumental in setting this work in motion.

Finally, as I conclude this master's journey with this thesis in Milan, I am filled with gratitude. Thanks to my friends and former colleagues—who were a call away across time zones and who even travelled to be by my side when my studies felt overwhelming. I also thank my family for their unconditional support and unfailing encouragement.