



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
IL FUTURO DELLA TERRA, CAMBIAMENTI CLIMATICI E SFIDE
SOCIALI

Ciclo 38

Settore Concorsuale: 02/C1 - ASTRONOMIA, ASTROFISICA, FISICA DELLA
TERRA E DEI PIANETI

Settore Scientifico Disciplinare: FIS/06 - FISICA PER IL SISTEMA TERRA E IL
MEZZO CIRCUMTERRESTRE

SYNERGISTIC USE OF LIMB AND NADIR OBSERVATIONS FOR STUDYING
OZONE STRATOSPHERIC INTRUSIONS IN THE HIMALAYAN REGION

Presentata da: Liliana Guidetti

Coordinatore Dottorato
Prof.ssa Silvana Di Sabatino

Supervisore
Dott.ssa Erika Brattich

Co-supervisor
Dott. Ugo Cortesi
Prof.ssa Michaela Hegglin
Dott.ssa Piera Raspollini
Dott. Simone Ceccherini
Dott.ssa Cecilia Tirelli
Dott. Nicola Zoppetti

Esame finale anno 2026

Abstract

Tropospheric ozone plays a key role in atmospheric chemistry, dynamics, and radiation, influencing both climate and air quality. Among its natural sources, stratosphere–troposphere exchange (STE) processes represent a major contributor to ozone variability in the upper troposphere, particularly over complex regions such as the Himalayas, where orography and meteorological dynamics favor stratospheric intrusions. However, observational limitations—especially in vertical resolution and spatial coverage, hamper a comprehensive characterization of these events.

This thesis addresses these challenges by developing and validating a novel ozone dataset for 2008–2011 obtained through the Complete Data Fusion (CDF) of limb observations from MIPAS and nadir observations from IASI. The fusion approach exploits the complementary strengths of both instruments, achieving improved vertical information and reduced uncertainties, especially in the Upper Troposphere–Lower Stratosphere (UTLS) region. Validation against ozonesonde profiles confirmed the enhanced performance of the fused product, particularly in terms of number of degrees of freedom and error reduction, demonstrating its potential as a reliable tool for studying ozone dynamics.

The second part of the work applies the fused dataset to investigate stratospheric intrusion events over the Himalayan region. A case study (28 January – 2 February 2008) was analyzed using a combination of satellite, model, and reanalysis data, revealing the capability of the fused dataset to capture key dynamical and chemical signatures—such as tropopause folding and the downward transport of ozone-rich air masses. Furthermore, an unsupervised classification framework based on Principal Component Analysis and k-means clustering was developed to objectively identify profiles associated with intrusion conditions. The approach successfully distinguished intrusion from non-intrusion profiles, highlighting the added value of the fused product for both process understanding and automated event detection.

Overall, this work demonstrates the scientific relevance of combining limb and nadir satellite observations to enhance the representation of atmospheric processes that are otherwise difficult to observe. The new dataset and methodological framework lay the foundation for large-scale detection and climatological analysis of stratospheric intrusions, providing a bridge between observation and modeling in atmospheric composition studies. Future extensions of this approach will support a more accurate investigation of stratosphere-to-troposphere exchange (STE) processes, contributing to an improved representation of ozone variability in atmospheric models and enhancing our understanding of its role in climate and air quality.

Contents

Introduction	15
1 Background	19
1.1 Atmospheric Ozone Overview	19
1.2 Ozone Stratospheric Intrusions	22
1.2.1 Introduction to Stratospheric Intrusion Events	22
1.2.2 Dynamical Processes and Drivers	24
1.2.3 Detection and Analysis Techniques	27
1.2.4 A Review of Stratospheric Ozone Intrusions over the Himalayas	31
1.3 Radiative Transfer Theory	34
1.3.1 Radiance and Spectral Flux	35
1.3.2 Radiative Transfer Equation	36
1.3.3 The Earth Emission Spectrum	37
1.4 The Inverse Problem	38
1.4.1 The Forward Model and Retrieval Process	39
1.4.2 Optimal Estimation Method	40
1.4.3 Uncertainties in the Retrieval Process	44
1.4.4 Diagnostics: Errors, Averaging Kernels, Degrees of Freedom, and Information Content	45
2 Production and Validation of a New 2008-2011 Ozone Dataset	49
2.1 Chapter Overview	49
2.2 Data	50
2.2.1 Remote Sensing Data	50
2.2.2 Ozonosondes Data	57
2.3 Methods	58
2.3.1 The Complete Data Fusion (CDF): Theory and Application	58
2.3.2 Validation Methodology	70
2.3.3 Comparison Methodology	72
2.4 Results and Discussion	72
2.4.1 Tuning Phase Results	73
2.4.2 V05: Comparison between Fused and L2	76
2.4.3 V05: Validation	78
2.4.4 About the Bias at 10km	80
2.5 Chapter Conclusions	83

3	Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset	85
3.1	Chapter Overview	85
3.2	Data	87
3.2.1	Nepal Climate Observatory Pyramid (NCO-P) and Intrusions Dataset	87
3.2.2	ERA5 Dataset	89
3.2.3	CAMS Dataset	89
3.2.4	EMAC Dataset	90
3.2.5	Considerations About Models	91
3.3	Methods	91
3.3.1	Case Study Definition	91
3.3.2	Model and Reanalysis Data Processing and Intercomparison	92
3.3.3	Profiles Characterization: Reanalysis, Model, Fused, and L2	93
3.3.4	Classification Methodology	94
3.4	Results and Discussion	97
3.4.1	Time Series Analysis Result for Case Study Definition	97
3.4.2	Visualizing the Case Study Event through ERA5, CAMS, and EMAC	99
3.4.3	Profiles Characterization	113
3.4.4	What We Learned from Classification: Towards O ₃ Intrusions De- tection	122
3.5	Chapter Conclusions	128
4	Conclusions and Future Developments	131
A	Basic Statistics	135
A.1	Mean, Variance, Covariance and the Gaussian Distribution	135
A.2	Errors and Propagation	136
B	Complete Data Fusion Background	139
B.1	Equivalence between CDF and Synergistic Retrieval	139
C	STE Events Background	141
C.1	Conserved Properties	141
C.1.1	Potential Temperature	141
C.1.2	Circulation, Vorticity and Potential Vorticity	142
D	MIPAS Validation	145
D.1	Validation Technique	145
D.2	Validation Results	146
E	In Situ NCO-P Data Analysis	149

List of Figures

1.1	Atmospheric ozone vertical distribution (M. I. L. A. Hegglin et al., 2015). This profile shows schematically how ozone changes with altitude in the tropics. Most ozone resides in the stratospheric “ozone layer.” Increases in ozone near the surface are a result of air pollutants released by human activities.	19
1.2	Earth absorption spectrum by atmospheric components, among them the ozone (second box). From NASA PACE.	21
1.3	Seasonally averaged stratosphere to troposphere (STT) mass flux for 1979–2011 (DJF: December, January, February; MAM: March, April, May; JJA: June, July, August; SON: September, October, November) derived from trajectory analysis using ERA-Interim reanalysis data as input. From Škerlak et al., 2014.	23
1.4	Dynamical aspects of stratosphere-troposphere exchange from Holton et al., 1995. Notably, according to present knowledge, tropical cumulonimbus clouds do not reach into the stratosphere (Stohl et al., 2003).	25
1.5	Himalayan region from Google Earth.	32
1.6	Simulated outgoing spectral radiance at the top of Earth’s atmosphere in clear and all sky conditions, showing the absorption at specific frequencies and the principle absorbers. For comparison, the red curves show the flux from a classic “blackbody” at 294 K, and the flux from the tropopause at 210 K.	38
1.7	Schematic representation of Bayes’ approach in a retrieval problem from Rodgers, 2000.	40
2.1	Optical layout of a Michelson interferometer.	50
2.2	Sinc function.	52
2.3	The MIPAS scanning geometry from ESA Earth Online official website (https://earth.esa.int/eogateway/instruments/mipas/description).	52
2.4	IASI observing mode. An instantaneous field of view of 4 circular pixels is taken every 50 km, over a swath width of 2x1100 km. Each circular pixel corresponds to a 12 km diameter footprint on the ground at nadir (Clerbaux et al., 2009).	54
2.5	Example of MIPAS (left) and IASI (right) spatial distribution for one day in 2008.	55
2.6	The left panels show an example of MIPAS (left-blue) and IASI (right-green) vertical grids. The right panels show MIPAS (right-above) and IASI (right-below) state vector scheme.	56

2.7	Example of diagonal elements of averaging kernel of MIPAS (left) and IASI (right) for the entire state vectors.	56
2.8	Global map of WOUDC stations (green markers). Red markers identify stations effectively used in this study.	57
2.9	Validation of the 9 versions of the CDF method in the tuning phase, performed in 2008 within the latitudinal band 5° - 55° N . The blue line represents the average percentage bias between the fused profiles and the smoothed ozonesondes with the AK of the fused profile. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), while the orange bars show the standard error of the mean. The background gray profiles represent the n averaged profiles. In the legend, DOFs mean u20 represents the average number of degrees of freedom below 20 km.	74
2.10	Mean total errors of fused products for each CDF version.	75
2.11	Comparison between fused, IASI, MIPAS and ozonesondes for version v05 over all years. The blue line represents the average percentage bias between fused or L2 and ozonesonde, where the ozonesonde has not been smoothed with the AK. The dashed red line represents the standard deviation of the mean at 2σ . The dashed black line represents the combined error (2σ), while the orange bars represent the standard error of the mean. The transparent gray profiles are the averaged profiles. The green profile represents N, which is the number of MIPAS measurements at each point of the vertical grid.	77
2.12	Comparison of the mean total errors of v05 fused products and L2 for each year.	78
2.13	Validation of version v05. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed with the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), while the black dashed line represents the combined error (2σ). The orange bars show the standard error of the mean. The transparent grey lines correspond to the individual averaged profiles. Here n indicates the total number of valid fused profiles considered.	79
2.14	Validation of v05 across three latitudinal bands. The blue line represents the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), and the orange bars show the standard error of the mean. The transparent grey lines correspond to the individual averaged profiles. Here n indicates the number of valid fused profiles considered. . .	79
2.15	Validation of IASI for 2008 in the 5 – 55° N region. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), and the orange bars correspond to the standard error of the mean. The transparent grey lines represent the individual averaged profiles. 20 profiles were found to perform the validation.	82

2.16	Validation of v05 with IASI bias correction applied for 2008. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), and the orange bars correspond to the standard error of the mean. The transparent grey lines represent the individual averaged profiles. Here n indicates the number of valid fused profiles.	82
2.17	Validation of version v05 with IASI bias correction for 2009–2010–2011. In blue is the mean percentage bias between fused profiles and ozonesondes, smoothed with the fused AK. The red dashed line indicates the standard deviation of the mean bias (2σ), while the black dashed line represents the composite error (2σ). Orange bars show the standard error of the mean. The semi-transparent grey profiles represent the individual averaged profiles. Here, n denotes the number of valid fused profiles considered. . . .	83
3.1	NCO-P geographical location (Cristofanelli et al., 2010).	87
3.2	Seasonal distribution of STEFLUX (left) and SIO (right) events. Bars represent the four seasons (DJF, MAM, JJA, SON), with stacked colors indicating the contribution of different years. The plots highlight the marked seasonality in the occurrence of events. Here the frequency is the 1-score occurrence. Bars are stacked chronologically: for each season, years are ordered from 2006 (bottom) to 2013 (top), following the same sequence shown in the legend.	98
3.3	ERA5 horizontal maps daily evolution at 500 hPa, 300 hPa, 150 hPa, and 50 hPa, at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by ERA5. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.	100
3.4	ERA5 horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by ERA5. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.	101
3.5	ERA5 vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by ERA5, while the black dashed line is the NCO-P location.	102
3.6	CAMS horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.	103

3.7	CAMS vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.	104
3.8	CAMS vertical distributions of stratospheric ozone SO_3 (meridional-left, zonal-right) on January 31st, 2008 at 12UTC. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.	105
3.9	CAMS vertical distribution of the ratio between stratospheric ozone O_3S and ozone O_3 for January 30th, 2008 at 12UTC.	106
3.10	EMAC horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by EMAC. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.	107
3.11	EMAC vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by EMAC, while the black dashed line is the NCO-P location.	108
3.12	EMAC vertical distributions of stratospheric ozone SO_3 (meridional-left, zonal-right) on January 31st, 2008 at 12UTC. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.	109
3.13	EMAC vertical distribution of the ratio between stratospheric ozone O_3S and ozone O_3 for January 30th, 2008 at 12UTC.	109
3.14	O_3 percentage differences between ERA5 and CAMS - horizontal distributions at 500 hPa, 300 hPa, 150 hPa, and 50 hPa (first and second row), and vertical distribution (third row), for January 30th,2008 at 12 UTC. Black point is the NCO-P location, while white shapes represent the mountain profile.	110
3.15	O_3 percentage differences between EMAC and CAMS - horizontal distributions at 500 hPa, 300 hPa, 150 hPa, and 50 hPa (first and second row), and vertical distribution (third row), for January 29th,2008 at 12 UTC. Black point is the NCO-P location, while white shapes represent the mountain profile.	112
3.16	O_3S percentage differences between EMAC and CAMS - vertical distributions on January 29th, 2008, at 12 UTC. Black dashed line is the NCO-P location.	113

3.17	ERA5 mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	114
3.18	CAMS mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	115
3.19	EMAC mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	116
3.20	ERA5 median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	118
3.21	CAMS median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	119
3.22	EMAC median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.	120
3.23	First row: mean and standard deviation (2σ) for fused and L2 event and non-event profiles. Second row: median and interquartile range (IQR) between the 25° and 75° percentile for fused and L2 event and non-event profiles.	121
3.24	Elbow method applied for all fused, MIPAS, and IASI case study profiles. .	123

3.25	PCA of ozone profiles from the fused, IASI, and MIPAS datasets of the case study. Left panels show the projection of the profiles onto the first two principal components (PC1 and PC2), coloured according the two clusters. Right panels display the corresponding PCA loadings or weights as a function of altitude.	124
3.26	Mean of the profiles per clusters and classification performances.	125
3.27	PCA scatterplots for fused (first row), IASI (second row), MIPAS (third row) products. First column shows scatterplots in the PC-plane. Second column shows scatterplots in the z_{max} , O_{3max} plane. Third column shows which of the profiles represented in the scatterplots lies in the two sub-regions selected in the Section 3.4.3 as sub-regions with event and non-event profile.	126
3.28	PCA loadings for fused, MIPAS and IASI products.	127
D.1	MIPAS validation against WOUDC ozononesondes in 2008. Left panel: validation using fixed-grid interpolation approach. Right panel: validation using binning approach.	147
D.2	MIPAS validation against WOUDC ozononesondes in 2008. Left panel: validation using fixed-grid interpolation approach and volume mixing ratio as unit measurements. Right panel: validation using fixed-grid interpolation approach and partial columns as unit measurements.	147
E.1	Distribution of daily event deviations of ozone (O_3) and relative humidity (RH) from the monthly mean for SIO and STEFLUX events. Each boxplot shows the median (black line in the box), interquartile range (the box), and outliers (the whiskers) for each month, with a reference red line at zero. . .	150
E.2	Annual histogram of SI events detected by the STEFLUX (light blue) and SIO (light red) methodologies. Each bar represents the total number of days per year in which events occurred, while overlaid labels indicate the number of distinct events and their average duration (in days). Events are defined as one or more consecutive days with a methodology value of 1, highlighting both the frequency and persistence of the phenomena.	151
E.3	Results of OLS regression, Seasonal Kendall, and Deseasonalized Y-P tests. Filled cells indicate statistical significance.	152
E.4	STL decomposition of the time series of stratospheric intrusion events frequency identified by the STEFLUX methodology at NCO-P for the period 2006–2013.	154
E.5	STL decomposition of the time series of stratospheric intrusion events frequency identified by the SIO methodology at NCO-P for the period 2006–2013.	155
E.6	STL decomposition of the time series of surface ozone values during stratospheric intrusion events identified by the STEFLUX methodology at NCO-P for the period 2006–2013.	156
E.7	STL decomposition of the time series of surface ozone values during stratospheric intrusion events identified by the SIO methodology at NCO-P for the period 2006–2013.	157

List of Tables

2.1	Summary of the nine selected CDF versions based on different choices of interpolation and coincidence covariance matrices (CM), use of off-diagonal reconstruction with a correlation length of 6 km, and application of a factor k on the coincidence CM.	69
2.2	DOFs and number of profiles containing negative values analysis. Header is composed of: the total number of fused profiles in 6 days (N_{tot}), the percentage of fused profiles with more than one negative value (N_{neg}), and the mean number of degrees of freedom both over the full profile and below 20 km, for both the fused ($Fused (tot \& u20 km)$) product and for L2 ($MIPAS (tot \& u20 km)$; $IASI (tot \& u20 km)$) products.	73
2.3	Mean DOFs for different altitude ranges (Δz_1 : 0–5km, Δz_2 : 5–20km, Δz_3 : 20–30km), and for each product and year.	76
3.1	Summary of datasets used for the case study, including variable, measurement unit, horizontal resolution and dataset type.	92

Introduction

Tropospheric ozone is a regulated air pollutant and a greenhouse gas with a central role in chemical, dynamical, and radiative processes in the atmosphere. Changes in its concentrations affect the climate through radiative and biogenic feedbacks and pose risks to human and ecosystem health (Fuhrer et al., 1997; Joiner et al., 2009; Fleming et al., 2018). Although it has no direct emission sources, it is produced through photochemical reactions involving sunlight, volatile organic compounds (VOCs), and nitrogen oxides (NO_x) emitted by natural and primarily anthropogenic activities (Crutzen and Lelieveld, 2001; J. Zhang et al., 2019; Donzelli and Suarez-Varela, 2024). Its short lifetime, spatial variability of sources and sinks, and dependence on sunlight result in significant spatial and temporal variability (Tarasick et al., 2010; Prather and Zhu, 2024).

A key natural source of tropospheric ozone is atmospheric circulation (Cristofanelli et al., 2015), particularly through Stratospheric-Tropospheric Exchange (STE) processes, which facilitate the bidirectional transfer of air masses, gases, and aerosols between the stratosphere and troposphere. STE dynamics, driven by mechanisms like Brewer-Dobson circulation, tropopause folding, and atmospheric waves, significantly influence tropospheric ozone concentrations (Holton et al., 1995; Stohl et al., 2003; Mohanakumar, 2008; Fueglistaler et al., 2009; Gettelman et al., 2011). Stratospheric contributions to tropospheric ozone account for 25-50% in the extratropics, with the strongest influence in the Southern Hemisphere (Banerjee et al., 2016). Low stratospheric ozone explains up to 70% of the variability in mid-tropospheric ozone at mid-latitudes (Hess and Zbinden, 2013). A global overall statistically significant increase in tropospheric ozone has been reported (Williams et al., 2019), and mainly attributed to changes in atmospheric circulation and the strengthening of the Brewer-Dobson Circulation (BDC) (Hegglin and Shepherd, 2009; Butchart, 2014; WMO, 2018).

The Himalayan region is particularly affected by stratospheric intrusions due to its high altitude, complex topography, and dynamic meteorological conditions (Stohl et al., 2003; Randel and Park, 2006). Processes such as tropopause folding, monsoonal convection, and interactions with the subtropical jet stream facilitate the downward transport of ozone-rich air masses. Studies have reported significant stratospheric contributions to surface-level ozone in this region, with a 25% contribution detected during winter (Cristofanelli et al., 2010). The rise in tropospheric ozone levels near the southern flank of the Himalayas has also been associated with increased stratospheric influence (Williams et al., 2019). Nevertheless, considerable uncertainty persists regarding the drivers of observed surface ozone variability, owing to the frequent failure of current models to capture the measured magnitude and temporal dynamics.

Our current understanding of the spatial and temporal climatology of tropospheric ozone is closely related to our ability to measure and simulate it. In this context, determin-

ing the stratosphere-to-troposphere exchange (STE) contribution to tropospheric ozone concentrations is crucial. Ozone observations are currently performed using ozonesondes and other in-situ and remote sensing instruments. These measurements are often combined with statistical models based on ozone tracers (e.g. Be, Pb, Rn)(Brattich et al.; Brattich et al., 2017, 2021). While in-situ measurements are limited in spatial and temporal coverage, remote sensing observations offer a broader perspective, despite the constraints introduced by the inversion techniques required for their retrieval. Dynamical and chemical models are also widely used to simulate STE events and are often compared with observations; however, lack vertical resolution and systematic biases between models and observations still exist (Hegglin et al., 2010; Williams et al., 2019). Understanding STE processes, particularly in regions such as the Upper Troposphere–Lower Stratosphere (UTLS), requires atmospheric observations with both high vertical resolution—due to the strong vertical gradients—and adequate horizontal resolution, since these events often occur over spatially confined regions. Satellite remote sensing instruments offer complementary strengths: limb-sounding sensors provide high vertical resolution but limited horizontal coverage, while nadir-viewing sensors offer dense horizontal sampling but lower vertical resolution. Combining coincident measurements from both observation geometries can help to produce ozone profiles that retain fine vertical structure while improving horizontal representation.

In this context, we combine limb observations from the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS), operating aboard ESA’s Envisat between 2002 and 2012 (Fischer and Oelhaf, 1996; Fischer et al., 2008), with nadir observations from the Infrared Atmospheric Sounding Interferometer (IASI), active since 2006 aboard EU-METSAT’s MetOp satellite series (Clerbaux et al., 2009; Razavi et al., 2009; Coheur et al., 2009). Both instruments operate in the mid-infrared (MIR) spectral range and provide valuable, complementary information on atmospheric composition. To optimally exploit their synergy, we apply the a posteriori Complete Data Fusion (CDF) methodology (Ceccherini et al., 2015), which has previously been used to merge nadir–nadir simulated measurements (Zoppetti et al., 2021; Tirelli et al., 2023), and limb–nadir simulated measurements (Tirelli et al., 2025), producing combined datasets with improved vertical information content. Here, we apply the latest version of the CDF method (Ceccherini et al., 2022) to real measurements.

In this work, we address two main objectives. First, we present the development and validation of a novel 2008–2011 ozone dataset created using the CDF methodology applied to MIPAS and IASI observations, and assess its quality using ozonesonde data. The main objective of this part is to evaluate and enhance the downward propagation of information from high-quality limb measurements in the stratosphere to the troposphere. In this lower region, MIPAS sensitivity decreases, retrieval uncertainties increase, or MIPAS data may be missing, whereas IASI provides more consistent coverage. This approach leverages the ability of MIPAS to resolve the UTLS (Upper Troposphere–Lower Stratosphere) region and the broader horizontal coverage and near-surface sensitivity of IASI. Notably, (Ceccherini et al., 2010) previously fused IASI and MIPAS data using the Measurement Space Solution (MSS) method (Ceccherini et al., 2009), which was later shown to be equivalent to the CDF method (Ceccherini, 2016), demonstrating the benefits of combining measurements from these two sensors.

The second objective is to use this dataset to investigate stratospheric ozone intrusion

events over the Himalayan region, considered a hotspot. Understanding these events is essential for characterizing their contribution to tropospheric ozone variability and for reducing current uncertainties in both observations and models. To achieve this, we lay the ground to develop an approach for detecting and characterizing stratospheric intrusion events, focusing on their occurrence frequency, depth, and vertical structure. The analysis exploits the improved vertical resolution of the fused dataset combined with a classification-based strategy for identifying ozone profiles associated with STE conditions.

The thesis is organized as follows. Chapter 1 presents the scientific background of this work, providing a brief overview of stratospheric ozone intrusion events and the fundamentals to understand CDF methodology. Then the manuscript is divided in two parts. The first part (Chapter 2) describes the development and validation of the novel ozone dataset, including the description of input datasets, methodology, and performance evaluation. The second part (Chapter 3) applies this dataset to the detection and characterization of stratospheric intrusions in the Himalayan region, always showing input data, methodology, and results. Finally, Chapter 4 brings together all the parts of the work with a brief summary, providing perspectives for future research.

Background

1.1 Atmospheric Ozone Overview

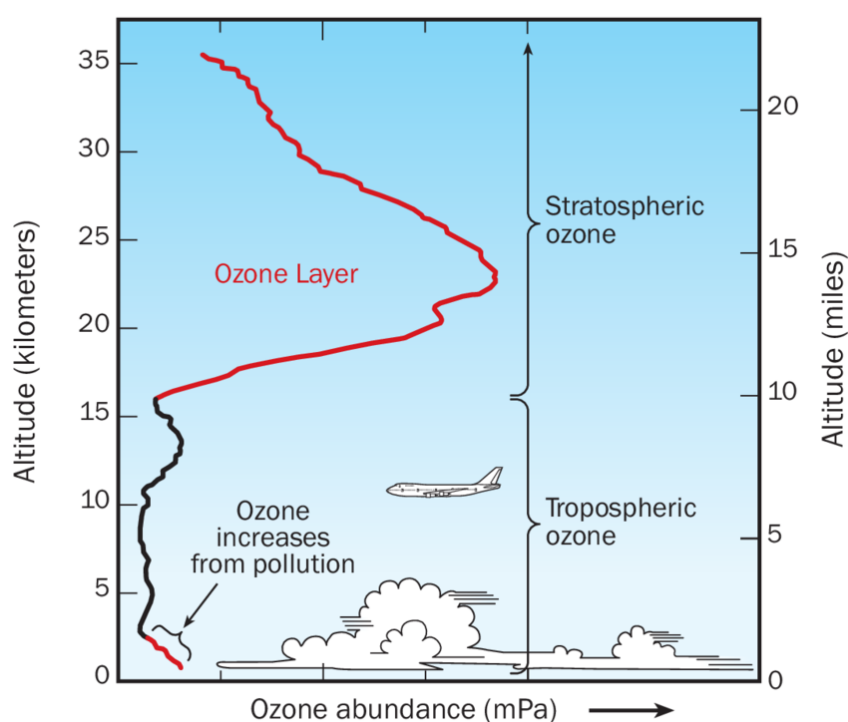


Figure 1.1: Atmospheric ozone vertical distribution (M. I. L. A. Hegglin et al., 2015). This profile shows schematically how ozone changes with altitude in the tropics. Most ozone resides in the stratospheric “ozone layer.” Increases in ozone near the surface are a result of air pollutants released by human activities.

Ozone is a gas that is naturally present in our atmosphere. Most ozone (about 90%) is found in the stratosphere, which begins approximately 10-15 kilometers above Earth's surface and extends up to about 50 km altitude. The region with the highest concentration of ozone, between 15 and 35 km, is commonly referred to as the *ozone layer* (Figure 1.1), which protects life on Earth by absorbing harmful ultraviolet radiation from the Sun (NASA, 2023a).

The ozone layer extends globally, although its altitude and thickness vary depending on latitude and season (US EPA, 2023). The remaining ozone (about 10%) resides in the troposphere, the lowest layer of the atmosphere, where it plays a role in air quality and climate (NOAA, 2018).

Ozone molecules are relatively scarce compared to other atmospheric gases like oxygen (O_2) and nitrogen (N_2). In the stratosphere near the peak of the ozone layer, there are typically a few thousand ozone molecules per billion air molecules. Near Earth's surface in the troposphere, ozone is even less abundant, typically ranging from 20 to 100 molecules per billion air molecules (NASA, 2023b).

Stratospheric ozone is formed naturally through a two-step photochemical process: ultraviolet sunlight breaks apart oxygen molecules (O_2) into individual atoms, which then react with other oxygen molecules to form ozone (O_3). This process occurs continuously in the presence of sunlight, with the highest production in the tropical stratosphere (NOAA, 2014). The production of stratospheric ozone is balanced by its destruction, which occurs through reactions with both naturally occurring and human-made chemicals, including nitrogen oxides, hydrogen oxides, and halogenated compounds like chlorine and bromine (US EPA, 2025). Some stratospheric ozone is transported downward into the troposphere, where it can influence ozone levels near the surface, especially in remote or mountainous regions (NOAA, 2023).

In the troposphere, ozone is produced by chemical reactions involving gases emitted from natural sources and human activities, such as hydrocarbons and nitrogen oxides. These reactions require sunlight and are most intense in polluted urban environments. Fossil fuel combustion is a major contributor to tropospheric ozone formation. Tropospheric ozone is also destroyed by chemical reactions and by deposition onto surfaces like soil and vegetation. Atmospheric circulation, through exchange processes of air masses between stratosphere and troposphere, can act as a natural source of tropospheric ozone, in addition to human activity contribution, as shown in the following section.

Ultimately, ozone levels in both the stratosphere and troposphere are governed by a dynamic balance between production and destruction processes, influenced by sunlight, temperature, altitude, and atmospheric circulation. This balance determines the global distribution of ozone and its variability over time (NOAA, 2022).

Ozone plays a dual role in the atmosphere: it acts as a greenhouse gas and a key agent in radiative forcing. While stratospheric ozone, as mentioned before, primarily shields the Earth from harmful ultraviolet radiation, ozone transported into the troposphere has a significant radiative effect, contributing to positive forcing and surface warming. Tropospheric ozone is the third most important greenhouse gas after CO_2 and CH_4 , with an estimated radiative forcing of $+0.4 \text{ W m}^{-2}$ (Forster et al., 2021). Its radiative significance

stems largely from its unique absorption characteristics across the ultraviolet (UV) spectrum. Indeed, ozone exhibits strong absorption in the ultraviolet region, particularly in three distinct bands:

- Hartley band (200–300 nm): the strongest absorption band, where ozone absorbs nearly all incoming UV-C radiation (100–280 nm), which is the most biologically harmful;
- Huggins band (310–350 nm): moderate absorption occurs here, affecting UV-B radiation (280–315 nm);
- Chappuis band (400–650 nm): weak absorption in the visible range, contributing slightly to ozone’s radiative properties.

Earth’s “Atmospheric Windows” to Solar Energy

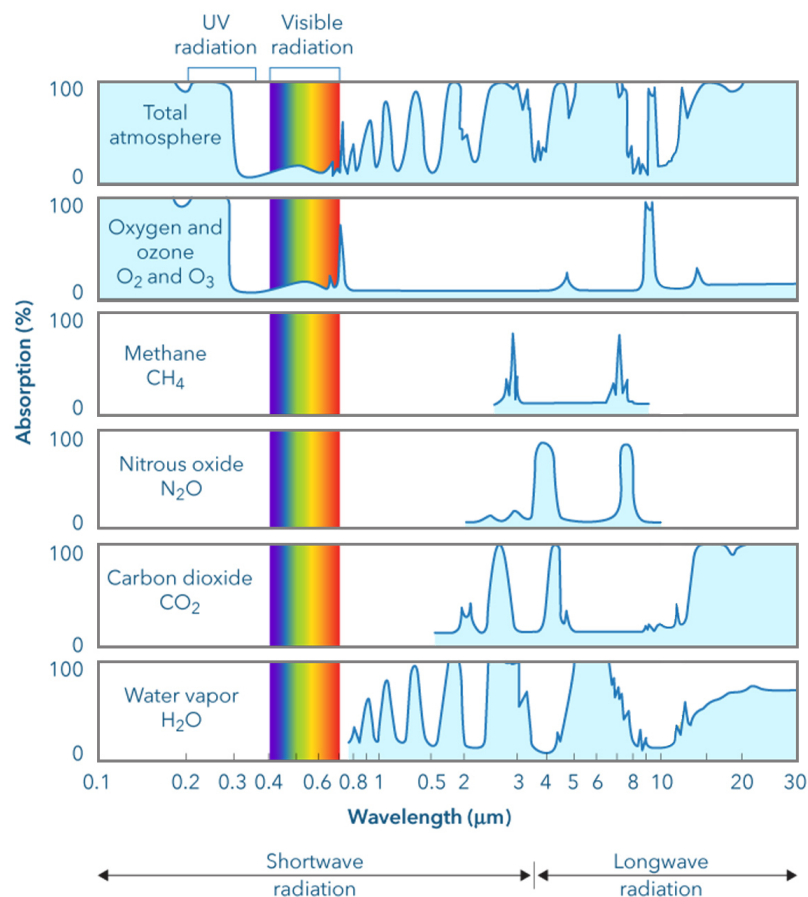


Figure 1.2: Earth absorption spectrum by atmospheric components, among them the ozone (second box). From NASA PACE.

Although ozone strongly absorbs UV radiation it also has distinct absorption features in the infrared, corresponding to the fundamental vibrational mode of the ozone molecule. The most important band in the IR, which is key for remote sensing in the thermal infrared, is the ν_3 one, around $9.6 \mu\text{m}$, followed by the ν_1 and ν_2 bands, around $7 \mu\text{m}$ and $14 \mu\text{m}$, respectively. Figure 1.2 shows the ozone absorption spectrum.

In addition to tropospheric ozone is a harmful pollutant. According to IPCC (Forster et al., 2021), it causes respiratory issues and aggravates asthma, damages vegetation and reduces crop yields, representing one of the major components of photochemical smog, particularly in summer.

1.2 Ozone Stratospheric Intrusions

1.2.1 Introduction to Stratospheric Intrusion Events

Stratospheric intrusion events are episodes during which stratospheric air, enriched in ozone and depleted in water vapor, penetrates into the troposphere through dynamic processes near the tropopause. This phenomenon is a major component of the broader stratosphere–troposphere exchange (STE), which governs the bidirectional transport of chemical species between the two atmospheric layers (Holton et al., 1995). These intrusions often occur via tropopause folding, frontal zones, and cut-off low systems, resulting in a downward transport of stratospheric ozone into the upper and middle troposphere (Stohl et al., 2003). A defining feature of intruding stratospheric air is its high ozone mixing ratio, typically between 100–500 ppbv, compared to 30–100 ppbv commonly observed in the upper troposphere under background conditions (Stohl et al., 2003; Mohanakumar, 2008). Additionally, stratospheric air masses are characterized by low relative humidity and high potential vorticity ($PV > 2$ PVU - see Appendix C about conserved quantities), making these properties key indicators for identification (Wernli and Bourqui, 2002, Mohanakumar, 2008).

The relevance of stratospheric intrusions stems from their contribution to the tropospheric ozone budget. Although photochemical production in the troposphere is the primary source of ozone in this layer, downward transport from the stratosphere represents a non-negligible component. According to Mohanakumar, 2008, ozone mixing ratios in stratospheric intrusions can reach 100–200 ppbv, compared to typical tropospheric values of 30–70 ppbv. This vertical and horizontal transport mechanism accounts for approximately 20–40% of the ozone observed in the free troposphere over certain regions (Lin et al., 2012). Global estimates suggest that stratosphere–troposphere exchange contributes between 300–600 Tg of ozone annually to the tropospheric ozone budget, with significant influence in the free troposphere — the portion of the troposphere above the planetary boundary layer, where atmospheric mixing is weaker and long-range transport processes dominate. (Holton and Lelieveld, 1996; Stevenson et al., 2006; Wild, 2007). These fluxes exhibit strong spatial and seasonal variability. Mid-latitude regions in both hemispheres experience the largest stratospheric contributions, particularly during spring when dynamic activity and tropopause folding frequency peak (Škerlak et al., 2014). In contrast, tropical regions display a relatively weak STE contribution due to the stronger stability of the tropopause and the dominant upward transport through deep convection. Quantifying STE remains challenging because of its episodic nature and complex mixing processes, which blur the distinction between stratospheric and tropospheric air masses. Despite these uncertainties, climatological studies consistently indicate that intrusions can locally enhance ozone concentrations in the upper troposphere influencing regional and global budgets.

Stratospheric intrusions are particularly relevant in the Upper Troposphere–Lower Stratosphere (UTLS) region, where small changes in ozone concentration can lead to substantial

alterations in radiative balance. By introducing stratospheric ozone into the troposphere, STE episodes can therefore enhance greenhouse forcing on synoptic to seasonal timescales, especially in mid-latitude regions where these events are frequent.

Beyond climate impacts, stratospheric intrusions can influence air quality at the surface, particularly when dynamical processes transport ozone-rich air masses downward through strong subsidence or mixing within high-pressure systems. Although the strongest signatures usually remain aloft, some intrusion events have been associated with near-surface ozone enhancements exceeding regulatory thresholds (Lin et al., 2012; Langford et al., 2009; Cristofanelli et al., 2010). These findings underscore the importance of accounting for STE in both climate and air quality assessments.

The frequency of stratospheric intrusions varies significantly across the globe, driven by meteorological dynamics and topographic influences. Mid-latitude regions exhibit the highest occurrence, typically associated with cyclonic systems and strong baroclinic activity. Climatological studies based on ERA-15 reanalysis indicate that cross-tropopause exchange occurs most frequently in the North Atlantic storm track and the subtropical jet region (Sprenger and Wernli, 2003). Topography also plays a key role (Schoeberl et al., 2019; Y. Zhang et al., 2024). Mountain ranges such as the Himalayas, the Alps, and the Rockies can enhance the likelihood of tropopause folding through orographic lifting and wave breaking. This makes regions like the Himalayas particularly relevant for detailed STE studies, as they combine strong dynamical forcing with complex terrain effects that promote stratosphere–troposphere exchange.

Figure 1.3 shows well all the concepts above.

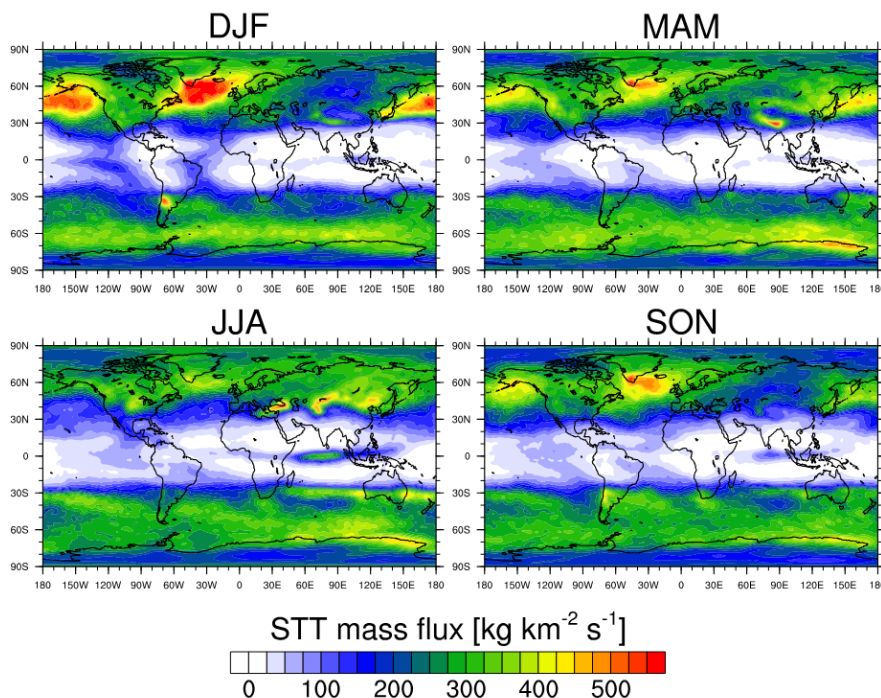


Figure 1.3: Seasonally averaged stratosphere to troposphere (STT) mass flux for 1979–2011 (DJF: December, January, February; MAM: March, April, May; JJA: June, July, August; SON: September, October, November) derived from trajectory analysis using ERA-Interim reanalysis data as input. From Škerlak et al., 2014.

In summary, the interplay between large-scale circulations such as the BDC (Brewer–Dobson Circulation, see Section 1.2.2) and localized features like tropopause folds governs the variability and intensity of STE events, thereby modulating their role in the global ozone budget.

The next section examines in detail the principal mechanisms responsible for episodic and localized ozone intrusions, starting from large-scale circulation and progressing to synoptic disturbances.

1.2.2 Dynamical Processes and Drivers

STE is primarily controlled by a combination of large-scale circulation systems and smaller-scale dynamical features that promote the downward transport of ozone-rich stratospheric air into the troposphere. These mechanisms operate across different spatial and temporal scales, from planetary waves modulating the mean circulation to synoptic disturbances that locally perturb the tropopause. Understanding these processes is fundamental to characterizing ozone intrusions and their contribution to the tropospheric ozone budget.

Before showing major exchange processes, it can be useful to understand the principal mechanism in which the tropopause plays a fundamental role. The tropopause is a dynamical and thermodynamical boundary that separates the well-mixed, convectively active troposphere from the more stable and stratified stratosphere. This boundary is characterized by a sharp increase in static stability, a strong gradient in PV, and typically acts as a quasi-impermeable surface under normal conditions. The PV, a conserved tracer under adiabatic and frictionless flow, provides a robust dynamical marker for distinguishing between air masses (see Appendix C). It is important to clarify that STE does not occur through vertical turbulent mixing, which is strongly inhibited by the stable stratification of the lower stratosphere. Instead, the apparent downward motion of stratospheric air arises from quasi-horizontal transport along isentropic surfaces, which can be dynamically depressed into the upper troposphere. Although the tropopause is dynamically stable, isentropic transport (constant potential temperature) enables STE through large-scale dynamical processes. During baroclinic wave activity and Rossby wave breaking, the jet stream undulates, leading to the folding of the tropopause. In such cases, isentropic surfaces (surfaces of constant potential temperature, see Appendix C) bend downward into the upper troposphere, allowing stratospheric air parcels to descend quasi-horizontally along these surfaces. These tropopause fold events occur frequently on the equatorward side of the jet stream, often downstream of cyclonic disturbances. The air brought down is typically dry, ozone-rich, and exhibits high PV values, distinct from the surrounding tropospheric air. In general the Brewer–Dobson circulation (BDC), a slow meridional overturning cell in the stratosphere, plays a crucial role in maintaining the global budget of STE. The BDC transports ozone-rich air poleward and downward on climatological timescales, establishing the background conditions for regions prone to intrusions. In the subtropics, diabatic descent associated with the BDC contributes to a steady, slow STE, in contrast to the rapid and localized transport during tropopause folds in the mid-latitudes. Notably, under climate change BDC is becoming faster (Garcia and Randel, 2008; Butchart, 2014). Figure 1.4 depicts the mechanisms of exchange.

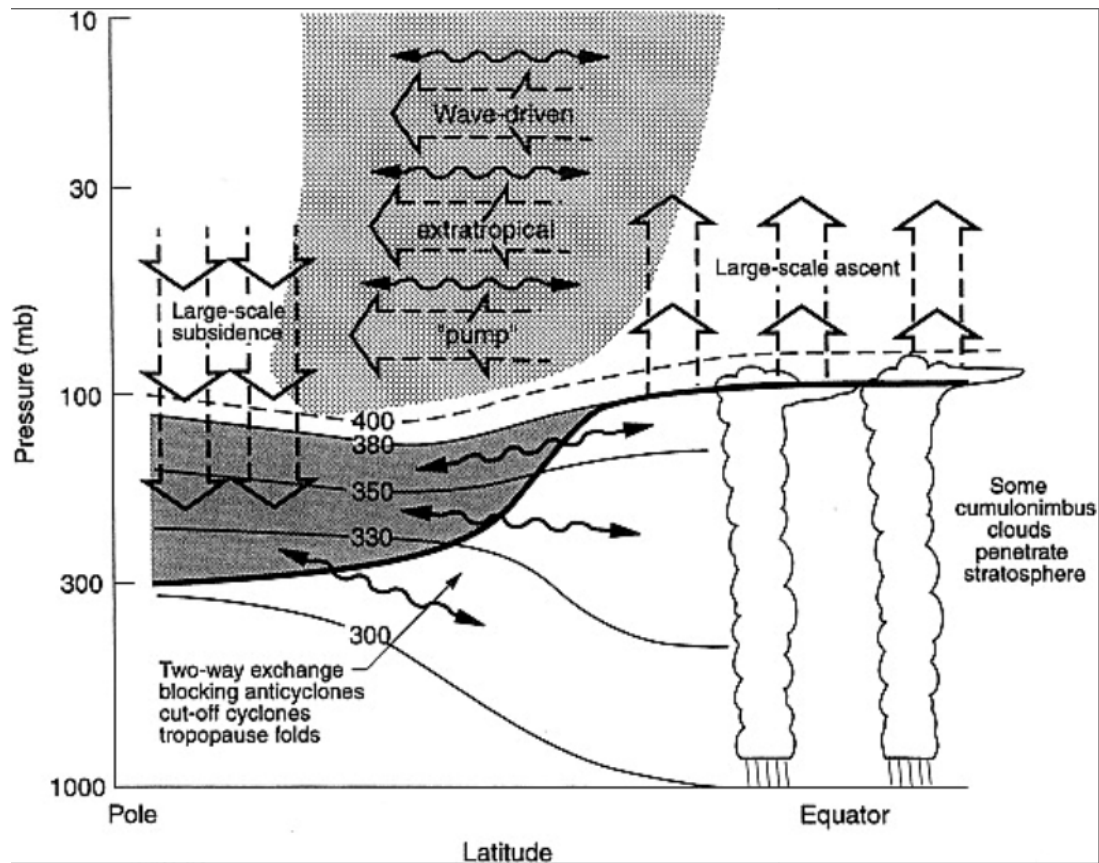


Figure 1.4: Dynamical aspects of stratosphere-troposphere exchange from Holton et al., 1995. Notably, according to present knowledge, tropical cumulonimbus clouds do not reach into the stratosphere (Stohl et al., 2003).

Once stratospheric air is intruded into the troposphere via folds or filaments, it often undergoes irreversible mixing through turbulent processes (instabilities, gravity waves breaking). These processes act on finer spatial and temporal scales than the synoptic dynamics and facilitate the final entrainment of ozone into the free troposphere.

Brewer-Dobson Circulation (BDC)

The Brewer–Dobson circulation represents the primary large-scale meridional overturning flow in the stratosphere, responsible for transporting air masses from the tropical tropopause upward and poleward before descending in mid- and high-latitude regions (Holton et al., 1995; Butchart, 2014). This circulation, driven by the breaking of planetary waves, sustains a gradual downward motion in the extratropical lower stratosphere, which in turn provides the background stratospheric ozone reservoir that can later be injected into the troposphere through dynamic disturbances. The BDC exhibits seasonal variability, being stronger in the winter hemisphere due to enhanced planetary wave activity. While the BDC alone rarely results in rapid mixing across the tropopause, it establishes the conditions for accumulation of ozone-rich air near the tropopause, particularly during late winter and spring when synoptic-scale processes become more active.

Quasi-Biennial Oscillation (QBO) and Wave–Mean Flow Interactions

The QBO is a dominant mode of variability in the equatorial stratosphere, characterized by alternating easterly and westerly wind regimes with a period of around 2 years. While its direct influence on STE is limited to equatorial regions, the QBO modulates planetary wave propagation and thus affects the strength of the BDC. During the westerly phase, planetary waves tend to propagate more poleward, enhancing extratropical descent and indirectly influencing the frequency of intrusions (Baldwin et al., 2001).

Tropopause Folding

Tropopause folding is the most prominent mechanism leading to direct and rapid stratospheric intrusions into the troposphere. A tropopause fold occurs when the dynamical tropopause (typically defined as the 2 PVU surface) becomes highly deformed due to differential advection by jet streams, creating an elongated filament of stratospheric air that penetrates the upper troposphere (Sprenger and Wernli, 2003). Folds typically extend downward from the mid-latitude tropopause into the upper troposphere and can occasionally reach altitudes as low as 700 hPa, particularly in intense baroclinic disturbances (Škerlak et al., 2014). Tropopause folding is most frequent in the mid-latitude storm tracks and downstream of jet streak exit regions, with maximum occurrence during spring and early summer in both hemispheres. Climatological analyses indicate a peak frequency over the North Atlantic and North Pacific basins, associated with Rossby wave breaking (Sprenger and Wernli, 2003). The most recognized mechanism for stratospheric intrusions is tropopause folding, which occurs when the dynamical tropopause (often defined by the 2 PVU surface) becomes distorted due to strong baroclinic activity near frontal zones (Danielsen, 1968; Shapiro, 1980). During cyclogenesis, the jet stream often exhibits sharp undulations that cause the stratospheric air to wrap around the upper-level trough, folding downward into the troposphere.

Role of Jet Streams and Rossby Wave Breaking

Rossby wave breaking (RWB) is a major precursor of tropopause folding. It occurs when large-amplitude planetary waves overturn and break, producing irreversible PV mixing and strong anomalies on the tropopause. RWB enhances the horizontal and vertical deformation of the tropopause and creates favorable conditions for folds. Anticyclonic breaking tends to transport stratospheric air equatorward. Cyclonic breaking usually results in poleward excursions of tropospheric air into the lower stratosphere.

Jet streaks, localized maxima in jet stream wind speed, further contribute by inducing strong ageostrophic circulations that promote stratospheric air descent on the equatorward side of the jet. These dynamical features typically occur over ocean basins and mountainous regions where baroclinic instability is active. Jet streams act as the primary dynamical boundary between stratospheric and tropospheric air masses. The polar and subtropical jets are regions of strong horizontal wind shear and high PV gradients, making them preferred sites for STE. Variations in jet intensity and position, often associated with Rossby wave breaking, can trigger large-scale intrusions.

Cut-off Lows

Cut-off lows represent another important synoptic-scale feature that promotes STE. These quasi-stationary cyclonic vortices, often detached from the main polar front jet, are characterized by cold-core structures and a lowered tropopause, which facilitates the mixing of stratospheric air into the troposphere. They are frequently observed in subtropical and Mediterranean regions, contributing to deep intrusions at lower latitudes (Nieto et al., 2008).

Although the dynamical mechanisms driving STE are well understood, quantifying the relative contribution of each process remains challenging. Climatological studies agree that tropopause folds and Rossby wave breaking represent the dominant pathway for rapid and episodic STE in the mid-latitudes, accounting for the majority of short-term downward ozone fluxes (e.g., Stohl et al., 2003; Škerlak et al., 2014). Cut-off lows also contribute significantly, especially in the subtropics and Mediterranean region, where they can generate deep intrusions reaching the middle troposphere (Nieto et al., 2008). In contrast, the Brewer–Dobson circulation provides a slow and continuous background descent, controlling the climatological supply of ozone-rich air near the extratropical tropopause (Butchart, 2014). Its role is not to create rapid intrusions, but to sustain the reservoir that synoptic disturbances later inject into the troposphere.

Overall, the literature indicates that synoptic-scale processes dominate short-term fluxes, while large-scale circulation governs long-term budgets. However, the exact partitioning among these mechanisms is still uncertain due to differences in tropopause definitions, model resolution, and tracer-based diagnostic methods.

The next section will be dedicated to existing intrusions detection and analysis techniques.

1.2.3 Detection and Analysis Techniques

A wide range of observational platforms and diagnostic methods have been developed to monitor intrusion processes, combining chemical tracers, dynamical indicators, and modeling approaches. This section provides a comprehensive overview of the main observation techniques and methodologies employed to identify and study these events, including in situ and remote sensing measurements, detection strategies, and the diagnostic use of chemical tracers.

In-Situ and Ground-Based Remote Sensing Observations

Ozonesondes are balloon borne instruments widely employed to monitor vertical ozone distribution and detect stratosphere to troposphere intrusions of ozone. They provide high vertical resolution (about 100-150 m) and simultaneous measurements of ozone, temperature, pressure, and relative humidity, making them particularly suitable for capturing the fine scale structure of ozone rich intrusions (Logan, 1999; Stohl et al., 2003). Ozone concentrations are typically reported in mixing ratios (ppbv) or partial pressure, obtained through electrochemical concentration cell (ECC) sensors during the balloon ascent up to altitudes of 30–35 km. Intrusions are identified in ozonesonde profiles by characteristic signatures: enhanced ozone mixing ratios (commonly >80 ppbv) in the upper or middle troposphere, coinciding with sharp decreases in relative humidity (typically $<20\%$). These dry, ozone-rich layers indicate the presence of stratospheric air that has crossed

the tropopause (Beekmann et al., 1997; Logan, 1999). To confirm the stratospheric origin, ozonesonde data are frequently combined with additional diagnostics such as PV analyses, tropopause height calculations, and Lagrangian back-trajectory models (e.g., FLEXPART, LAGRANTO), which trace the air mass history and verify its stratospheric source (Stohl et al., 2003; Cui et al., 2009; Sprenger and Wernli, 2015; Chen et al., 2021). Long term ozonesonde networks, including the World Ozone and Ultraviolet Radiation Data Center (WOUDC) and the Southern Hemisphere Additional Ozonesondes (SHADOZ) program, have been extensively used to quantify the frequency, intensity, and seasonal variability of ozone intrusions across different regions (Thompson et al., 2003; Tarasick et al., 2019). For instance, Beekmann et al., 1997 documented strong ozone enhancements linked to tropopause folds over Europe, while Stohl et al., 2003 analyzed ozonesonde data in combination with trajectory models to identify intrusion pathways and their role in tropospheric ozone budgets. These measurements remain critical for validating satellite observations and chemical transport models, offering high resolution ground truth for stratospheric air mass detection.

In addition, also LIDAR systems and ground-based spectrometers can be useful. Ground-based LIDAR (Light Detection and Ranging) systems complement ozonesonde observations by providing quasi-continuous measurements of ozone profiles, often from about 2 km to 15–20 km above sea level, depending on the system. Differential absorption lidar (DIAL) instruments are capable of detecting ozone with high vertical resolution (100–300 m), allowing detailed observation of the downward propagation of ozone layers associated with intrusions (Di Girolamo et al., 2009; Gronoff et al., 2021). Lidar systems are often deployed at mountain stations (e.g., Jungfraujoch in the Swiss Alps; Larchevêque et al., 2002), where stratospheric intrusions frequently occur due to the influence of jet streams and tropopause folds, but key is the fact that the high elevation makes the identification and occurrence more frequent. Despite their high accuracy and temporal resolution, lidar observations are geographically limited and require clear-sky conditions, making them unsuitable for global-scale studies.

Ultraviolet (UV) and infrared (IR) spectrometers, including Fourier Transform Infrared (FTIR) instruments, provide additional means of monitoring ozone in the lower stratosphere and upper troposphere. While they offer good sensitivity and can retrieve partial column information, their ability to resolve fine vertical structures associated with intrusions is limited compared to sondes and lidar.

Satellite Remote Sensing Techniques

Satellite-based remote sensing plays a fundamental role in the observation and characterization of stratospheric intrusions and related ozone transport processes. Unlike ground-based measurements, which are limited in spatial coverage, satellite instruments provide global and continuous observations, enabling the study of ozone dynamics over large regions and remote areas.

Satellite ozone retrievals are obtained primarily from infrared (IR) and ultraviolet (UV) spectrometers. Among currently operating sensors, IASI (Infrared Atmospheric Sounding Interferometer), OMI (Ozone Monitoring Instrument), and TROPOMI (TROPOspheric Monitoring Instrument) provide long and consistent records for investigating ozone variability in the UTLS. Other instruments, such as MIPAS (Michelson Interferometer for Passive Atmospheric Sounding), were active in the past and remain valuable for historical

analyses and climatological assessments. IR instruments like MIPAS and IASI offer - together with complementary measurements thanks to the different observation geometries (limb and nadir, respectively) - good vertical sensitivity in the UTLS by exploiting thermal emission, whereas UV instruments such as OMI and TROPOMI rely on backscattered solar radiation and are primarily sensitive to the stratosphere, with more limited vertical resolution in the troposphere. Several studies have applied these datasets to identify and characterize stratospheric intrusions. For example, Eremenko et al., 2008 used IASI observations to detect ozone enhancements in the UTLS region associated with tropopause folding. Gaudel et al., 2018 provided a global assessment of tropospheric ozone trends and variability using multi-satellite products, highlighting the contribution of STE to observed patterns. These works demonstrate the ability of satellite products not only to detect individual events but also to provide long-term climatologies for STE activity.

One of the major strengths of satellite remote sensing is its ability to detect large-scale patterns and seasonal variability of intrusions and quantify their contribution to the tropospheric ozone budget at a global scale. For example, climatologies derived from satellite data have been widely used to investigate the frequency of tropopause folds and the associated ozone enhancements across mid-latitude storm tracks (Stohl et al., 2003; Škerlak et al., 2014; Gaudel et al., 2018). Satellite observations have also been crucial in assessing interannual variability linked to large-scale circulation anomalies, such as ENSO or the North Atlantic Oscillation (Monks et al., 2015).

Despite these advantages, satellite observations have limitations that must be considered. Vertical resolution in the troposphere is typically coarse (5–10 km), which makes it challenging to resolve the fine-scale structure of intrusion layers observed by ozonesondes or aircraft. Furthermore, retrieval sensitivity decreases near the surface and can be affected by clouds and aerosols. For this reason, satellite-based studies often rely on combined approaches, such as integrating remote sensing data with in situ measurements and chemical transport models (CTMs) or Lagrangian models to improve vertical detail and attribution of stratospheric air masses. The integrated approach of these tools with satellite and in situ observations compensates for the limitations of individual methods and represents a robust strategy for understanding ozone transport processes in the UTLS region.

In this study, we make use of satellite data only, rather than combining different observational systems or model-based analyses. Specifically, we rely on Level-2 ozone products from IASI (nadir-viewing thermal infrared) and MIPAS (limb-emission infrared), which provide complementary information on ozone variability in the UTLS. These datasets allow us to characterize the vertical structure of ozone intrusions and their temporal evolution without performing retrievals or data assimilation. The techniques reviewed above therefore serve as a general context for stratosphere–troposphere exchange studies, while the methodology adopted here is based exclusively on the analysis of satellite-derived ozone profiles and columns.

Model-Based Approaches and Reanalyses

Numerical models and reanalysis products are essential tools for investigating stratospheric intrusions, as they provide dynamically consistent three-dimensional fields of ozone, meteorological variables, and transport tracers on a global scale. Unlike observations alone, these datasets enable the reconstruction of the temporal evolution of STE events, their spatial extent, and their impact on tropospheric ozone budgets.

Modern reanalyses assimilate a wide range of observations into global models, combining

information from satellites, radiosondes, and ground-based networks to provide physically consistent fields. The ERA5 reanalysis (Hersbach et al., 2020) and MERRA-2 (Gelaro et al., 2017) are widely used for STE studies, particularly for diagnosing tropopause folds and quantifying stratospheric ozone intrusions in the UTLS (Sprenger and Wernli, 2015; Greenslade et al., 2017; M. Wang and Fu, 2021). The Copernicus Atmosphere Monitoring Service (CAMS) provides operational reanalyses and forecasts of atmospheric composition, including ozone, which have been applied to monitor STE activity and evaluate its contribution to surface ozone variability (Inness et al., 2019). CAMS reanalysis products offer the advantage of combining chemical data assimilation with a global chemistry transport model, allowing for improved ozone distribution in the UTLS and better attribution of stratospheric sources.

In addition to reanalyses, global chemistry–climate models (CCMs) and chemical transport models (CTMs) provide a unique framework for sensitivity experiments and process attribution. Specifically, CTMs are Eulerian models that simulate the evolution of atmospheric constituents on a fixed spatial grid, providing a detailed description of chemistry and transport processes. The EMAC model (ECHAM/MESSy Atmospheric Chemistry -), for instance, or GEOS-Chem from NASA, have been extensively used to quantify the impact of stratospheric intrusions on tropospheric ozone budgets and their seasonal and interannual variability (Hofmann et al., 2016; Akritidis et al., 2016; Hua et al., 2024). Such models can separate contributions from dynamics and chemistry and simulate long-term trends under varying climate and emission scenarios. Advanced implementations also include ozone tagging techniques and stratospheric tracers, which allow the quantification of the relative contributions of STE, long-range transport, and in situ photochemical production.

Furthermore, Lagrangian transport models, such as FLEXPART or LAGRANTO (Cui et al., 2009) or CLaMS (McKenna et al., 2002), complement Eulerian approaches by tracking air parcels along trajectories, and resulting particularly effective in identifying transport pathways and diagnosing the origin of stratospheric intrusions.

However, the combination of reanalysis products, global models, and satellite observations to build STE climatologies, identify regional hotspots, and assess impacts on tropospheric composition and surface air quality remains widely used. These integrated approaches are essential for advancing our understanding of the processes governing STE and its response to climate variability.

Diagnostic Tracers and Indicators

The identification and analysis of stratospheric ozone intrusions rely on a suite of diagnostic tracers and advanced detection algorithms that integrate dynamical and chemical indicators. Among the most robust tracers, PV serves as a key dynamical marker distinguishing stratospheric air, which typically exhibits high PV values (above 1.5–2 PVU - see Appendix C), from tropospheric air. This separation is crucial because stratospheric intrusions often manifest as filaments or folds of high-PV air penetrating into the UTLS region. Complementing PV, specific humidity acts as a valuable tracer since stratospheric air is characteristically much drier than its tropospheric counterpart, providing a humidity-based discriminator. Furthermore, correlations between ozone and carbon monoxide (CO) are commonly exploited: stratospheric intrusions are generally rich in ozone but low in CO, contrasting with polluted tropospheric air, which tends to show elevated CO and comparatively lower ozone concentrations. Isentropic analysis, which

tracks air parcels along surfaces of constant potential temperature, provides insightful perspectives on cross-tropopause transport pathways, allowing researchers to identify the timing and pathways of stratosphere-to-troposphere exchange events. In addition to the commonly used dynamical and chemical tracers, cosmogenic isotopes such as the ones from beryllium (Be), and its nuclides, such as Be-7 and Be-10 (Koch and Rind, 1998; Jordan et al., 2003; Brattich et al., 2017; Ajtić et al., 2017; X. Liu et al., 2024), serve as valuable indicators of stratospheric air masses. These isotopes are produced in the upper atmosphere by spallation reactions between cosmic rays and atmospheric molecules, atoms and nuclei and have relatively short half lives, making them effective tracers for recent stratosphere-to-troposphere exchange events. Elevated atmospheric concentrations of the cosmogenic nuclides Be-7 and Be-10 serve as sensitive tracers of stratosphere-to-troposphere transport. Quantified from aerosol filters—Be-7 by gamma spectrometry and Be-10 by AMS—their peaks enable the timing of stratospheric intrusions and the analysis of vertical mixing to be determined. Such isotopic measurements, typically obtained from ground-based aerosol sampling or aircraft campaigns, complement the dynamical and chemical tracers by providing independent evidence of stratospheric influence, especially in regions or periods where direct chemical observations are sparse.

Detection algorithms synthesize these diagnostics to effectively identify intrusion events. Threshold methods often apply combined cutoffs in PV and ozone concentration - generally 2 PVU and 0.1 ppmv, respectively - from model or reanalysis fields to objectively delineate stratospheric air masses. Meanwhile, Lagrangian trajectory models offer backward or forward tracking of air parcels, confirming the stratospheric origin of observed ozone enhancements by reconstructing their transport history. These models help to differentiate genuine stratospheric intrusion events from local photochemical ozone production or pollution transport. In addition, composite analyzes and detailed case studies leverage multiple observational platforms - including satellite remote sensing, ozonesondes, and ground-based instruments - along with model output to validate and characterize intrusion events across spatial and temporal scales. This approach improves confidence in event detection, elucidates the underlying transport dynamics, and advances understanding of the role of stratospheric intrusions in shaping tropospheric ozone variability.

1.2.4 A Review of Stratospheric Ozone Intrusions over the Himalayas

Orography and Meteorology

The Himalayan region (Figure 1.5), stretching over 2400 km across the northern edge of the Indian subcontinent, represents one of the most prominent orographic features on Earth. This mountain range, characterized by towering peaks exceeding 7000 meters above sea level, acts as a geographical barrier that separates the tropical and subtropical climatic zones of the Indian subcontinent from the mid-latitude continental air masses to the north and northwest. The region's complex topography and varied elevation gradients critically influence the local and regional meteorology, creating a dynamic environment with distinctive atmospheric circulation patterns.

Meteorologically, the Himalayan region is subject to the pronounced influence of the South Asian monsoon system, a large-scale seasonal circulation characterized by strong



Figure 1.5: Himalayan region from Google Earth.

moisture-laden southwesterly winds during the summer months and dry northeasterly winds in winter. The monsoon imposes intense convective activity, leading to substantial precipitation and profound vertical motions in the atmosphere (Gadgil, 2003). This convective environment fosters turbulent mixing and enhances the vertical transport of trace gases and aerosols. During the winter and pre-monsoon periods, the region experiences relatively stable stratification but is influenced by mid-latitude westerly disturbances and the presence of high-pressure systems over Central Asia. These systems modulate synoptic-scale dynamics and can trigger wave activity, including planetary-scale Rossby waves and smaller-scale gravity waves. Additionally, due to the steep and elevated terrain, the Himalayas serve as an active region for stratosphere to troposphere exchange processes, which are driven by a complex interplay of orographically forced atmospheric waves, synoptic scale disturbances, and seasonal meteorological variability. The mountainous topography induces strong vertical motions and gravity wave breaking, which can lead to localized thinning of the tropopause and facilitate the downward intrusion of stratospheric air enriched in ozone and other trace gases into the upper troposphere.

The tropopause, defined as the boundary separating the troposphere and stratosphere, exhibits significant spatiotemporal variability over the Himalayan region. Studies have documented a pronounced lowering of the tropopause during spring and pre-monsoon seasons (Srivastava et al., 2002; D. Li et al., 2018), coinciding with enhanced wave activity and increased dynamical forcing. This lowering creates conditions conducive to enhanced STE events, during which stratospheric ozone can penetrate deeper into the troposphere. Intrusions are particularly impacting in South Asia, where there is a high level of aerosol and pollutant gases. Over the Himalayas, particularly, a knowledge gap on concerning

the typical levels of O_3 still exists, and contribution of intrusions, besides influencing the mountain ecosystem air quality, can strongly contribute to the atmospheric heating of absorbing aerosols (Ramanathan and Carmichael, 2008).

Evidence of Stratospheric Intrusions

Mountainous regions are particularly suitable for monitoring background tropospheric ozone due to their unique geographic and atmospheric characteristics. High-altitude stations are typically located far from major anthropogenic emission sources, thereby minimizing the influence of local pollution and allowing measurements to be more representative of large-scale background conditions. Furthermore, their elevation often places them within or near the free troposphere, above the diurnal boundary layer, which reduces variability associated with local surface processes. This positioning also enhances the sensitivity to long range transport and STE, both of which are significant contributors to background ozone levels Henne et al., 2008. Thanks to these factors, several studies focused on the Himalayan region and over the Tibetan Plateau attesting the contribution of stratospheric intrusions. T. Wang et al., 2006 detected signal of stratospheric air masses by analysing ozone and carbon monoxide recorded during the north-eastern Qinghai-Tibetan Plateau campaign. Here, also Ding and Wang, 2006 detected stratospheric intrusions. The role of the orography in this region has been investigated looking at the Mt Everest. Several studies (Moore and Semple, 2004; Zhu et al., 2006) showed how the Plateau interact with large scale circulation favoring stratospheric intrusions, which at the same time, are able to trigger extreme weather events. They highlighted how katabatic winds act to pump down ozone from upper air levels.

More recent studies confirmed the importance of stratospheric ozone intrusions as a significant natural contribution to background tropospheric ozone. Cristofanelli et al., 2009 analyzed surface ozone measurements from the WMO GAW¹ (Global Atmosphere Watch programme of the World Meteorological Organization) station Nepal Climate Observatory Pyramid (NCO-P) founding that during 25 intrusion events days the daily ozone mixing ratio increased by 9.3% compared to seasonal values. Always Cristofanelli et al., 2010 detected up to a 25% contribution of stratospheric intrusions to surface tropospheric ozone across the Himalaya, highlighting a pronounced seasonal cycle with mean ozone concentrations of 61 ppbv in the pre-monsoon period and 39 ppbv during the monsoon season, and documenting how deep STE events can dominate the signal during certain periods. Similarly, satellite observations (MLS, TES) and MOZART-2 model simulations by Fadnavis et al. (2010) revealed recurrent ozone intrusions (100–200 ppbv) into the upper troposphere during winter and pre-monsoon seasons, with variations in PV and tropopause height across the Indian regional scale. Finally also Yin et al., 2023 synthesized surface measurements, reanalysis data, and chemical models (WRF-Chem, CAM-chem), identifying the Tibetan Plateau as a global hotspot for stratospheric intrusions, with direct impacts on surface ozone during summer and correlations with tropopause folding events.

¹The GAW programme is a global initiative coordinated by the WMO to monitor the chemical composition of the atmosphere and support climate and air quality research. It includes over 40 Global stations, more than 400 Regional stations, and around 100 Contributing stations worldwide. These stations provide high-quality data on ozone, greenhouse gases, aerosols, and other key atmospheric constituents.

Resuming, stratospheric intrusions over the Himalayas are primarily driven by large-scale dynamical processes. The most frequently reported mechanisms include Rossby wave breaking and associated tropopause folding, which create pathways for stratospheric air to descend into the upper troposphere (Sprenger and Wernli, 2003; Stohl et al., 2003). In addition, western disturbances, which are synoptic-scale systems affecting the region during winter and pre-monsoon seasons, contribute significantly to dynamical instabilities and tropopause variability (Fadnavis et al., 2010). At smaller scales, orographically induced gravity waves generated by the steep Himalayan terrain can enhance vertical motions and thin the local tropopause, facilitating deep intrusions. These processes show a marked seasonal pattern: intrusions occur more frequently in winter and pre-monsoon when the subtropical jet is strongest and the tropopause is lower, while they are less frequent during the monsoon season due to strong convective activity, which dominates vertical transport and reduces the stability of folding structures (Cristofanelli et al., 2010; Yin et al., 2023).

Knowledge Gap and Actual Limitations

Despite considerable progress in understanding stratosphere-to-troposphere exchange (STE), significant methodological and observational limitations persist for the Himalayan region. In-situ observations at high altitudes are extremely scarce due to the logistical challenges of operating instruments above 4000–5000 m, and balloon-borne ozonesondes, which offer high vertical resolution, are deployed only sporadically and mainly at a few stations (Cristofanelli et al., 2010). Satellite sensors, while providing broad spatial coverage, are limited by their coarse vertical resolution, making it difficult to resolve thin tropopause folds or small-scale intrusions (Fadnavis et al.; Yin et al., 2010, 2023). On the modeling side, reanalysis datasets and global chemistry-climate models underestimate orographic effects and fail to fully capture gravity-wave activity and associated fine-scale tropopause dynamics, primarily due to grid resolution constraints (Sprenger and Wernli, 2003). Furthermore, disentangling stratospheric signals from tropospheric contributions remains challenging because commonly used diagnostic tools (e.g., PV thresholds, tracer-tracer correlations) are sensitive to the choice of parameters and suffer from large uncertainties in this complex orographic setting (Stohl et al., 2003). These limitations hinder accurate quantification of intrusion frequency, depth, and chemical impact, leaving the role of STE in regional ozone variability and climate forcing highly uncertain.

1.3 Radiative Transfer Theory

The Radiative Transfer theory (RT) describes how electromagnetic radiation propagates through a medium while undergoing absorption, emission, and scattering by atmospheric gases, aerosols, and clouds. RT theory forms the physical foundation of remote sensing, providing the Forward Model—a mathematical framework that connects atmospheric states to measured satellite radiances. Its Inverse Problem aims to retrieve these atmospheric variables from the data. While these concepts will be explored later, their core definitions are introduced here for clarity. First, however, some theoretical background is needed to understand the primary product of a satellite instrument.

1.3.1 Radiance and Spectral Flux

The spectral radiance is defined as the energy emitted per unit area, per unit time, per unit solid angle, and per unit wavenumber or wavelength. For a blackbody at thermodynamic equilibrium, the spectral radiance is described by the Planck law, which can be expressed either as a function of frequency ν or wavelength λ :

$$\begin{aligned} B_\nu(T) &= \frac{2h\nu^3}{c^2} \cdot \frac{1}{\exp\left(\frac{h\nu}{k_B T}\right) - 1}, \\ B_\lambda(T) &= \frac{2hc^2}{\lambda^5} \cdot \frac{1}{\exp\left(\frac{hc}{\lambda k_B T}\right) - 1}. \end{aligned} \quad (1.1)$$

where $h = 6.62 \cdot 10^{-34} Js$ is the Planck constant, $c = 2.99 \cdot 10^8 m/s$ is the light speed, $k = 1.3110^{-23} J/K$ is the Boltzmann constant, and T is the emitting body temperature. The spectral radiance L_ν emitted by a body of uniform temperature T at a wavenumber ν is described by the product of the black body radiance B and the emissivity ϵ :

$$L_\nu = B_\nu \epsilon_\nu. \quad (1.2)$$

The radiation emitted by a black body can be converted into another useful quantity, the Brightness Temperature (BT), defined as the temperature of the body emitting the radiance L_ν at the wavenumber ν . The BT (i.e T_ν in the following equation) can be obtained by inverting 1.1:

$$T_\nu = \frac{hc\nu}{k \ln\left(1 + \frac{2hc^2\nu^3}{L_\nu}\right)}. \quad (1.3)$$

From the spectral radiance, it is possible to compute the spectral flux F_ν , defined as the integral of the radiance over a hemisphere of solid angle:

$$F_\nu = \int_{\Omega} L_\nu(\theta, \phi) \cos\theta d\Omega, \quad (1.4)$$

where $d\Omega$ is the infinitesimal element of the solid angle Ω , the hemispheric domain of integration. In spherical coordinates, it is represented as $d\Omega = \sin(\theta)d\theta d\phi$, where θ is the zenith angle and ϕ is the azimuth angle:

$$F_\nu = \int_0^{2\pi} d\phi \int_0^{\frac{\pi}{2}} d\theta L_\nu(\theta, \phi) \cos\theta \sin\theta. \quad (1.5)$$

As a consequence, if the radiance field is isotropic, the spectral flux can be simply obtained by the following expression:

$$F_\nu = \pi L_\nu. \quad (1.6)$$

Most of the solar energy is confined in the spectral range between 0.2 and 4 μm and is termed short-wave (SW) radiation, including near-infrared, visible and ultra-violet spectral ranges. Conversely, the Earth's emitted energy falls between approximately 3 and 100 μm (from about 100 to 3000 cm^{-1}) and is known as Long-Wave (LW) radiation (Far-Infrared and Mid-Infrared spectral ranges).

The total flux can be obtained by integrating the spectral flux over frequency and, under the assumption of isotropic radiation, the total flux of a black body computed over the spectral range from 0 to ∞ is equal to:

$$F = \sigma T^4. \quad (1.7)$$

where $\sigma = 5.67 \cdot 10^{-8} \text{ W m}^{-2} \text{ K}^4$. The formula, known as Stefan-Boltzmann law, demonstrates that the energy per unit time and unit area emitted by a black body depends on the fourth power of the temperature.

1.3.2 Radiative Transfer Equation

A beam of radiation propagating through a medium undergoes three fundamental processes: absorption, emission, and scattering. Neglecting scattering and emission, the reduction of a pencil of radiation follows Lambert's law:

$$\frac{dL_\nu}{L_\nu} = -\rho k_{a,\lambda} ds, \quad (1.8)$$

where ρ is the mass density of the medium, s is the distance travelled by the beam, and $k_{a,\lambda}$ is the absorption cross section (area per mass). Equation 1.8 may also be expressed replacing the product $\rho k_{a,\lambda}$ with $n\tilde{k}_{a,\lambda}$, where n is the particle number density and $\tilde{k}_{a,\lambda}$ represents the absorbing area of a single particle. Integrating Eq. 1.8 along the propagation path yields:

$$L_\nu(s) = L_\nu(0) e^{-\int_0^s \rho k_{a,\lambda} ds'}. \quad (1.9)$$

It is convenient to introduce the optical depth:

$$\tau_\nu(s) = \int_0^s \rho k_{a,\lambda} ds', \quad (1.10)$$

which quantifies the attenuation of radiation along the path. With this definition, Lambert's law becomes:

$$L_\nu(s) = L_\nu(0) e^{-\tau_\nu(s)}. \quad (1.11)$$

A related quantity is the transmissivity $T_\nu(s)$, defined as the ratio between the emerging and incident radiance:

$$T_\nu(s) = \frac{L_\nu(s)}{L_\nu(0)} = e^{-\tau_\nu(s)}. \quad (1.12)$$

We invoke the Local Thermodynamic Equilibrium (LTE) assumption in the following derivation to simplify the radiative transfer equation. Under LTE, the source function can be replaced by the Planck function, which yields a physically intuitive description of atmospheric emission and absorption. It is crucial to note that this is a didactic choice. Infrared sounders such as MIPAS (limb) or IASI (nadir) can operate in regions where non-LTE effects are significant (typically above 60 km). However, the operational Level 2 products utilized in this work are derived from retrieval algorithms that already incorporate non-LTE physics where required. Therefore, the LTE framework is presented solely to establish a compact theoretical foundation for the reader, not as an assumption underlying our analysis.

In radiative equilibrium a light-absorbing body must also emit radiation. Under LTE conditions a substance emits and absorbs with the same efficiency, so that the absorptivity

a_ν equals the emissivity ϵ_ν , according to Kirchhoff's law. LTE requires that collisional processes dominate over radiative ones, ensuring that the local temperature characterizes the energy level populations. This condition is typically met in the troposphere and stratosphere (below ~ 60 km). From Kirchhoff's law, the differential change in spectral radiance satisfies:

$$\frac{dL_\nu}{B_\nu(T)} = d\epsilon_\nu = da_\nu, \quad (1.13)$$

where $B_\nu(T)$ is the Planck function. In LTE and in the absence of scattering, the source function J_ν coincides with $B_\nu(T)$, so that

$$dL_\nu = B_\nu(T) \rho k_{a,\lambda} ds. \quad (1.14)$$

Combining absorption and emission along a single propagation direction $\mathbf{n}$, the differential form of the radiative transfer equation becomes:

$$\frac{dL_\nu}{ds} = -\rho k_{a,\lambda} L_\nu + \rho k_{a,\lambda} B_\nu(T). \quad (1.15)$$

For the infrared spectral region relevant to this study, absorption processes are predominant, while scattering by molecules and aerosols is negligible. Consequently, our radiative transfer formulation neglects scattering contributions.

Using the optical depth, Eq. 1.15 may be written in the standard form:

$$\frac{dL_\nu}{d\tau_\nu} = -L_\nu + B_\nu(T). \quad (1.16)$$

The formal solution, known as the Schwarzschild equation, is:

$$L_\nu(S) = L_\nu(0) e^{-[\tau_\nu(S) - \tau_\nu(0)]} + \int_0^S B_\nu(s) e^{-[\tau_\nu(S) - \tau_\nu(s)]} d\tau_\nu(s), \quad (1.17)$$

where the first term describes the attenuation of the incident radiation, and the second term represents the thermal emission from each layer, attenuated by the overlying optical depth.

Modern passive and active remote-sensing instruments exploit these principles of radiative transfer to infer atmospheric state variables from measured radiances.

1.3.3 The Earth Emission Spectrum

At the Top of the Atmosphere (TOA), Earth gains energy from the solar radiation (Short-wave Radiation SR), which is partly reflected back to space, while it loses energy through the Outgoing Longwave Radiation (OLR). The OLR at TOA comes from the small portion of energy emitted by the surface, which freely passes through the atmosphere, and by the radiative cooling of the atmosphere and clouds. The interaction of infrared radiation emitted by the Earth with optically active gases, such as water vapour, carbon dioxide, methane and ozone, produces several absorption lines due to the vibrational and rotational transitions of these molecules, as shown in Figure 1.6.

The $9.6 \mu\text{m}$ ozone absorption band is a critical spectral feature for infrared remote sensing. In the nadir geometry employed by instruments like IASI, this band is used to retrieve total column and lower-to-middle stratospheric ozone, leveraging its strong absorption

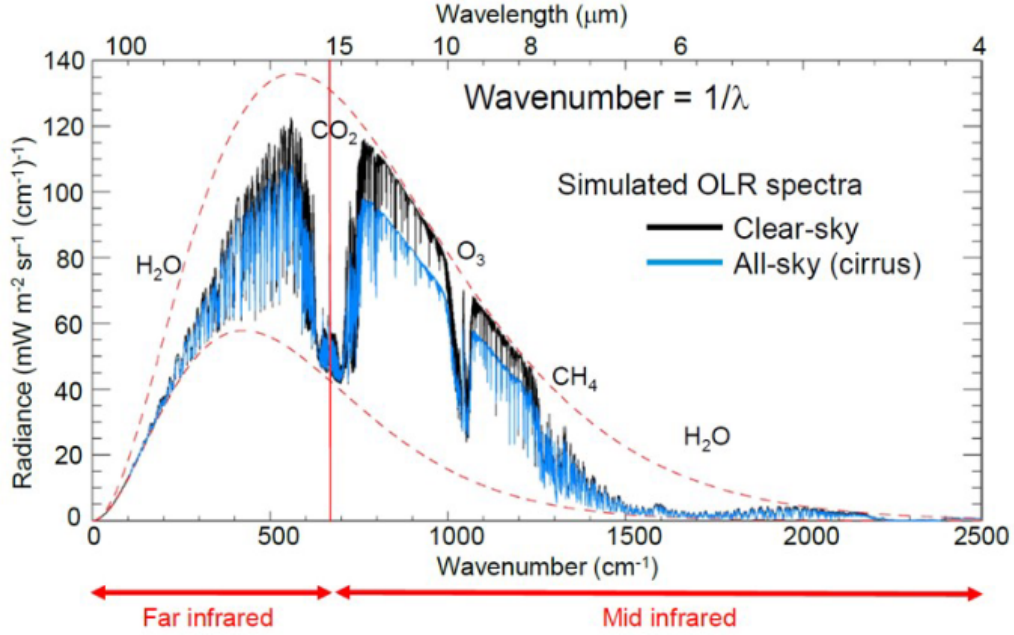


Figure 1.6: Simulated outgoing spectral radiance at the top of Earth's atmosphere in clear and all sky conditions, showing the absorption at specific frequencies and the principle absorbers. For comparison, the red curves show the flux from a classic "blackbody" at 294 K, and the flux from the tropopause at 210 K.

signature and good thermal contrast. In contrast, the limb-sounding geometry used by MIPAS exploits the same band's long atmospheric pathlength to achieve high vertical sensitivity in the middle and upper stratosphere, enabling accurate ozone profiling through an enhanced effective optical depth. Thus, the 9.6 μm band is pivotal for both nadir and limb observations, with the primary distinction between the two geometries being their region of peak altitude sensitivity.

1.4 The Inverse Problem

The *Inverse Theory* is a generic theory to solve an *Inversion Problem* (IP), which occurs when an indirect measurement is performed, and the measurement is a more or less complicated function of quantities or parameters to be obtained. Sensors installed on satellite can help us in determining the state of the atmosphere, however we are not actually observing what we want to know. These instruments measure a change in voltage, the number of photons passing into a detector over a certain period, a Doppler shift in radar or lidar frequency, and the intensity of the back-scattered energy, whereas we are interested in very specific atmospheric variables, e.g., the distribution of temperature and atmospheric constituents. Therefore, to solve an IP in our case means to extract very specific information from the remote sensing observations that are typically only partially related to the variable of interest. Associated to the solution state of an IP are questions to determine the information content of the measurements and the relationship between the true state and the retrieved one, and to perform an overall error analysis on the system (Rodgers, 2000).

1.4.1 The Forward Model and Retrieval Process

To formalize the Inverse Problem we need firstly introduce a *direct problem* defining a relation mapping the quantities we are interested into our measurements. Therefore, having a vector $\mathbf{y} = [y_1, \dots, y_m]$ of measurements, defined in the measurements vector space, we want to find the vertical profile $\mathbf{x} = [x_1, \dots, x_n]$ of the variable defined in the states vector space, through the inversion of a function f , named *forward function*, such that:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}). \quad (1.18)$$

Because of the measurements noise and the impossibility to know the exact physics of the problem it is only possible to define a model, called *forward model* $\mathcal{F}$, such as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \boldsymbol{\epsilon}, \quad (1.19)$$

where $\boldsymbol{\epsilon}$ is the *measurement noise*. In general, if we have a measurement that has some sensitivity to the atmospheric variable we desire, we can compute the signal that we would observe with our remote sensor using the forward model. Hence, if we know the atmospheric state, which includes here all variables affecting the observed signal, we can reproduce the measurement. To extract the atmospheric state vector $\mathbf{x}$ we have to invert 1.19, i.e. to perform a *retrieval process*. This is complicated by many factors: we have to invert a set of noisy observations using an imperfect forward model where multiple discrete descriptions of the atmospheric state could reproduce the measurements. Taking into account the errors due to the instrumental noise and model uncertainties, indeed, means that is no guarantee that there is only a single $\mathbf{x}$ that maps to the observation $\mathbf{y}$, on the contrary there is an infinite set of solutions consistent with the measurement vector and its error. For this reason, additional prior information on the atmospheric state (e.g., derived from climatology or other sources) is introduced and combined with the forward model to constrain the solution. This prior knowledge helps to address the ill-posed nature of the problem by favoring physically plausible states among the many possible solutions. The accuracy of the retrieved state depends on how representative this prior information is of the actual conditions; if the true state is well captured by the prior, the retrieval is expected to be more accurate, otherwise the accuracy of the retrieved product will be poor. Another factor that complicates the problem is that $\mathcal{F}$ is usually non-linear and usually not possible to be inverted analytically. To face this problem, the forward model is linearized around a reference state $\mathbf{x}_0$:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\epsilon}. \quad (1.20)$$

where $\mathbf{K}$ is the the Jacobian matrix of the Frechet derivative of the model with respect to the state vector, also called *weighting function matrix*:

$$\mathbf{K} = \frac{\partial \mathcal{F}}{\partial \mathbf{x}}. \quad (1.21)$$

In atmospheric science this can be justified by the fact that several problems are physically governed by exponential and power laws without discontinuities, which therefore can be considered moderately linear.

A wide range of techniques has been developed to perform retrievals (Maahn et al., 2020). The paragraph 1.4.2 will describe one of the most used algorithm, classified as a “physical”

retrieval algorithm. Such a method exploits the fact that observations, forward models, and the assumptions are affected by inherent uncertainty and can be represented by probability distributions. These algorithms typically use the Bayes's theorem.

1.4.2 Optimal Estimation Method

Theoretical base

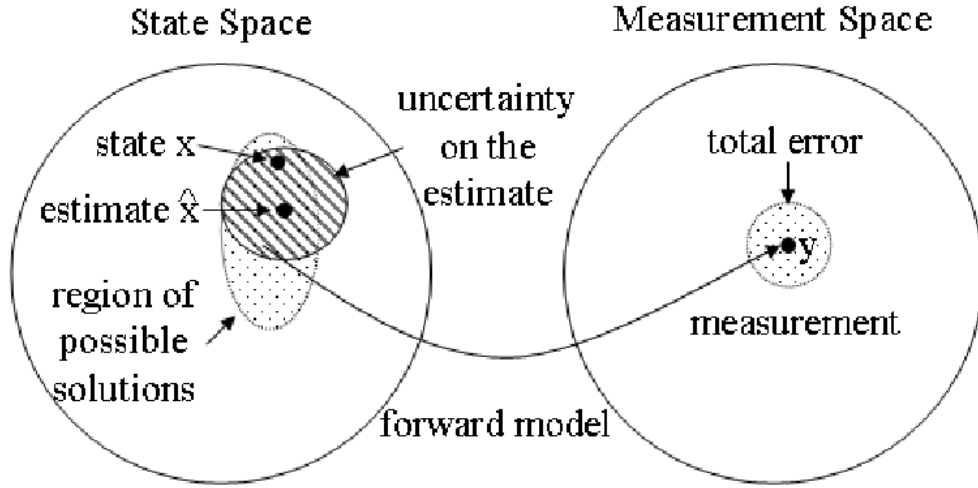


Figure 1.7: Schematic representation of Bayes' approach in a retrieval problem from Rodgers, 2000.

The *Optimal Estimation* (OE) is an inversion method of the forward model based on Bayes' theorem, and on the assumption that observations have inherent uncertainties, characterized by Gaussian distribution (see Appendix A). According to the Bayes' theorem, it is true that:

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}, \quad (1.22)$$

where P indicates a probability density function (pdf). The 1.22 tells us that the *a posteriori* pdf of the state $\mathbf{x}$ when the measurement $\mathbf{y}$ is given ($P(\mathbf{x}|\mathbf{y})$), is equal to the ratio between the following two quantities: the product of the pdf of the measurement $\mathbf{y}$ given the state $\mathbf{x}$ ($P(\mathbf{y}|\mathbf{x})$) with the prior pdf of the state $\mathbf{x}$ ($P(\mathbf{x})$), divided by the prior pdf of the measurement $\mathbf{y}$ before it is made ($P(\mathbf{y})$). Since Normal distribution is assumed, we have that the m -dimensional $\mathbf{y}$ vector pdf is:

$$P(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{S}_{\mathbf{ny}}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y} - \langle \mathbf{y} \rangle)^T \mathbf{S}_{\mathbf{ny}}^{-1} (\mathbf{y} - \langle \mathbf{y} \rangle)}, \quad (1.23)$$

where $\langle \mathbf{y} \rangle$ is the expected value of the measurement $\mathbf{y}$ having dimension m , and $\mathbf{S}_{\mathbf{ny}}$ is its noise covariance matrix. In our case the expected value is that provided by the model, hence $\mathcal{F}(\mathbf{x})$. The same reasoning can be done with the prior pdf. Being $\mathbf{y}$ and $\mathbf{x}_a$ completely independent, we have that the joint probability is the product of the pdfs:

$$P(\mathbf{y}, \mathbf{x}_a) = \frac{1}{(2\pi)^{\frac{m+n}{2}} |\mathbf{S}_{\mathbf{ny}}|^{\frac{1}{2}} |\mathbf{S}_a|^{\frac{1}{2}}} e^{-\frac{1}{2}[(\mathbf{y} - \mathcal{F}(\mathbf{x}))^T \mathbf{S}_{\mathbf{ny}}^{-1} (\mathbf{y} - \mathcal{F}(\mathbf{x})) + (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a)]}. \quad (1.24)$$

To obtain an explicit retrieval, only one state had to be chosen from the ensemble described by $P(x|y)$. This state is the expected value or the most probable state, together with a measure of the width of the pdf. It is simple to demonstrate that, in the linear case, the retrieved state ($\hat{x}$) and its covariance matrix ($\hat{S}$) are (Rodgers, 2000):

$$\hat{x} = x_a + (K^T S_{ny}^{-1} K + S_a^{-1})^{-1} K^T S_{ny}^{-1} (y - K x_a), \quad (1.25a)$$

$$\hat{S} = (K^T S_{ny}^{-1} K + S_a^{-1})^{-1}. \quad (1.25b)$$

Proof. For theoretical demonstration, we will refer to the linear problem, using logarithmic expressions, which simplify calculations. In the logarithmic expression, the pdfs become:

$$-2\ln P(y|x) = (y - Kx)^T S_{ny}^{-1} (y - Kx) + c_1 \quad (1.26)$$

$$-2\ln P(x) = (x - x_a)^T S_a^{-1} (x - x_a) + c_2 \quad (1.27)$$

where c_1 and c_2 are constants. Since $P(y)$ is in practice only a normalization factor, we have that:

$$-2\ln P(x|y) = (y - Kx)^T S_{ny}^{-1} (y - Kx) + (x - x_a)^T S_a^{-1} (x - x_a) + c_3 \quad (1.28)$$

that is a quadratic form in x and can be generalized as:

$$-2\ln P(x|y) = (x - \hat{x})^T \hat{S}^{-1} (x - \hat{x}) + c_4 \quad (1.29)$$

Comparing eq.s 1.28 and 1.29, and therefore the quadratic form, we find:

$$x^T \hat{S}^{-1} x = x^T K^T S_{ny}^{-1} K x + x^T S_a^{-1} x. \quad (1.30)$$

From which:

$$\hat{S}^{-1} = K^T S_{ny}^{-1} K + S_a^{-1}. \quad (1.31)$$

Comparing eq.s 1.28 and 1.29, and therefore the linear form in x^T , we find:

$$x^T \hat{S}^{-1} \hat{x} = x^T K^T S_{ny}^{-1} y + x^T S_a^{-1} x_a. \quad (1.32)$$

From which:

$$\hat{x} = x_a + (K^T S_{ny}^{-1} K + S_a^{-1})^{-1} K^T S_{ny}^{-1} (y - K x_a). \quad (1.33)$$

□

In general the optimal estimate solution $\hat{x}$ can be found by minimizing the *cost function* $c(x)$ with a Gaussian constraint on the state:

$$c(x) = (y - \mathcal{F}(x))^T S_{ny}^{-1} (y - \mathcal{F}(x)) + (x - x_a)^T S_a^{-1} (x - x_a). \quad (1.34)$$

Minimizing 1.34 means that the solution is such that it provides the observations if the model is applied, and it converges to the constrained profile. The first term represents the agreement between the simulated spectrum $\mathcal{F}(x)$ and the measurements y . However, due to the ill-conditioning of the problem, the minimization of the first term cannot be carried over alone and a constraint is needed. To this aim the second term is added. This term is called *regularization* or *penalization term*, and is a special case of the more

general Tikhonov theory of regularization. The addition of this term steers the minimization towards results that are within the atmospheric variability of a model, or within the expected error of a correlated measurement.

Manipulating the equation 1.25a and considering the linear case, we can rewrite it as:

$$\hat{\mathbf{x}} = \mathbf{x}_a + \mathbf{G}\mathbf{K}(\mathbf{x} - \mathbf{x}_a) + \mathbf{G}\boldsymbol{\epsilon}. \quad (1.35)$$

where:

$$\mathbf{G} = (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \quad (1.36)$$

is called *gain matrix* $\mathbf{G}$ and provides the sensitivity of the retrieved profile to the observations. The quantity $\mathbf{G}\mathbf{K}$ in 1.35 can be explicated as:

$$\boxed{\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}_{\text{true}}} = (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K}.} \quad (1.37)$$

It is called *averaging kernel matrix* $\mathbf{A}$ and provides the sensitivity of the retrieval to the true profile ($\mathbf{x}_{\text{true}}$). Therefore, 1.35 can be written as:

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{x}_{\text{true}} + (\mathbf{I} - \mathbf{A})\mathbf{x}_a + \mathbf{G}\boldsymbol{\epsilon}. \quad (1.38)$$

The retrieval can be expressed as the modulation of the true profile plus a modulated term due to the constraint, which represents the fraction of the retrieval value originating from the *a priori* rather than from the signal. The retrieval can be expressed as the modulation of the true profile plus a modulated term due to the constraint. The modulation factor of the true profile is the averaging kernel matrix. If it is the identity matrix we find that $\hat{\mathbf{x}} \simeq \mathbf{x}_{\text{true}}$. Therefore, each row of $\mathbf{A}$ is a smoothing function, called averaging kernel, which maps the true space into the retrieval one. Its width is a measure of the retrieval vertical resolution.

Performing an error analysis of the retrieved profile it is found that the difference between the retrieved and the true profile is:

$$\hat{\mathbf{x}} - \mathbf{x}_{\text{true}} = (\mathbf{A} - \mathbf{I})(\mathbf{x}_{\text{true}} - \mathbf{x}_a) + \mathbf{G}\boldsymbol{\epsilon}. \quad (1.39)$$

From 1.39 we can see that the total error has two contributions.

The first term in 1.39 is called smoothing error and can be estimated in the following way. Since the true state cannot be known, statistics of the error are needed to obtain mean and covariance over some appropriate ensemble of states. These statistics may be described by $\mathbf{x}_a$ and $\mathbf{S}_a$. In this case, if the mean is $\langle \mathbf{x} \rangle$, we have that:

$$(\mathbf{A} - \mathbf{I})(\langle \mathbf{x} \rangle - \mathbf{x}_a) = 0, \quad (1.40)$$

and the associated covariance is:

$$\begin{aligned} \mathbf{S}_s &= (\mathbf{A} - \mathbf{I}) \langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^T \rangle (\mathbf{A} - \mathbf{I})^T = \\ &= (\mathbf{A} - \mathbf{I})\mathbf{S}_a(\mathbf{A} - \mathbf{I})^T \end{aligned} \quad (1.41)$$

Putting the expression of the averaging kernel matrix (eq 1.37) in 1.41, we find that:

$$\boxed{\mathbf{S}_s = (\mathbf{K}^T \mathbf{S}_{ny}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{S}_a^{-1} (\mathbf{K}^T \mathbf{S}_{ny}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1}.} \quad (1.42)$$

The second term in 1.39 is called retrieval noise. It is a random quantity that contributes to the covariance as:

$$\mathbf{S}_n = \mathbf{G} \mathbf{S}_{ny} \mathbf{G}^T \quad (1.43)$$

Putting the expression of the gain matrix (eq. 1.36) in 1.43, we find that:

$$\boxed{\mathbf{S}_n = (\mathbf{K}^T \mathbf{S}_{ny}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{K}^T \mathbf{S}_{ny}^{-1} \mathbf{K} (\mathbf{K}^T \mathbf{S}_{ny}^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1}.} \quad (1.44)$$

Summing $\mathbf{S}_n$ and $\mathbf{S}_s$ the total covariance $\hat{\mathbf{S}}$ (eq. 1.25b) is obtained.

Numerical Implementation

To perform the method numerically, a discretization of the state vector is required. The iterative process typically begins with an initial guess $\mathbf{x}_0$ of the atmospheric state. In the absence of model output or previous retrievals, the a priori profile $\mathbf{x}_a$ is used as the first guess. The forward operator $\mathcal{F}$ is then applied to $\mathbf{x}_0$ to simulate the corresponding observation in measurement space ($\mathbf{y}_0 = \mathcal{F}(\mathbf{x}_0)$). The difference between the simulated observation $\mathbf{y}_0$ and the actual measurement $\mathbf{y}$ is used to update the state vector, producing a new estimate $\mathbf{x}_1$. This update is weighted by the error covariance matrices: $\mathbf{S}_y$ for the observation space and $\mathbf{S}_0$ (or $\mathbf{S}_a$ if $\mathbf{x}_0 = \mathbf{x}_a$) for the state space. These matrices ensure that the correction accounts for the relative uncertainties in both the measurements and the prior information. This step requires $\mathcal{F}$ inversion, which is possible only if it is moderately linear at least. The linearization is performed computing the Jacobian $\mathbf{K}$. The process has to be iterated until $\mathbf{x}$ converges to a stable, optimal solution, computing $\mathbf{K}$ for each iteration. The retrieval is considered convergent when the change in the state vector between two iterations i and $i + 1$ is smaller than the uncertainty of $\mathbf{x}$, given by the covariance of $i - th$ step.

Finding the solution minimizing the cost function also need an iterative process. This is usually carried out using the *Gauss-Newton method* (GN) which consists in finding at each step the minimum of a quadratic approximation of the target function. In our case, to perform a quadratic expansion of the cost function:

$$c(\mathbf{x}) = c(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla c(\mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \nabla^2 c(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0). \quad (1.45)$$

The minimum of the function corresponds to the point where the gradient vanishes, i.e., $\nabla c(\mathbf{x}) = 0$. Using a second-order Taylor expansion of the gradient around a point $\mathbf{x}_0$, we obtain:

$$\nabla c(\mathbf{x}) \approx \nabla c(\mathbf{x}_0) + \nabla^2 c(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0), \quad (1.46)$$

Setting this expression to zero and solving for $\mathbf{x}$ yields the Newton update:

$$\mathbf{x} = \mathbf{x}_0 - (\nabla^2 c(\mathbf{x}_0))^{-1} \nabla c(\mathbf{x}_0). \quad (1.47)$$

In a general case in which $\mathcal{F}(\mathbf{x})$ is not linear in $\mathbf{x}$ an iterative process has to be set, in which the i th + 1 term is:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - (\nabla^2 c(\mathbf{x}_i))^{-1} \nabla c(\mathbf{x}_i). \quad (1.48)$$

The sequence will converge to the minimum. Substituting the expression of the cost function (1.34), and exploiting the definition of the Jacobian matrix (1.21), we can find the form of the solution:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + (\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \mathbf{K}_i + \mathbf{S}_a^{-1})^{-1} [\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} (\mathbf{y} - \mathcal{F}(\mathbf{x}_i)) + \mathbf{S}_a^{-1} (\mathbf{x}_a - \mathbf{x}_i)]. \quad (1.49)$$

Many tools have been implemented to achieve better convergence. One of the most common is the *Levenberg-Marquardt (LM)* technique (Rodgers, 2000). This technique may be seen as a variable Tikhonov regularization in the parameter space with the a priori set as the solution at the previous iteration. In the Marquardt variant, the Gauss-Newton's iteration gets modified as follows:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + (\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \mathbf{K}_i + \mathbf{S}_a^{-1} + \lambda_i \text{diag}(\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \mathbf{K}_i))^{-1} \cdot [\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} (\mathbf{y} - \mathcal{F}(\mathbf{x}_i)) + \mathbf{S}_a^{-1} (\mathbf{x}_a - \mathbf{x}_i)]. \quad (1.50)$$

where λ_i is the *Marquardt* parameter depending on the iteration and $\text{diag}(\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \mathbf{K}_i)$ is the diagonal matrix filled with the diagonal elements of $\mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \mathbf{K}_i$.

The LM technique addresses two issues:

- the damping term λ_i helps in the inversion of the principal part of the Hessian of the cost function, which is necessary for the computation of the subsequent iteration of the state vector with the Gauss-Newton iterative method. In fact, adding a positive diagonal term to a matrix reduces the ill-conditioning;
- the Gauss-Newton technique is applied to a model which is not linear. Therefore, it is not guaranteed that the minimum predicted at each iteration, which is calculated considering a linearization of the model, brings us to a point where the cost function is less than that at the previous step. In the Levenberg formulation, we add a term in the opposite direction to the gradient. This is by definition a descent direction. Thus, for a sufficiently large λ_i the value of the cost function decreases at each step of the iteration.

However, the drawback of the method is that if the term λ_i is large, the inverse is small and the distance between two iterates becomes small. Moreover the value of the parameter has to be set, and it depends on the problem. One of the methods to choose it is to calculate the new value of the cost function at each iteration. If the cost function is reduced, we accept the iteration and set $\lambda_{i+1} = d\lambda_i$ for some constant $d < 1$. If the cost function increases, we reject the iteration and try again with a larger value of λ_i . We repeat the increase of λ_i until the cost function reduces or we perform a maximal number of cycles, exiting the Gauss-Newton sequence with an error code.

1.4.3 Uncertainties in the Retrieval Process

As shown in the previous paragraphs more sources of uncertainties appear in the retrieval process mainly due to three factors: the a priori constraint, the forward operator, and the measurement, each described by the related covariance error matrix $\mathbf{S}_{\text{err}}$. These sources of error deserve an in-depth comments.

Prior Uncertainties

The a priori information is a strong constraint on the retrieval, so the uncertainty in the retrieval must be smaller than the uncertainty in the prior. Otherwise, the retrieval does not add any information from the observations. Therefore, the wrong choice of the a priori constraint can produce biased retrieval products. However, when the information content of observation is high enough, the prior does not serve as a serious constraint. It is therefore crucial choosing a good $\mathbf{S}_a$ to constrain the solution, but it is not always simple to derive the $\mathbf{S}_a$ matrix from existing observations.

Forward Model Uncertainties

Limitations in our understanding of the physics of the systems and in the computational efficiency make the forward model itself a source of error. $\mathcal{F}$ not only depends on $\mathbf{x}$, but also on the parameter vector $\mathbf{b}$ ($\mathcal{F}(\mathbf{x}, \mathbf{b})$), which includes additional input variables for the forward operator that are not retrieved by the algorithm and are considered fixed, for example additional profiles necessary for the retrieval (e.g. pressure) or other parameters of the radiative transfer model. An error covariance matrix $\mathbf{S}_b$ is associated to the parameter vector $\mathbf{b}$. The presence of uncertainties in the forward model is often neglected. However, considering a total covariance error matrix given by the sum of the measurement noise error matrix and the forward model error matrix could be more rigorous.

Measurement Uncertainties

A measurement $\mathbf{y}$ is typically multidimensional due to the presence of observations from multiple channels of a single instrument, or multiple instruments used in the retrieval. There will be random uncorrelated errors for each particular element of $\mathbf{y}$, as well as correlated errors between the different elements. All this is represented by the covariance matrix $\mathbf{S}_{ny}$.

1.4.4 Diagnostics: Errors, Averaging Kernels, Degrees of Freedom, and Information Content

Smoothing Error

The smoothing error, represented by the variance-covariance matrix $\mathbf{S}_s$ (1.42), is a diagnostic quantity which represents the expected deviation of the retrieval from the true state due to the a priori constraint term, which can have the effect of smoothing the true state. Directly comparable to other retrieval errors, e.g. noise, it is the quantity used to characterize the uncertainty due to the difference between the actual atmospheric state and the prior. It is to take into account that the linear estimate presented in eq 1.41 holds only if $\mathbf{x}_a = \langle \mathbf{x} \rangle$, where $\langle \rangle$ denotes the expectation value. More precisely, it is required that $\mathbf{S}_a$ represents the covariance around $\langle \mathbf{x} \rangle$, and not the covariance around $\mathbf{x}_a$ if the latter happens not to be chosen to equal $\langle \mathbf{x} \rangle$.

Retrieval Noise

The retrieval noise is the easiest component to evaluate. It directly derives from the measurement noise which is usually randomic, unbiased and uncorrelated between spectral

channels. Its variance-covariance matrix is usually known.

Averaging Kernel Matrix

The Averaging Kernel Matrix (AKM), $\mathbf{A}$ (eq.1.37), relates the most probable state to the true state. It can be interpreted as the weights of a weighted mean of the true state and an a priori estimate that gives rise to the retrieval. The rows of the AKM are the averaging kernels or smoothing functions that map the true state into retrieval space. The width of the kernel is a measure of retrieval resolution. When the AKM is approximately the identity matrix, the retrieval can be considered accurate and can be interpreted as a measure of the fraction of the retrieved value that originates from the signal and not from the a priori.

Number of Degrees of Freedom

In general, given a data sample, the *degrees of freedom* d is the scalar quantity representing the maximum number of logically independent values, which may vary in the sample. For instance, consider to have three choices corresponding to potentially three degrees of freedom. Adding a constraint such as three numbers are required to add up to some number t , then the first two numbers can be selected at random, but the third must be chosen so that it makes the total equal to t , thus the number of degrees of freedom is now two. In the Gaussian framework of the retrieval, if the state vector $\mathbf{x}$ consists of n normally distributed variables with mean zero and unit variance, the cost function can be expressed as the sum of the squared components of $\mathbf{x}$, i.e., $\mathbf{x}^T \mathbf{x}$. The expected value of this sum is n , which corresponds to the number of degrees of freedom in the system. For different mean and standard deviation of the Normal distribution, the number of degrees of freedom will change: the largest number of possible degrees of freedom is determined by the number of elements in the measurement vector m or the state vector n , whichever is the smaller (Rodgers, 2000). The number of degrees of freedom can be divided into degrees of freedom from the signal d_s and degrees of freedom from noise d_n , expressed as follows:

$$d_s = \text{tr}(\mathbf{A}), \quad (1.51)$$

$$d_n = \text{tr}(\mathbf{S}_{\text{ny}}(\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_{\text{a}}^{-1})^{-1}). \quad (1.52)$$

If the degrees of freedom for signal at the solution is lower than n it may be that there is less information in the measurement about one of the state vector elements, or that there are correlated state vector elements.

Information Content

The information content of a measurement is a quantitative diagnostic parameter to understand how much the measurement has improved our knowledge about a quantity. One method to measure the information content of a measurement is using the entropy, as defined in the *Shannon Information* content (SI). The entropy represents the disorder in a system, hence the SI measures how much the entropy is reduced by the measurement, i.e. the difference between entropy of the state before and after the measurement. The

entropy is defined as follows (Rodgers, 2000):

$$S(P) = \sum_{i=1}^{\infty} p_i \ln p_i, \quad (1.53)$$

where p_i is the probability that the system will be in the state i . The logarithm on base two is usually used instead of natural logarithm when the information units are bits. The information of two independent measurements is additive. Therefore, the Shannon Information content H quantifies the reduction in entropy due to the measurement. It is computed as the difference between the entropy of the probability density function (pdf) of the state before the measurement (the prior) and the entropy of the pdf after the measurement (the posterior):

$$H = S[P(\mathbf{x}_a)] - S[P(\mathbf{x}|\mathbf{y})] \quad (1.54)$$

In a linear Gaussian case, we have that the entropy is proportional to the natural logarithm of the determinant of the variance-covariance matrix. Hence the SI will be:

$$H = \frac{1}{2} \ln |\mathbf{S}_a| - \frac{1}{2} \ln |\hat{\mathbf{S}}| = \frac{1}{2} \ln |\hat{\mathbf{S}}^{-1} \mathbf{S}_a| = -\frac{1}{2} \ln |\mathbf{I} - \mathbf{A}|. \quad (1.55)$$

Therefore, The information content of a retrieval is intrinsically related to the inverse of the variance-covariance matrix and the properties of the averaging kernel. It is maximized when the averaging kernel matrix is close to the identity, corresponding to a retrieval with high accuracy and resolution. Operationally, it quantifies the uncertainty reduction provided by the measurement: a large information content indicates that the posterior covariance is much smaller than the prior covariance. This implies that the retrieval is strongly influenced by the measurement, with minimal dependence on the a priori constraint.

Physical Interpretation for This Study

Although the retrieval theory presented in this section is not applied directly in this work—since we make use of operational Level-2 products—it provides the conceptual basis for interpreting the characteristics of satellite ozone profiles. In particular, the averaging kernels (AKs) and the Degrees of Freedom for Signal (DOFS) offer an intuitive measure of the vertical sensitivity of different remote-sensing observation.

From a physical point of view, an averaging kernel describes how a perturbation in the true atmospheric state at a given altitude affects the retrieved state. A narrow AK with a peak close to unity indicates good vertical resolution and a retrieval dominated by the measurement, whereas a broad kernel reflects a smoothed representation of the true profile, more strongly influenced by the a priori. Therefore, AKs can be interpreted as the “effective vertical resolution” of the retrieval system.

The DOFS quantify the number of independent pieces of information that the measurements contribute to the retrieved profile. Typical values differ substantially between nadir and limb geometries. Nadir thermal-infrared sounders such as IASI generally provide about 3 DOFS for ozone, with a coarse vertical sensitivity and a the retrieval of one or two independent layers. In contrast, limb-emission instruments such as MIPAS provide about 8 DOFS, with much finer vertical resolution especially in the UTLS, and enhanced sensitivity along the limb path.

These typical values, together with the physical interpretation of AKs, are essential for understanding the capabilities and limitations of the Level-2 products used in this study and for correctly interpreting the ozone vertical structures.

Production and Validation of a New 2008-2011 Ozone Dataset

2.1 Chapter Overview

This chapter presents the first part of the thesis, focusing on the development and ozonesonde-based validation of a new synergistic ozone dataset for the period 2008–2011. The dataset is created by applying the Complete Data Fusion (CDF) methodology to observations from two complementary sensors: the limb-viewing Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) and the nadir-viewing Infrared Atmospheric Sounding Interferometer (IASI). Validation data are obtained from the World Ozone and Ultraviolet Radiation Data Centre (WOUDC) network.

The core rationale for this fusion lies in the complementary vertical sensitivities of limb and nadir geometries. Limb emission instruments like MIPAS view the atmosphere along a tangential path, providing high effective optical depth and excellent stratospheric vertical resolution but limited sensitivity in the lower troposphere. Conversely, nadir sounders like IASI measure upwelling radiance from below, yielding greater sensitivity to the lower and middle troposphere but coarser vertical resolution aloft. By merging these sensitivities, the fused product achieves increased vertical information content and reduced uncertainties, offering a more comprehensive ozone profile than either instrument could provide independently. A key objective of this analysis is to generate a dataset with enhanced vertical resolution and reduced error, particularly in regions of frequent stratosphere-troposphere exchange (STE)—where, as established in the preceding chapter, such improvements are critical. In the contemporary landscape of multi-instrument satellite observations, exploiting such complementarity is essential for extracting maximal atmospheric information.

The chapter is structured as follows: the *Data* section describes the datasets used in this work; the *Methodology* section outlines the fusion approach and processing steps; the *Results and Discussion* section presents the main findings and interprets them; and finally, the *Conclusions* section summarizes the results and discusses their implications,

including the link to the second part of this thesis.

This first part is foundational, serving as the primary input for the subsequent analysis and thus should not be considered in isolation. It builds directly upon a manuscript submitted to Atmospheric Measurement Techniques (<https://doi.org/10.5194/amt-19-167-2026>): “Development and validation of a new ozone dataset using Complete Data Fusion of MIPAS and IASI observations: a step towards understanding stratospheric ozone intrusions”. While some figures and text elements are adapted from the manuscript, additional explanations and clarifications have been included to ensure consistency and completeness within the broader context of this thesis.

2.2 Data

2.2.1 Remote Sensing Data

Basic Principles of a Fourier Transform Spectrometer

Both MIPAS and IASI are based on the principle of Fourier Transform Spectrometer (FTS), which relies on interferometric measurements of infrared radiation to retrieve atmospheric spectra with high spectral resolution. To better understand the measurement principle, here basic concepts of a Fourier spectrometer will be shown (see Figure 2.1).

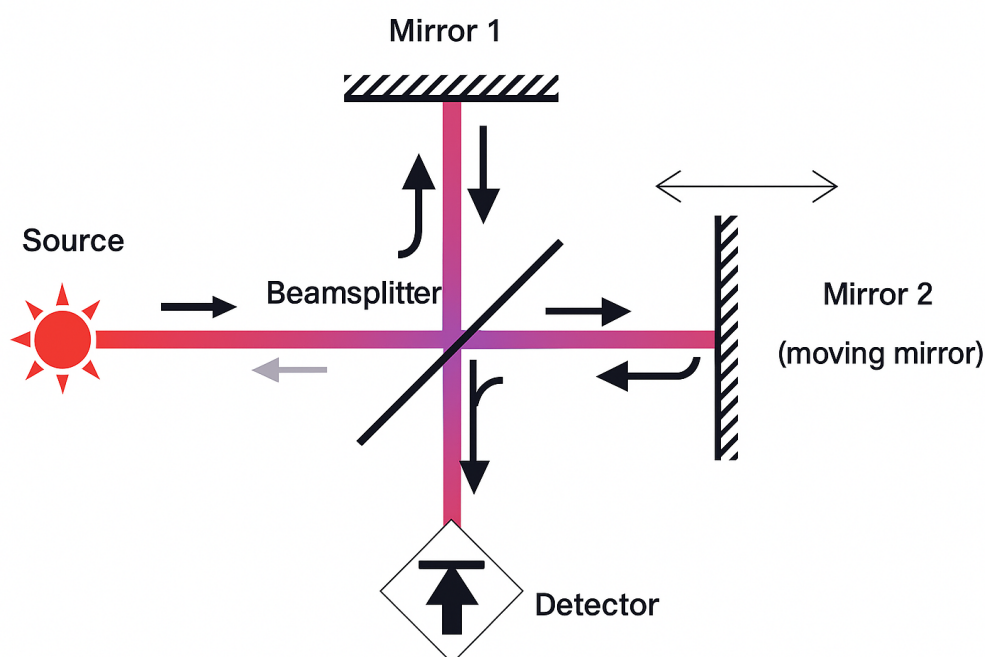


Figure 2.1: Optical layout of a Michelson interferometer.

A source of monochromatic light crosses a semi-reflective mirror (Beam splitter) and it is divided in two different beams that are reflected again by a movable mirror (Mirror 1) and a fixed mirror (Mirror 2), respectively. The two paths crossed by the lights have different lengths in order to produce a phase shift. The beams cross again the beam splitter and meet the detector R.

The intensity of a monochromatic source measured by the detector is:

$$I = E^2 \epsilon [1 + \cos(2\pi\sigma x)] \quad (2.1)$$

Where in equation 2.1 E is the sum of the electric fields of the two beams, ϵ is the efficiency of the beam splitter, σ is the wavenumber and x is the optical path difference between the two beams.

When the source is not monochromatic and is characterized by a continuum spectral distribution $B(\nu)$ we can define the relation:

$$dE^2(\sigma) = B(\sigma) d\sigma \quad (2.2)$$

We can now introduce the interferogram I , defined as:

$$dI = \epsilon B(\sigma) \cos(2\pi\sigma x) d\sigma \quad (2.3)$$

Thus, the total interferogram is obtained by an integration over the wavenumber σ :

$$I(x) = \int_0^{+\infty} \epsilon B(\sigma) \cos(2\pi\sigma x) d\sigma \quad (2.4)$$

The spectrum $B(\sigma)$ is obtained from the inverse Fourier transform of 2.4 :

$$B(\sigma) = \int_{-\infty}^{+\infty} I(x) \cos(2\pi\sigma x) dx \quad (2.5)$$

This solution is based on the assumption that the optical path difference between the beams, that is to say the distance covered by the movable mirror, can be infinite. But, in a real case, $x \in (-L, +L)$, where L is the Maximum Optical Path Difference (MOPD) for which the interferogram has been measured.

The truncation of this integral is equivalent to multiplying the integral in 2.5 by a function $F(x) = 1$ for $x \in (-L, +L)$ and 0 elsewhere.

Based on the properties of Fourier transform, this operation is equivalent to convolute the spectrum obtained from the interferogram with the "sinc" function:

$$B'(\sigma) = B(\sigma) * \text{sinc}(2\pi\sigma L) \quad (2.6)$$

The function $\text{sinc}(y) \equiv \sin(y)/y$, also called instrument response function (or instrument line-shape) is shown in Figure 2.2.

The first intersection of the function with the x-axis in $\sigma = 1/2L$ provides the *spectral resolution* of the instrument.

Compared to other types of spectrometers, the Fourier transform spectrometers are usually characterized by higher spectral resolution in the IR region, are easier to radiometrically calibrate, tend to be more sensitive and can have a larger field of view.

MIPAS

The Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) was launched aboard ESA's Envisat (2002–2012) (Fischer and Oelhaf, 1996). Operating in a Sun-synchronous orbit at an altitude of approximately 800 km, it allowed dense vertical (about 3 km) and horizontal (about 500 km) coverage, providing data relevant for the study of

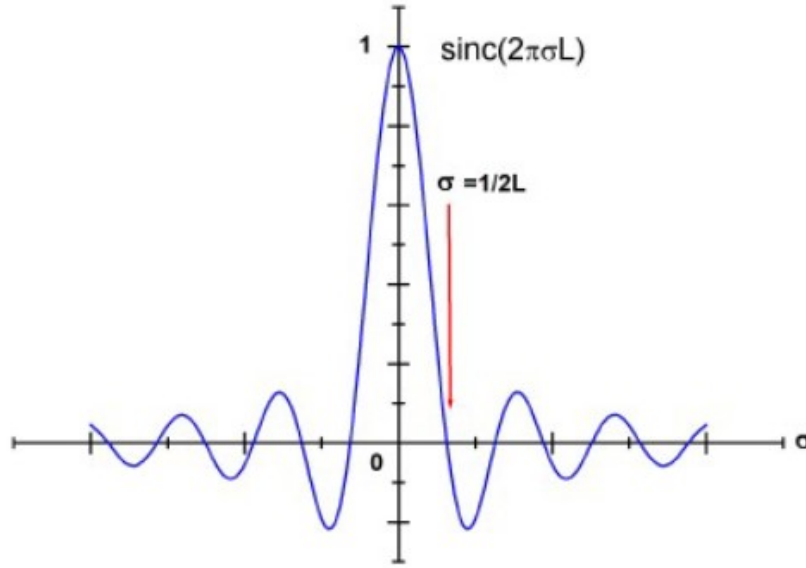
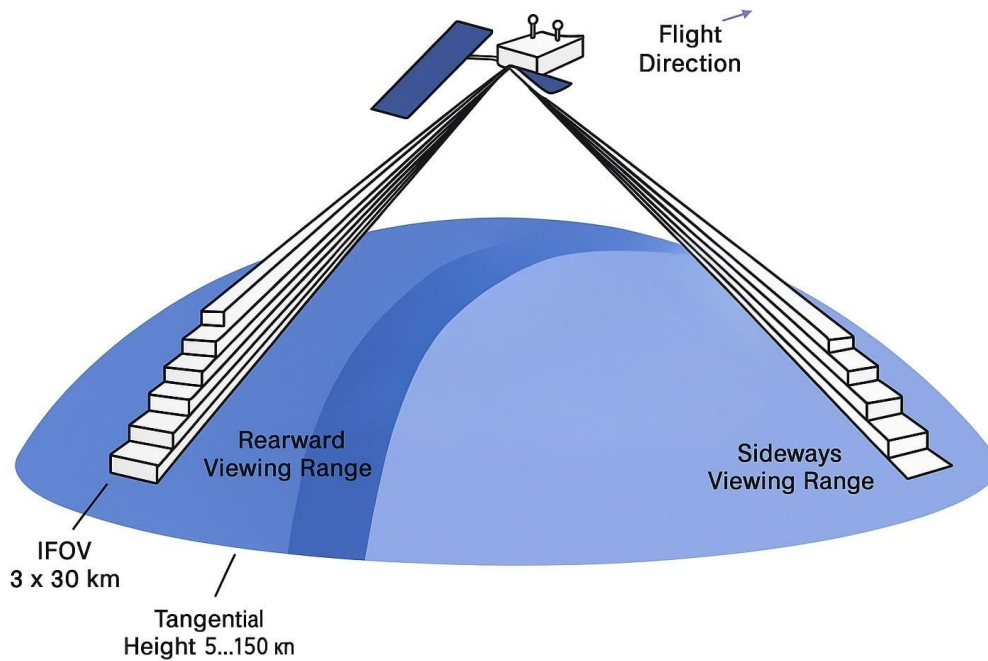


Figure 2.2: Sinc function.

Figure 2.3: The MIPAS scanning geometry from ESA Earth Online official website (<https://earth.esa.int/eogateway/instruments/mipas/description>).

chemical and dynamical atmospheric processes, and for stratospheric-tropospheric interactions. By retrieving vertical profiles of several atmospheric trace species, MIPAS addressed diverse scientific objectives, especially in the UTLS (Chauhan et al., 2009; Höpfner et al., 2015; Sembhi et al., 2012). Despite orbit adjustments in 2010, the mission continued until Envisat's abrupt loss in April 2012.

The instrument was a Fourier transform spectrometer designed to detect limb emission

spectra from the middle and upper atmosphere (Fischer et al., 2008) as shown in Figure 2.3. Operating in the mid-infrared range (from $4.15\ \mu\text{m}$ to $14.6\ \mu\text{m}$, or $2410 - 685\ \text{cm}^{-1}$), MIPAS provided high spectral resolution. Initially, MIPAS operated at full resolution (FR) with a step size of up to $0.025\ \text{cm}^{-1}$, and with 17 limb views per scan (covering altitudes from 6 to 68 km), providing a vertical sampling of approximately 3 km in the troposphere and 3-8 km in the stratosphere. Following technical issues in 2004, MIPAS switched to optimized resolution (OR) in 2005 with a step size of up to $0.0625\ \text{cm}^{-1}$, increasing the number of views to 27 per scan (covering altitudes from 3 to 70 km) and allowing a vertical sampling of about 1.5 km. MIPAS maintained a 100% duty cycle during FR but initially dropped to 30% in OR, gradually recovering to full operation by 2007. Horizontally, the instrument had a field of view of about 30 km, with a sampling distance of nearly 500 km along the orbit, providing 72-75 vertical profiles per orbit (Dinelli et al., 2021).

Level 2 (L2) products from MIPAS spectral measurements serve as input for the CDF algorithm. Retrieval of these products is performed using the Optimised Retrieval Model (ORM) v8.22, which is the Level 2 processor developed for the final full-mission reanalysis of MIPAS data (Raspollini et al., 2022). The retrieval process follows an iterative approach where the difference between the radiance computed by the forward model for a given atmospheric state and the observations is minimised, using the Jacobian matrix, describing the sensitivity of the radiance to variations in the retrieved parameters. Regularization techniques are applied to stabilize the inversion and reduce uncertainties. The operational dataset includes vertical profiles of temperature, pressure, and 21 trace gases, such as H_2O (water), O_3 (ozone), HNO_3 (nitric acid), CH_4 (methane), N_2O (nitrous oxide), NO_2 (nitrogen dioxide), CFC-11 (trichlorofluoromethane), ClONO_2 (chlorine nitrate), N_2O_5 (dinitrogen pentoxide), CFC-12 (dichlorodifluoromethane), COF_2 (carbonyl fluoride), CCl_4 (carbon tetrachloride), CF_4 (carbon tetrafluoride), HCFC-22 (chlorodifluoromethane), C_2H_2 (acetylene), CH_3Cl (methyl chloride), COCl_2 (phosgene), C_2H_6 (ethane), OCS (carbonyl sulfide), and HDO (semi-heavy water). The retrieval methodology accounts for different observational phases of the mission, including Full Resolution (FR) (2002–2004) and Optimized Resolution (OR) (2005–2012), ensuring consistency across the entire dataset (Dinelli et al., 2021).

In this study, we reprocessed ozone MIPAS OR measurements from 2008 to 2012 using ORM v8.22, configured with the OE method as required by the CDF framework. The initial guess consists on the IG2 as described by Remedios et al., 2007, then the algorithm considers values of the previous scan as described in Raspollini et al., 2022. This reprocessed version, indeed, differs from the operational ESA Level 2 (L2 v8.22) dataset, which is retrieved with a Levenberg–Marquardt scheme combined with a posteriori Tikhonov regularization. The inputs for the CDF algorithm includes ozone state vectors, variance-covariance matrices (VCMs), averaging kernels (AKs), and the a priori information for each product, the a priori profiles. State vectors are composed of volume mixing ratio (VMR) ozone profiles, water continuum and offsets values. The reprocessed MIPAS ozone dataset has been validated to ensure its scientific reliability. Results of the validation are shown in Appendix D.

IASI

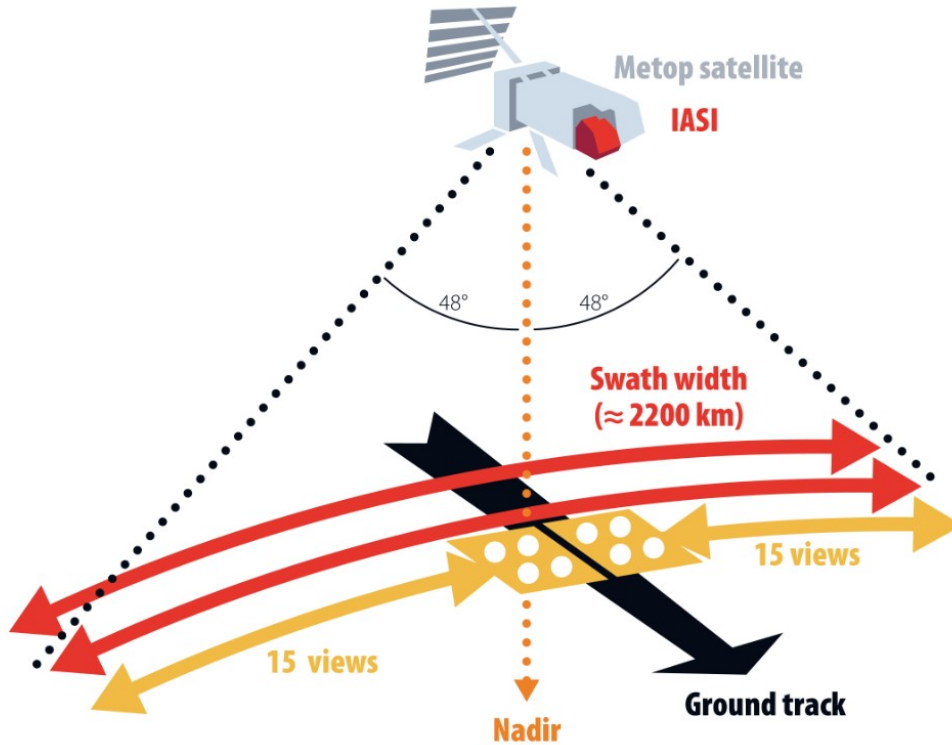


Figure 2.4: IASI observing mode. An instantaneous field of view of 4 circular pixels is taken every 50 km, over a swath width of 2×1100 km. Each circular pixel corresponds to a 12 km diameter footprint on the ground at nadir (Clerbaux et al., 2009).

The Infrared Atmospheric Sounding Interferometer (IASI) is a multi-purpose sounding instrument composed of a Fourier transform spectrometer based on Michelson interferometer (Blumstein et al., 2004). IASI has been providing continuous data since October 2006, when it was first launched on board MetOp-A. It was followed by IASI-B (MetOp-B), launched in 2012 and IASI-C (MetOp-C), launched in 2018. The three instruments (IASI -A, -B and -C) cover the spectral range $2760 - 645 \text{ cm}^{-1}$, with a spectral sampling of 0.25 cm^{-1} , for a total of 8461 spectral channels. IASI is a nadir viewing instrument, that scans the atmosphere symmetrically across-track with respect to the nadir direction in 30 scan steps, for a total swath width of ± 1100 km. The Field of Regard (FOR) of IASI comprises a 2×2 matrix of circular samples Instantaneous Field of View (IFOV). At the nadir direction, and a satellite altitude of 819 km, each IFOV corresponds to a ground resolution of approximately 12 km, although this size increases varying with the scan angle (ellipse-shaped). Each FOR is centered on the viewing direction and corresponds to a footprint of about $50 \text{ km} \times 50 \text{ km}$ at the nadir. The instrument scans in a step and stare mode and is able to acquire a total of 30×120 interferograms per scan line. IASI delivers vertically resolved atmospheric profiles of temperature and humidity and retrieves the concentrations of key greenhouse gases. Its data are critical for a wide range of applications, including improving numerical weather prediction (NWP), monitoring atmospheric composition, assessing air quality, and studying long-term climate variability (Clerbaux et al., 2009; Razavi et al., 2009; Coheur et al., 2009). Additionally, IASI has

been widely recognized for its capability to provide accurate radiance measurements in clear-sky conditions, making it a reference dataset for validating and evaluating climate models.

The ozone data used in this study come from the OZOS IASI-A archive, available on the IASI AERIS Data Portal. This dataset was generated using the Fast Optimal Retrievals on Layers for IASI (FORLI-O₃) processor. The FORLI processor is based on the OE (Optimal Estimation) method to derive atmospheric ozone profiles from IASI spectral data in the thermal infrared region (1000–1070 cm⁻¹), using the ozone climatology by McPeters et al., 2007 as a priori information (Hurtmans et al., 2012; Boynard et al., 2016). The operational dataset provides vertical ozone profiles in partial column density units (Dobson Units, DU) on 40 layers between surface and 40 km, with an extra layer from 40 to 60 km, the top of the atmosphere. In addition to ozone profiles, the product includes key diagnostic information, such as VCMs, AKs, and a priori partial column profile. However, the a priori VCM is provided separately and must be reconstructed as described in the Product User Manual (PUM) for Near real-time IASI total O₃ and O₃ profile. FORLI O₃ data have been available since 2008; in this study, we used data corresponding to the MIPAS activity period (2008–2012).

Key Differences in MIPAS and IASI Observations and Products

This section better highlights some key operational differences between MIPAS and IASI data in order to better understand the processing steps and methodological choices applied in this work. These differences concern aspects that strongly affect retrieval characteristics and their subsequent use in data fusion.

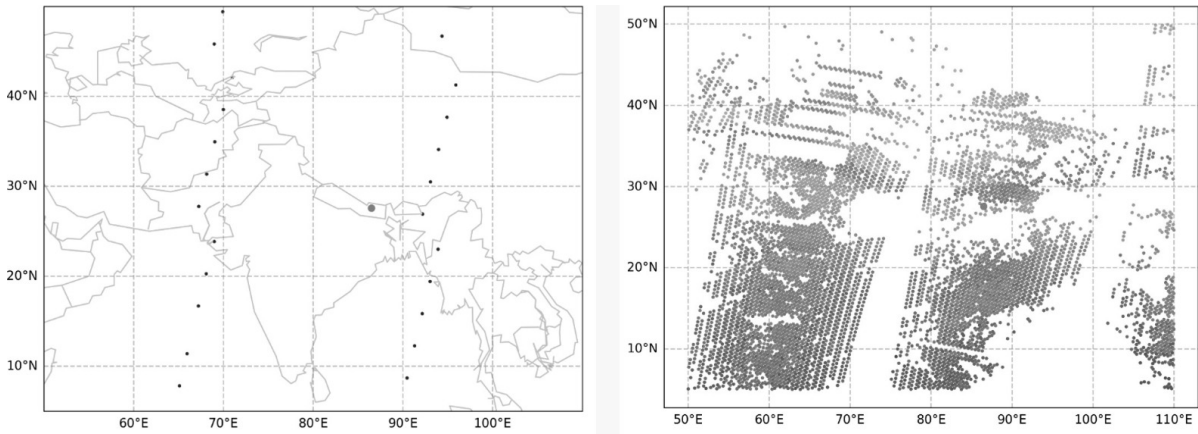


Figure 2.5: Example of MIPAS (left) and IASI (right) spatial distribution for one day in 2008.

The first and most fundamental difference between MIPAS and IASI concerns their viewing geometry. MIPAS operates in limb-emission mode, observing the atmosphere along slant paths. Conversely, IASI is a nadir-viewing instrument that measures the upwelling infrared radiance from below. These distinct geometries lead to markedly different vertical sensitivity of the two instruments, having MIPAS higher resolution in the upper troposphere and above, while IASI having higher resolution in the lower-middle troposphere. The second difference lies in the vertical grid and the structure of the state vectors. The MIPAS state vector consists of ozone in VMR, the continuum — additional parameters in-

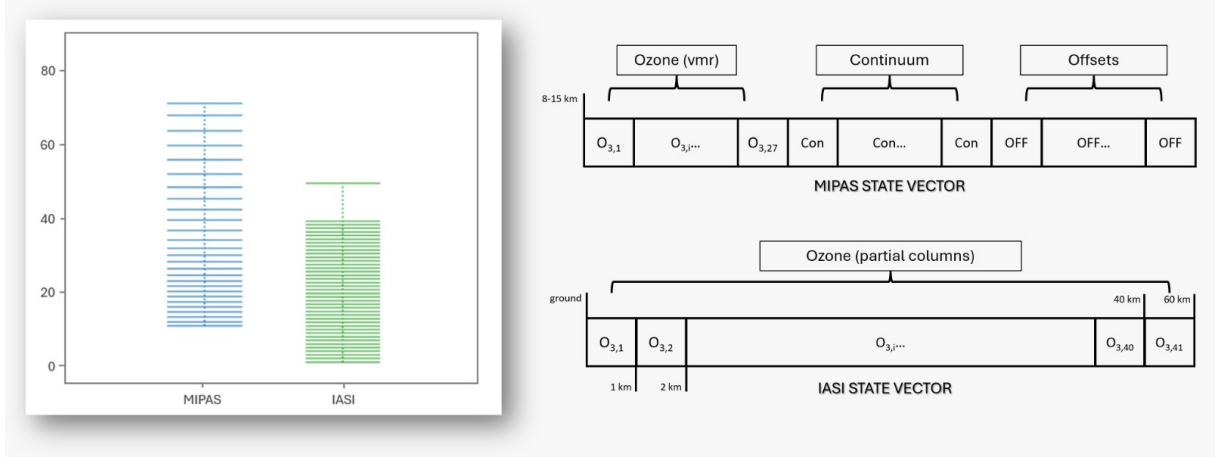


Figure 2.6: The left panels show an example of MIPAS (left-blue) and IASI (right-green) vertical grids. The right panels show MIPAS (right-above) and IASI (right-below) state vector scheme.

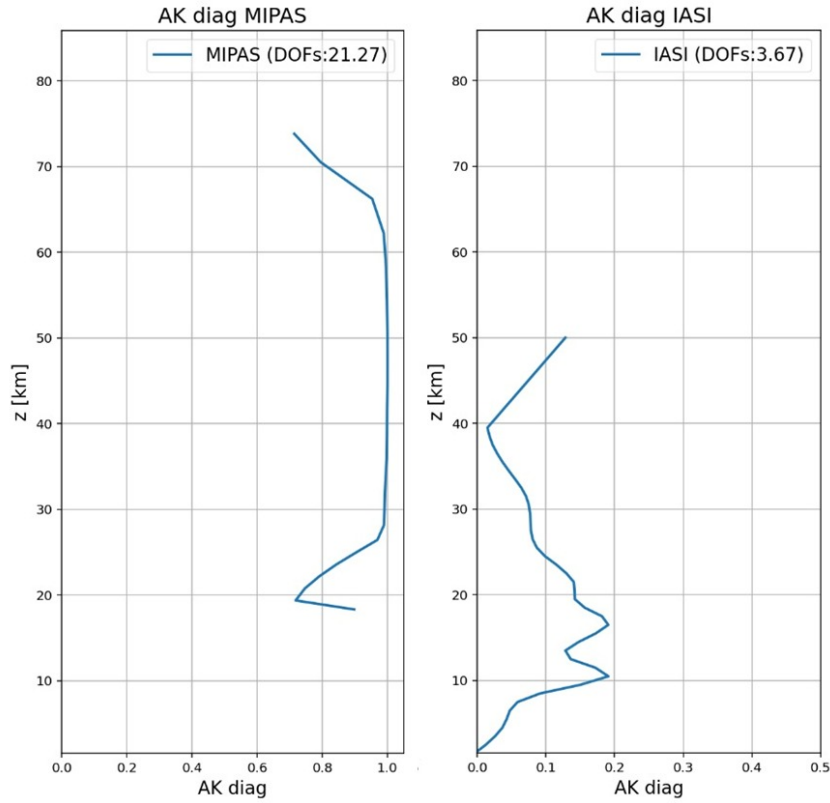


Figure 2.7: Example of diagonal elements of averaging kernel of MIPAS (left) and IASI (right) for the entire state vectors.

roduced to model broadband residual signals in the measured spectra, such as calibration uncertainties, aerosol contributions, and other spectrally smooth features not accounted for by line-by-line absorption calculations — and the offsets — parameters that account for systematic radiance biases by introducing an additive correction term that is constant across the spectral range of a given band. The ozone-related portion of the vector typically includes 27 elements, arranged on a non-uniform grid, with the profile generally

starting between 5 and 15 km (MIPAS does not measure below 5 km). The structure of the MIPAS grid and state vector is shown in Figure 2.6. The same figure also displays the IASI grid and its state vector, which is provided on a regular grid and consists solely of ozone partial column elements.

The third operational difference concerns horizontal sampling and spatial resolution. As shown in Figure 2.5, IASI provides dense global coverage with high horizontal sampling frequency, while MIPAS limb scans are more sparsely distributed, following the instrument's vertical scanning sequence along the orbit track. These differences have practical implications for co-location, fusion, and the handling of representativity errors.

A final observation regards the information content carried by the vertical profiles of the two instruments and how it is distributed vertically. Figure 2.7 illustrates the averaging kernel diagonals for a coincident pair of MIPAS and IASI state vectors. A clear difference can be seen in the number of degrees of freedom across the entire vector. In this case, the MIPAS profile started around 18 km (likely due to the presence of clouds) and shows its maximum information content in the stratosphere, where values approach unity. Conversely, IASI reaches down to the surface with peaks of information below 20 km; however, the corresponding values remain confined below 0.2.

2.2.2 Ozonsondes Data

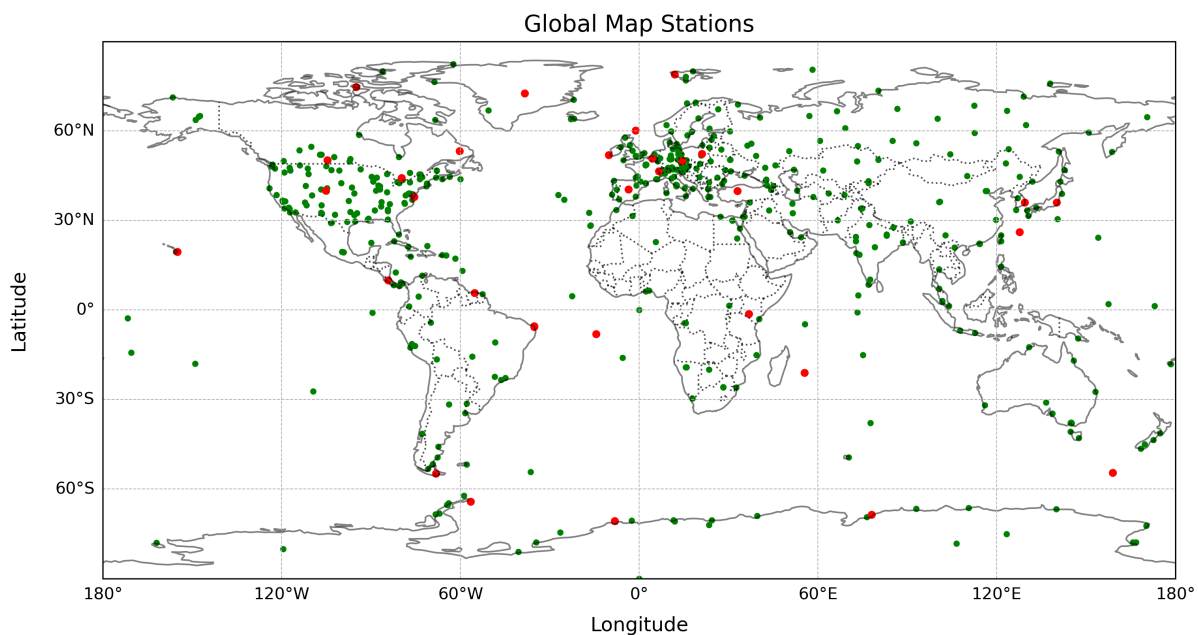


Figure 2.8: Global map of WOUDC stations (green markers). Red markers identify stations effectively used in this study.

The ozonsonde data used in this study to validate the newly developed ozone dataset were obtained from the World Ozone and Ultraviolet Radiation Data Centre (WOUDC), which collects and archives ozone profile measurements from ozonsondes launched at various locations worldwide. These data are publicly available in the WOUDC archive and can be downloaded by following the instructions provided on the Data Access webpage.

A total of 33 ozonsonde stations were selected for this analysis based on their spatiotemporal coincidence with the fused ozone profiles and the individual MIPAS and IASI

datasets (± 200 km and ± 1 h). These stations are highlighted in red in Figure 2.8. As shown, the stations spatial distribution is primarily concentrated between 30°N and 60°N , with a higher density over Europe and the United States.

2.3 Methods

2.3.1 The Complete Data Fusion (CDF): Theory and Application

This section provides a comprehensive description of the CDF method, organized into seven subsections. The first five subsections introduce the method and its theoretical basis, while the sixth presents some practical considerations and application criteria. The last two subsections describe the CDF setup and the tuning carried out in this study. To reserve the 2009–2011 period for validation, both setup and tuning were performed using 2008 data only. These two phases are presented separately to clearly distinguish the setup—essential for the implementation of CDF and closely tied to input data quality—from the tuning, which enhances the method by allowing parameter adjustments.

Method Motivations and Advantages

The growing number of satellite missions, as well as airborne and ground-based campaigns, has increased the availability of data to estimate the vertical profiles of atmospheric quantities and their distributions. A large amount of data has required the development of algorithms to fuse data together in order to obtain the greatest amount of information. When two or more instruments sound the same portion of the atmosphere and observe the same species in different spectral regions or with different geometries, two approaches can be used to estimate the best vertical profile: the synergistic retrieval and the a posteriori data fusion methods, which includes the *Complete Data Fusion* (CDF). The synergistic retrieval uses all the observations acquired by different instruments as input of the retrieval algorithm, which provides a single profile, with enhanced accuracy and information with respect to the retrieval of the observation acquired by a single instrument. Using the data fusion method, retrievals are done from the observation acquired by the single instrument, and then, a fusion algorithm combines a posteriori the retrieved profiles into a single one. The first approach provides the best estimate of the profile because it takes into account all the interactions among different inputs. However, it is difficult to implement since it requires a model able to simulate all the observations and makes use of a lot of data. On the contrary, the a posteriori methods are less accurate than synergistic retrieval but overcome the operational limits of managing high volumes of data, and are of simpler implementation. The CDF method is based on the a posteriori approach and uses standard retrieval products obtained with the Optimal Estimation method. Over time, the original CDF formulas have been modified, in order to generalize them, overcoming limits due to singularities in matrices. The 'Complete' term means that CFD takes into account all the features of the measurements, by using the variance-covariance matrix (VCM) and the averaging kernel matrix (AKM) of each retrieved profile.

Mathematical Approach

We suppose to have N sets of radiance measurements ($\mathbf{y}_i$; $i=1, \dots, N$) from different instruments Ceccherini et al., 2015. These measurements are independent, simultaneous, and referred to the same geolocation. We can perform N retrievals, obtaining retrieved vertical profile vectors ($\hat{\mathbf{x}}_i$; $i=1, \dots, N$) of a certain atmospheric component. Therefore, we have N independent estimates of the profile on a common grid. Each profile vector is characterized by:

- the covariance matrix indicated as $\mathbf{S}_i$, *i.e.* the mean value of the $\boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T$, where $\boldsymbol{\sigma}_i$ is the vertical profile error:

$$\mathbf{S}_i = \langle \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T \rangle. \quad (2.7)$$

The expression of $\mathbf{S}_i$ is the one in 1.25b and we have seen that it is the sum of two terms. The first term is the covariance matrix of the noise errors $\mathbf{S}_{ni}$ 1.43. The second term is the covariance matrix of the smoothing error $\mathbf{S}_{si}$ 1.42;

- the averaging kernel matrix $\mathbf{A}_i$, *i.e.* the derivatives of the components of the retrieved profiles with respect to the components of the true profile 1.37:

$$\mathbf{A}_i = \frac{\partial \hat{\mathbf{x}}_i}{\partial \mathbf{x}_{\text{true}}}. \quad (2.8)$$

Considering a linear problem or a moderately non-linear one, the solution will be linear too, so we can expand each retrieved profile around the a priori profile, exploiting the definition of the averaging kernel matrix, and rewriting the equation (1.38):

$$\hat{\mathbf{x}}_i = \mathbf{x}_{ai} + \mathbf{A}_i(\mathbf{x}_{\text{true}} - \mathbf{x}_{ai}) + \boldsymbol{\sigma}_{ni}. \quad (2.9)$$

From which, developing calculations, we can define the following quantities:

$$\boxed{\boldsymbol{\alpha}_i \equiv \hat{\mathbf{x}}_i - (\mathbf{I} - \mathbf{A}_i)\mathbf{x}_{ai} = \mathbf{A}_i\mathbf{x}_{\text{true}} + \boldsymbol{\sigma}_{ni}}. \quad (2.10)$$

The quantity $\boldsymbol{\alpha}_i$ is the retrieved profile reduced by the modulation of the a priori one, so we are removing the effect of the a priori constraint. It is a measure of the true profile modulated by the AKM and, when $\mathbf{A}_i$ is the identity matrix, we have that $\boldsymbol{\alpha}_i = \mathbf{x}_{\text{true}} + \boldsymbol{\sigma}_{ni}$. Moreover, $\boldsymbol{\alpha}_i$ can be calculated from the available retrieval products and is affected by the same error $\boldsymbol{\sigma}_{ni}$.

At this point we have to find the $\mathbf{x}_{\text{true}}$ that satisfies the following system:

$$\begin{cases} \boldsymbol{\alpha}_1 \simeq \mathbf{A}_1 \mathbf{x} \\ \boldsymbol{\alpha}_2 \simeq \mathbf{A}_2 \mathbf{x} \\ \dots \\ \boldsymbol{\alpha}_N \simeq \mathbf{A}_N \mathbf{x} \end{cases} \quad (2.11)$$

To find $\mathbf{x}$, here called $\mathbf{x}_f$, we require it to minimize the cost function $c(\mathbf{x})$, defined as:

$$c(\mathbf{x}) = \left[\sum_{i=1}^N (\boldsymbol{\alpha}_i - \mathbf{A}_i \mathbf{x})^T \mathbf{S}_{ni}^{-1} (\boldsymbol{\alpha}_i - \mathbf{A}_i \mathbf{x}) \right] + (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a). \quad (2.12)$$

Minimizing $c(\mathbf{x})$ we find $\mathbf{x}_f$, and therefore, the associated $\mathbf{S}_{nf}$ and $\mathbf{A}_f$:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} x_a \right), \quad (2.13a)$$

$$\mathbf{S}_{nf} = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1}, \quad (2.13b)$$

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i \right). \quad (2.13c)$$

Proof.

- To find $\mathbf{x}_{nf}$ it is necessary minimizing the cost function 2.12.

$$\frac{\partial c(\mathbf{x})}{\partial \mathbf{x}} = 0. \quad (2.14)$$

Developing the derivative, we have:

$$\left[\sum_{i=1}^N -2\mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\alpha}_i + 2\mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i \mathbf{x} \right] - 2\mathbf{S}_a^{-1} x_a + 2\mathbf{S}_a^{-1} \mathbf{x} = 0. \quad (2.15)$$

Therefore, we can write:

$$\left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right) \mathbf{x} = \sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} x_a. \quad (2.16)$$

Form which

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} x_a \right). \quad (2.17)$$

- To find $\mathbf{S}_{nf}$ it is first of all necessary to propagate the error into $\mathbf{x}_f$, calculating:

$$\Delta \mathbf{x}_f = \frac{\partial \mathbf{x}_f}{\partial \boldsymbol{\alpha}} \boldsymbol{\sigma}_\alpha, \quad (2.18)$$

where, being $\boldsymbol{\alpha} = \mathbf{A}\mathbf{x} + \boldsymbol{\sigma}_{ni}$, we have that $\boldsymbol{\sigma}_\alpha = \boldsymbol{\sigma}_{ni}$. Developing the 2.32, we obtain that:

$$\Delta \mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\sigma}_{ni} \right). \quad (2.19)$$

Now we are able to calculate $\mathbf{S}_{nf}$ as:

$$\mathbf{S}_{nf} = \langle \Delta \mathbf{x}_f \Delta \mathbf{x}_f^T \rangle =$$

$$= \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1} \left(\sum_{i,j=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} < \boldsymbol{\sigma}_{\mathbf{n}i} \boldsymbol{\sigma}_{\mathbf{n}j}^T > \mathbf{S}_{\mathbf{n}j}^{-1} \mathbf{A}_j^T \right) \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1}. \quad (2.20)$$

Being $< \boldsymbol{\sigma}_{\mathbf{n}i} \boldsymbol{\sigma}_{\mathbf{n}j}^T > = 0$ if $i \neq j$, and $< \boldsymbol{\sigma}_{\mathbf{n}i} \boldsymbol{\sigma}_{\mathbf{n}i}^T > = \mathbf{S}_{\mathbf{n}i}$, the 2.20 becomes:

$$\mathbf{S}_{\mathbf{n}f} = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1}. \quad (2.21)$$

- $\mathbf{A}_f$ is the sensitivity of the retrieved profile to the true profile. Therefore:

$$\mathbf{A}_f = \frac{\partial \mathbf{x}_f}{\partial \mathbf{x}_{\text{true}}}. \quad (2.22)$$

Substituting the expression for $\mathbf{x}_f$, we find:

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \frac{\partial \boldsymbol{\alpha}_i}{\partial \mathbf{x}_{\text{true}}} \right). \quad (2.23)$$

Being $\frac{\partial \boldsymbol{\alpha}_i}{\partial \mathbf{x}_{\text{true}}} = \mathbf{A}_i$, we have that:

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{A}_i^T \mathbf{S}_{\mathbf{n}i}^{-1} \mathbf{A}_i \right). \quad (2.24)$$

□

It is possible to demonstrate (see Appendix B) that the solution obtained with the CDF is the same as the solution obtained with the synergistic retrieval method if the forward model is linear in $\mathbf{x}$, within the range of variability of the single retrieval Ceccherini et al., 2015.

Generalization of the CDF formulas

The inverse of the noise VCMs ($\mathbf{S}_{\mathbf{n}i}$) in the CDF formulas in eqs 2.13 can be a problem if matrices are singular Ceccherini et al., 2022. In this case, formulas can be applied only considering the generalized inverse instead of the inverse of the noise VCMs. The generalized inverse, however, is only an approximation, and, in some cases, the noise is amplified or there is a loss of information. Nevertheless, it is possible to rewrite eqs 2.13 in a generalized form, using the total VCMs, in the form of eq 1.25b, instead of the noise VCMs. For each retrieval $i = 1, \dots, N$, we have that the noise CM can be written as:

$$\mathbf{S}_{\mathbf{n}i} = \mathbf{S}_i \mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}i}^{-1} \mathbf{K}_i \mathbf{S}_i. \quad (2.25)$$

Recalling eq 1.37, the i -th AKM can be written as:

$$\mathbf{A}_i = \mathbf{S}_i \mathbf{K}_i^T \mathbf{S}_{\mathbf{ny}i}^{-1} \mathbf{K}_i, \quad (2.26)$$

the i -th VCM can be written as:

$$\mathbf{S}_{ni} = \mathbf{A}_i \mathbf{S}_i. \quad (2.27)$$

Assuming that the noise VCMs are non-singular, we have that:

$$\mathbf{S}_i^{-1} = \mathbf{A}_i^T \mathbf{S}_{ni}^{-1}. \quad (2.28)$$

Replacing the eq 2.28 in the eqs 2.13, we find:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right), \quad (2.29)$$

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i \right). \quad (2.30)$$

$$\mathbf{S}_{nf} = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1}. \quad (2.31)$$

In order to calculate the total VCM of the fused profile we have to sum the fused noise VCM to the fused smoothing VCM.

The smoothing error VCM of the fused profile ($\mathbf{S}_{sf}$) can be obtained as made for the noise error VCM of the fused profile. Firstly, we need to propagate the error from $\mathbf{x}_a$ to $\mathbf{x}_f$, calculating:

$$\Delta \mathbf{x}_{sf} = \frac{\partial \mathbf{x}_f}{\partial \mathbf{x}_a} \boldsymbol{\sigma}_a = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \mathbf{S}_a^{-1} \boldsymbol{\sigma}_a. \quad (2.32)$$

where we have used eq 2.29. Now we are able to calculate $\mathbf{S}_{sf}$ as:

$$\begin{aligned} \mathbf{S}_{sf} &= \langle \Delta \mathbf{x}_{sf} \Delta \mathbf{x}_{sf}^T \rangle = \\ &= \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \mathbf{S}_a^{-1} \langle \boldsymbol{\sigma}_a \boldsymbol{\sigma}_a^T \rangle \mathbf{S}_a^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} = \\ &= \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \mathbf{S}_a^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1}. \end{aligned} \quad (2.33)$$

The CM of the total error associated with the fusion profile will be:

$$\mathbf{S}_f = \mathbf{S}_{nf} + \mathbf{S}_{sf} = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1}. \quad (2.34)$$

Therefore, we have the generalized CDF formula:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right), \quad (2.35a)$$

$$\mathbf{S}_f = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1}, \quad (2.35b)$$

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{S}_i^{-1} \mathbf{A}_i \right). \quad (2.35c)$$

The equations in 2.35 do not require the hypothesis that noise VCMs are non-singular anymore, hence we can conclude that these equations are always valid.

Error Analysis: Interpolation and Coincidence Errors

As presented in the previous Section, the CDF method works well if measurements refer to the same true profile and are retrieved on the same vertical grid. If these conditions are not verified, the CDF algorithm has to be corrected in order to manage these cases Ceccherini et al., 2018. We will refer to the effect of the first case as *coincidence error*, and to the effect of the second case as *interpolation error*. The coincidence error affects the method when profiles, which do have not the same space-time location, are fused. The interpolation error is due to the fact that the averaging kernel matrices (AKMs) on the fusion grid are an interpolation of the original AKMs, hence an approximation of the real AKMs on the fusion grid.

The introduction of the CDF method equation was made considering the assumption of measurements referring to the same geolocation and on a common grid. Therefore, the CDF formulas, eq. 2.13, involve a summation of quantities that require to have the AKMs of different retrievals represented on the same grid. If the profiles are retrieved on different vertical grids, even their AKMs are defined on different vertical grids. In this case, a resampling of each AKM is needed. For each retrieval, we can define a matrix $\mathbf{H}_i$, which interpolates the profiles from the retrieval grid to the fusion grid. Therefore, we can use this matrix to represent the AKM on the fusion grid:

$$\mathbf{A}'_i = \mathbf{A}_i \mathbf{H}_i^\# \quad (2.36)$$

where $\mathbf{H}_i^\#$ is the generalized inverse matrix of $\mathbf{H}_i$. This interpolation is performed only for AKMs, and not for $\boldsymbol{\alpha}_i$ or VCMs. Considering this transformation we have that the equation (2.13a) becomes:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \mathbf{A}_i \mathbf{H}_i^\# + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T \mathbf{S}_{ni}^{-1} \boldsymbol{\alpha}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right). \quad (2.37)$$

From here in after, we will use the following notation: $\mathbf{x}_i^{(i)}$ to indicate the true profile related to the i -th measurement on its starting grid; $\mathbf{x}_i^{(f)}$ to indicate the true profile related to the i -th measurement on the fusion grid, here indicated by (f) ; $\mathbf{x}_i$ to indicate the true profile related to the i -th measurement represented on a fine grid, which includes all the levels of the fusion grid and of each i -th grid. Defining $\mathbf{C}^{(i)}$ the sampling matrix from the fine grid to the i -th one, and $\mathbf{C}^{(f)}$ the sampling matrix from the fine grid to the fusion one, we can write:

$$\begin{aligned} \mathbf{x}_i^{(i)} &= \mathbf{C}^{(i)} \mathbf{x}_i, \\ \mathbf{x}_i^{(f)} &= \mathbf{C}^{(f)} \mathbf{x}_i \end{aligned} \quad (2.38)$$

Recalling the equation 2.10:

$$\boldsymbol{\alpha}_i = \mathbf{A}_i \mathbf{x}_i^{(i)} + \boldsymbol{\sigma}_{ni}, \quad (2.39)$$

we can write it as:

$$\boldsymbol{\alpha}_i = \mathbf{A}_i \mathbf{H}_i^\# \mathbf{x}_i^{(f)} + \mathbf{A}_i (\mathbf{x}_i^{(i)} - \mathbf{H}_i^\# \mathbf{x}_i^{(f)}) + \boldsymbol{\sigma}_{ni}, \quad (2.40)$$

and therefore:

$$\alpha_i = \mathbf{A}_i \mathbf{H}_i^\# \mathbf{C}^{(f)} \mathbf{x}_i + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \mathbf{x}_i + \sigma_{ni}, \quad (2.41)$$

As previously observed, the coincidence error affects measurements made in different space-time locations. In this case, they are fused only if lying within a given coincidence criterion. Different space-time areas present different true profiles $\mathbf{x}_i$, and with the data fusion we can aim to determine the mean value of these true profiles, here indicated by $\bar{\mathbf{x}}$. Every single true profile can be written as:

$$\mathbf{x}_i = \bar{\mathbf{x}} + \sigma_{i,\text{coin}}, \quad (2.42)$$

where $\sigma_{i,\text{coin}}$ is the difference between $\mathbf{x}_i$ and the mean true profile.

We can now rewrite the equation 2.41:

$$\begin{aligned} \alpha_i &= \mathbf{A}_i \mathbf{H}_i^\# \mathbf{C}^{(f)} \bar{\mathbf{x}} + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \bar{\mathbf{x}} + \mathbf{A}_i \mathbf{C}^{(i)} \sigma_{i,\text{coin}} + \sigma_{ni} = \\ &= \mathbf{A}_i \mathbf{H}_i^\# \bar{\mathbf{x}}^{(f)} + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \bar{\mathbf{x}} + \mathbf{A}_i \mathbf{C}^{(i)} \sigma_{i,\text{coin}} + \sigma_{ni}. \end{aligned} \quad (2.43)$$

Correction of the CDF formulation

Considering the mean true profile in equation 2.42 as the sum of the a priori profile and the difference from it:

$$\bar{\mathbf{x}} = \mathbf{x}_a + \sigma_a \quad (2.44)$$

the equation 2.43 can be written as:

$$\alpha_i = \mathbf{A}_i \mathbf{H}_i^\# \bar{\mathbf{x}}^{(f)} + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \mathbf{x}_a + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \sigma_a + \mathbf{A}_i \mathbf{C}^{(i)} \sigma_{i,\text{coin}} + \sigma_{ni}. \quad (2.45)$$

We can define the quantity $\tilde{\alpha}_i$, measurement of the mean true profile evaluated on the fusion grid, as:

$$\begin{aligned} \tilde{\alpha}_i &\equiv \alpha_i - \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \mathbf{x}_a = \\ &\mathbf{A}_i \mathbf{H}_i^\# \bar{\mathbf{x}}^{(f)} + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \sigma_a + \mathbf{A}_i \mathbf{C}^{(i)} \sigma_{i,\text{coin}} + \sigma_{ni}. \end{aligned} \quad (2.46)$$

Each quantity defined in 2.46 is a measurement of the mean true profile on the fusion grid, smoothed by the re-sampling of the AKM $\mathbf{A}_i \mathbf{H}_i^\#$, plus error terms.

The first error term is the interpolation error. We can calculate its covariance matrix $\mathbf{S}_{i,\text{int}}$:

$$\begin{aligned} \mathbf{S}_{i,\text{int}} &= \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) < \sigma_a \sigma_a^T > (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)})^T \mathbf{A}_i^T = \\ &= \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)}) \mathbf{S}_a (\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)})^T \mathbf{A}_i^T. \end{aligned} \quad (2.47)$$

The second error term is the coincidence error. The covariance matrix associated to the coincidence error $\mathbf{S}_{i,\text{coin}}$ will depend on the variability of the true profile with the location, being $\bar{\mathbf{x}}$ the mean true profile. Therefore, we introduce the covariance matrix $\mathbf{S}_{\text{coin}}$, which is associated to $\sigma_{i,\text{coin}}$, and obtain the following expression of $\mathbf{S}_{i,\text{coin}}$:

$$\mathbf{S}_{i,\text{coin}} = \mathbf{A}_i \mathbf{C}^{(i)} < \sigma_{i,\text{coin}} \sigma_{i,\text{coin}}^T > \mathbf{C}^{(i)T} \mathbf{A}_i^T = \mathbf{A}_i \mathbf{C}^{(i)} \mathbf{S}_{\text{coin}} \mathbf{C}^{(i)T} \mathbf{A}_i^T. \quad (2.48)$$

The coincidence uncertainty reflects the fact that two profiles measured at different locations or times may not correspond to the same true atmospheric state. Its covariance matrix ($\mathbf{S}_{\text{coin}}$) explicitly accounts for the natural variability of the atmosphere within the coincidence window. A large variability indicates that the fused profile cannot perfectly represent both measurements, requiring the algorithm to balance their relative contributions. Similarly, the interpolation error arises because the averaging kernels from different instruments must be projected on a common fusion grid. Any mismatch between the original and fused representations contributes to $\mathbf{S}_{i,\text{int}}$ which captures the information loss introduced by this necessary regridding step. These error sources—coincidence and interpolation—are incorporated into the generalized noise covariance matrix. Consequently, they directly influence the fusion outcome: elevated uncertainties from either effect diminish the instrument's weighting in the solution and elevate the final uncertainty of the fused profile.

In conclusion, the final formula of CDF, which takes into account coincidence and interpolation errors is the equation (2.37), which we recall with α_i replaced by $\tilde{\alpha}_i$, and $\mathbf{S}_{ni}$ replaced by $\tilde{\mathbf{S}}_{ni}$:

$$x_{\mathbf{f}} = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T \tilde{\mathbf{S}}_{ni}^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_{\mathbf{a}}^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T \tilde{\mathbf{S}}_{ni}^{-1} \tilde{\alpha}_i + \mathbf{S}_{\mathbf{a}}^{-1} x_{\mathbf{a}} \right), \quad (2.49)$$

where:

$$\tilde{\alpha}_i \equiv \alpha_i - \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^{\#} \mathbf{C}^{(f)}) x_{\mathbf{a}}, \quad (2.50)$$

and

$$\tilde{\mathbf{S}}_{ni} = \mathbf{S}_{ni} + \mathbf{S}_{i,\text{int}} + \mathbf{S}_{i,\text{coin}}. \quad (2.51)$$

Both effects of coincidence and interpolation errors enter the generalized noise covariance $\tilde{\mathbf{S}}_{ni}$ and directly influence the fused product: higher coincidence or interpolation uncertainties reduce the weight of a given instrument in the fusion and increase the final uncertainty of the fused profile.

Correction of the generalized CDF formulation

In this section, we will broaden the interpolation and coincidence errors discussion to the generalized CDF formulas 2.35 Ceccherini et al., 2022. The equation 2.49 contains the problem of the inverse of the $\tilde{\mathbf{S}}_{ni}$ matrix, shown in eq 2.52 (exploiting eqs 2.51, 2.47, and 2.48), when it is singular:

$$\tilde{\mathbf{S}}_{ni} = \mathbf{S}_{ni} + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^{\#} \mathbf{C}^{(f)}) \mathbf{S}_{\mathbf{a}} (\mathbf{C}^{(i)} - \mathbf{H}_i^{\#} \mathbf{C}^{(f)})^T \mathbf{A}_i^T + \mathbf{A}_i \mathbf{C}^{(i)} \mathbf{S}_{\text{coin}} \mathbf{C}^{(i)T} \mathbf{A}_i^T. \quad (2.52)$$

In order to overcome this limit, we define the matrix $\tilde{\mathbf{S}}_i$:

$$\tilde{\mathbf{S}}_i = \mathbf{S}_i + \mathbf{A}_i (\mathbf{C}^{(i)} - \mathbf{H}_i^{\#} \mathbf{C}^{(f)}) \mathbf{S}_{\mathbf{a}} (\mathbf{C}^{(i)} - \mathbf{H}_i^{\#} \mathbf{C}^{(f)})^T + \mathbf{A}_i \mathbf{C}^{(i)} \mathbf{S}_{\text{coin}} \mathbf{C}^{(i)T}. \quad (2.53)$$

Since $\mathbf{S}_{ni} = \mathbf{S}_i \mathbf{A}_i^T$, exploiting eq 2.52, we can write that $\tilde{\mathbf{S}}_{ni} = \tilde{\mathbf{S}}_i \mathbf{A}_i^T$.

Therefore, replacing $\tilde{\mathbf{S}}_{ni}$ with $\mathbf{S}_i \mathbf{A}_i^T$, eq 2.49 can be written as:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T (\tilde{\mathbf{S}}_i \mathbf{A}_i^T)^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T (\tilde{\mathbf{S}}_i \mathbf{A}_i^T)^{-1} \tilde{\boldsymbol{\alpha}}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right), \quad (2.54)$$

If $\tilde{\mathbf{S}}_{ni}$ is nonsingular, $\tilde{\mathbf{S}}_i$ and $\mathbf{A}_i$ are also non singular, and the eq (2.55) becomes the new CDF formula for the fused profile:

$$\mathbf{x}_f = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \tilde{\boldsymbol{\alpha}}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right). \quad (2.55)$$

Since in this equation $\tilde{\mathbf{S}}_{ni}^{-1}$ does not appear anymore, we will assume valid eq 2.55 also in the case that $\tilde{\mathbf{S}}_{ni}$ is singular.

The new AKM formula can be obtained by exploiting its definition:

$$\mathbf{A}_f = \frac{\partial \mathbf{x}_f}{\partial \bar{\mathbf{x}}} = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \frac{\partial \tilde{\boldsymbol{\alpha}}_i}{\partial \bar{\mathbf{x}}}. \quad (2.56)$$

where we remind that $\bar{\mathbf{x}}$ is the unknown true profile of the data fusion, which can be the mean value of the true profiles of the measurements that are fused. Using eq 2.46, we write the new AKM formula as:

$$\mathbf{A}_f = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#}. \quad (2.57)$$

In conclusion, we write the new CM formula associated with the fused profile. It can be demonstrated that it has a form close to the eq 2.34:

$$\mathbf{S}_f = \mathbf{S}_{nf} + \mathbf{S}_{sf} = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1}. \quad (2.58)$$

Proof. The expression for $\mathbf{S}_{nf}$ in the presence of coincidence and interpolation errors can be found following the proof of eq 2.13b. Firstly, it is necessary to propagate the error into $\mathbf{x}_f$, calculating:

$$\Delta \mathbf{x}_f = \sum_{i=1}^N \frac{\partial \mathbf{x}_f}{\partial \tilde{\boldsymbol{\alpha}}_i} \sigma_{\tilde{\boldsymbol{\alpha}}_i} \quad (2.59)$$

Exploiting that the noise VCMs associated with $\tilde{\boldsymbol{\alpha}}_i$ are $\tilde{\mathbf{S}}_{ni}$, and recalling that $\mathbf{S}_{nf} = < \Delta \mathbf{x}_f \Delta \mathbf{x}_f^T >$, we can find:

$$\begin{aligned} \mathbf{S}_{nf} = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \tilde{\mathbf{S}}_{ni} (\tilde{\mathbf{S}}_i^{-1})^T \mathbf{H}_i^{\#} \\ \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \mathbf{A}_i^T (\tilde{\mathbf{S}}_i^{-1})^T \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1}. \end{aligned} \quad (2.60)$$

The product $\tilde{\mathbf{S}}_i^{-1}\tilde{\mathbf{S}}_{ni}(\tilde{\mathbf{S}}_i^{-1})^T$ is symmetric, hence, knowing that $\tilde{\mathbf{S}}_{ni} = \mathbf{A}_i\tilde{\mathbf{S}}_i^T$, we can write:

$$\tilde{\mathbf{S}}_i^{-1}\tilde{\mathbf{S}}_{ni}(\tilde{\mathbf{S}}_i^{-1})^T = \tilde{\mathbf{S}}_i^{-1}\mathbf{A}_i\tilde{\mathbf{S}}_i^T(\tilde{\mathbf{S}}_i^{-1})^T = \tilde{\mathbf{S}}_i^{-1}\mathbf{A}_i = \mathbf{A}_i^T(\tilde{\mathbf{S}}_i^{-1})^T. \quad (2.61)$$

where the last equality is obtained by making the transpose and exploiting the fact that the matrix $\tilde{\mathbf{S}}_i^{-1}\tilde{\mathbf{S}}_{ni}(\tilde{\mathbf{S}}_i^{-1})^T$ is symmetric. Replacing eq 2.61 in eq 2.60, we find:

$$\mathbf{S}_{nf} = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1}. \quad (2.62)$$

In the same way, it is possible to obtain the smoothing error $\mathbf{S}_{sf}$. Firstly, we calculate:

$$\Delta \mathbf{x}_{sf} = \frac{\partial \mathbf{x}_f}{\partial \mathbf{x}_a} \boldsymbol{\sigma}_a \quad (2.63)$$

Therefore:

$$\begin{aligned} \mathbf{S}_{sf} &= \langle \Delta \mathbf{x}_{sf} \Delta \mathbf{x}_{sf}^T \rangle = \\ &= \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1} \mathbf{S}_a^{-1} \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1}. \end{aligned} \quad (2.64)$$

The sum of $\mathbf{S}_{nf}$ and $\mathbf{S}_{sf}$ will be $\mathbf{S}_f$:

$$\mathbf{S}_f = \left(\sum_{i=1}^N \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^{\#} + \mathbf{S}_a^{-1} \right)^{-1}. \quad (2.65)$$

□

CDF in this Work and Some Considerations

In this work, we applied the generalized CDF formulation to spatially and temporally coincident pairs of IASI and MIPAS Level 2 profiles. Coincidence was defined by thresholds of 200 km and 1 hour. As no prior studies have applied CDF to real MIPAS-IASI ozone retrievals, there are no established coincidence criteria for this specific limb–nadir combination. The chosen thresholds were therefore guided by physical and methodological considerations rather than predefined standards. In particular, STE events typically extend over several hundred kilometres and evolve over timescales of a few hours. A 200 km spatial window is narrow enough to avoid pairing unrelated air masses, while remaining compatible with the intrinsic MIPAS limb footprint and its 400 km along-track sampling. Similarly, a 1-hour temporal window ensures the atmosphere can be considered quasi-stationary in the UTLS without excessively reducing the number of valid matchups. It is important to note that a MIPAS limb measurement does not correspond to a single point but to an extended line-of-sight, whose horizontal extent varies with tangent altitude and latitude. The geolocation in the Level 2 product represents the tangent point—the centre of the sampled air mass. Consequently, the adopted coincidence window should not be interpreted as requiring strict volumetric overlap, but as a physically meaningful criterion to ensure broad representativeness for limb–nadir fusion.

Because MIPAS has coarser spatial sampling than IASI, we centered the fusion on the IASI profiles. As a result, each MIPAS profile can be fused multiple times with several nearby

IASI profiles. This strategy is non-trivial: it is designed to preserve the spatial density of IASI while still leveraging the vertical information content of MIPAS. This approach has both benefits and drawbacks. On the one hand, it enhances the spatial representativity of the fused dataset, allowing for denser spatial sampling around MIPAS tracks. On the other hand, it introduces some redundancy in the stratospheric contribution from MIPAS, which is replicated across multiple fused profiles. Nonetheless, given the scientific objectives of this work, this choice remains well justified.

Due to the non-coincident nature of the profiles, we adopted the generalized formulation of the CDF method, which accounts for both interpolation and coincidence errors (Ceccherini et al., 2022, 2018, 2022).

As equations above suggest, the performance of the CDF method depends on the quality and consistency of the input profiles. However, it offers several advantages. Firstly, it allows for greater control over the a priori constraint imposed on the fused product, which can theoretically be chosen independently of the a priori constraints of the Level 2 products being merged. Secondly, it handles profiles on different grids, and allows for fusion onto an arbitrary target grid. By combining information from the Level 2 products while removing their individual constraints, the method enables the generation of a product with potentially higher information content and lower uncertainty. As a diagnostic method for the fusion process compared to each L2 input, the number of degrees of freedom (DOFs), given by the trace of the AK, can serve as a measure of vertical information content. Moreover, the diagonal elements of the AK can be analyzed to assess whether and where vertical resolution increases, and errors can be evaluated as the square root of the diagonal elements of the variance-covariance matrix. Finally, the CDF method inherently requires consistency among input data, thereby also providing a means for verifying their quality.

CDF Setup

The MIPAS and IASI datasets differ not only in their vertical resolution and retrieval characteristics but also in their underlying state vectors and covariance structures. To ensure a meaningful and numerically stable fusion, a preliminary harmonization step is required. This setup aims to bring the two datasets into a common framework, both in terms of variables and representation. Therefore, we harmonized the data through two main steps:

State-vector extension.

We extended the IASI state vector to include the water continuum and offset parameters present in the MIPAS state vector but not directly retrieved by IASI. Following Tirelli et al., 2021, we appended zero-value entries to IASI's AK and added zeros on the off-diagonal and 10^{-2} -values on the diagonal outside the profile. This choice reflects the structure of the MIPAS a priori CM and avoids numerical instabilities. The extension ensures that both datasets share the same state vector dimensionality, which is a prerequisite for applying the fusion algorithm consistently. Without this adjustment, the mismatch in variable definitions would prevent the computation of meaningful error covariances and would invalidate the fused output.

Partial-column conversion.

We converted MIPAS profile values to moles per unit volume and then integrated them vertically between consecutive levels using the Levels As Layers (LAL) algorithm on a regular z grid (van Peet, 2018). Interpolation matrices were constructed on partial columns, assuming constant density within each layer and computing the fractional overlap between original and new layers.

CDF Tuning

CDF version	Interpolation CM Source	Off-diagonal	Coincidence CM Source	Off-diagonal	k
v00	FUSED a priori CM		Fused a priori CM		0.05
v01	FUSED a priori CM	correlation length	Fused a priori CM		
v02	A priori L2		A priori L2		
v03	A priori L2		A priori L2		
v04	A priori L2		A priori L2	correlation length	
v05	A priori L2		Fused a priori CM		
v06	A priori L2		Fused a priori CM	correlation length	
v07	A priori L2	correlation length	A priori L2		0.025
v08	A priori L2		A priori L2	correlation length	0.01

Table 2.1: Summary of the nine selected CDF versions based on different choices of interpolation and coincidence covariance matrices (CM), use of off-diagonal reconstruction with a correlation length of 6 km, and application of a factor k on the coincidence CM.

The CDF algorithm offers flexibility in tuning several parameters. One of the strengths of the fusion process lies in its ability to define both the a priori constraint and the vertical grid for the fused product. Prior to performing the fusion, we conducted a set of tests in which we modified both the a priori constraint and the vertical grid of the input datasets. These tests showed that such modifications resulted in stable outputs for MIPAS, while for IASI, they introduced instabilities, likely due to the characteristics of the corresponding covariance (CM) and averaging kernel (AK) matrices. In this study, we selected the IASI vertical grid as the fusion grid and adopted the IASI a priori constraint for the fused product. This choice is scientifically justified for several reasons. First, since the study focuses on the UTLS and lower troposphere, the IASI grid — by design optimized for these atmospheric regions — offers a natural basis for fusion. Secondly, as IASI is one of the input datasets, retaining its native grid ensures compatibility with its associated AKs and CMs, minimizing the introduction of numerical instabilities observed in prior tests. To ensure consistency, the IASI a priori constraint — composed of both the a priori profile and its associated CM — was extended by including the off-profile elements from the MIPAS a priori state vector. For the corresponding CM, the off-diagonal elements

were set to zero, and the diagonal elements outside the profile were assigned the values from the corresponding diagonal elements in the MIPAS a priori CM.

The other two key configurable parameters are the interpolation and coincidence errors. These form the basis for tuning the fusion setup. To systematically evaluate the impact of these structural choices on the final retrieval, we tested nine distinct CDF configurations (v00–v08). These configurations differ in the construction of the CDF operators, specifically in the formulation of the interpolation covariance matrix and the coincidence covariance matrix. The matrices were implemented using either the fused a priori covariance matrix or the original Level 2 a priori covariance matrices, and were tested both in their native form and with reconstructed off-diagonal elements. The nine versions are summarized in Table 2.1. For each version, we applied the CDF algorithm for the year 2008. The table reports the sources of the covariance matrices (CM) used to model coincidence and interpolation errors. When the Fused a priori CM is indicated, the same a priori CM from the fused profiles was used for all inputs to model additional errors. In contrast, 'A priori L2' refers to the use of the individual a priori CM specific to each L2 product. In some versions, off-diagonal elements were reconstructed by applying a correlation function with a fixed correlation length of 6 km. This reconstruction was implemented using the following formulation:

$$CM_{ij} = \exp\left(-\frac{|z_i - z_j|}{\text{len_corr}}\right) \sigma_i \sigma_j \quad (2.66)$$

where σ_i and σ_j are the standard deviation at levels i and j , z_i and z_j are the midpoints of the interpolation grid, and $\text{len_corr} = 6$ km. Conversely, when not specified, CM matrix is considered without modification. The choice of a 6 km correlation length is consistent with previous studies applying the Complete Data Fusion method to ozone profiles (Ceccherini et al., 2018; Tirelli et al., 2021). This value reflects the typical vertical scale of ozone variability and is commonly used to model the correlation structure of retrieval errors in satellite-based profile measurements.

Finally, some versions apply a factor k to the coincidence covariance matrix, effectively scaling it with different strengths.

To identify the most suitable configuration, we considered the results of validation and comparison against L2 products for the year 2008. Metrics, including degrees of freedom (DOFs) for the full profile and below 20 km, total retrieval errors, and the diagonal elements of the averaging kernel matrix (AKM), were evaluated over a large dataset of approximately 7000 profiles.

2.3.2 Validation Methodology

To assess the reliability and accuracy of the fused ozone profiles, we used ozonesonde observations as an independent reference. Ozonesondes provide high-vertical-resolution measurements and are commonly considered a standard for ozone profile validation. Ozonesondes data from WOUDC (World Ozone and Ultraviolet Radiation Data Centre) stations of 2008 were used to tune the method, while data from the 2009–2012 period were used to validate the entire dataset. Ozonesondes profiles were converted into partial columns, and we assumed a 15% uncertainty for each sonde profile in all our metrics. This value is consistent with previous studies that report typical uncertainties ranging from 5%

to 20%, depending on altitude, instrument type, and data processing methods. In particular, Sterling et al., 2018 analyzed NOAA ozonesonde data and found that uncertainties in the troposphere can reach up to 20%, while remaining below 6% in the stratosphere. Therefore, a 15% uncertainty represents a conservative and realistic estimate, especially for tropospheric ozone. Moreover, the validation was performed using coincident fused-ozonesonde profiles, after smoothing the ozonesonde profiles with the AK of the fused profile:

$$\mathbf{x}_{sonde}^{smoothed} = \mathbf{x}_a + \mathbf{A}(\mathbf{x}_{sonde} - \mathbf{x}_a). \quad (2.67)$$

where $\mathbf{x}_a$ is the a priori profile, $\mathbf{A}$ is the AK, and $\mathbf{x}_{sonde}$ is the ozonesonde profile. For the tuning phase, only stations between 5°N and 55°N were considered. This latitudinal range provides a sufficient number of coincidences for the fusing step, while avoiding outliers due to large latitudinal differences. Spatial and temporal thresholds for coincidences between fused profiles and ozonesondes were set at maximum distance of 200 km from the station, and temporal interval of 3 hours before and after the starting and the ending time of the sounding. These thresholds represent a deliberate compromise, balancing the atmospheric representativeness of the paired observations against the practical necessity of obtaining a sufficient number of matchups. They are fully consistent with standard validation practice for satellite ozone products, which commonly employ spatial windows of 100–300 km and temporal windows of ± 3 hours. The chosen 200 km distance is further motivated by the intrinsic characteristics of MIPAS limb measurements. The instrument's limb footprint naturally extends over several hundred kilometers, and its along-track sampling distance is approximately 400 km. Consequently, a narrower spatial criterion would not be physically meaningful for this geometry and would drastically reduce the number of usable coincidences. For each coincident fused-sounding profile, we determined the absolute and percentage bias. The metrics used for the validation are the mean percentage bias over the entire latitudinal band, the mean standard deviation, the mean standard error, and the composite error. This last one should be directly comparable with the standard deviation, and is obtained as the quadratic sum of the noise of the fused product and the error on the ozonesonde. For the sake of clarity, we summarize the relevant metrics below. The mean percentage bias can be determined as the mean absolute bias divided by the mean ozonesonde profile:

$$\langle bias_{percent,i} \rangle = 100 \times \frac{\langle x_{f,i} - x_{sonde,i} \rangle}{\langle x_{sonde,i} \rangle}, \quad (2.68)$$

where $\langle \rangle$ stands for arithmetic mean, and index i refers to each layer. The percentage mean standard deviation is calculated as:

$$\langle \sigma_{percent,i} \rangle = 100 \times \frac{\langle \sigma_i \rangle}{\langle x_{sonde,i} \rangle}, \quad (2.69)$$

where $\langle \sigma_i \rangle$ is the mean standard deviation of the i -th layer. The variance deriving from the composite error is the square of:

$$\langle err_i^{comp} \rangle = \sqrt{\mathbf{S}_{f,ii}^{noise} + err_{sonde,i}^2}, \quad (2.70)$$

where $\mathbf{S}_{f,ii}^{noise}$ is the diagonal noise element of the fused CM, and $err_{sonde,i}$ is the error of the ozonesonde profile. Then the associated percentage error can be obtained as:

$$\langle err_{percent,i}^{comp} \rangle = 100 \times \frac{\langle err_i^{comp} \rangle}{\langle x_{sonde,i} \rangle}. \quad (2.71)$$

For the validation over the 2009–2011 period, the same criteria and metrics were applied. In this case, ozonesonde stations were grouped into latitudinal bands in addition to the 5°N–55°N band. The aim of this grouping was to investigate the performance of the fused product across different latitudinal zones and to provide a spatial characterization of the dataset.

The choice of the latitudinal bands was driven by the number of profile coincidences available within each band. Due to the limited number of coincidences, the following broad latitudinal bands were selected to ensure a minimum level of statistical representativeness in each zone: (–75°S to –25°S), (–25°S to 25°N), and (25°N to 75°N). This latitudinal segmentation also allowed us to evaluate the geographical robustness of the fusion method, especially in regions with known variability in ozone vertical structure such as the tropics and mid-latitudes.

2.3.3 Comparison Methodology

Comparative analyses were carried out to assess the performance of the fused product against each individual Level 2 input dataset, with a specific focus on the tuning performed for the year 2008, and extended more generally to the other years.

We compared the fused product and each L2 dataset with ozonesonde measurements. The procedure was similar to the validation process. Indeed, operationally, the distinction between validation and comparison lies in the application of the instrument-specific averaging kernel (AK). For the validation of the fused product, ozonesonde profiles are smoothed with the fused AK so that the reference and the satellite share the same vertical sensitivity. However, in a multi-sensor comparison, applying different AKs to the same ozonesonde profile would produce three different “versions” of the reference dataset, shaped by the vertical sensitivity of each instrument rather than by the atmosphere itself. This would contaminate the intercomparison with method induced differences. This choice avoids the introduction of instrument-dependent artifacts and ensures that all comparisons are made on the same vertical resolution baseline. This step was important for assessing the quality of each L2 input dataset and the fused product, and for applying potential corrections if needed.

2.4 Results and Discussion

This section presents and discusses the results of the ozone profile fusion analysis, with the aim of evaluating both the performance of the proposed method and the impact of bias corrections. Results are reported following the logical sequence of the workflow, beginning with the tuning of the fusion parameters, continuing with the validation against ozonesonde measurements, and with a comparative assessment of different fusion versions. Discussion is integrated throughout the section to provide immediate interpretation of the findings in the context of previous studies and the physical processes involved.

2.4.1 Tuning Phase Results

As previously mentioned, the tuning phase was performed using data from 2008 over the 5°–55°N latitude band. This phase focused on the initial validation of the different versions of the fusion method against ozonesonde profiles, on the analysis of DOFs, and on the retrieval errors of the fused and L2 products.

Figure 2.9 shows the validation results for the nine proposed versions of the fusion method. All versions exhibit a bias peak around 10 km, that will be later discuss in detail; however, this peak is reduced (by less than 25%) in version v05. Bias alone is not sufficient to determine the best-performing version. Due to the different constraints applied in each version, some profiles exhibit oscillatory behavior or even negative values. Therefore, the number of profiles containing negative values was also considered as an additional criterion for evaluating the quality of each version. As a result, each version may consequently have a different number of valid profiles. The quantity n in Figure 2.9 shows, for each version, the number of profiles contributing to the mean. More strongly constrained versions generally exhibit fewer negative values. However, a stronger constraint typically reduces the information content of the profiles, and thus the number of DOFs. To improve the statistical robustness of the analysis, this aspect together with DOFs was also assessed using a larger dataset, selecting six random days in 2008 and including all fused profiles within the region, rather than only those coincident with ozonesondes.

Version	N tot	N neg (%)	Fused (tot & <20 km)	MIPAS (tot & <20 km)	IASI (tot & <20 km)
v00	7067	22.9	8.81 (3.50)	8.09 (2.75)	3.29 (1.85)
v01		0.8	8.13 (3.28)	7.47 (2.53)	
v02		0.0	6.29 (3.12)	6.31 (2.45)	
v03		0.4	7.66 (3.55)	8.74 (2.96)	
v04		1.5	7.78 (3.68)	8.74 (2.96)	
v05		11.6	9.41 (3.64)	8.74 (2.96)	
v06		10.6	9.21 (3.71)	8.74 (2.96)	
v07		4.0	8.78 (3.35)	8.74 (2.96)	
v08		0.3	7.17 (3.71)	6.31 (2.45)	

Table 2.2: DOFs and number of profiles containing negative values analysis. Header is composed of: the total number of fused profiles in 6 days (N_{tot}), the percentage of fused profiles with more than one negative value (N_{neg}), and the mean number of degrees of freedom both over the full profile and below 20 km, for both the fused ($Fused (tot \& u20 km)$) product and for L2 ($MIPAS (tot \& u20 km)$; $IASI (tot \& u20 km)$) products.

Table 2.2 presents, for each version, the total number of profiles considered, the percentage of fused profiles containing at least one negative value, and the mean number of DOFs, both over the full vertical range and below 20 km. These metrics are reported for the fused product as well as for the original Level 2 datasets.

Since the original Level 2 products are defined on different vertical grids and may rely on different a priori information, their DOFs are not directly comparable. To enable a consistent comparison, MIPAS profiles were interpolated onto the common fusion grid and a shared a priori using the CDF algorithm. This step harmonizes vertical resolution and alignment, reducing discrepancies due to differing retrieval configurations and enabling a

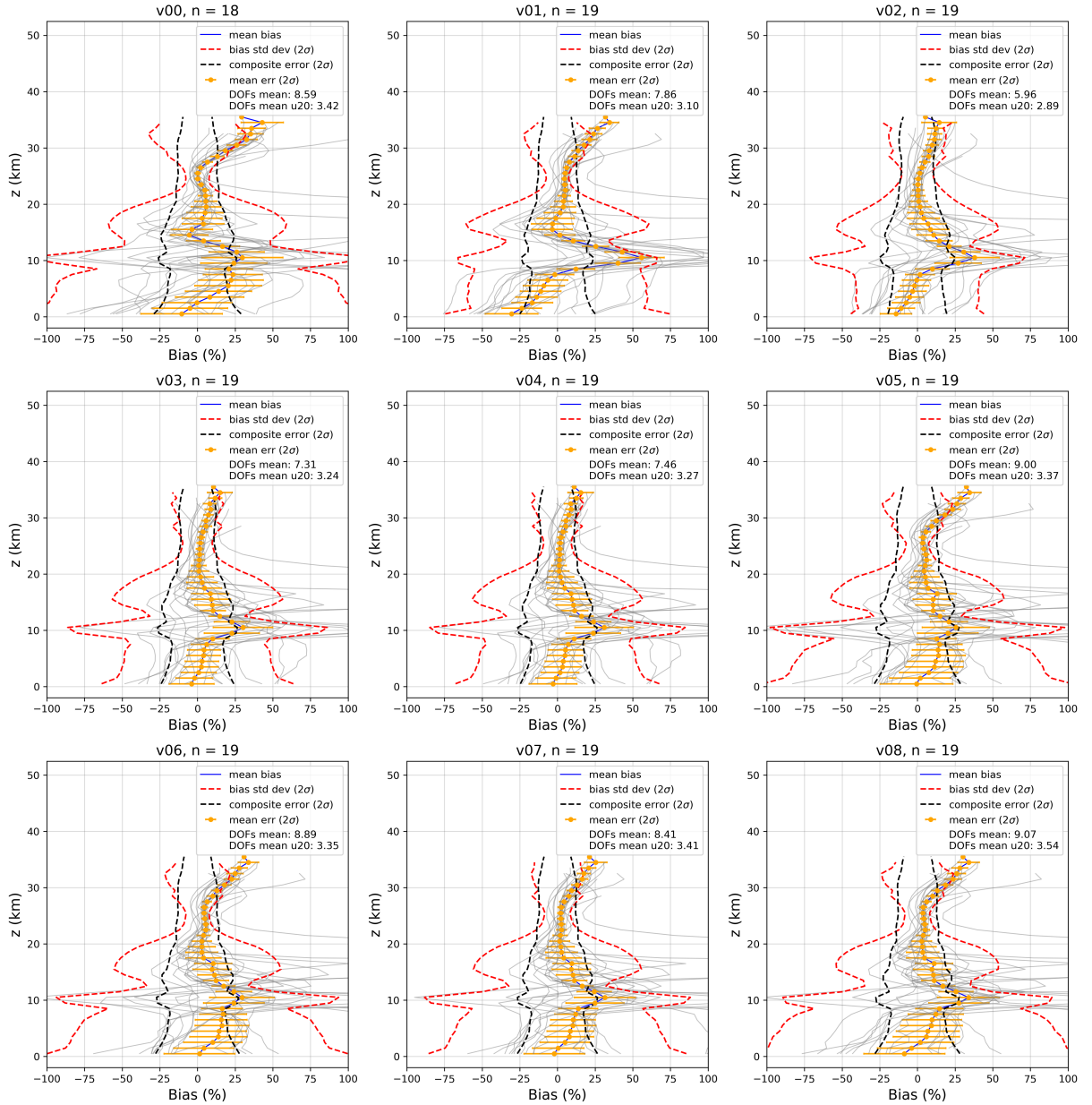


Figure 2.9: Validation of the 9 versions of the CDF method in the tuning phase, performed in 2008 within the latitudinal band 5° - 55° N . The blue line represents the average percentage bias between the fused profiles and the smoothed ozonesondes with the AK of the fused profile. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), while the orange bars show the standard error of the mean. The background gray profiles represent the n averaged profiles. In the legend, DOFs mean u20 represents the average number of degrees of freedom below 20 km.

meaningful comparison of the DOF values across datasets.

Among all versions, v00, followed by v05 and v06, shows the highest percentages of profiles with at least one negative value. Nevertheless, these versions, together with v01, perform best in terms of DOFs and agreement with MIPAS and IASI. This is especially evident for v05, which shows the highest DOFs across the full vertical range and a notable

gain in DOFs (+0.67) over MIPAS, while also offering a good compromise in the lower atmosphere (+ 0.68 below 20 km). For MIPAS, values below 5 km are reconstructed using the CDF algorithm, since the instrument does not provide measurements below this altitude. Therefore, the actual improvement in vertical information content should be primarily assessed with respect to IASI, which provides 3.29 degrees of freedom over the full profile, of which 1.85 are below 20 km.

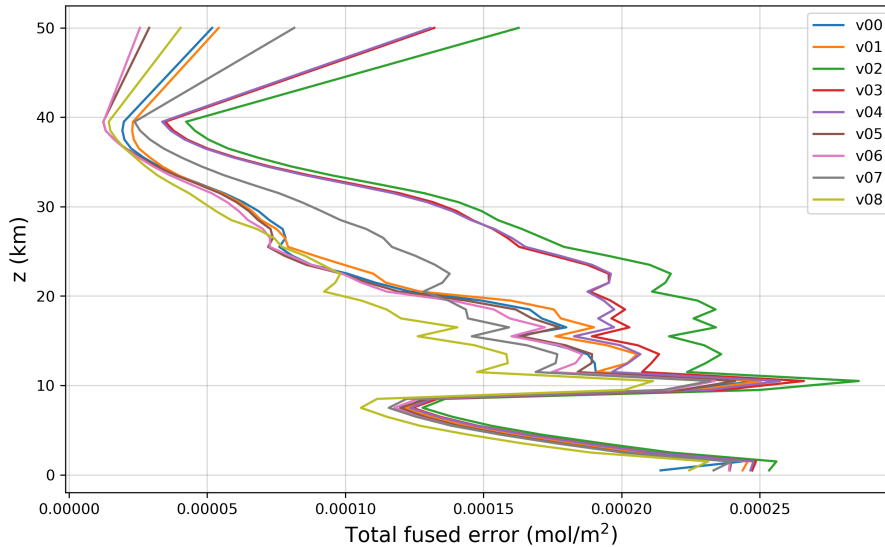


Figure 2.10: Mean total errors of fused products for each CDF version.

These results clearly demonstrate how limb measurements can significantly improve nadir observations by increasing the vertical information content. Notably, the increase in DOFs is observed not only in the altitude range directly sampled by MIPAS, but also in regions where MIPAS does not provide measurements. This is due to the vertical correlations introduced by the fusion process, which propagate information from the limb instrument across the vertical profile. This outcome is fully consistent with one of the key objectives of our work, namely to exploit the complementary sensitivities of limb and nadir sensors to enhance the vertical information content of the resulting product, especially in the troposphere and in the UTLS.

Fusing individual retrievals from MIPAS and IASI enables us to quantify the precise information contributed by each instrument, detailing how the synergy between their viewing geometries improves the overall result. The gain in Degrees of Freedom (DOFs) for the fused product demonstrates a more efficient exploitation of their complementary vertical sensitivities. Fundamentally, the algorithm achieves this by merging the instruments' averaging kernels, thereby reducing the solution's dependence on separate a priori constraints and synthesizing a vertically more informative profile.

Finally, looking at the fused total errors in Figure 2.10, an error peak is visible around 10 km. From version v05 onwards (through v08), the errors are lower throughout the entire profile compared to earlier versions.

Based on all the considerations discussed above, version v05 was selected for further analysis, as it offers the most balanced trade-off, especially with regard to bias, degrees of

freedom, and retrieval errors.

2.4.2 V05: Comparison between Fused and L2

Version v05 was subsequently applied to the remaining years.

Figure 2.11 shows the comparison between fused and L2 for 2008-2011. As can be seen, the fused product shows an improvement over the L2 products across all years, starting from the fact that it provides information from the surface to high altitudes. Based on a range-dependent analysis, the following considerations can be drawn.

Above 30 km, the difference with ozonesondes increases, reaching a positive bias exceeding 100% in all cases. However, at these altitudes, the accuracy of ozonesonde measurements decreases due to the very low pressure, and this bias should not be considered fully reliable (Stauffer et al., 2014).

Between 15 and 30 km, the fused profile shows smaller differences with ozonesondes than MIPAS, and even more so compared to IASI. In this altitude range, the fusion is mostly driven by MIPAS, leading to improved performance relative to IASI. Notably, the fused DOFs are equal to those of MIPAS, indicating that the fusion does not increase the retrieval sensitivity in this range. However, the influence of IASI is still evident in the comparison with ozonesondes, indicating a potential impact on the profile shape or bias characteristics.

Between 20 and 30 km, MIPAS contributes the largest number of measurements. Below this range, its decreases, as shown in the right panels of Figure 2.11.

Below 15 km, the contribution of IASI to the fused product increases, and the fused profile tends to follow the IASI profile. Below 5 km, the information comes almost entirely from IASI since MIPAS provide no measurements. Nonetheless, even below 15 km, the information propagated from higher layers, where MIPAS data is available, improves the IASI-based retrieval.

All these aspects become clearer and more quantifiable when examining the DOFs distributed over multiple altitude ranges for the fused and L2 products over years (Table 2.3).

Product	2008			2009			2010			2011		
	Δz_1	Δz_2	Δz_3	Δz_1	Δz_2	Δz_3	Δz_1	Δz_2	Δz_3	Δz_1	Δz_2	Δz_3
Fused	0.25	3.24	2.27	0.29	3.39	2.22	0.36	3.59	2.27	0.3	3.01	2.29
MIPAS	0.00	2.66	2.28	0.00	2.88	2.22	0.00	3.25	2.27	0	2.41	2.29
IASI	0.19	1.63	0.74	0.25	1.66	0.74	0.32	1.65	0.74	0.27	1.56	0.73

Table 2.3: Mean DOFs for different altitude ranges (Δz_1 : 0–5km, Δz_2 : 5–20km, Δz_3 : 20–30km), and for each product and year.

Finally, the comparison between L2 and fused mean total errors (Figure 2.12) shows that the fused product exhibits lower errors than each L2 product in the UTLS region, and more generally below 18 km in all years. Above 18 km, the fused error is much lower than that of IASI and comparable to that of MIPAS, where MIPAS performs best. The slightly higher fused error compared to MIPAS is likely due to additional uncertainties

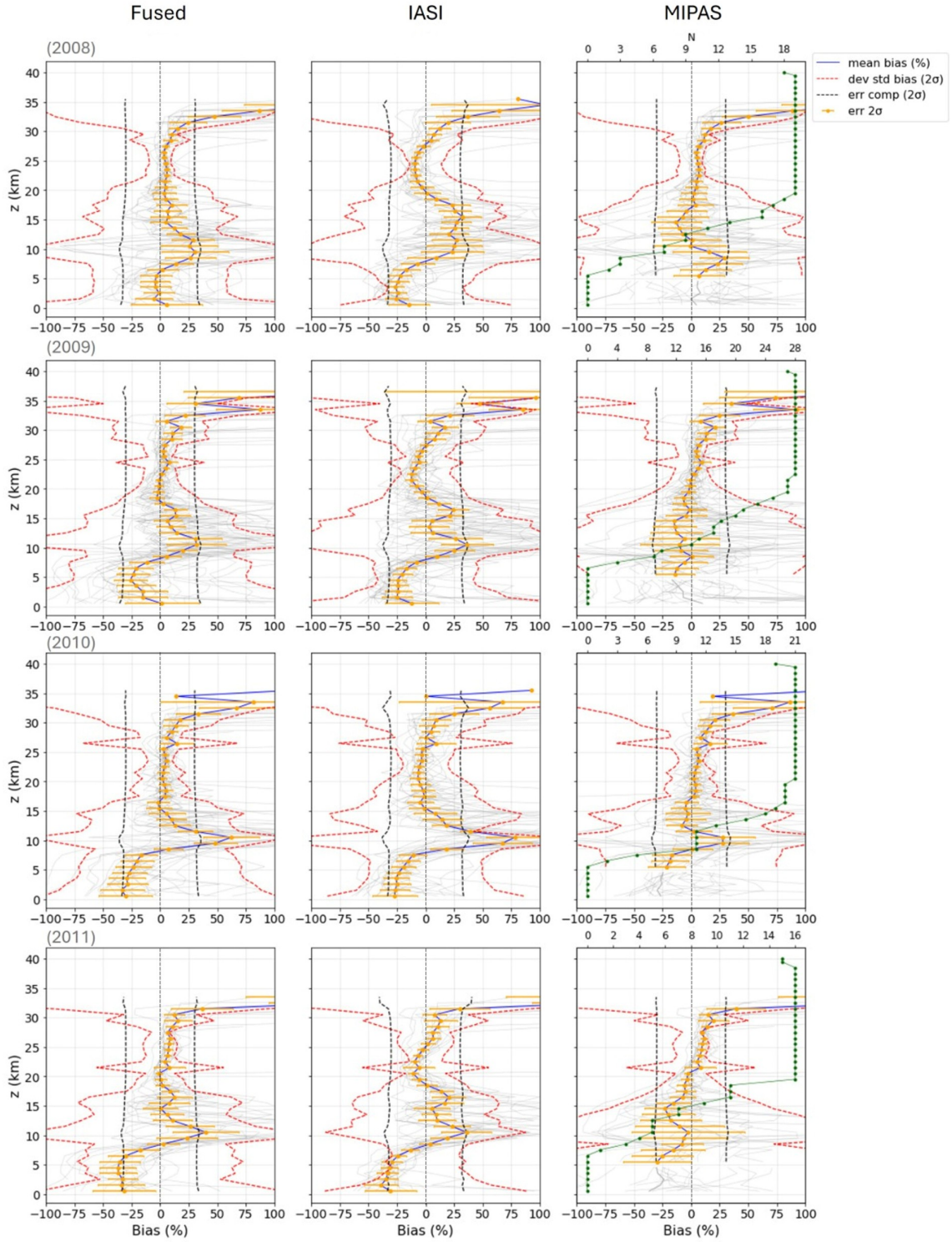


Figure 2.11: Comparison between fused, IASI, MIPAS and ozonsondes for version v05 over all years. The blue line represents the average percentage bias between fused or L2 and ozonsonde, where the ozonsonde has not been smoothed with the AK. The dashed red line represents the standard deviation of the mean at 2σ . The dashed black line represents the combined error (2σ), while the orange bars represent the standard error of the mean. The transparent gray profiles are the averaged profiles. The green profile represents N , which is the number of MIPAS measurements at each point of the vertical grid.

introduced during CDF processing, such as co-location and interpolation errors.

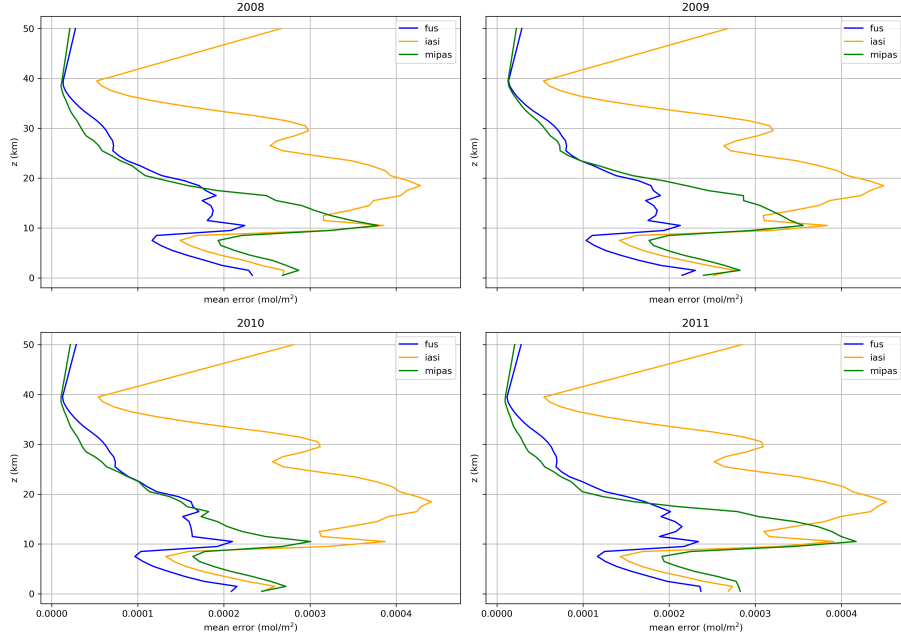


Figure 2.12: Comparison of the mean total errors of v05 fused products and L2 for each year.

2.4.3 V05: Validation

Validation over the 5–55°N latitudinal band for the remaining years 2009–2011 confirms the key features of the fused dataset already highlighted previously (Figure 2.13). The fused product exhibits a significant bias across most altitude levels. Between 10 and 30 km, the bias is positive and always below 25%, peaking around 10 km. Below 8 km, the fused product shows a significant negative bias, also remaining within 25%. It should be noted that the total number of coincidences is reduced by the subsampling strategy adopted for IASI: to limit the computational load while ensuring monthly representativeness, only the single day per month with the highest number of global ozonesonde launches was processed. As a consequence, the 2009–2011 validation period includes 65 coincident ozonesonde–fused profiles.

The validation across three latitudinal bands for the same years (Figure 2.14) allows assessment of the spatial variability of the bias in the fused product, although this analysis is limited by the number of coincident observations available in each region.

In the latitude band -75° to -25° South, which covers most of the Southern Hemisphere including Antarctica and sub-Antarctic regions, the statistics based on 28 profiles show a positive bias of less than 25% throughout the vertical range, although it is not significant below 8 km and between 20–30 km. A significant peak of approximately 35% is observed around 10 km.

In the -25° to 25° latitude band, which includes tropical and subtropical regions, the statistics rely on 10 coincidences, and the bias shows much stronger variability. It is negative below 12 km and becomes significant around 10 km, reaching a minimum of

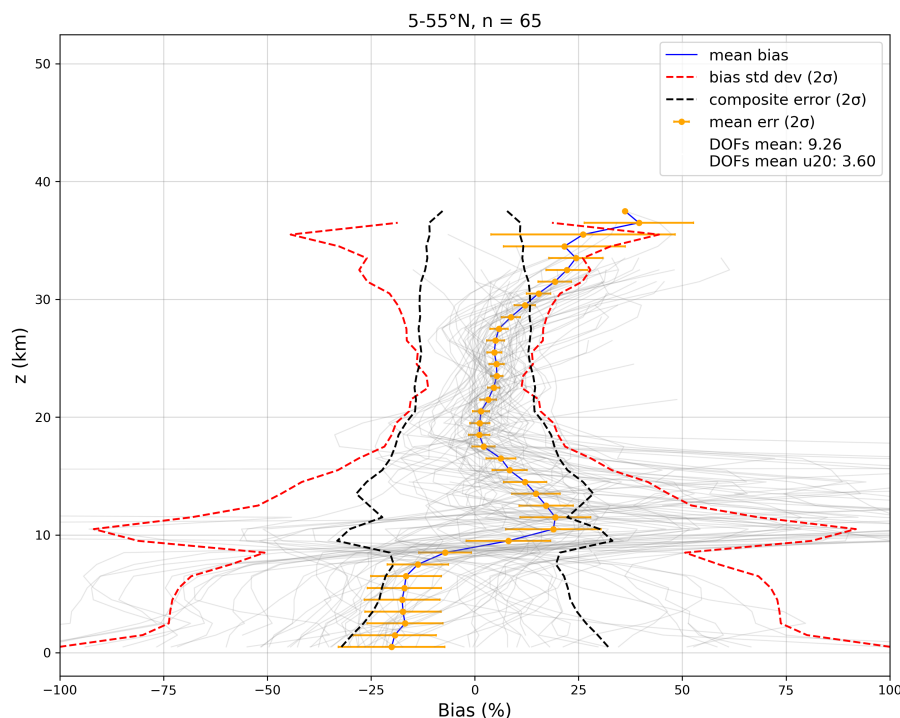


Figure 2.13: Validation of version v05. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed with the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), while the black dashed line represents the combined error (2σ). The orange bars show the standard error of the mean. The transparent grey lines correspond to the individual averaged profiles. Here n indicates the total number of valid fused profiles considered.

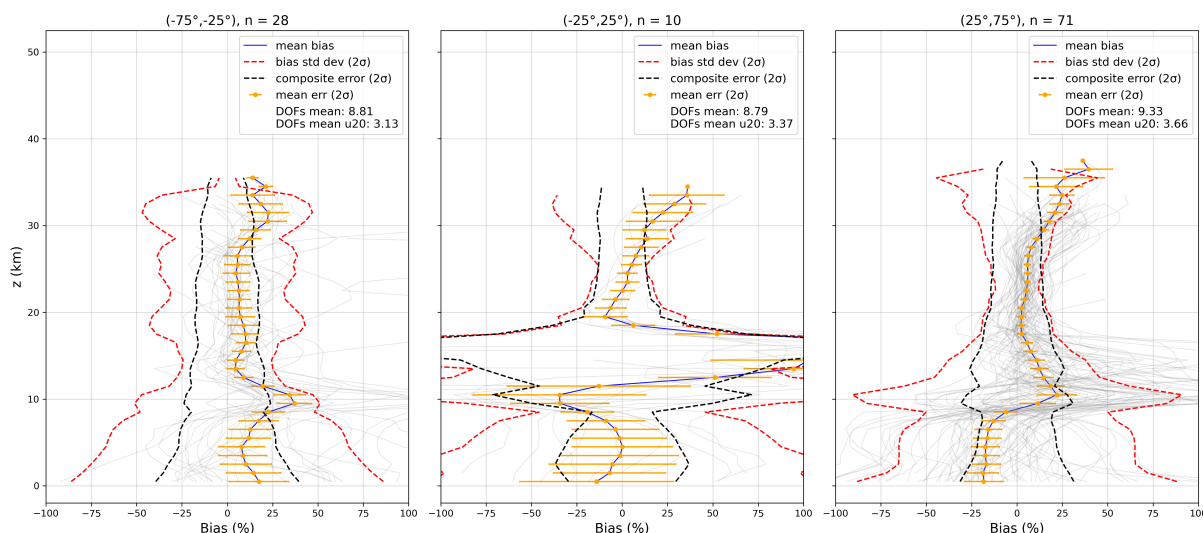


Figure 2.14: Validation of v05 across three latitudinal bands. The blue line represents the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product's AK. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), and the orange bars show the standard error of the mean. The transparent grey lines correspond to the individual averaged profiles. Here n indicates the number of valid fused profiles considered.

approximately -35%. Between 12 km and 18 km, The bias exceeds 100%, becoming very large. Above 18 km, the bias is no longer significant. However, due to the limited number of coincidences, no definitive conclusions can be drawn

In the 25° to 75° North band, representing the Northern Hemisphere (temperate and sub-Arctic regions), the number of coincidences increases to 71. The statistics indicate that the bias is not significant across most of the vertical profile: it is negative and below 25% below 8 km, and positive above 8 km, with a peak of about 25% near 10 km.

These results show that the fused product improves upon the individual products in terms of bias over the entire vertical profile.

2.4.4 About the Bias at 10km

A distinctive and recurrent feature emerging from the validation analysis is a pronounced bias in the vicinity of 10 km altitude. This layer lies within the Upper Troposphere–Lower Stratosphere (UTLS) transition zone, a region strongly influenced by stratosphere-to-troposphere exchange (STE) processes. Because of its potential impact on ozone variability and trend assessments, understanding and mitigating this bias is crucial for improving the accuracy of the fused dataset. The following section investigates its origin and quantifies the improvement achieved through a dedicated bias correction.

As can be observed in Figure 2.11, the bias appears to being originated from IASI, an observation consistent with the findings of Boynard et al., 2016, who reported a similar bias in the FORLI IASI dataset. Indeed, there authors carried out a comprehensive global validation of IASI vertical ozone profiles retrieved with the FORLI algorithm over the 2008–2014 period. The validation made use of a broad set of independent observations, including UV–visible satellite data from GOME-2A, ground-based measurements from Brewer and Dobson spectrophotometers, SAOZ instruments, and vertical profiles from ozonesondes. Their results show that, in the troposphere, IASI generally underestimates ozone by about 5–15 %, with the exact magnitude depending on latitude. In contrast, in the stratosphere—particularly in the 10–25 km altitude range—the instrument tends to overestimate ozone by roughly 10–40 %, with the largest positive biases occurring in the tropical tropopause region. This stronger bias in the tropics is likely related to the low ozone mixing ratios in that region, which tend to amplify relative errors. The analysis also revealed a marked latitude dependence, with smaller biases at mid- and high-latitudes and larger relative biases in the tropics, especially just above the tropopause. Across all validations, a consistent positive bias of about 25% is observed at 10 km. While the exact magnitude may vary from the above mentioned study due to methodological differences, it is broadly consistent with reported biases for nadir infrared retrievals in the UTLS. Notably, it is somewhat higher than the 10–20% bias found for UV-VIS sensors like TROPOMI (Keppens et al., 2024). This finding directly motivates the bias-correction strategy applied to the IASI dataset in this study.

Firstly, it has to be noticed that CDF is robust to input profiles correction. Hence, we analyze the effect of a linear scaling of the retrieved profiles on the fused solution. Let

the scaled profiles be defined as:

$$\hat{\mathbf{x}}'_i = (1 - p_i)\hat{\mathbf{x}}_i, \quad (2.72)$$

where p_i is a fixed fraction subtracted from the original retrievals.

By definition, the original vectors are $\boldsymbol{\alpha}_i = \hat{\mathbf{x}}_i - (\mathbf{I} - \mathbf{A}_i)\mathbf{x}_{a,i}$ and $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i - \mathbf{A}_i(\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)})\mathbf{x}_a$. After scaling, the vectors become:

$$\boldsymbol{\alpha}'_i = \hat{\mathbf{x}}'_i - (\mathbf{I} - \mathbf{A}_i)\mathbf{x}_{a,i} = (1 - p_i)\hat{\mathbf{x}}_i - (\mathbf{I} - \mathbf{A}_i)\mathbf{x}_{a,i}, \quad (2.73)$$

$$\tilde{\boldsymbol{\alpha}}'_i = \boldsymbol{\alpha}'_i - \mathbf{A}_i(\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)})\mathbf{x}_a = (1 - p_i)\hat{\mathbf{x}}_i - (\mathbf{I} - \mathbf{A}_i)\mathbf{x}_{a,i} - \mathbf{A}_i(\mathbf{C}^{(i)} - \mathbf{H}_i^\# \mathbf{C}^{(f)})\mathbf{x}_a. \quad (2.74)$$

Subtracting the original alpha tilde, we find:

$$\tilde{\boldsymbol{\alpha}}'_i - \tilde{\boldsymbol{\alpha}}_i = -p_i\hat{\mathbf{x}}_i. \quad (2.75)$$

Recalling the fusion equation (eq 2.55):

$$\mathbf{x}_f = \left(\sum_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i \mathbf{H}_i^\# + \mathbf{S}_a^{-1} \right)^{-1} \left(\sum_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \tilde{\boldsymbol{\alpha}}_i + \mathbf{S}_a^{-1} \mathbf{x}_a \right).$$

and substituting $\tilde{\boldsymbol{\alpha}}'_i$, the term in the second parentheses becomes:

$$\sum_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \tilde{\boldsymbol{\alpha}}'_i + \mathbf{S}_a^{-1} \mathbf{x}_a = \sum_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \tilde{\boldsymbol{\alpha}}_i - \sum_i p_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \hat{\mathbf{x}}_i + \mathbf{S}_a^{-1} \mathbf{x}_a. \quad (2.76)$$

It can be observed that the matrices $\mathbf{A}_i$, $\mathbf{H}_i^\#$, and $\tilde{\mathbf{S}}_i$, as well as the a priori term $\mathbf{S}_a$, are independent of the absolute values of $\hat{\mathbf{x}}_i$. Consequently, the structure of the fusion, i.e. the combination of all retrieved profiles and the a priori information, remains unchanged. The only modification is a term linear in the input profiles, $-\sum_i p_i \mathbf{H}_i^{\#T} \tilde{\mathbf{S}}_i^{-1} \hat{\mathbf{x}}_i$, which does not alter the weighting matrices.

Therefore, while the fused solution $\mathbf{x}_f$ is not strictly invariant, the effect of a linear scaling of the input profiles is entirely captured by a predictable linear term. In this sense, the fusion method can be considered formally robust with respect to linear corrections of the input profiles: the structure of the information combination remains intact, and the result is modified only by a linear contribution proportional to the scaling. If the term p_i is small enough, the final formula can be considered unchanged.

Moreover, $\mathbf{A}_f$ and $\mathbf{S}_f$, depending on $\mathbf{A}_i$, $\mathbf{H}_i^\#$, $\tilde{\mathbf{S}}_i$, and $\mathbf{S}_a$, are independent of $\hat{\mathbf{x}}_i$, remaining completely unchanged.

Following the reasoning above, as an additional part of this work, we focused on 2008 and on 5–55°N region, and we applied a correction to each IASI profile by subtracting the percentage bias identified in the IASI validation over the same area and year. Figure 2.15 shows the result of IASI validation, performed using the same methodology described in the Methodology section. It can be seen that there is a bias between 10 km and 20 km as documented by Boynard et al., 2016, that recall the same bias in the fused validation case.

Figure 2.16 shows now the new validation of the fused profile after having applied the bias correction. Compared to the results presented in Figure 2.9, correcting the IASI bias leads to an improvement in the fused product: the peak near 10 km is now greatly reduced and

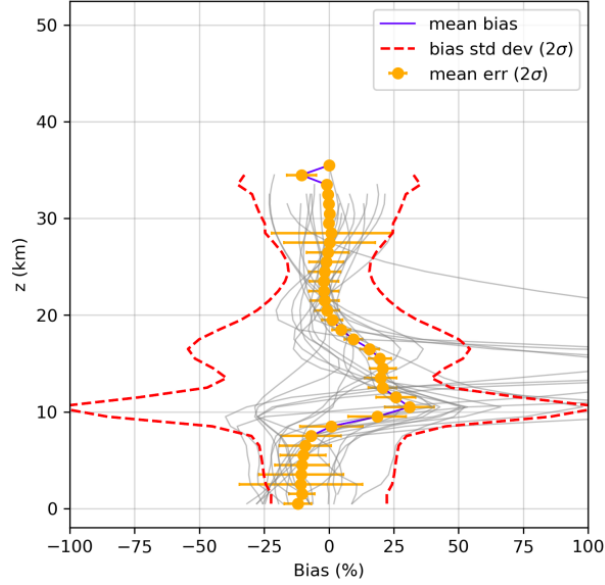


Figure 2.15: Validation of IASI for 2008 in the 5–55°N region. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product’s AK. The red dashed line indicates the standard deviation of the mean bias (2σ), and the orange bars correspond to the standard error of the mean. The transparent grey lines represent the individual averaged profiles. 20 profiles were found to perform the validation.

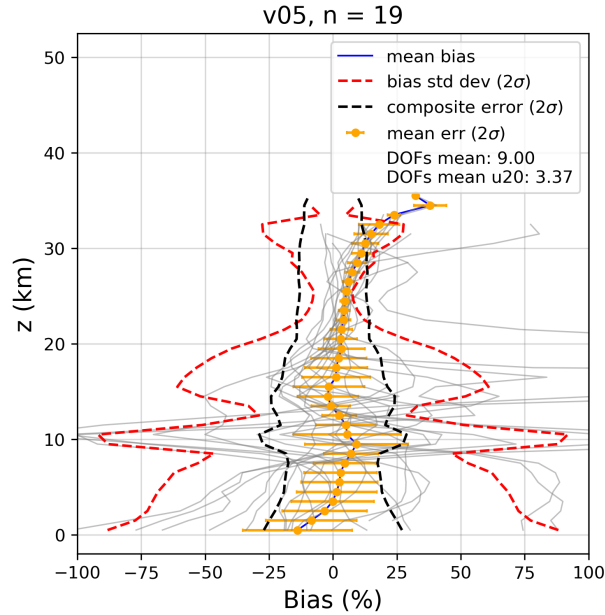


Figure 2.16: Validation of v05 with IASI bias correction applied for 2008. The blue line shows the mean percentage bias between the fused profiles and the ozonesondes smoothed using the fused product’s AK. The red dashed line indicates the standard deviation of the mean bias (2σ), the black dashed line represents the combined error (2σ), and the orange bars correspond to the standard error of the mean. The transparent grey lines represent the individual averaged profiles. Here n indicates the number of valid fused profiles.

becomes statistically non-significant.

To briefly understand the behavior of this correction, we tried to apply the same IASI bias correction derived for 2008 to the remaining years, obtaining the results shown in Figure 2.17. It can be seen that the 2008 IASI bias also improves the fused product for the other years, although to varying degrees. This suggests that the IASI bias changes from year to year, as found by Boynard et al., 2016.

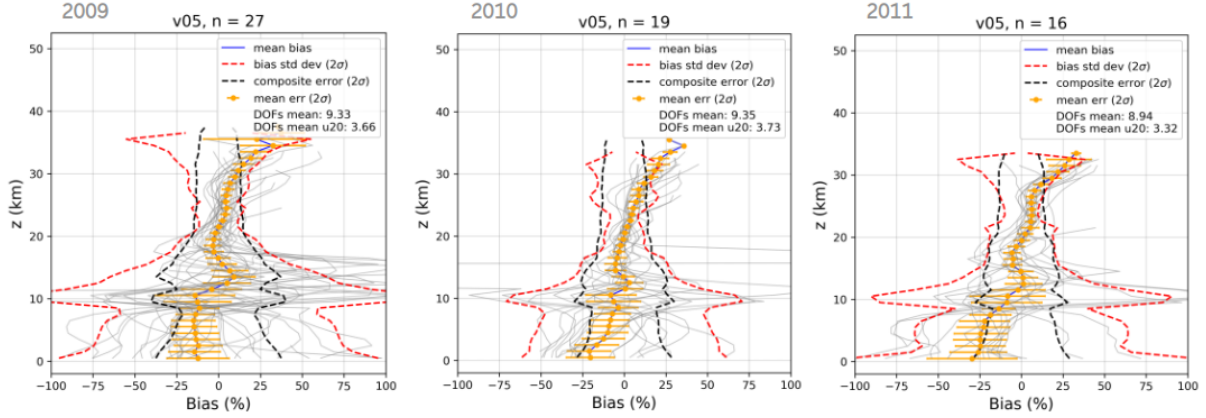


Figure 2.17: Validation of version v05 with IASI bias correction for 2009–2010–2011. In blue is the mean percentage bias between fused profiles and ozonesondes, smoothed with the fused AK. The red dashed line indicates the standard deviation of the mean bias (2σ), while the black dashed line represents the composite error (2σ). Orange bars show the standard error of the mean. The semi-transparent grey profiles represent the individual averaged profiles. Here, n denotes the number of valid fused profiles considered.

Nevertheless, this correction preserves the key characteristics of the fused product in terms of DOFs and retrieval errors, while considerably reducing the bias—especially in the UTLS region, where it becomes non-significant.

2.5 Chapter Conclusions

In this chapter, the development and validation of a new 2008-2011 ozone dataset is presented. The dataset is obtained through the application of the CDF algorithm combining high-quality MIPAS limb observations with the broader coverage of IASI nadir measurements. This fused dataset is particularly well suited for studying stratospheric intrusions, as it provides enhanced vertical resolution in the Upper Troposphere–Lower Stratosphere (UTLS) and improved coverage in regions where individual instruments have reduced sensitivity or do not measure (MIPAS case), such as the lower troposphere. Focusing on the period 2008–2011 and on the Himalayas region, the dataset demonstrates the capability of propagating stratospheric information downward while reducing retrieval uncertainties. Validation against WOUDC ozonesondes confirmed that the fused profiles increase the degrees of freedom for signal and reduce uncertainties relative to the individual MIPAS and IASI inputs. While some biases remain, particularly in the UTLS, their magnitude can be mitigated through input corrections, highlighting both the robustness of the fusion

approach and the importance of accounting for temporal and spatial variability in the retrievals. The bias-correction experiment performed in this work was used solely as a diagnostic test to attribute the 10 km feature to the known IASI bias; although the correction proved feasible and consistent with previous findings, it was not applied in the generation of the final fused dataset. The publicly released version of the dataset will therefore remain uncorrected, ensuring full transparency and avoiding the introduction of processing steps that are not yet fully consolidated.

The resulting L2-like dataset includes partial-column ozone profiles with associated covariance matrices, averaging kernels, and a priori information, all mapped onto the IASI grid. This comprehensive structure not only allows for rigorous validation and uncertainty assessment but also provides a foundation for potential assimilation in atmospheric reanalyses, such as ERA5 and CAMS, improving the representation of ozone dynamics in both the troposphere and stratosphere.

Importantly, this work sets the stage for the next part of the thesis, which focuses on the application of the fused ozone dataset to study stratospheric ozone intrusions. By bridging the gap between high-resolution limb measurements and broader nadir coverage, the dataset enables more accurate analyses of stratosphere–troposphere exchange and regional ozone variability, supporting both observational and modeling studies.

Overall, the development and validation of this fused dataset demonstrate the value of combining complementary satellite observations, providing a tool that paves the way for the analyses presented in the following Chapter.

Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

3.1 Chapter Overview

This chapter describes the second part of this thesis. It builds upon the newly developed fused ozone dataset presented in Chapter 2 and explores its potential for the detection and characterization of stratosphere-to-troposphere exchange events over the Himalayas. The ultimate goal is to develop a profile-based classification framework capable of distinguishing and quantifying different types of exchange events based on their dynamical and chemical features.

Over the past decades, several studies (J. Liu et al., 2020; Mkololo et al., 2020; Seguel et al., 2025) have investigated stratosphere-to-troposphere exchange (STE) events using ozone profile measurements, primarily from ozonesondes. These instruments provide high vertical resolution and have been widely used to detect and characterize ozone intrusions based on threshold criteria (e.g., elevated ozone and low humidity) or more quantitative methods such as bandpass filtering and Brunt–Väisälä frequency analysis, as demonstrated by Greenslade et al., 2017. Ozonesonde data have also been used to build long-term climatologies of tropospheric ozone concentrations variability and STE frequency (e.g., Zang et al., 2024). In parallel, STE detection has often relied on the use of dynamical tracers such as potential vorticity and reconstruction of air mass history, for instance by means of back-trajectories computed using Lagrangian or hybrid transport models such as LAGRANTO Putero et al., 2016 or Hysplit. In spite these methods are effective, they require auxiliary meteorological data and are not always applicable to large-scale satellite datasets. More recently, machine learning techniques have been explored for ozone profile classification, but mainly in the context of pollution forecasting or model correction (e.g., Dhomse et al.,

2021). To date, no extensive study has applied classification methods directly to ozone profiles retrieved from remote sensing instruments for the detection and characterization of STE events. In this context, the present study aims to apply a classification methodology of ozone profiles obtained from MIPAS and IASI instruments, as well as from the fused dataset obtained in the previous Chapter 2. By applying a classification approach to these profiles, this work aims to provide a scalable and automated tool for identifying and characterizing stratospheric ozone intrusions from satellite data, which can offer a broader spatial and temporal coverage compared to ozonesondes, contributing to the development of global monitoring capabilities and supporting future climatological analyses.

To this end, a case study of an intrusion event (January 28th – February 2nd, 2008) is used as a pilot analysis. The use of a single case study at this stage allows us to manage a reduced amount of data while laying the groundwork for a more extensive analysis. In our case, this would mean extending the study over the entire Himalayan region and over a longer time span (2008–2011), by exploiting the full fused dataset. The datasets employed in this chapter include the in situ observations, the fused and original satellite retrievals (IASI and MIPAS), and data from reanalysis, the ECMWF Reanalysis v5 and the Copernicus Atmosphere Monitoring Service reanalysis (ERA5, CAMS), and the ECHAM/MESSy Atmospheric Chemistry model (EMAC). The analysis is structured around three main steps: (1) first, model data are used to map and characterize the spatial and vertical features of the selected intrusion and to identify representative intrusion and non-intrusion profiles; (2) second, a qualitative comparison between model profiles and remote sensing data (IASI, MIPAS, fused) is carried out; (3) finally, these insights are used to design an exploratory classification approach aimed at distinguishing intrusion from non-intrusion profiles in both models and remote sensing data. Particular attention is paid to assessing whether the fused profiles provide an added value over the individual satellite products, especially in terms of vertical resolution.

The main objective of this second part is therefore to develop and test a profile-based methodology for detecting stratospheric ozone intrusions, using the pilot case study as a proof of concept. While this classification is not intended as a definitive solution, it serves as a foundational step toward a transferable framework that can be extended to other cases in the future. It is important to emphasize that STE events encompass a wide range of dynamical configurations and chemical signatures. Consequently, the performance of the proposed classification framework is expected to vary across different event types and cannot be fully assessed from a single six-day episode. In this case study, we demonstrate that distinct groups of profiles emerge, and that the fused product enhances the separability between intrusion and non-intrusion conditions. However, the identification of these groups and the robustness of the classification strategy must be validated against a broader set of STE occurrences with varying intensities, geometries, and meteorological drivers. Future work will therefore extend the methodology to the full multi-year fused dataset (2008–2011) over the Himalayan region. This will allow us to assess its generality, identify recurrent patterns, and determine whether the same classification scheme can reliably capture the diversity of STE phenomena.

This chapter is structured as follows: the Data section introduces the observational and model datasets used; the Methodology section outlines the workflow and the classification

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

approach; the Results and Discussion section presents the main findings from the case study analysis; and the Conclusions section summarizes the results and their implications, highlighting the potential of the fused dataset for STE studies in the Himalayan region.

3.2 Data

3.2.1 Nepal Climate Observatory Pyramid (NCO-P) and Intrusions Dataset

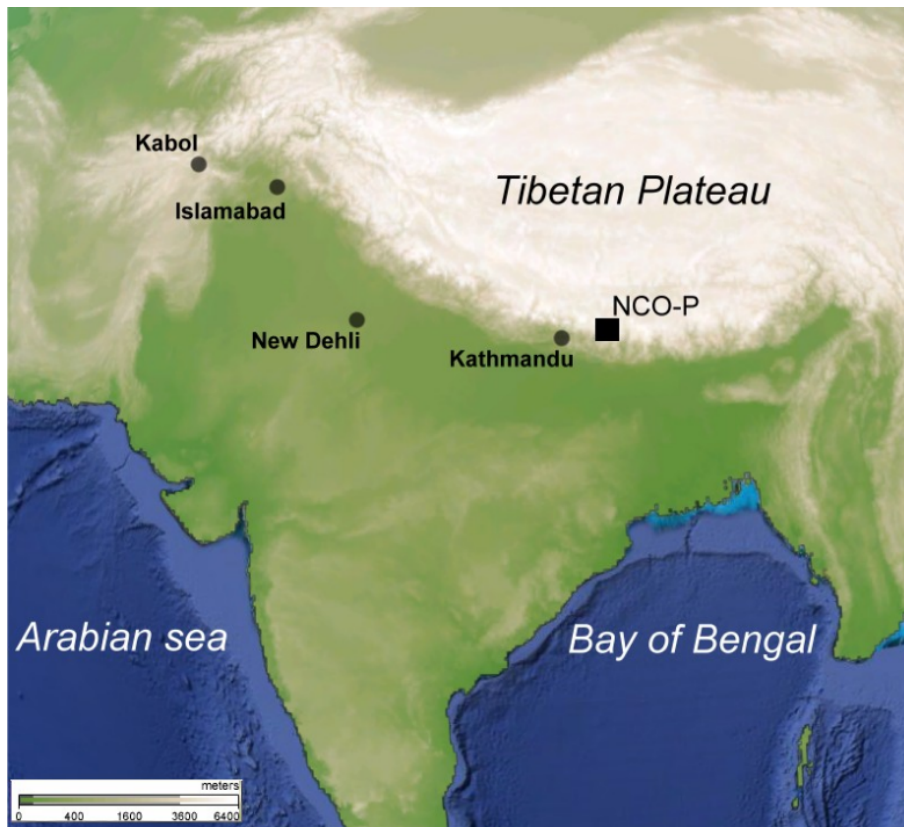


Figure 3.1: NCO-P geographical location (Cristofanelli et al., 2010).

The Nepal Climate Observatory - Pyramid (NCO-P, 27.95 N, 86.82 E; 5079 m a.s.l.) is located in the southern Himalayan region at the confluence of the secondary valley of Lobuche (oriented NNW-SSE) and the main Khumbu Valley. The station was placed not far from the Pyramid International Laboratory/Observatory and in proximity to the Mt. Everest base camp (Nepal Climate Observatory - Pyramid, 2011). Figure 3.1 shows the location of the station. The NCO-P, operating within the WMO/GAW programme, continuously collects a wide range of measurements, including trace gases (e.g., ozone, carbon monoxide, nitrogen oxides, greenhouse gases), aerosol physical (optical coefficients, size distribution) and chemical (e.g., concentration of black carbon, inorganic and carbonaceous species) properties, and basic meteorological parameters (atmospheric temperature, barometric pressure, relative humidity, wind intensity and direction, radiation). These long-term observations are designed to investigate the transport of atmospheric pollution from South Asia to the Himalayas and its interactions with background atmospheric

composition and climate processes. In particular, surface ozone is measured in situ using ground-based UV absorption instruments, providing continuous time series of near-surface concentrations at this unique high-altitude site.

In the context of this work, we used a specific dataset developed by D. Putero and P. Cristofanelli from the Institute of Atmospheric Sciences and Climate of the Italian National Research Council (CNR-ISAC) for the NCO-P station. This dataset provides a binary classification (event/non-event) on a daily basis, according to two complementary methodologies (Putero et al., 2016). The first method is based on the analysis of in situ ozone measurements at the station, while the second relies on a Lagrangian transport model to identify stratospheric intrusion episodes.

More specifically, as described by Putero et al., 2016, the first method identifies days potentially influenced by stratospheric intrusion (SI) events at NCO-P by applying a dedicated statistical methodology - called SIO (Stratospheric Intrusion Observations)- to the time series of both observed and modeled variables. The identification method is based on a combination of in situ measurements (ozone; atmospheric pressure; and relative humidity), satellite observations (total column ozone, TCO, from the Ozone Monitoring Instrument, OMI), and NWP (Numerical Weather Prediction)-based back-trajectories computed with LAGRANTO (Sprenger and Wernli, 2015). The methodology relies on four distinct criteria, and a day is classified as potentially affected by a stratospheric intrusion if at least one of these criteria is satisfied: (1) significant variations in daily surface pressure (P) and the presence of back-trajectories with potential vorticity (PV) values greater than 1.6 PVU; (2) significant daily increases in total column ozone (TCO) and the presence of back-trajectories with PV greater than 1.6 PVU; (3) simultaneous significant variations in daily pressure P and TCO values; (4) relative humidity (RH) values below 60%, combined with a significant negative correlation between ozone and RH, a daily ozone maximum exceeding the seasonal average, and significant variation in at least one among P, PV, or TCO. This last criterion accounts for the possible influence of downward valley winds transporting stratospheric air masses to the surface.

Conversely, the second approach employs a Lagrangian transport model, known as STE-FLUX (Stratosphere-to-Troposphere Exchange Flux), developed by Putero and Cristofanelli. STEFLUX exploits the climatology of stratosphere-to-troposphere exchange (STE) derived from reanalysis data, such as ERA-Interim, to identify intrusion events. The model calculates the flux of ozone from the stratosphere into the troposphere, allowing the detection of periods when stratospheric ozone reaches the surface. This method provides a quantitative and spatially distributed estimate of intrusion events, complementing the in situ observations.

This product therefore serves as a reference for detecting intrusion events at the site. From this dataset, we selected a six-day event consistently identified by both methods, occurring between January 28th and February 2nd, 2008, which serves as the reference case study for the subsequent analysis. Further details about the case study selection are provided in the following Section.

3.2.2 ERA5 Dataset

ERA5 represents the fifth generation of atmospheric reanalyses produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) within the Copernicus Climate Change Service (C3S) (from ERA5 Copernicus website). It provides hourly data on a wide range of atmospheric, land-surface, and oceanic variables, with a spatial resolution of approximately $0.25^\circ \times 0.25^\circ$ and a temporal coverage extending from 1940 to the present. In this study, we used hourly data from 1979 to present.

ERA5 is based on the Integrated Forecasting System (IFS) version 41r2, operational at ECMWF during March–November 2016. The IFS model assimilates observations from multiple sources, including satellites, radiosondes, buoys, and aircraft measurements, to generate a consistent and comprehensive estimate of the global atmospheric state. This assimilation improves the quality of the atmospheric estimates, reducing errors and uncertainties associated with individual observational datasets.

It is important to note that ERA5 does not include an explicit atmospheric chemistry model; therefore, ozone is treated as a passive variable, largely influenced by atmospheric dynamics. Ozone data are available on 37 standard pressure levels, from 1000 hPa to 1 hPa, on a regular latitude–longitude grid.

In the context of this study, ERA5 is used as a reference to analyze the dynamical conditions associated with stratospheric ozone intrusions over the Himalayan region. Although ERA5 does not include direct information nor specific modules to simulate atmospheric chemical reactions, its estimates of dynamical variables, such as potential vorticity and relative humidity, are highly useful for identifying and characterizing intrusion events. In particular, maps of potential vorticity and relative humidity provide insights into dynamic features, such as tropopause folding, which are typically associated with stratospheric ozone intrusions.

3.2.3 CAMS Dataset

CAMS stands for Copernicus Atmosphere Monitoring Service, which provides a global reanalysis dataset (CAMS global reanalysis | ECMWF website) that offers comprehensive insights into atmospheric composition. This dataset is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and is part of the Copernicus Programme. It delivers three-dimensional, time-consistent fields of atmospheric composition, including aerosols, chemical species, and greenhouse gases, covering the period from 2003 to December 2024.

The CAMS reanalysis employs a 4D-Var data assimilation system within ECMWF’s Integrated Forecasting System (IFS), Cycle 42r1. This system integrates satellite observations of atmospheric composition, such as ozone, carbon monoxide, nitrogen dioxide, and aerosol optical depth, to constrain the atmospheric model and produce accurate analyses (see Satellite observations | Copernicus). The IFS operates with 60 hybrid sigma-pressure model levels, extending up to 0.1 hPa, and provides data on 25 standard pressure levels ranging from 1000 hPa to 1 hPa. Compared to ERA5 the horizontal grid of CAMS is coarser, of $0.75^\circ \times 0.75^\circ$.

In the CAMS reanalysis, unlike ERA5, ozone is treated as a chemical species within the atmospheric model, allowing for the simulation of its distribution and variability based on atmospheric dynamics and chemical processes. Importantly, CAMS does not rely on model dynamics and chemistry alone: ozone is strongly constrained by the assimilation of satellite observations, which directly shape the analysed ozone distribution. This approach enables the generation of detailed vertical profiles and horizontal distributions of ozone concentrations, which are essential for studying phenomena such as stratosphere-to-troposphere exchange (STE) events. The reanalysis provides ozone data on multiple pressure levels, facilitating in-depth analyses of its behavior across different atmospheric layers.

The CAMS reanalysis is particularly valuable for investigating stratospheric ozone intrusions, as it offers high-resolution data on ozone concentrations and related atmospheric variables. Studies (Akritidis et al., 2016; Akritidis et al., 2018; Roy et al., 2023) have demonstrated the utility of CAMS reanalysis in identifying and analyzing STE events, such as tropopause folds, also by examining the temporal and spatial patterns of ozone anomalies. These analyses are crucial for understanding the dynamics of ozone transport and its implications for air quality and climate.

3.2.4 EMAC Dataset

The European Centre for Medium-Range Weather Forecasts (ECMWF) Hamburg (ECHAM) Modular Earth Submodel System (MESSy) Atmospheric Chemistry (EMAC) model is a global numerical chemistry–climate model that simulates the interaction between dynamical and chemical processes from the surface up to the middle atmosphere (Brenninkmeijer and Cai, 2016). Specifically, in this study results from the REF-D1 specific dynamics (SD) hindcast experiment were used (Jöckel et al., 2024), with the model being nudged towards ECMWF ERA5 reanalysis data through Newtonian relaxation of temperature, divergence, vorticity, and logarithm of surface pressure. The REF-D1 setup prescribes observed sea surface temperatures and sea ice, as well as external forcings such as solar flux, greenhouse gases, ozone-depleting substances, stratospheric aerosols, and the quasi-biennial oscillation (QBO affects the the distribution and transport of ozone, see Section 1.2.2), in order to reproduce the historical atmospheric state as accurately as possible.

The simulations are performed at a T42 spectral resolution ($2.8^\circ \times 2.8^\circ$ on a Gaussian grid), with 90 hybrid sigma–pressure vertical levels extending from the surface to 0.01 hPa, and a time step of 720 seconds. EMAC incorporates several chemical submodels, including MECCA (Sander et al., 2011) for gas and aqueous phase chemistry, and SCAV (Tost et al., 2006) for scavenging processes, allowing for a detailed representation of tropospheric and stratospheric chemistry (O_3 , CH_4 , HO_x , NO_x , NMHCs, halogens, and sulfur compounds). The model calculates natural emissions such as lightning NO_x and soil NO_x interactively (Tost et al., 2007; Kerkweg et al., 2006), and includes online photolysis (JVAL, SSander et al., 2014) and shortwave radiation (FUBRAD, Tarasick et al., 2019), with explicit representation of the 11-year solar cycle (Morgenstern et al., 2017). Anthropogenic emissions follow the CCMI-2020 protocol for REF-D1 hindcast simulations.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

In addition, EMAC provides the diagnostic tracer O_3S , which directly quantifies stratosphere-to-troposphere exchange (STE) of ozone (Brenninkmeijer and Cai, 2016; Williams et al., 2024). This tracer allows for the separation of stratospheric (O_3S) and tropospheric ozone (O_3T) contributions: tropospheric ozone is computed as the difference between total and stratospheric ozone (Roelofs and Lelieveld, 1997).

3.2.5 Considerations About Models

While ERA5, CAMS, and EMAC are all used in this study, it is important to highlight their conceptual and intrinsic differences. ERA5 is a purely meteorological reanalysis, providing dynamically consistent fields of atmospheric variables but without explicit atmospheric chemistry; ozone in ERA5 is therefore only indirectly represented through dynamical processes and assimilated observations. CAMS, instead, extends the reanalysis framework to include atmospheric composition: it assimilates satellite retrievals of ozone and other trace gases together with meteorological observations, thereby offering realistic global distributions of chemical species. EMAC, by contrast, is not a reanalysis but a chemistry–climate model. In the configuration used here, it is nudged with ERA5 meteorological data to reproduce the observed atmospheric state, while its chemical tracers, including diagnostic species such as O_3S , are explicitly simulated by the model’s chemistry schemes without direct assimilation. This distinction implies that, whereas ERA5 and CAMS rely heavily on observation assimilation to constrain their fields, EMAC provides a process-based and internally consistent representation of the interactions between dynamics and chemistry.

3.3 Methods

3.3.1 Case Study Definition

The dataset covering the period from 1 March 2006 to 31 December 2013 and described in Section 3.2.1 was employed, including daily binary indices (1 for event, 0 for no event) for the identification of stratospheric intrusions based on STEFLUX (SI-STEFLUX) and statistical methods (SI-SIO), briefly described in Section 3.2.1, surface ozone mixing ratios ($\text{O}_3\text{-NCO-P}$), relative humidity (RH-NCO-P), and black carbon concentrations (BC-NCO-P). Both a trend analysis and a time series analysis were performed on the database. Here only the main results of the time series analysis are shown to justify the choice of the case study. However, other analyses and the trends are documented in Appendix E. To explore the temporal distribution of events, I derived seasonal and monthly frequencies for both STEFLUX and SIO indices. Seasonal frequencies were calculated by grouping the data by year and season, and then represented through stacked bar charts, either emphasizing the contribution of seasons within each year or the contribution of years within each season. In parallel, average monthly frequencies were computed across the full time period to highlight intra-annual variability. After that, potential stratospheric intrusion events were identified by selecting days for which both SI-STEFLUX and SI-SIO indicators were equal to 1, corresponding to conditions favorable for stratosphere-to-troposphere transport of ozone. According to these criteria, an event lasting six days, from January, 28th to February 2nd, 2008, was selected as the reference case study. This episode was

characterized by enhanced surface ozone values relative to surrounding periods, making it particularly suitable for investigating the role of stratospheric intrusions.

3.3.2 Model and Reanalysis Data Processing and Intercomparison

Table 3.1: Summary of datasets used for the case study, including variable, measurement unit, horizontal resolution and dataset type.

Dataset	Variable	Unit	Horizontal resolution	Vertical resolution
ERA5	O ₃	MMR	0.25° × 0.25°	37 levels (1000–1 hPa)
ERA5	PV	(K m ² kg ⁻¹ s ⁻¹)	0.25° × 0.25°	37 levels (1000–1 hPa)
ERA5	RH	%	0.25° × 0.25°	37 levels (1000–1 hPa)
CAMS	O ₃	MMR	0.75° × 0.75°	25 levels (1000–1 hPa)
CAMS	O ₃ S	MMR	0.75° × 0.75°	25 levels (1000–1 hPa)
CAMS	PV	(K m ² kg ⁻¹ s ⁻¹)	0.75° × 0.75°	25 levels (1000–1 hPa)
CAMS	RH	%	0.75° × 0.75°	25 levels (1000–1 hPa)
EMAC	O ₃	ppmv	2.8° × 2.8°	90 levels (surf–0.01 hPa)
EMAC	O ₃ S	ppmv	2.8° × 2.8°	90 levels (surf–0.01 hPa)
EMAC	PV	(K m ² kg ⁻¹ s ⁻¹)	2.8° × 2.8°	90 levels (surf–0.01 hPa)
EMAC	RH	%	2.8° × 2.8°	90 levels (surf–0.01 hPa)

This step of the analysis aims to evaluate whether the models are able to represent the stratospheric intrusion, and if so, how they capture its timing, intensity, and vertical structure. To this end, the extended region 5–55°N / 50–110°E around the station was selected. For this region, horizontal maps were produced at selected pressure levels (e.g., 500, 300, 150, and 50 hPa), along with vertical distributions of the variables listed in Table 3.1. Vertical profiles are presented for latitudinal and meridional sections passing through the station, with the closest grid point from each model selected for comparison. Given the differing native temporal resolutions of ERA5 (1-hourly), CAMS (3-hourly), and EMAC (5-hourly), the datasets were harmonized through temporal matching. Fields from ERA5 and CAMS were aligned using shared timestamps, whereas EMAC data were matched to the closest available output time. This results in a temporal offset of only a few hours, a trade-off that preserves the transient signature of the event by avoiding temporal averaging and ensures sufficient synchronicity for intercomparison.

Ozone concentrations provided in mass mixing ratio (kg kg⁻¹) were converted to volume mixing ratio (ppmv) to allow a consistent comparison across datasets. The conversion was performed using the molar masses of air and ozone according to:

$$\text{O}_3 \text{ [ppmv]} = \text{O}_3 \text{ [kg/kg]} \times \frac{M_{\text{air}}}{M_{\text{O}_3}} \times 10^6$$

where $M_{\text{air}} = 28.97 \text{ g/mol}$ and $M_{\text{O}_3} = 48 \text{ g/mol}$. This procedure ensures that all datasets are expressed in the same units, facilitating a direct comparison of ozone values between ERA5, CAMS, and EMAC.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

Where data were provided in model levels, the conversion to pressure levels was applied. The pressure at model levels was computed from the surface pressure and the standard hybrid coefficients a_k and b_k according to the formula:

$$p_k = (a_k + b_k p_{\text{surf}}) \times 10^{-2},$$

where p_{surf} is the surface pressure at each grid point and time step, and p_k is the pressure at the k -th model level in hPa. This approach allows the transformation from the hybrid sigma-pressure coordinates used in the original dataset to physical pressure levels, facilitating intercomparison between different datasets and computation of vertical profiles. Model outputs often provide values below the actual surface, due to the vertical discretization of model levels. To avoid nonphysical values in these regions, all variables at pressures exceeding the surface pressure were masked, ensuring that only atmospheric values above ground level are considered in the analysis.

Finally, to evaluate model and reanalysis differences in case study representation, interpolation on the same vertical and horizontal grid was applied. For our purpose, the CAMS horizontal and vertical grids were used. This choice was motivated by the different horizontal resolutions of the models (as shown in Table 3.1) and by the fact that the CAMS grid provides sufficient spatial detail to capture the structure of intrusions, while also offering a practical compromise between the finer horizontal resolution of ERA5 and the coarser resolution of EMAC. For this reason, two distinct interpolations were applied, namely a bilinear interpolation in the horizontal (latitude-longitude) plane, and a linear interpolation on the logarithm of pressure levels in the vertical, which preserves the atmospheric structure in log-pressure coordinates.

3.3.3 Profiles Characterization: Reanalysis, Model, Fused, and L2

Prior to applying the clustering analysis, within the case study region, two subregions were defined based on maps from reanalyses and model outputs (ERA5, CAMS, EMAC). The first subregion (Lat: 5-20°N, Lon: 50-70 °E) was characterized by the absence of intrusions-related features in all reanalysis and model, while the second subregion (Lat: 25-50°N, Lon: 80-110 °E) corresponded to areas where none of the three exhibited intrusion features. Ozone profiles from reanalyses, model simulations, the fused dataset, and Level 2 products were extracted for both subregions in order to characterize the typical structure of STE versus non-STE conditions. For the reanalyses and model data, which contained a large number of profiles, mean profiles with their corresponding standard deviations were calculated. In addition, median profiles and interquartile ranges (25th–75th percentiles) were computed, as the median provides a more robust measure against outliers. Vertical profiles were further analyzed across multiple layers to highlight differences between event and non-event conditions. Mean and median profiles were also derived for the fused dataset and Level 2 products to facilitate a direct comparison across datasets, providing a foundation for the subsequent automated classification.

In this part for simplicity, we refer to profiles in the subregion with event features as event profiles, while we refer to profiles in the sub-region without event features as non-event profiles.

3.3.4 Classification Methodology

To explore the potential of the fused ozone dataset for the detection of stratosphere-to-troposphere exchange (STE) signatures, a preliminary classification was applied directly to the vertical ozone profiles. The aim of this approach is to identify intrusion events through typical vertical profile structures associated with them, such as sharp upper-tropospheric enhancements or multi-layered features, and to distinguish them from background or purely tropospheric conditions. A data-driven clustering technique (k-means) was adopted as a first step toward an objective identification of intrusion-like profiles. However, this approach is not intended to serve as a definitive methodology. Rather, it represents an initial analysis that will inform and guide future developments.

Before introducing the k-means methodology, it can be useful knowing that different types of classification methodologies exist. They are divided into supervised and unsupervised methodologies.

Supervised classification relies on the availability of labeled data, i.e. samples for which the correct class or category is already known. The algorithm learns from these examples to predict the class of new, unseen samples. Typical supervised methods include Support Vector Machines (SVMs), Random Forests, and Neural Networks (Boateng et al., 2020). In atmospheric sciences, supervised techniques are often used when reference datasets or prior classifications are available, for example, to train models to distinguish stratospheric and tropospheric air masses using labelled observations or model outputs (Blume et al., 2012; Runde et al., 2016).

By contrast, unsupervised classification (or clustering) does not require pre-labelled data. Instead, it seeks to discover natural groupings or patterns in the dataset based solely on similarities or distances between samples. Common unsupervised methods include k-means clustering, hierarchical clustering, and Self-Organizing Maps (SOMs) (Serra and Tagliaferri, 2019; Naeem et al., 2023). In atmospheric and climate research, these methods are widely used to objectively identify regimes or patterns, such as different types of stratospheric intrusions, weather states, or air-mass clusters, when no prior classification exists (Philippopoulos et al., 2014; Richards, 2022; Fahrin et al., 2022).

In this work an unsupervised approach was chosen. This choice can be motivated by several factors. First, all used datasets are relatively short and the only dataset that can be used for labeling is the NCO-P dataset. However, it contains only a binary indicator of intrusions occurrence for specific days, without detailed labels for the vertical ozone profiles themselves. This makes supervised learning infeasible, as there are no target classes to train on (Fahrin et al., 2022; Chang et al., 2025). Moreover, the small number of labeled events (days with intrusions occurrences), a supervised approach would risk overfitting and might not generalize well to other profiles (Wu et al., 2021). Even when considering vertical profiles from alternative datasets such as ERA5, EMAC, or CAMS, the unsupervised approach remains appropriate, since these datasets do not provide labeled intrusion events profiles for each observation. Second, unsupervised methods are well suited for exploratory analysis, allowing the identification of natural groupings of profiles based on their morphological features, such as vertical distribution and intensity of ozone enhancements. Third, the multidimensional nature of the profiles makes clustering particularly appropriate, as distances between profiles can be used to define similarity. Finally, using an unsupervised method provides a first proof-of-concept for detecting characteristic

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

STE-like structures, laying the groundwork for future analyses or more sophisticated approaches. Nonetheless, in the following we will refer to the information provided by the models as feedback for the considerations made on the basis of the unsupervised analysis performed on the remote measurements.

Among unsupervised techniques, the k-means algorithm was selected for its simplicity, robustness, and computational efficiency (Xiao, 2024). K-means partitions the data into a predefined number of clusters by minimizing intra-cluster variance, making it suitable for capturing the main morphological patterns of ozone profiles. This first implementation serves as a preliminary exploration, providing insight into typical structures associated with STE without requiring complex assumptions.

K-means algorithm (Ahmed et al., 2020; Oti et al., 2021; Xiao, 2024) is designed to partition a set of N observations into k clusters, such that observations within the same cluster are more similar to each other than to those in other clusters. In the context of this study, each observation corresponds to a vertical ozone profile, represented as a multidimensional vector of ozone values at different pressure levels. Mathematically, the algorithm seeks to minimize the total intra-cluster variance, expressed as the sum of squared distances between each observation and the centroid of its assigned cluster:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2 \quad (3.1)$$

where C_j denotes the set of observations assigned to cluster j , μ_j is the centroid of cluster j , and $\|\mathbf{x}_i - \mu_j\|^2$ represents the squared Euclidean distance between the profile x_i and the cluster centroid.

The algorithm iteratively performs two main steps until convergence:

1. Each observation is assigned to the nearest centroid based on distance:

$$c_i = \arg \min_j \|\mathbf{x}_i - \mu_j\|^2 \quad (3.2)$$

2. Centroids are updated as the mean of the observations in each cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} \mathbf{x}_i \quad (3.3)$$

The process continues until cluster assignments no longer change, indicating convergence to a minimum of the objective function J . The final output consists of k clusters of profiles with similar vertical structures.

Once the clustering mechanism is defined, a crucial aspect is the selection of the number of clusters, k , which directly influences the granularity and interpretability of the results. The selection of k is a critical step in k-means clustering, as it directly affects the ability of the algorithm to capture meaningful patterns in the data. Choosing too few clusters may result in oversimplified groupings that fail to distinguish important variations in ozone profiles, while too many clusters may fragment natural structures and reduce interpretability. Techniques such as the elbow method and silhouette analysis are commonly used to guide this choice.

The elbow method is based on evaluating the total intra-cluster variance, often referred to as the sum of squared errors (SSE) or within-cluster sum of squares (WCSS), as a function of k :

$$\text{SSE}(k) = \sum_{j=1}^k \sum_{x_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2 \quad (3.4)$$

Here, C_j denotes the set of profiles assigned to cluster j , and μ_j is the centroid of cluster j . The procedure involves running k-means for a range of k values (e.g., $k = 1$ to 10), computing SSE for each, and plotting SSE versus k . The optimal number of clusters is identified at the *elbow point*, where further increases in k lead to minimal reductions in SSE.

Silhouette analysis quantifies how well each profile fits within its assigned cluster compared to other clusters. For each profile i , the silhouette coefficient $s(i)$ is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.5)$$

where $a(i)$ is the average distance between profile i and all other profiles in the same cluster, and $b(i)$ is the minimum average distance between profile i and all profiles in any other cluster. The silhouette coefficient ranges from -1 to 1, with values close to 1 indicating well-clustered profiles, values near 0 indicating profiles near cluster boundaries, and negative values suggesting possible bad assignment. By computing the mean silhouette coefficient across all profiles for different values of k , the optimal number of clusters can be selected as the value of k that maximizes the mean silhouette score.

Principal Component Analysis (PCA) was applied prior to clustering to reduce the dimensionality of the ozone profile dataset and facilitate the identification of dominant structural patterns. PCA (DataScienceBase Link) is a linear transformation technique that projects the original data onto a new set of orthogonal axes (principal components), ordered by the amount of variance they capture in the data. This transformation allows for a more compact representation of the profiles, preserving the most relevant information while discarding noise and redundancy. In case of classification taking into account ozone profiles, each profile is represented as a vector of ozone values across pressure levels. PCA transforms these vectors into a new coordinate system where the first few components capture the majority of the variability in the dataset. This dimensionality reduction is particularly useful when dealing with high-dimensional data, as it improves the efficiency and stability of clustering algorithms such as k-means, and helps mitigate the curse of dimensionality. Mathematically, PCA seeks to find a set of orthogonal vectors (principal components PC) that maximize the variance of the projected data. Given a centered data matrix X of size $N \times p$, where N is the number of profiles and p the number of pressure levels, PCA solves the eigenvalue problem:

$$X^T X v = \lambda v \quad (3.6)$$

where v are the eigenvectors (principal components) and λ the corresponding eigenvalues, which quantify the amount of variance explained by each component. The first few principal components are then used as input features for the k-means algorithm. This approach not only reduces computational cost but also enhances the interpretability of the clustering results, as the components often correspond to meaningful modes of variability

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

in the vertical ozone distribution. In addition to dimensionality reduction, PCA provides insight into the dominant patterns of variability in the dataset. By examining the loadings (weights of the linear combination of PCs) and explained variance of each component, it is possible to identify typical structures such as upper-tropospheric enhancements or multi-layered features, which are relevant for the detection of STE-like profiles.

In addition to applying PCA and clustering to the full dataset including vertical ozone profiles, a complementary analysis was performed in this work using a set of physically meaningful features derived from each profile on the basis of the previous analysis. These features include the altitude of the ozone maximum, the value of the columnar maximum, and the ratio between stratospheric and tropospheric ozone columns, used in two-steps: firstly altitude and value of ozone maximum and then adding also the columns ratio. In particular, to compute the ratio between the stratospheric and tropospheric ozone columns, the tropopause height provided in the IASI dataset was used as a reference. This value allows for the separation of the vertical ozone profile into stratospheric and tropospheric portions, enabling the sum of partial columns above and below the tropopause level. Each observation will be, therefore, represented by a set of two or three feature, and the analysis is based on their combination. This approach allows for a more interpretable characterization of the profiles, focusing on key indicators relevant to stratospheric intrusion processes. Applying PCA to this reduced feature space serves two purposes. First, it enables the identification of dominant modes of variability among the profiles based on physically interpretable parameters. Second, it facilitates clustering by reducing redundancy and emphasizing the most informative combinations of features. Compared to clustering on full profiles, this method is less sensitive to local fluctuations and noise, and more focused on capturing structural differences that are potentially linked to intrusion events. The resulting clusters can be directly interpreted in terms of physical characteristics, such as profiles with high ozone maxima at upper-tropospheric levels or enhanced stratospheric contributions.

Here, the classification is applied to the fused dataset and the Level 2 datasets as part of a case study, with the aim of setting up and testing a method, and evaluating the added value of the fused dataset. However, the analysis is still ongoing and remains incomplete; further refinement and consolidation are needed to fully assess the robustness of the approach.

3.4 Results and Discussion

3.4.1 Time Series Analysis Result for Case Study Definition

A significant result that guided the selection of the case study is shown in Figure 3.2, where the seasonal analysis highlights a clear recurrence of events and reveals a marked seasonality in their distribution. The two methods show well distinct behaviors, however, both agree in indicating the winter (DJF) seasonal maximum and the summer (JJA) seasonal minimum. In addition to the presence of a well recognizable seasonal cycle, the plots suggest the presence of interannual variability, with certain years showing peak frequencies in the number of events (2008, 2012). However, due to the limitedness of the dataset, far from being representative of a climatological period (normally, 30 year period), this apparent variability should be interpreted with caution, and it is briefly

discussed in Appendix E.

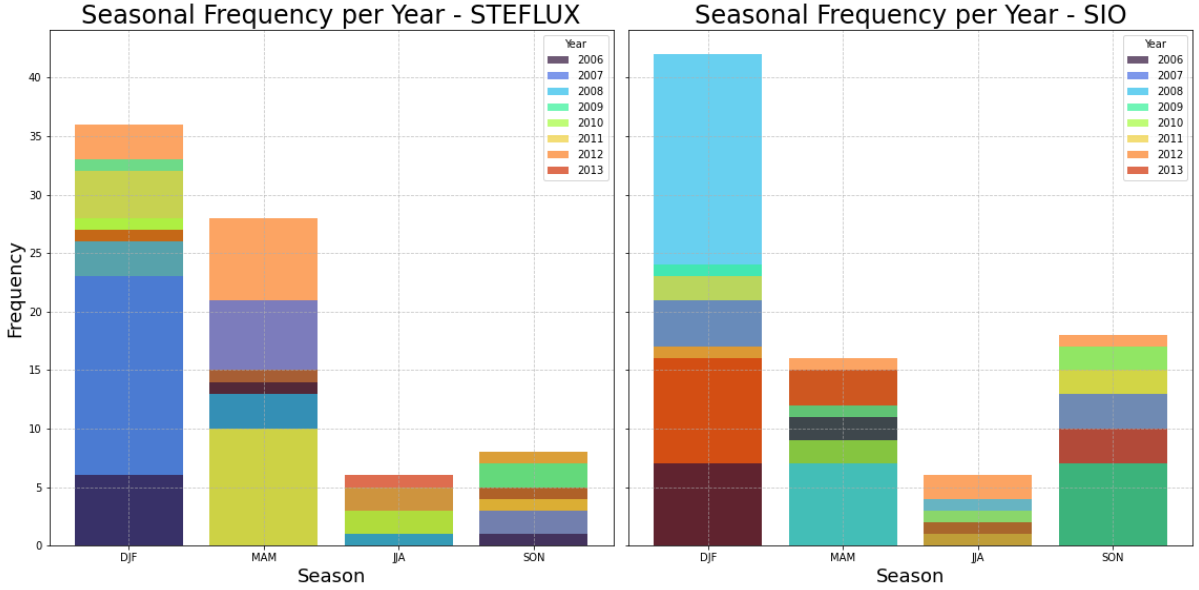


Figure 3.2: Seasonal distribution of STEFLUX (left) and SIO (right) events. Bars represent the four seasons (DJF, MAM, JJA, SON), with stacked colors indicating the contribution of different years. The plots highlight the marked seasonality in the occurrence of events. Here the frequency is the 1-score occurrence. Bars are stacked chronologically: for each season, years are ordered from 2006 (bottom) to 2013 (top), following the same sequence shown in the legend.

Several studies have demonstrated that stratospheric intrusions over the Himalayas and the Tibetan Plateau exhibit a marked seasonal cycle, with higher frequency and intensity during winter and pre-monsoon periods. This behavior is mainly related to the lower winter tropopause height and the enhanced activity of the subtropical jet, which favor tropopause folding and the downward transport of stratospheric air masses. Cristofanelli et al., 2010 reported that at the NCO-P station the contribution of stratospheric intrusions to surface ozone is particularly relevant in winter, reaching up to 24% of the measured values. Fadnavis et al., 2010 also highlighted that enhanced upper-tropospheric ozone of stratospheric origin occurs predominantly in winter and pre-monsoon seasons over India. More recently, studies on tropopause folds over the Tibetan Plateau have confirmed that medium and deep folds occur mainly in winter, with multiple tropopause events reaching very high frequencies during certain winters (Škerlak et al., 2014; Luo et al., 2019).

The time series analysis and the context described above justify the selection of a wintertime event. Since 2008 was a peak year and the analysis in this thesis is restricted to the period 2008–2011, the January–February 2008 event was chosen as a case study for deeper investigation. During this period, both STEFLUX and SIO identified an event of stratospheric intrusion, which was characterized by anomalously high surface ozone concentrations compared to the surrounding periods.

3.4.2 Visualizing the Case Study Event through ERA5, CAMS, and EMAC

In this section, horizontal and vertical distribution maps for each model are shown. The aim is to describe the intrusion event and investigate its dynamics from the perspective of each distinct reanalysis and models utilized in this work.

Horizontal maps are provided to show the variability of O_3 concentrations for each day at 12UTC, with a zoom on January 30th and 31st. Vertical distributions are provided for all the variables (O_3 , O_3S when available, relative humidity RH, and potential vorticity PV), only on January 30th and 31st, at 12UTC.

The choice of these two days is motivated by the fact that they are located in the middle of the intrusion event, when the intrusion is expected to be already well developed and, consequently, all its main features can be more clearly captured. Conversely, to ensure temporal consistency across the comparison, all maps are shown at 12UTC (or as close as possible to this time in the case of EMAC). This choice allows a direct comparison among the two reanalyses and the model by minimizing the influence of the diurnal cycle. This time of the day can provide a representative snapshot of the dynamical conditions, avoiding biases associated with nocturnal or late evening extremes.

ERA5

Figure 3.3 shows the ozone horizontal distribution and evolution over the six days of the event. At 50 hPa, well within the stratosphere, the maps appear saturated, while at 500 hPa, on the opposite side, the signal is much weaker. In contrast, the levels at 300 hPa and 150 hPa, located closer to the tropopause region, provide the most informative representation of the intrusion events. At these altitudes (better visible in Figure 3.4), it is clearly possible to see an ozone mass ranging between 0.1 ppmv and 3 ppmv, which moves over the station spot.

The observation of horizontal maps alone does not clearly determine the origin of the observed ozone mass either in the stratosphere or from horizontal transport. However, by also examining the potential vorticity (PV) contours at 2 PVU and 3.5 PVU—typical stratospheric values as suggested by Wernli and Sprenger, 2007—it becomes evident how these contours outline the ozone masses. This is particularly visible at 300 hPa, whereas at 150 hPa the PV contours indicate a sharp gradient. The same pattern is confirmed by the 0.1 ppmv ozone isoline, which has been used in several studies (Singh et al., 1980; K. Li et al., 2025) as a threshold between background tropospheric and stratospheric ozone. At 150 hPa, the contours show a clear division between regions above 30°N and lower latitudes, corresponding to the location of the station.

This observation reflects a strong gradient in both atmospheric stability and wind speed, resulting in the formation of a dynamic front in potential vorticity (PV). Such PV fronts typically arise under two main dynamical conditions: tropopause folding and jet stream activity (Martius et al., 2010; Hitchman and Rowe, 2021). Tropopause folds occur when stratospheric air, characterized by high PV and elevated ozone concentrations, is advected downward into the troposphere—often in association with upper-level troughs or cut-off lows. This descent creates narrow PV filaments with steep gradients, which appear as sharp boundaries in PV and ozone maps. Similarly, the jet stream, due to its strong wind shear and baroclinic structure, generates intense PV gradients along its flanks. When per-

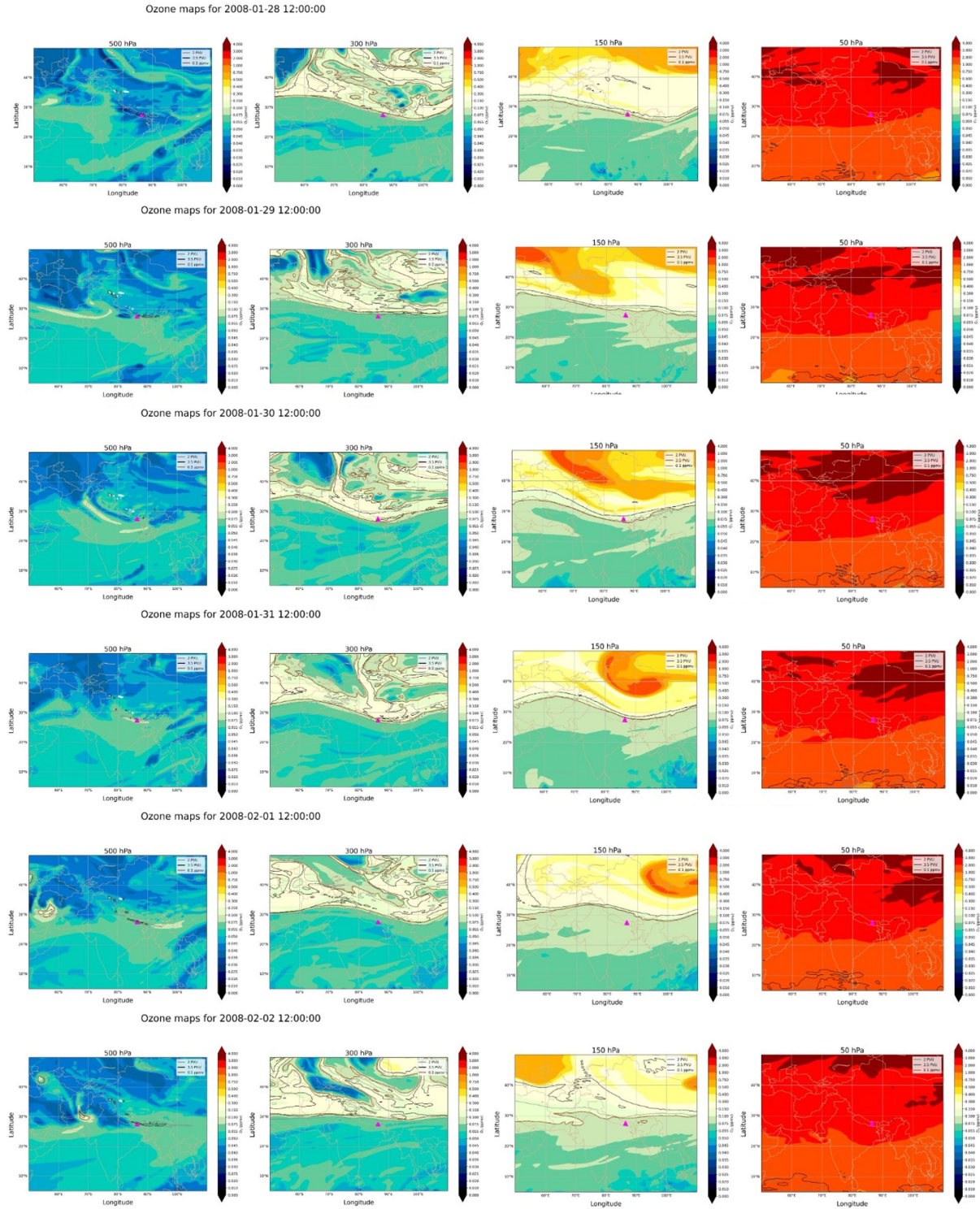


Figure 3.3: ERA5 horizontal maps daily evolution at 500 hPa, 300 hPa, 150 hPa, and 50 hPa, at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by ERA5. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

Ozone maps for 2008-01-31 12:00:00

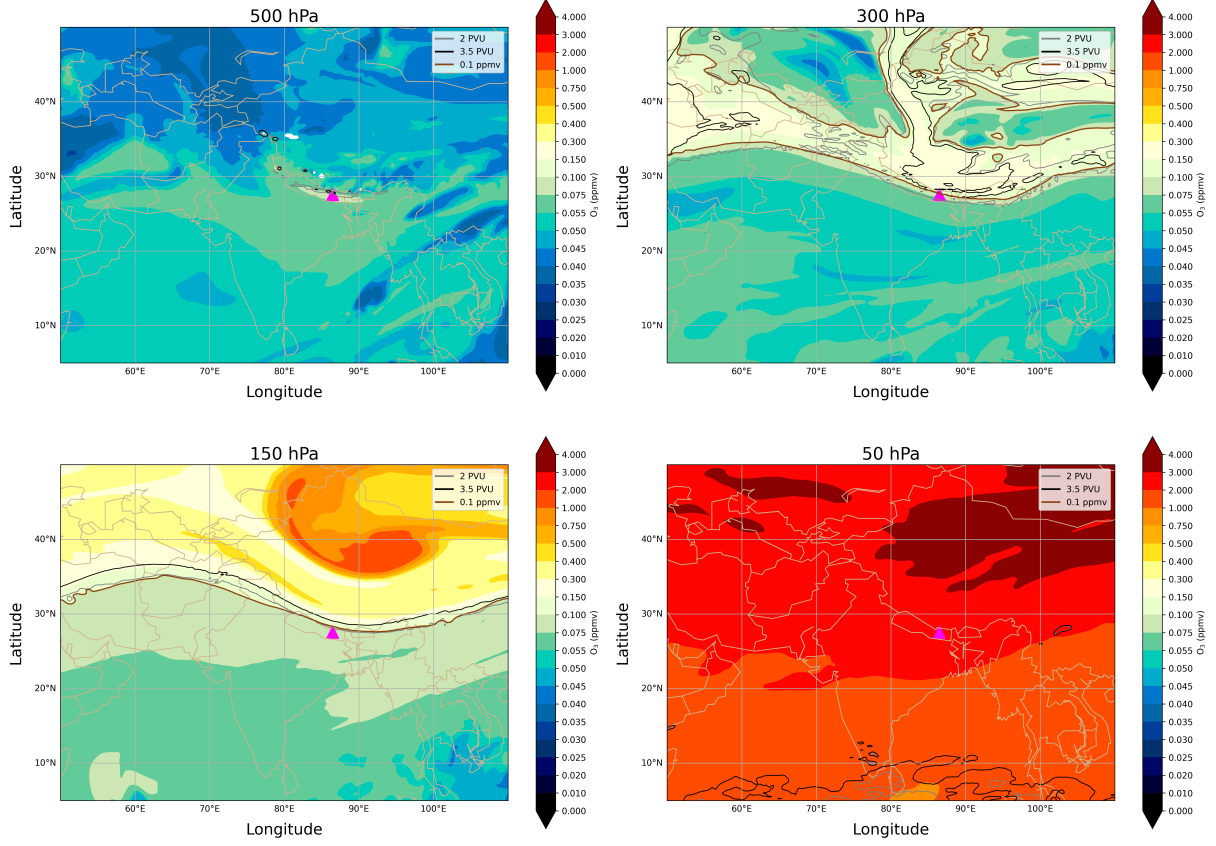


Figure 3.4: ERA5 horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by ERA5. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppbv isoline. The pink triangle represents NCO-P location.

turbed by planetary waves or isolated low-pressure systems (cut-off lows), the tropopause can deform significantly, enhancing the front-like structure of PV contours and facilitating stratosphere–troposphere exchange.

All of these features become even more evident when examining the vertical distributions, which clearly show the stratospheric origin of the ozone intrusions and the associated PV structures throughout the troposphere and lower stratosphere. Figure 3.5 shows the meridional and zonal distribution in coincidence with the station of O₃, PV, and RH.

It is possible to observe here an ozone mass exceeding 0.1 ppbv that descends from the stratosphere down to about 400 hPa, surrounded by potential vorticity (PV) contours at 2 PVU and 3.5 PVU. This feature is particularly evident in the meridional cross section, where a sharp front is clearly visible around 30°N, according to what is observed in horizontal maps. The zonal distribution shows lower ozone values, although still above 0.1 ppbv. Another characteristic of this intrusion also emerges: zonally, a narrowing occurs between 300 hPa and 200 hPa, leading to the detachment of the ozone mass from the upper layers. This narrowing is typical of tropopause folding and PV filaments: when a fold extends downward, stratospheric air masses can detach from the upper stratosphere and penetrate into the middle and upper troposphere. In the literature, this process is

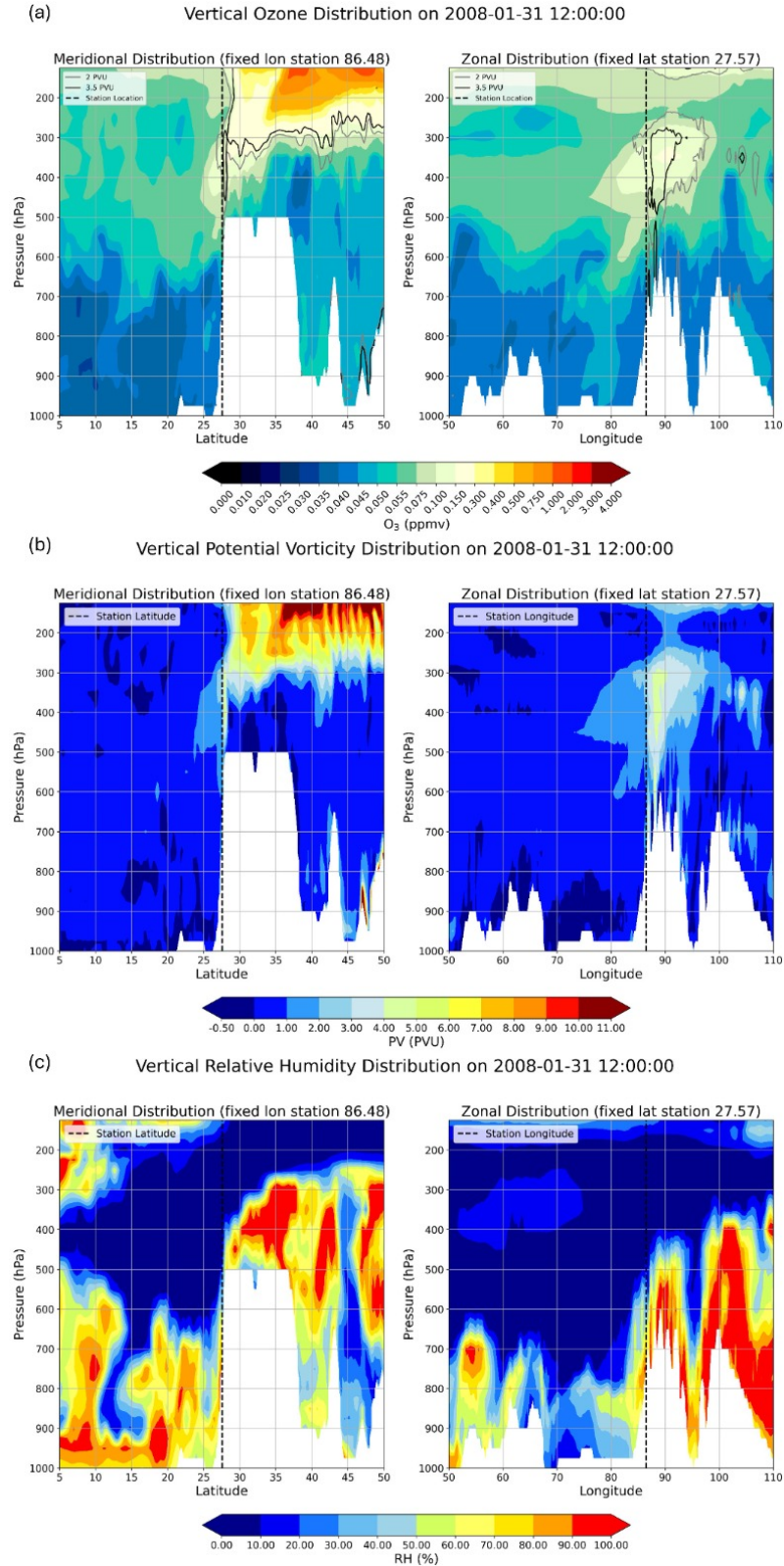


Figure 3.5: ERA5 vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by ERA5, while the black dashed line is the NCO-P location.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

often described as a “cut-off stratospheric intrusion” or “PV filament detachment” (Stohl et al., 2003; Wernli and Sprenger, 2007; Kunz et al., 2015), occurring as a result of planetary wave activity, zonal jets, and tropospheric vortices. Under such conditions, stratospheric air masses can detach and form filaments or isolated nuclei, which subsequently move independently toward lower latitudes. This is clearly visible in the PV vertical distribution in perfect agreement with what is observed in ozone distributions. A core of values higher than 6 PVU detaches and moves downward. The opposite occurs for relative humidity: it is high near the surface but drops to very low values, on the order of 10%, in correspondence with the intrusion. This behavior is typical of stratospheric intrusions, which transport very dry air downward into the troposphere, where it encounters a much more humid environment near the surface.

CAMS

Ozone maps for 2008-01-31 12:00:00

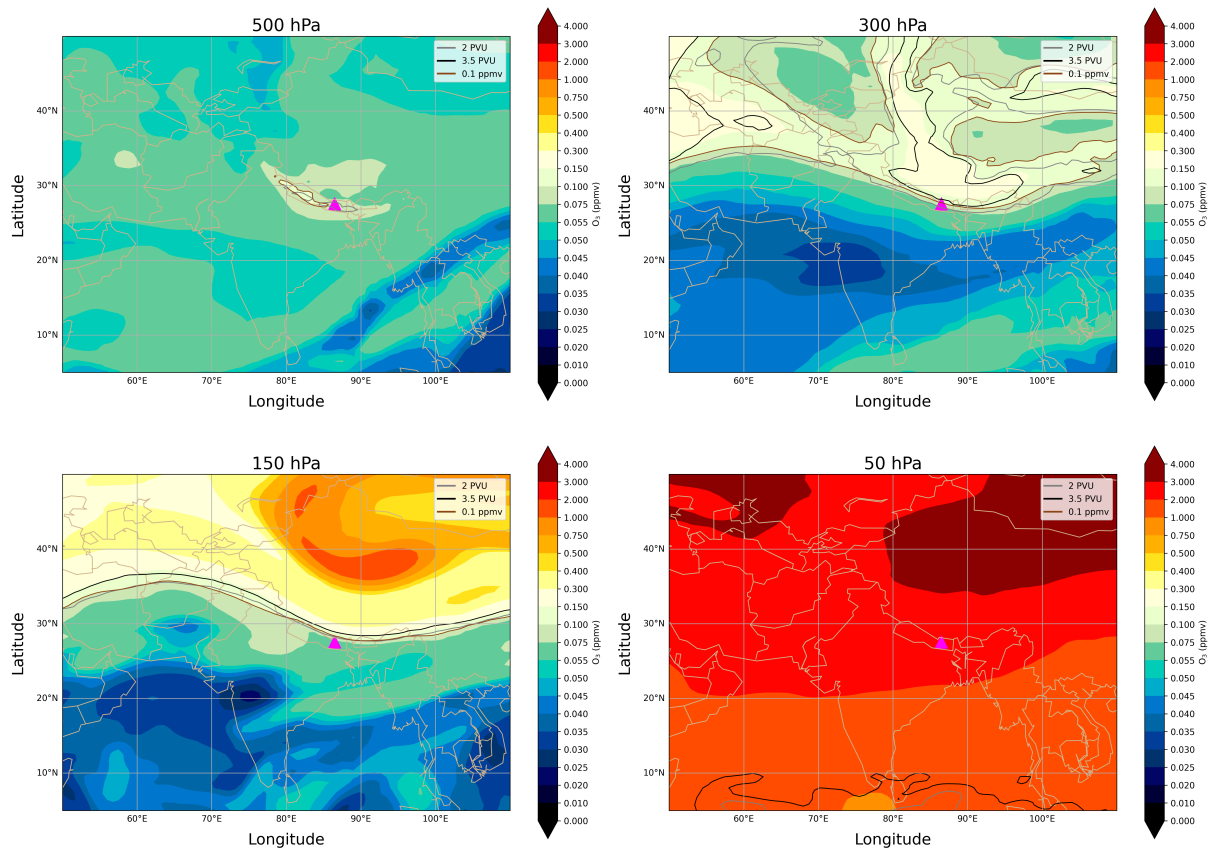


Figure 3.6: CAMS horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.

Figure 3.6 shows the ozone horizontal distribution. In the case of CAMS, similar observations can be made as for ERA5, also for vertical distributions (Figure 3.7). Both reanalyses capture the same features, although CAMS has a lower horizontal resolution. In fact, when compared to the ERA5 maps, some details of the structures appear less

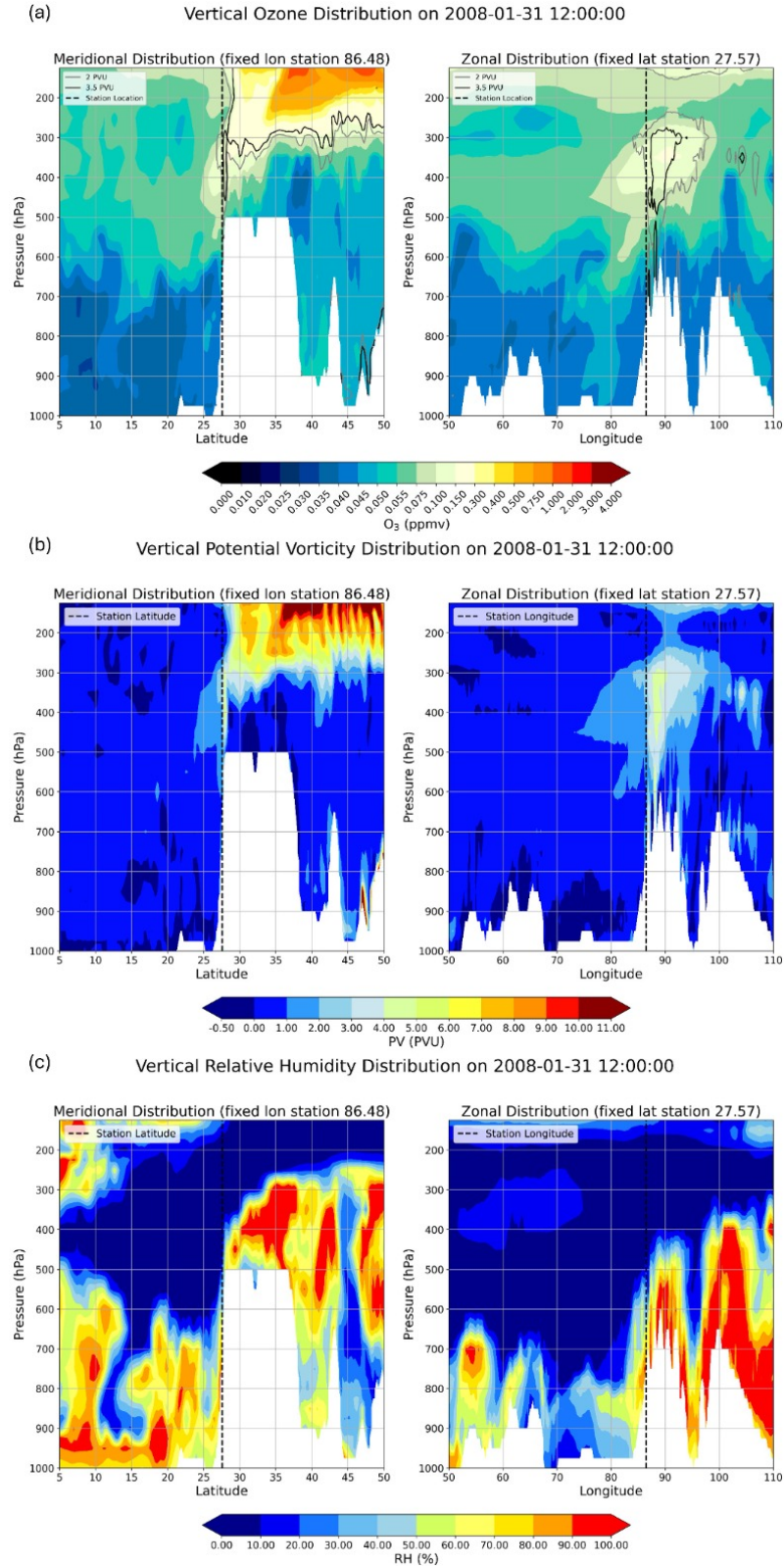


Figure 3.7: CAMS vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

evident. Nevertheless, CAMS seems to capture the essential characteristics of the events. Considering that CAMS includes chemical reactions involving ozone, ERA5 does not appear to show significant differences at first glance.

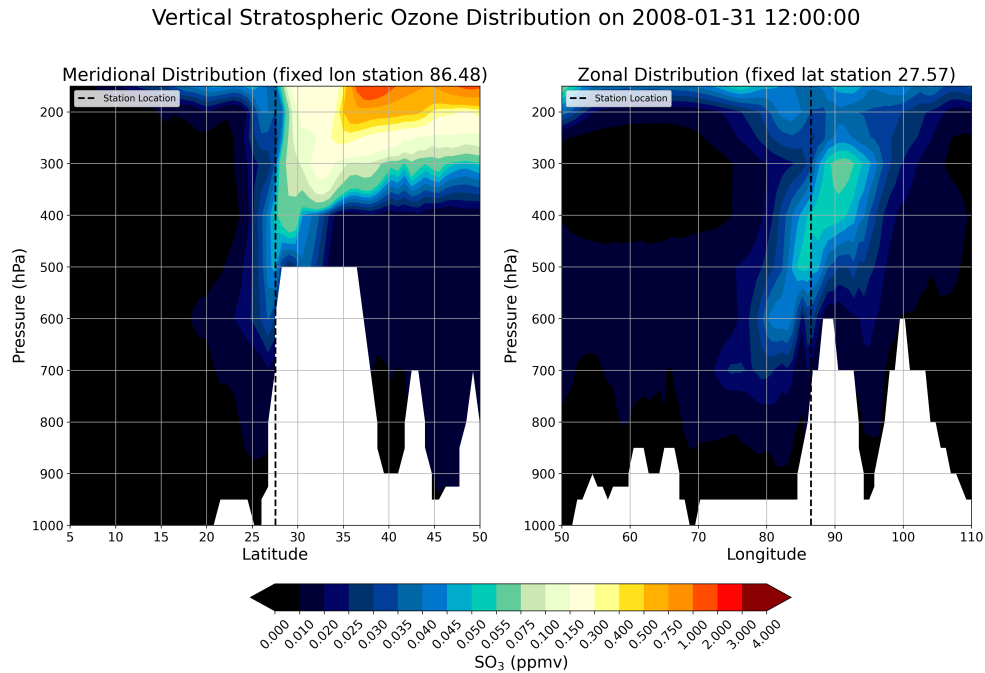


Figure 3.8: CAMS vertical distributions of stratospheric ozone SO₃ (meridional-left, zonal-right) on January 31st, 2008 at 12UTC. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.

What CAMS adds compared to ERA5 is the explicit representation of stratospheric ozone, as shown in Figure 3.8. The vertical maps of stratospheric ozone further confirm all the considerations previously made for ERA5. This allows for some considerations on the contribution of stratospheric ozone to the total ozone. As shown in the Figure 3.9, along the mountain profile near the surface, at higher elevations around 500 hPa, stratospheric ozone accounts for about 30%–40% of the total. At lower altitudes, such as around 800 hPa, its contribution still reaches about 10%.

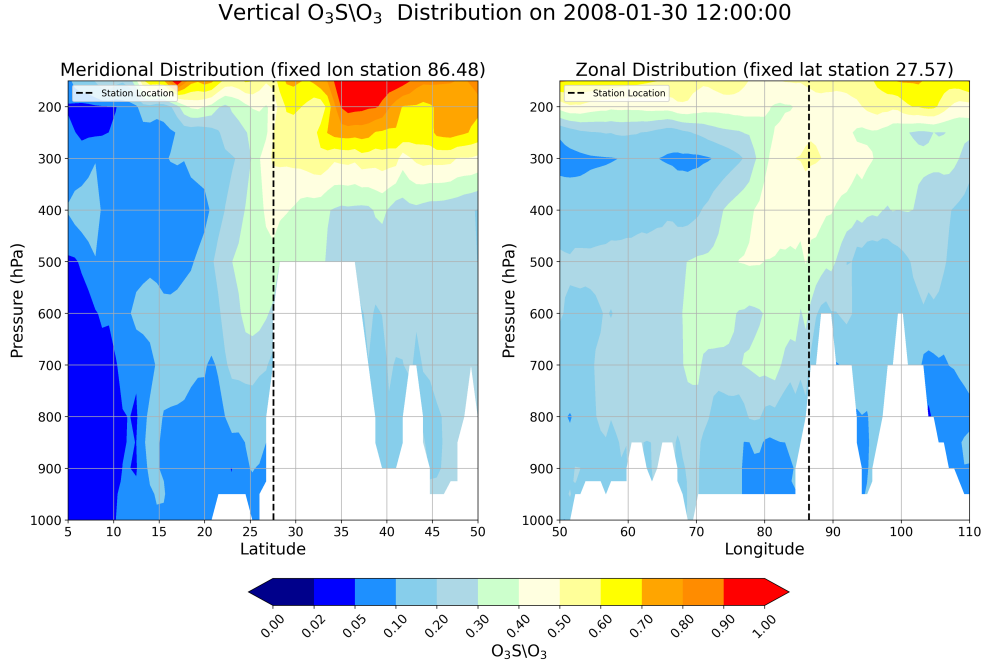


Figure 3.9: CAMS vertical distribution of the ratio between stratospheric ozone O_3S and ozone O_3 for January 30th, 2008 at 12UTC.

EMAC

Figure 3.10 and Figure 3.11 show the ozone horizontal and ozone, PV, and RH vertical distributions for EMAC, respectively. It is also able to capture the event, although its lower horizontal resolution, compared to CAMS and ERA5, makes some features less visible and more smoothed or averaged out. It should be noted that the EMAC figures are not exactly at 12 UTC but vary, since EMAC provides output every 5 hours in order to capture all times of the day. This means that EMAC times are either ahead of or behind those shown for CAMS and ERA5, and therefore the visible structures are slightly different, although the overall evolution of the event can be considered largely comparable.

Although the horizontal resolution of EMAC is lower than that of the reanalyses, the same cannot be said for the vertical resolution. The essence of the phenomenon remains the same as that depicted by ERA5 and CAMS, but in this case the structures appear more defined, as shown in the Figure 3.11. Again, all the presented variables behave consistently with each other and with the phenomenon. Stratospheric ozone as seen by EMAC (Figure 3.12) is higher than in the reanalyses (i.e., CAMS), while the opposite is true for potential vorticity (Figure 3.11-(b)). Relative humidity (Figure 3.11-(c)), on the other hand, shows a less pronounced gradient.

Similar to what was done for CAMS, when focusing on the fraction of stratospheric ozone relative to the total (Figure 3.13), the EMAC results highlight even more clearly than CAMS that almost 100% of the structures visible in the vertical ozone distributions are of stratospheric origin, with a stratospheric contribution exceeding 50% from the surface up to 800 hPa.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

Ozone maps for 2008-01-31 14:00:00

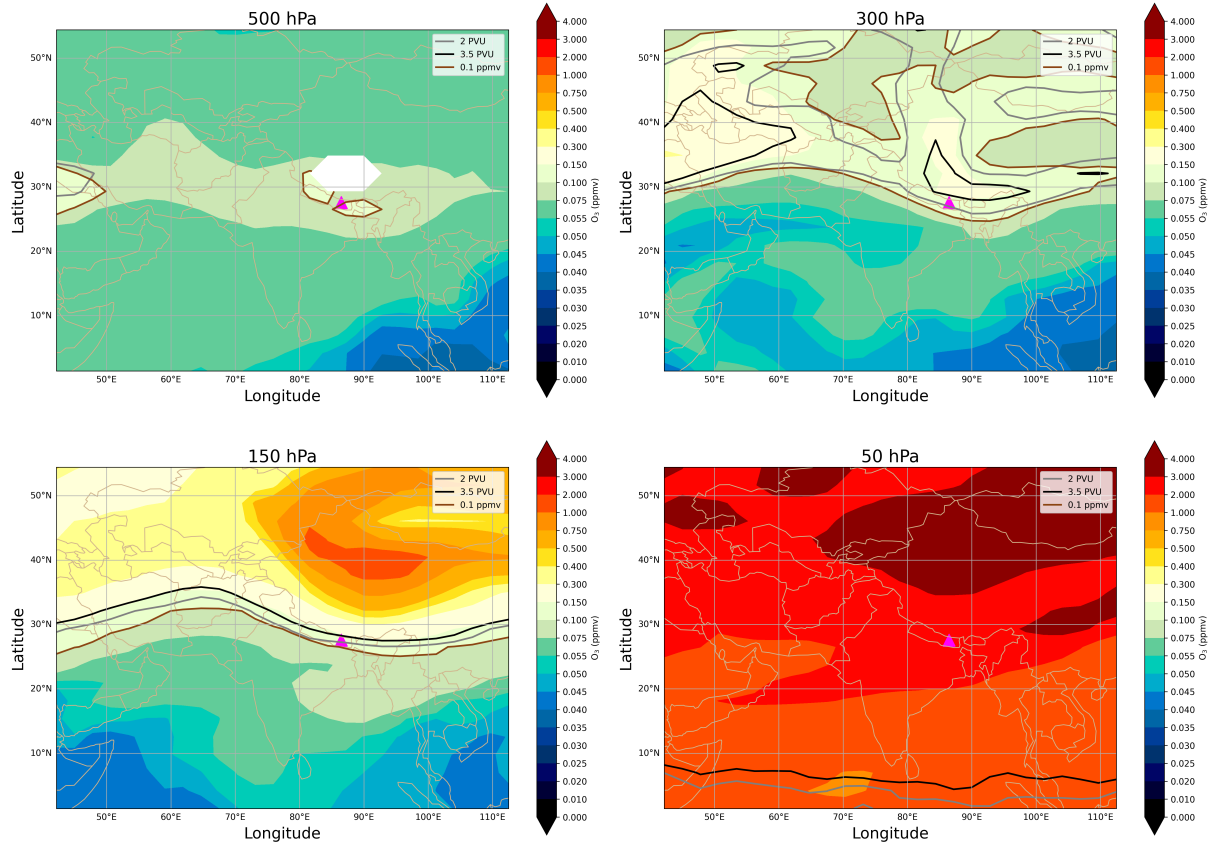


Figure 3.10: EMAC horizontal maps at 500 hPa, 300 hPa, 150 hPa, and 50 hPa. Zoom on January 31st 2008 at 12UTC. White shapes at 500hPa represent the peak of the mountain, as seen by EMAC. Black and grey isolines are 3.5 PVU and 2 PVU, respectively, and the brown one is the 0.1 ppmv isoline. The pink triangle represents NCO-P location.

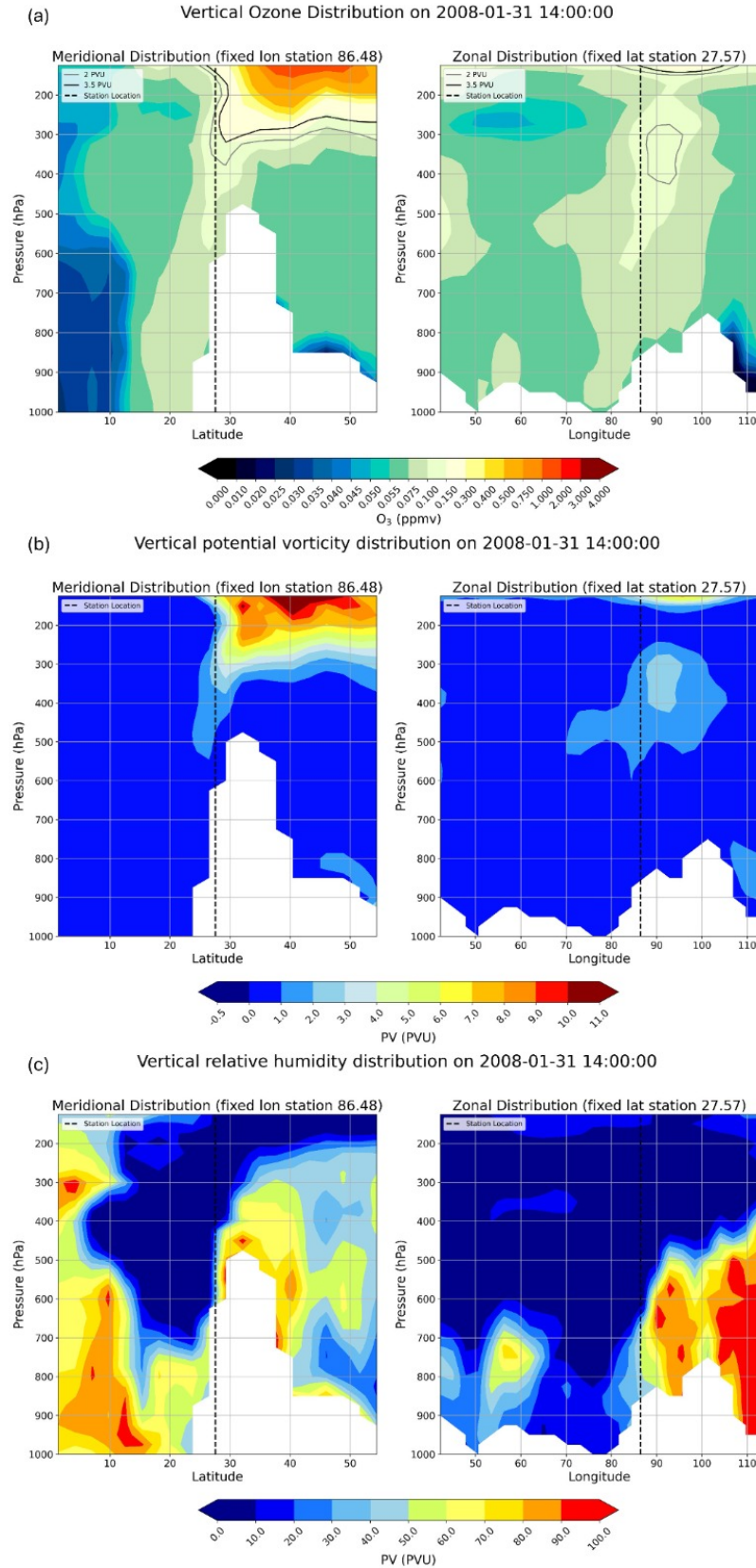


Figure 3.11: EMAC vertical distributions of ozone O_3 (meridional-left, zonal-right) (a), potential vorticity PV (b), and relative humidity RH (c). All the variable are presented for January 31st, 2008 at 12UTC. For ozone, PV isolines at 2 PVU (grey) and 3.5 PVU (black) are shown. The white shape is the mountain profile as seen by EMAC, while the black dashed line is the NCO-P location.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

Vertical stratospheric ozone distribution on 2008-01-31 14:00:00

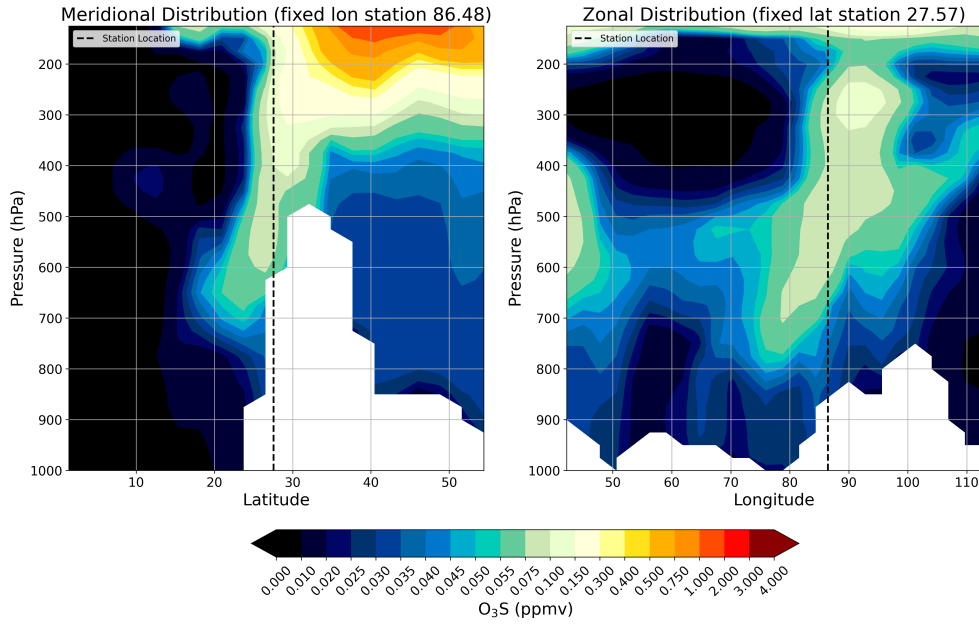


Figure 3.12: EMAC vertical distributions of stratospheric ozone SO_3 (meridional-left, zonal-right) on January 31st, 2008 at 12UTC. The white shape is the mountain profile as seen by CAMS, while the black dashed line is the NCO-P location.

Vertical O_3S/O_3 distribution on 2008-01-30 13:00:00

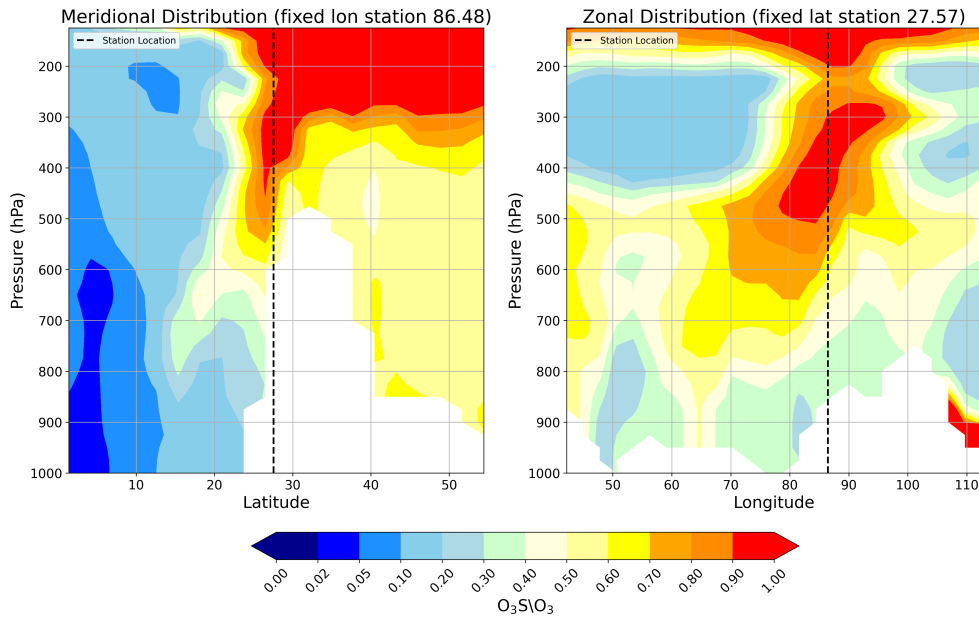


Figure 3.13: EMAC vertical distribution of the ratio between stratospheric ozone O_3S and ozone O_3 for January 30th, 2008 at 12UTC.

Model and Reanalysis Intercomparison

Figure 3.14 upper panels show horizontal maps of $\delta O_3\%$ between ERA5 and CAMS at selected pressure levels (500, 300, 150, and 50 hPa). Red shades indicate higher ozone

Ozone $\Delta(\text{ERA5} - \text{CAMS})$ maps on CAMS grid for 2008-01-30 12:00:00

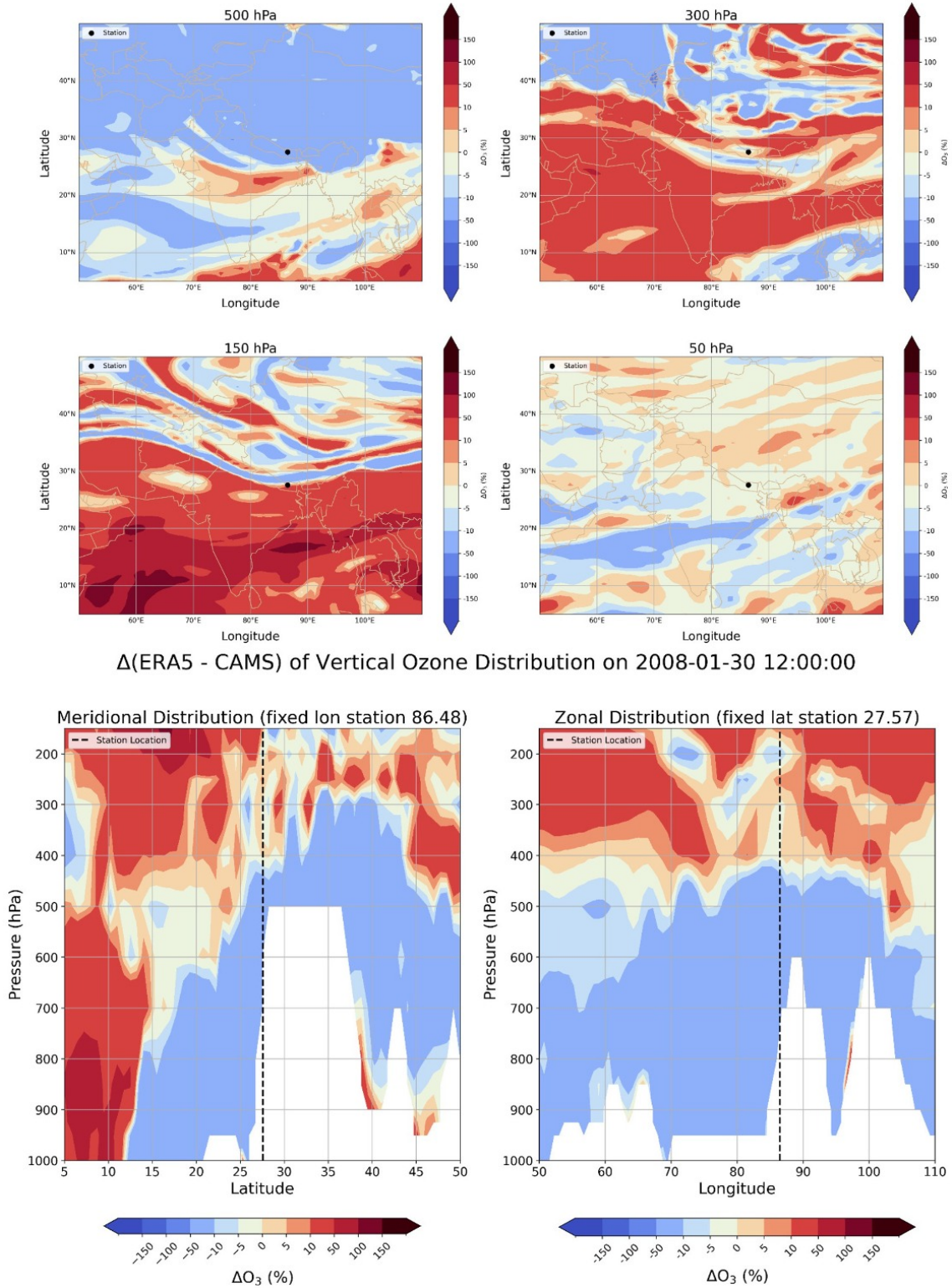


Figure 3.14: O_3 percentage differences between ERA5 and CAMS - horizontal distributions at 500 hPa, 300 hPa, 150 hPa, and 50 hPa (first and second row), and vertical distribution (third row), for January 30th, 2008 at 12 UTC. Black point is the NCO-P location, while white shapes represent the mountain profile.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

values in ERA5 relative to CAMS. At mid-tropospheric levels (500 hPa), differences are generally small, with minor positive anomalies over specific southern regions. In the upper troposphere and lower stratosphere (300–150 hPa), ERA5 systematically exhibits higher ozone concentration values than CAMS across mid to high latitudes. At 50 hPa, differences are more heterogeneous, with alternating positive and negative patterns.

The lower panels present meridional (left) and zonal (right) vertical sections through the station. Differences are most pronounced in the stratosphere, where ERA5 shows substantially higher ozone than CAMS. In the troposphere, relative differences are smaller; however, this can be due to the low ozone concentrations in this region rather than a closer agreement between the models. Sharp δ O₃% gradients are evident near the tropopause, indicating that the largest relative discrepancies are confined to the upper atmospheric layers.

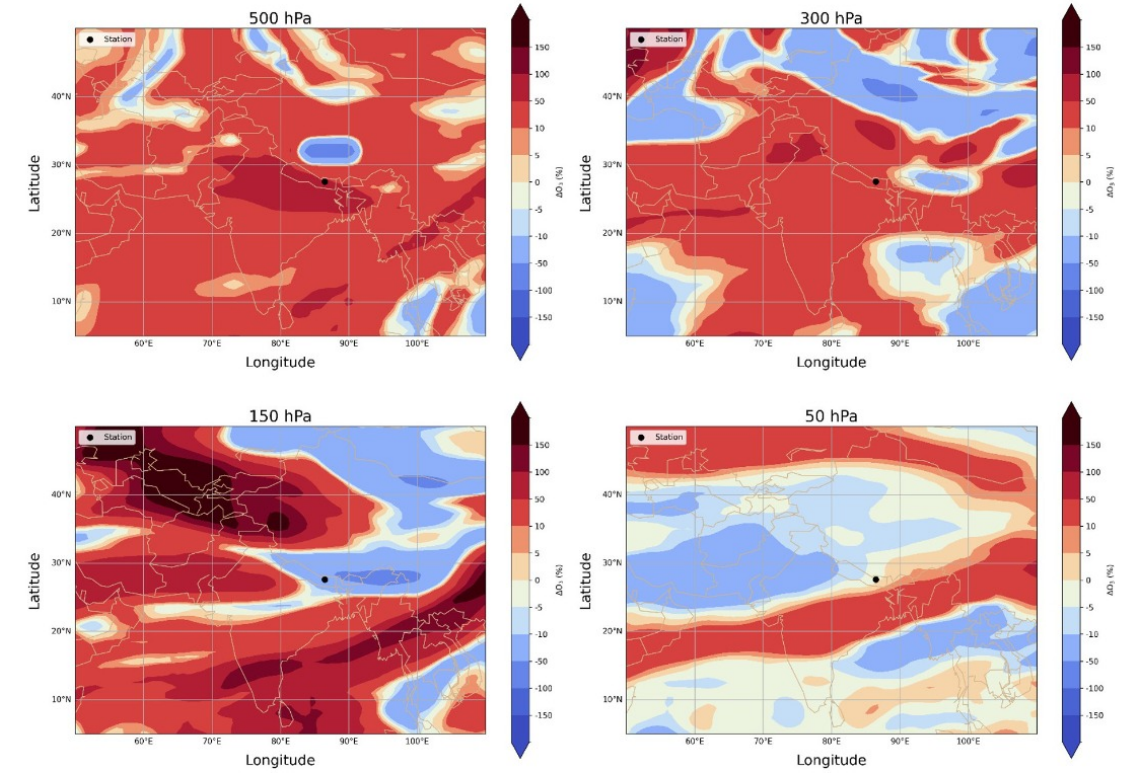
Overall, the figure demonstrates that ERA5 generally overestimates ozone relative to CAMS in the stratosphere, whereas both datasets are more consistent in the lower troposphere. Spatial heterogeneity is evident, emphasizing regional variability in the ERA5–CAMS differences.

Figure 3.15 upper panels show horizontal maps of δ O₃% between EMAC and CAMS at selected pressure levels (500, 300, 150, and 50 hPa). Similar reasoning can be done for this case. At mid-tropospheric levels (500 hPa), EMAC systematically and uniformly shows higher ozone values compared to CAMS. This uniformity is lost at higher levels - in the upper troposphere and lower stratosphere (300–150 hPa) - which, instead, appear more heterogeneous. At 50 hPa, differences are more heterogeneous, with alternating positive and negative patterns, but higher compared to the ERA5–CAMS case.

The lower panels present meridional (left) and zonal (right) vertical sections through the station. Differences are always most pronounced in the stratosphere, where EMAC shows higher ozone than CAMS. However, this time also tropospheric differences remain high, meaning that EMAC generally simulates higher total ozone values with respect to CAMS. Larger differences in stratospheric ozone are observed between EMAC and CAMS (Figure 3.16). EMAC tends to overestimate stratospheric ozone relative to CAMS, with a noticeable contribution also in the troposphere.

It is important to note that these considerations are based only on a six-day case study and may not be generally valid. More general conclusions would require a statistical and climatological analysis of multiple events, which is beyond the scope of the present work.

Ozone $\Delta(\text{EMAC} - \text{CAMS})$ maps on CAMS grid for 2008-01-29 12:00:00



$\Delta(\text{EMAC} - \text{CAMS})$ of Vertical Ozone Distribution on 2008-01-29 12:00:00

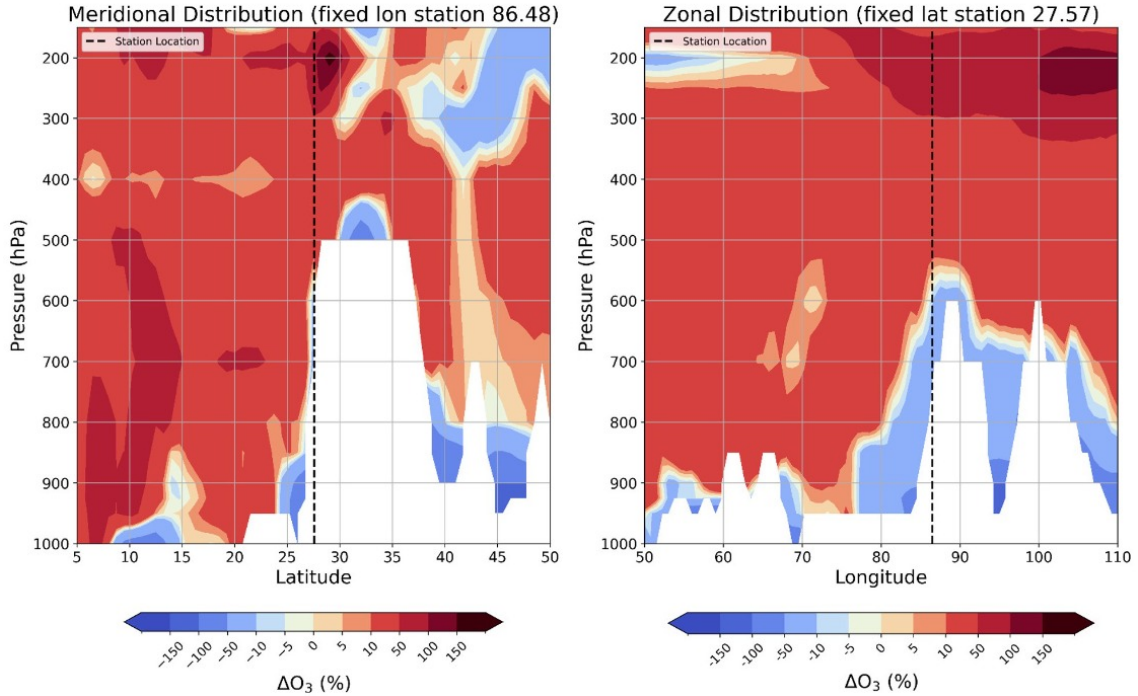


Figure 3.15: O₃ percentage differences between EMAC and CAMS - horizontal distributions at 500 hPa, 300 hPa, 150 hPa, and 50 hPa (first and second row), and vertical distribution (third row), for January 29th, 2008 at 12 UTC. Black point is the NCO-P location, while white shapes represent the mountain profile.

$\Delta(\text{EMAC} - \text{CAMS})$ of Vertical Stratospheric Ozone Distribution on 2008-01-29 12:00:00

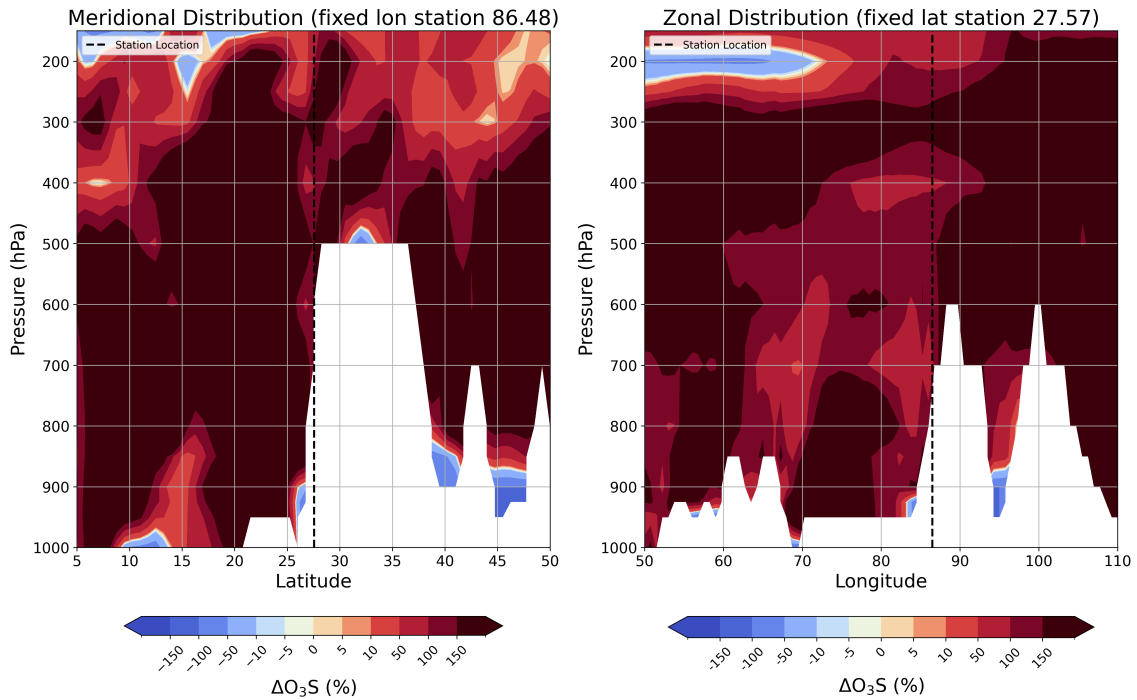


Figure 3.16: O_3S percentage differences between EMAC and CAMS - vertical distributions on January 29th, 2008, at 12 UTC. Black dashed line is the NCO-P location.

3.4.3 Profiles Characterization

The guiding question for this part of the work is: how do event and non-event profiles appear in the selected case study? To address this, as explained in Section 3.3.3, two areas were selected from the reanalysis and model results (Section 3.4.2): one where an intrusion feature was present (Lat: 25-50°N, Lon: 80-110 °E), and one where it was absent (Lat: 5-20°N, Lon: 50-70 °E).

Figure 3.17, Figure 3.18, and Figure 3.19 show how ERA5, CAMS, and EMAC capture event profiles and background or non-event profiles, respectively. Both reanalysis and models in ppmv capture two distinct peaks: the main peak is between 30 km and 40 km corresponding to the stratospheric ozone maximum, while the second peak is just above the surface and below 10 km. Same characteristics are also visible in partial columns. However, in this case, peaks are distributed across more vertical layers for event profiles, and are no more at the same level between mean event profiles and non-event profiles. Generally, the main peak of non-event profiles is larger than event profiles, while non-events profiles show lower values below about 25 km. Notably, the variability associated with event profiles is significantly larger compared to that of non-event profiles. This increased spread reflects the more complex and heterogeneous nature of ozone structures during intrusion events, which often involve multi-layered features and varying intensities across different altitudes.

To manage the possible presence of outliers, in addition to the mean and standard deviation, the median and interquartile range (IQR) were also computed to provide a more robust and resistant statistical indicator of the ozone profiles. Indeed, while the mean

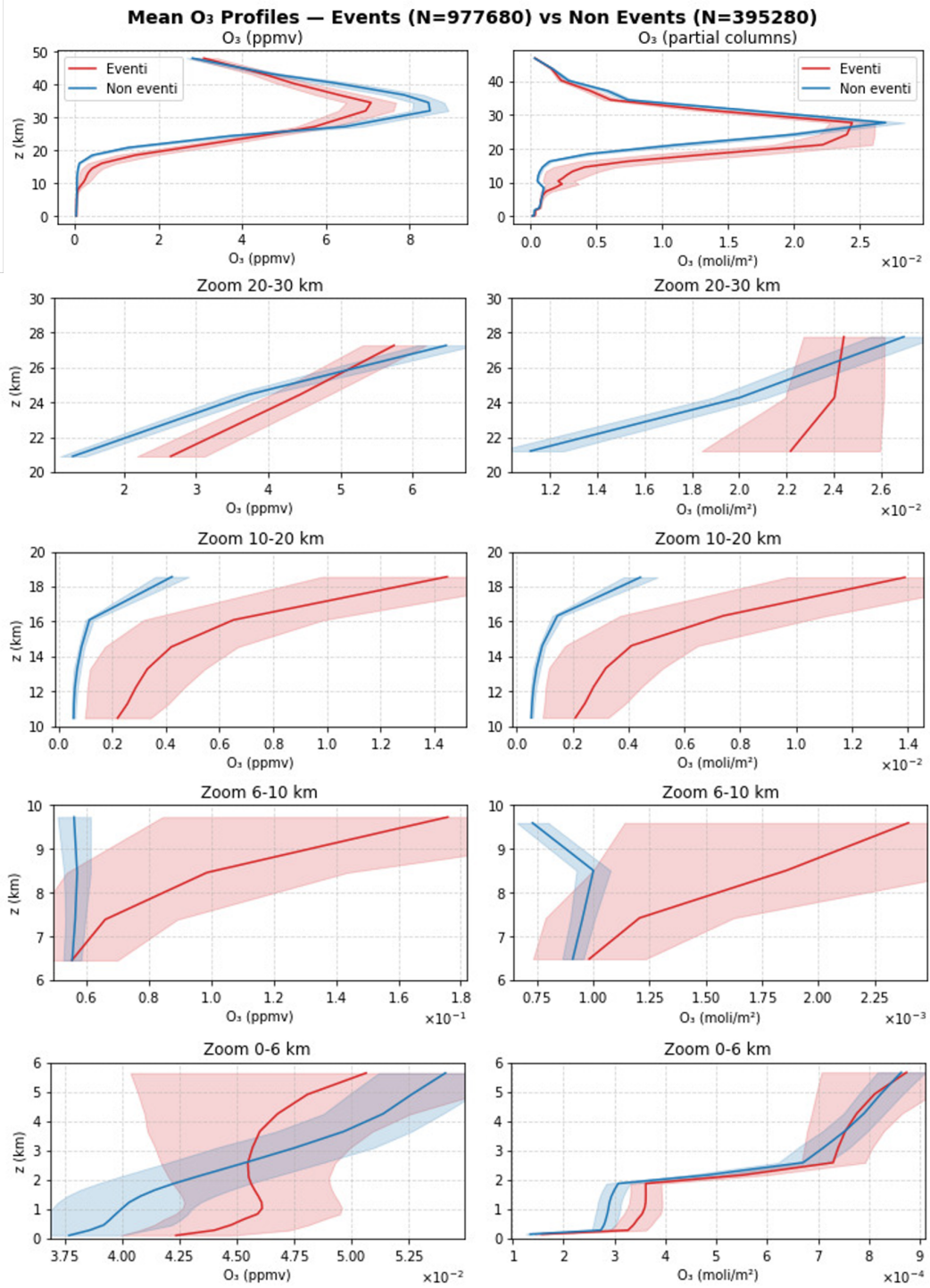


Figure 3.17: ERA5 mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

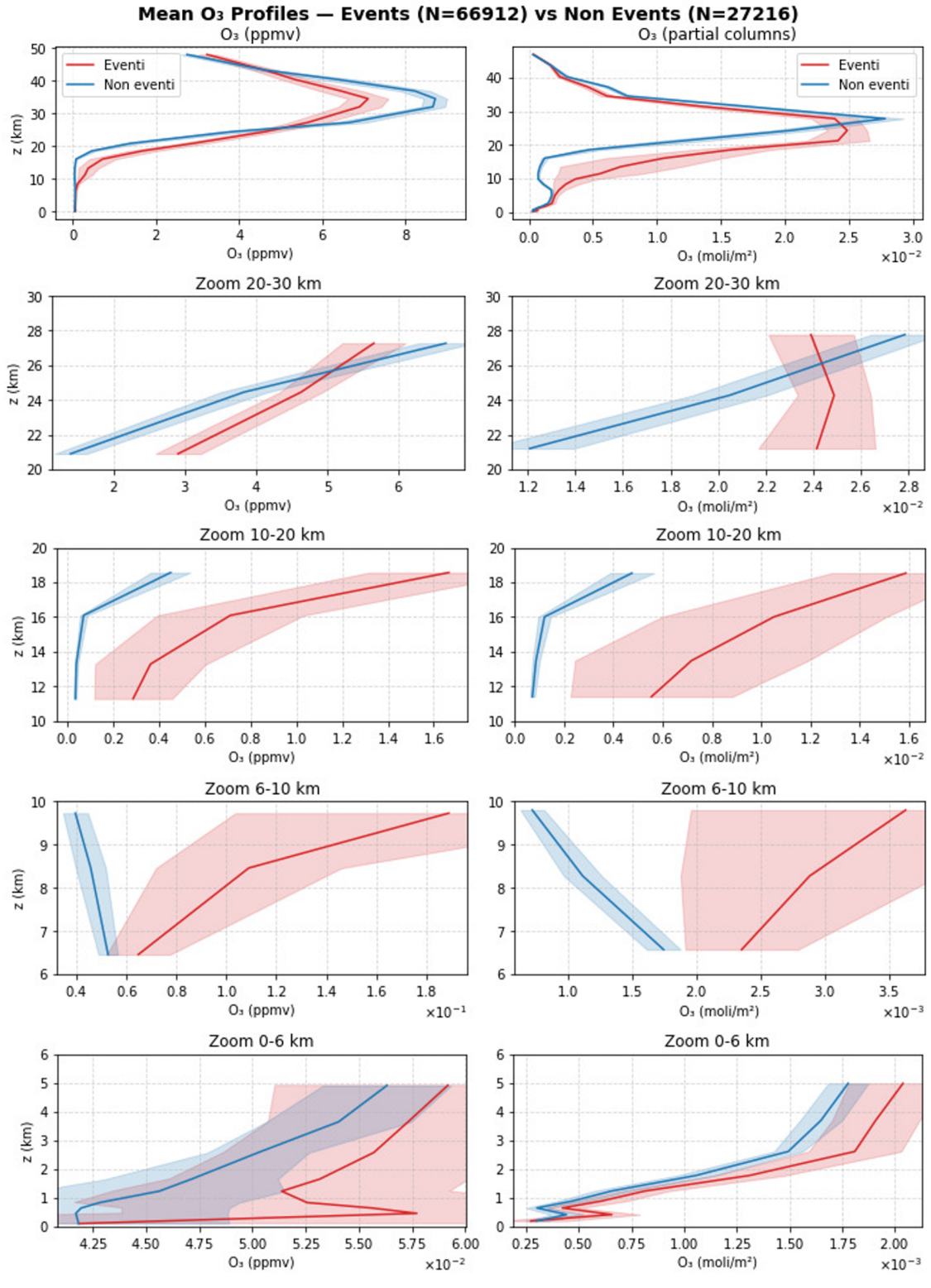


Figure 3.18: CAMS mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

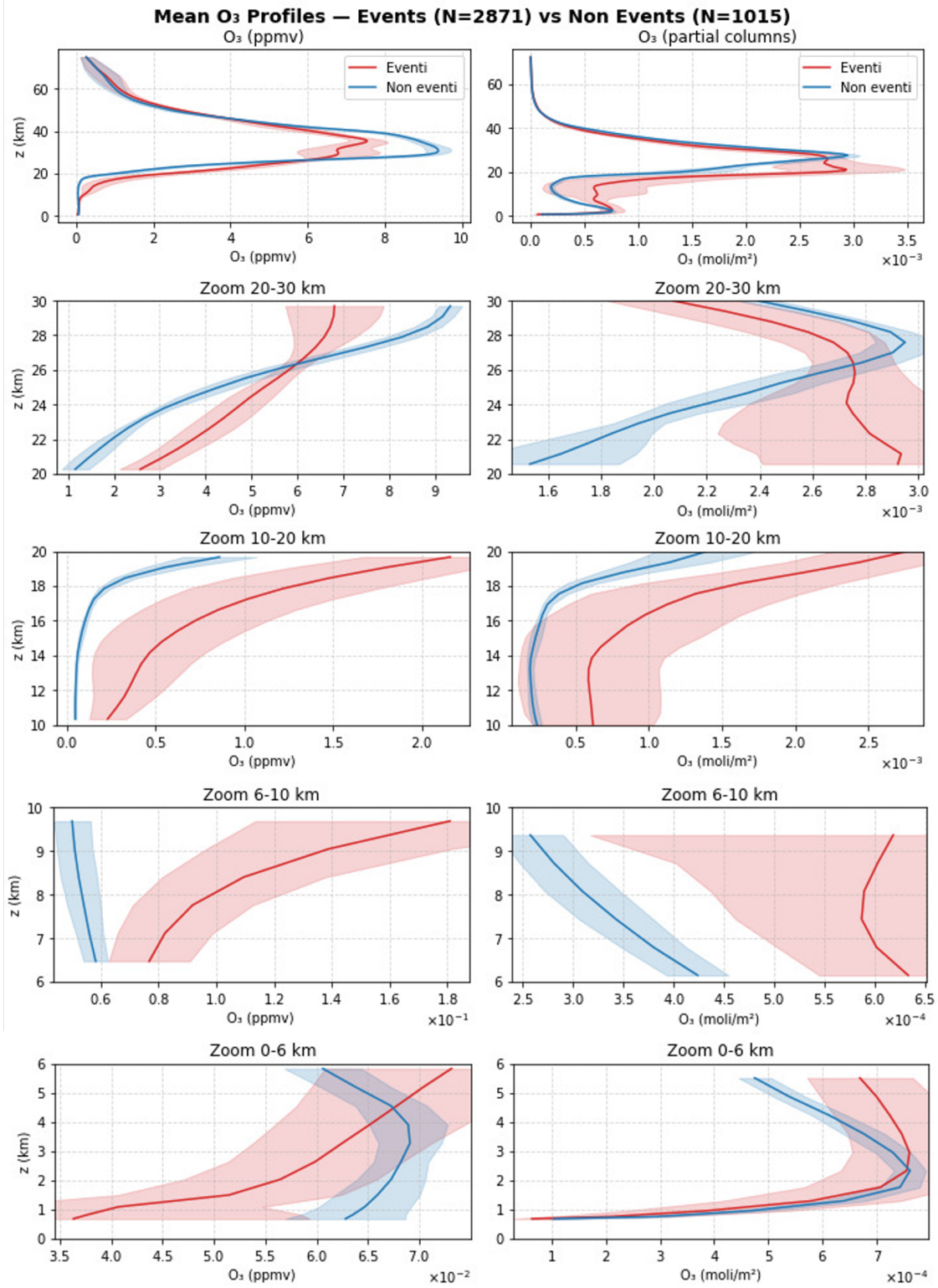


Figure 3.19: EMAC mean profiles and standard deviation (1σ) in region with event features (red), and non-event features (blue). The first row shows the entire mean profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

and standard deviation are sensitive to the presence of outliers and skewed distributions very diverging from the Gaussian assumption, the median and IQR metrics offer complementary insights by capturing the central tendency and spread of the data. Figure 3.20, Figure 3.21, and Figure 3.22 show the same figures described above, but obtained for median of profiles in the two subregions, and the interquartile range as the difference between the 25th to the 75th percentile for ERA5, CAMS, and EMAC, respectively.

In both statistics, intrusion events exhibit enhanced ozone concentrations in the upper troposphere (approximately 6–12 km) and a concurrent depletion in the lower stratosphere (around 12–20 km) compared to non-events. This opposite anomaly pattern clearly indicates a downward transport of stratospheric air rich in ozone into the troposphere, consistent with intrusion processes that locally perturb the tropopause region.

In the lowest troposphere (0–6 km), differences are relatively small and mostly confined to a few kilometers above the surface, suggesting that only the most intense intrusions are able to penetrate this deep. Between 6 and 10 km, however, the contrast between events and non-events becomes strongest, marking the typical altitude range where the descending stratospheric filaments mix with tropospheric air. In the 10–20 km layer, ozone values are systematically lower during events, reflecting the partial “drainage” of ozone from the lower stratosphere toward the troposphere. Above 20 km, the two profiles converge, indicating that large-scale stratospheric dynamics dominate and the influence of individual intrusions becomes negligible.

The large shaded envelopes, especially in the upper-tropospheric region, reveal a substantial variability among individual events, in line with the episodic and spatially heterogeneous nature of intrusion events. Comparing mean and median profiles highlights the skewness of the ozone distribution: the mean emphasizes the contribution of the upper extreme highly affected by the high-ozone intrusions, while the median provides a more robust representation of the typical event structure. Accordingly, the interquartile range (IQR) narrows the uncertainty by excluding outliers, showing a smoother transition across the tropopause. Partial-column profiles confirm the same behavior in both statistics, with enhanced tropospheric ozone loadings and slightly reduced lower-stratospheric values during events, evidencing a net downward redistribution of total ozone mass. Overall, the two statistics together suggest that ozone intrusions are intermittent but high-impact phenomena, capable of substantially modifying the vertical ozone structure through the transport.

Finally, Figure 3.23 shows the mean and median profiles, respectively, for the fused and levels 2 profiles within the two sub-regions.

Compared with ERA5, CAMS, and EMAC, fused, MIPAS and IASI ozone profiles display distinct vertical structures that better reflect the observational nature of the data. In the reanalyses and model simulation, event and non-event profiles generally exhibit main peak between 25 and 35 km (looking partial columns), which is stronger for non-events than for events. In contrast, the observational datasets reveal an opposite but physically consistent pattern: non-event profiles show a weaker and higher-altitude ozone maximum, while event profiles exhibit a stronger peak located at slightly lower altitudes. Below about 25 km, the event profiles remain consistently higher than the non-event ones, confirming that real intrusion episodes are associated with an actual downward transport of ozone-rich air into the upper and middle troposphere.

This downward shift of the ozone maximum during events in the fused and MIPAS profiles suggests that the models may underestimate the vertical displacement of ozone associated

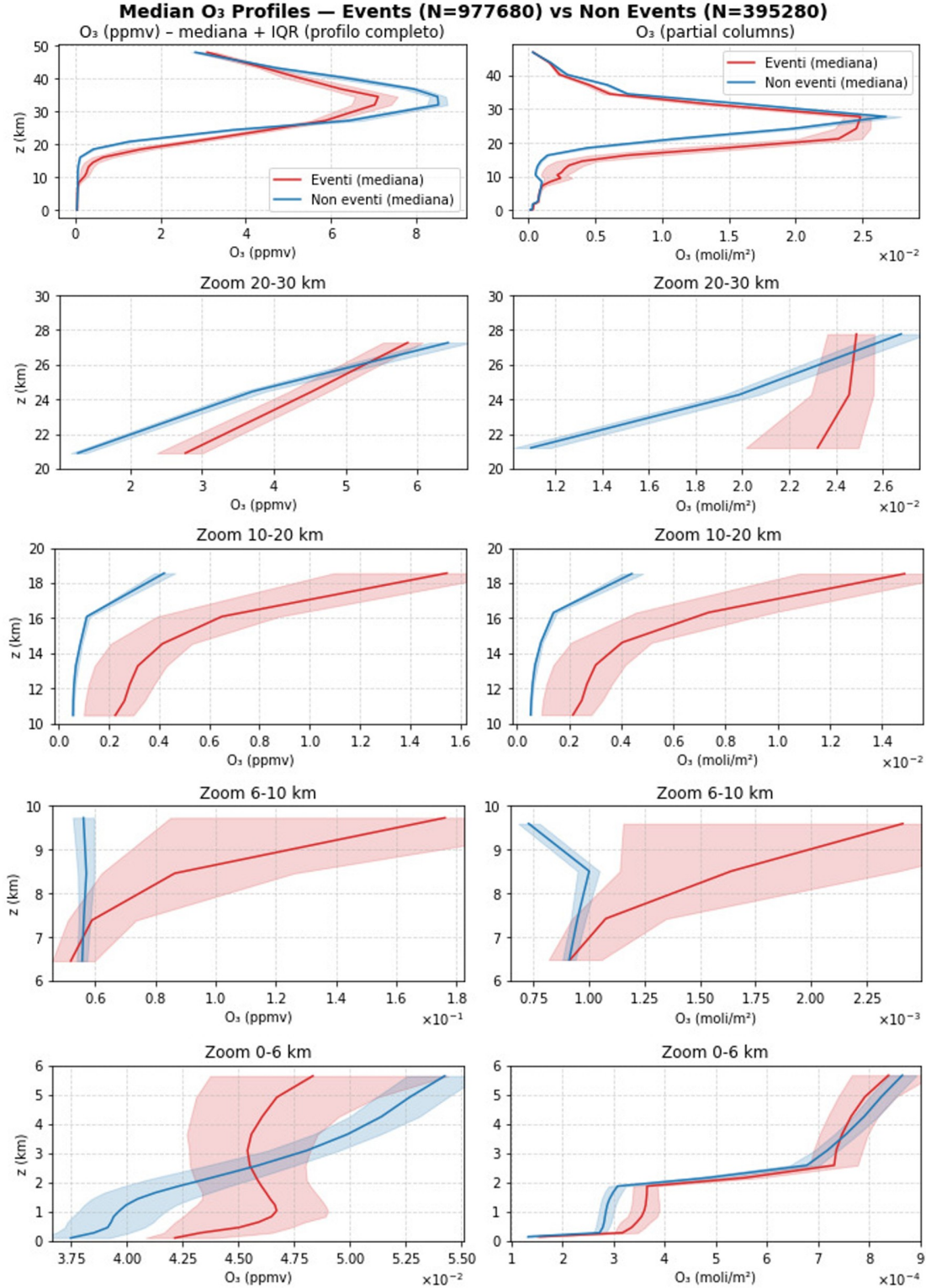


Figure 3.20: ERA5 median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

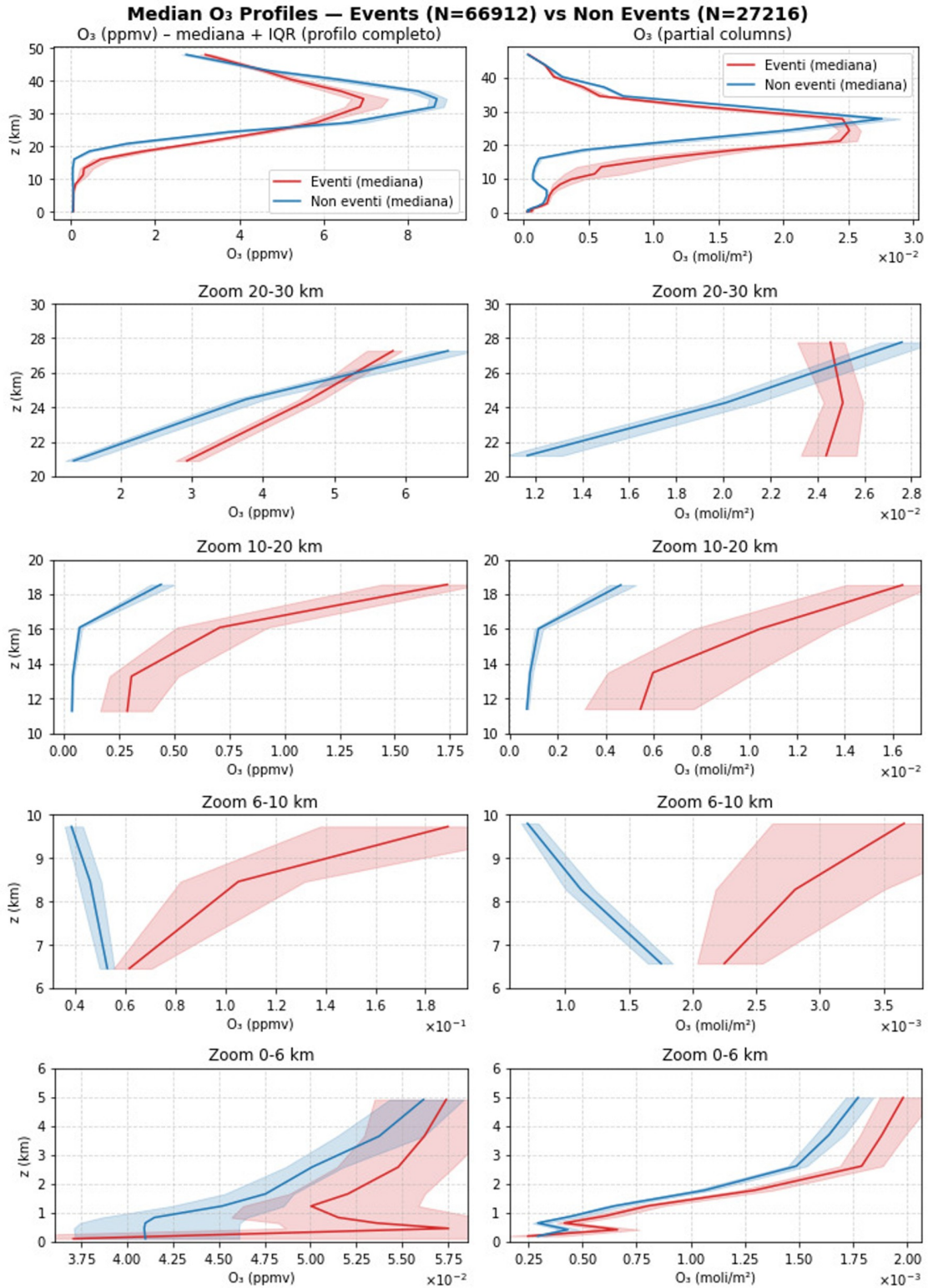


Figure 3.21: CAMS median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

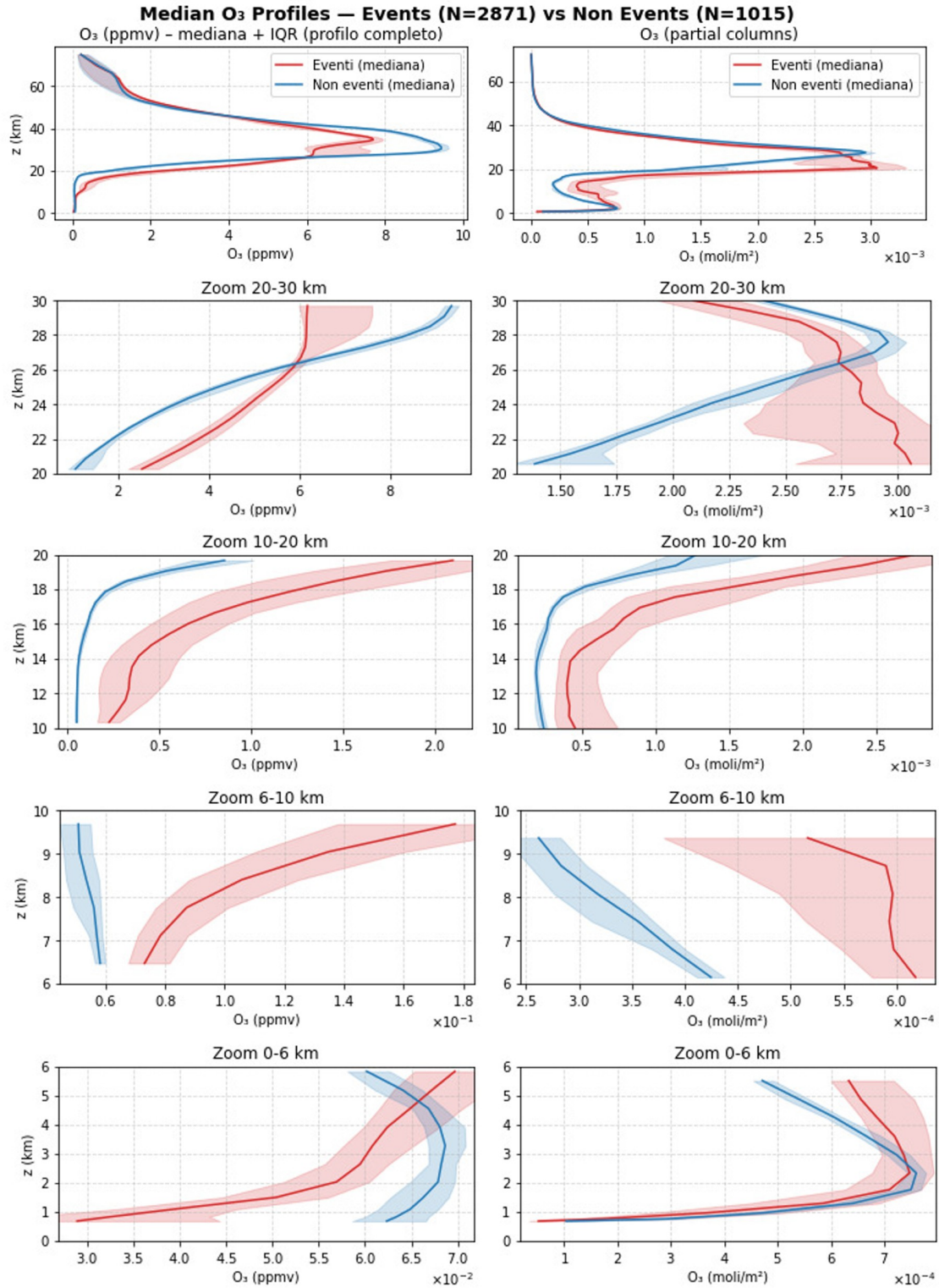


Figure 3.22: EMAC median profiles and interquartile range (IQR) as the difference between the 25° and 75° percentiles in region with event features (red), and non-event features (blue). The first row shows the entire median profile, the following 4 rows are a zoom for different atmospheric layers in the troposphere and stratosphere: 0-6km, 6-10km, 10-20km, 20-30km. The first column shows ozone in ppmv, the second column shows ozone in partial columns.

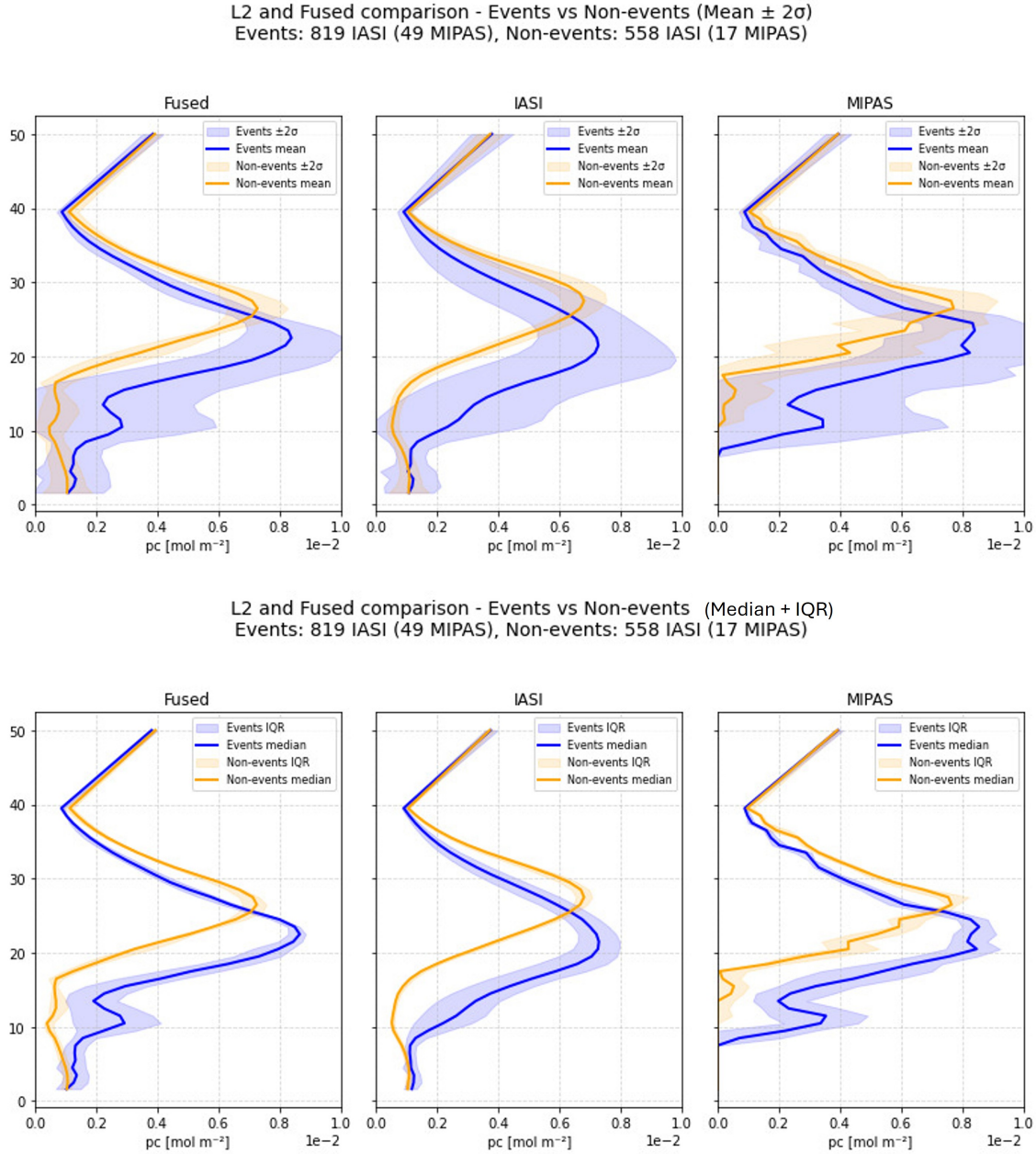


Figure 3.23: First row: mean and standard deviation (2σ) for fused and L2 event and non-event profiles. Second row: median and interquartile range (IQR) between the 25° and 75° percentile for fused and L2 event and non-event profiles.

with stratospheric intrusions, but a more extended study is needed to investigate this point. The smoother and more vertically confined peaks in ERA5, CAMS, and EMAC likely can arise from the effects of coarser vertical resolution and the use of assimilated fields constrained by transport schemes that damp sharp gradients. In contrast, the observational datasets, especially MIPAS and the fused product, retain more of the fine-scale vertical variability typical of real ozone filaments descending through the tropopause. MIPAS shows overall more irregular and jagged vertical structures, consistent with its higher vertical sensitivity in the upper troposphere–lower stratosphere (UTLS) but limited coverage below about 8 km, due to the absence of measurements. IASI, by contrast, provides smoother profiles with weaker signatures of ozone enhancement, reflecting its broader averaging kernels and lower vertical sensitivity in the UTLS region. The fused product combines the strengths of both sensors, producing profiles that are vertically more resolved than IASI alone and extending the information content downward compared to MIPAS. Notably, the fused dataset captures an additional ozone enhancement for event cases slightly above 10 km—also visible, though less pronounced, in MIPAS but almost absent in IASI alone, indicating that the fusion effectively reconstructs intermediate layer features linked to intruding ozone.

As for variability, both the standard deviation and the interquartile range (IQR) are larger for event profiles than for non-events, reflecting the inherently heterogeneous and episodic nature of STE events. The fused dataset also tends to exhibit greater variability in the lower atmosphere compared to IASI, likely because the fusion procedure enhances sensitivity where both instruments contribute complementary information, or maybe due to additional error introduced within the fusion method. The median profiles show only moderate differences from the mean, but the narrower IQR bands emphasize that most of the variability arises from a subset of particularly outlier profiles.

Overall, these observational results support and refine the model-based interpretation: ozone intrusions are intermittent and spatially variable processes that can significantly reshape the vertical ozone structure. However, the observations reveal that the actual vertical displacement and local intensity of ozone enhancements are stronger and more structured than those simulated by reanalyses and models, highlighting the added value of using synergistic satellite datasets to characterize the fine-scale vertical features of stratosphere-to-troposphere exchange.

3.4.4 What We Learned from Classification: Towards O₃ Intrusions Detection

As a last step of this work the k-means classification was applied to the entire fused and L2 profiles for the 6-days (January 28th–February 2nd, 2008) event in the entire selected region over the Himalayas (Lat: 5–50°N; Lon: 50–110°E), first. The guiding questions we want to answer in this section can be summarized as follows: (1) do the characteristics of the observational data in the selected area support the possibility of classifying them according to the presence of intrusive phenomena? (2) How important is this criterion in representing the variability of the observed data? (3) Is it possible to correlate the responses obtained with unsupervised clustering with those obtained using information gleaned from model observations, and in particular from the selection of two areas, one in which the models indicate intrusion and one in which the models agree in indicating its absence?

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

At this step, a number of cluster $k=2$ was used. Indeed, Figure 3.24 shows the result of the elbow method which was used to confirm the choice of the k number as explained below.

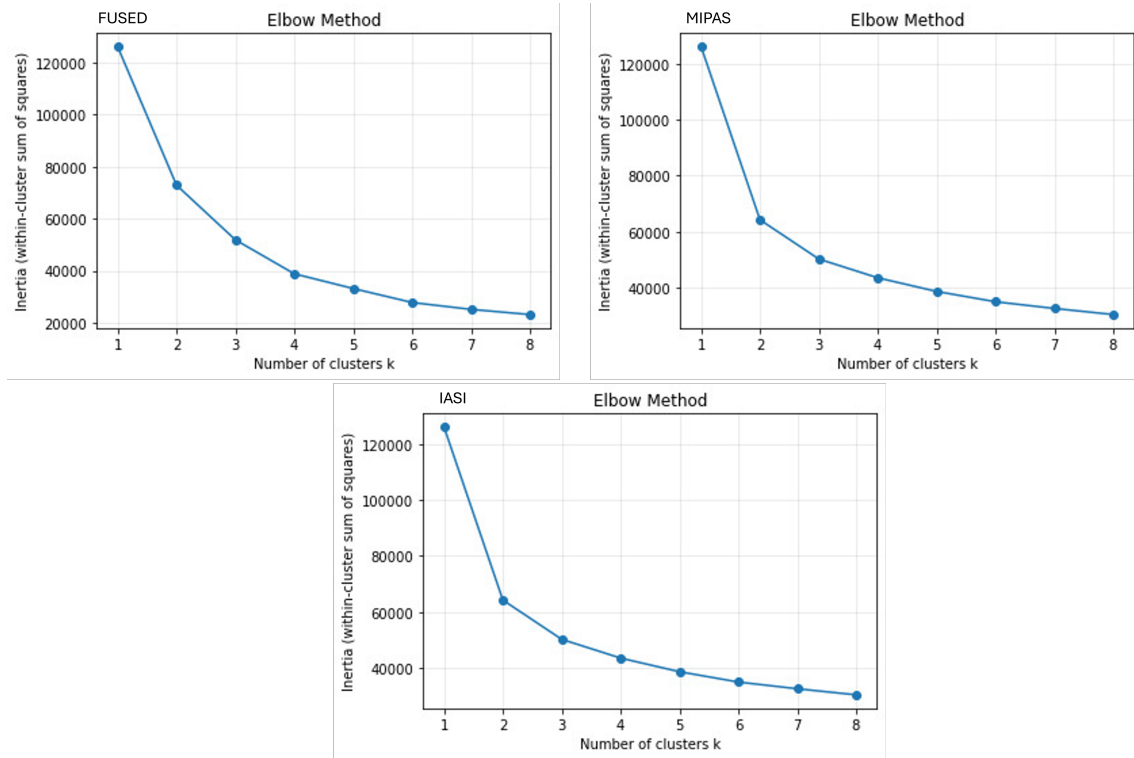


Figure 3.24: Elbow method applied for all fused, MIPAS, and IASI case study profiles.

The high number of features in this case (about 3000 profiles per 41 levels each one) results in high inertia value. However, the focus is how much the inertia decreases with increasing k , and if the plot flattens out, reaching the elbow. In case of MIPAS and IASI this happens starting from $k=2$, while for the fused product which shows more information, and the inertia starts flattening out from $k=3-4$. However, we initially used $k=2$ for all products in order to avoid minor variations or noise captured by the fused product. Moreover, increasing the number of clusters resulted in halving the silhouette score.

To analyse again the choice of the number of clusters and explore the main pattern of variability in the fused ozone dataset to reduce data dimensionality, we applied the Principal Component Analysis (PCA). The results of the PCA are shown in Figure 3.25. The scatterplots on the left column of the figure display the distribution of the profiles projected onto the first two PC, which explain a large fraction of the total variance (around 75–80%), indicating that the main modes of variability are effectively captured. The fused product (top panels) exhibits two clearly separable clusters in the PC1–PC2 space, similarly to IASI and MIPAS, confirming that the fusion preserves the dominant variability patterns of the original datasets.

The PCA loadings (right panels) describe the vertical structure of the variability captured by each component. In the fused dataset, PC1 (58.2% of the variance) shows loadings with opposite signs - large and positive in the middle and upper stratosphere (20–45 km) and

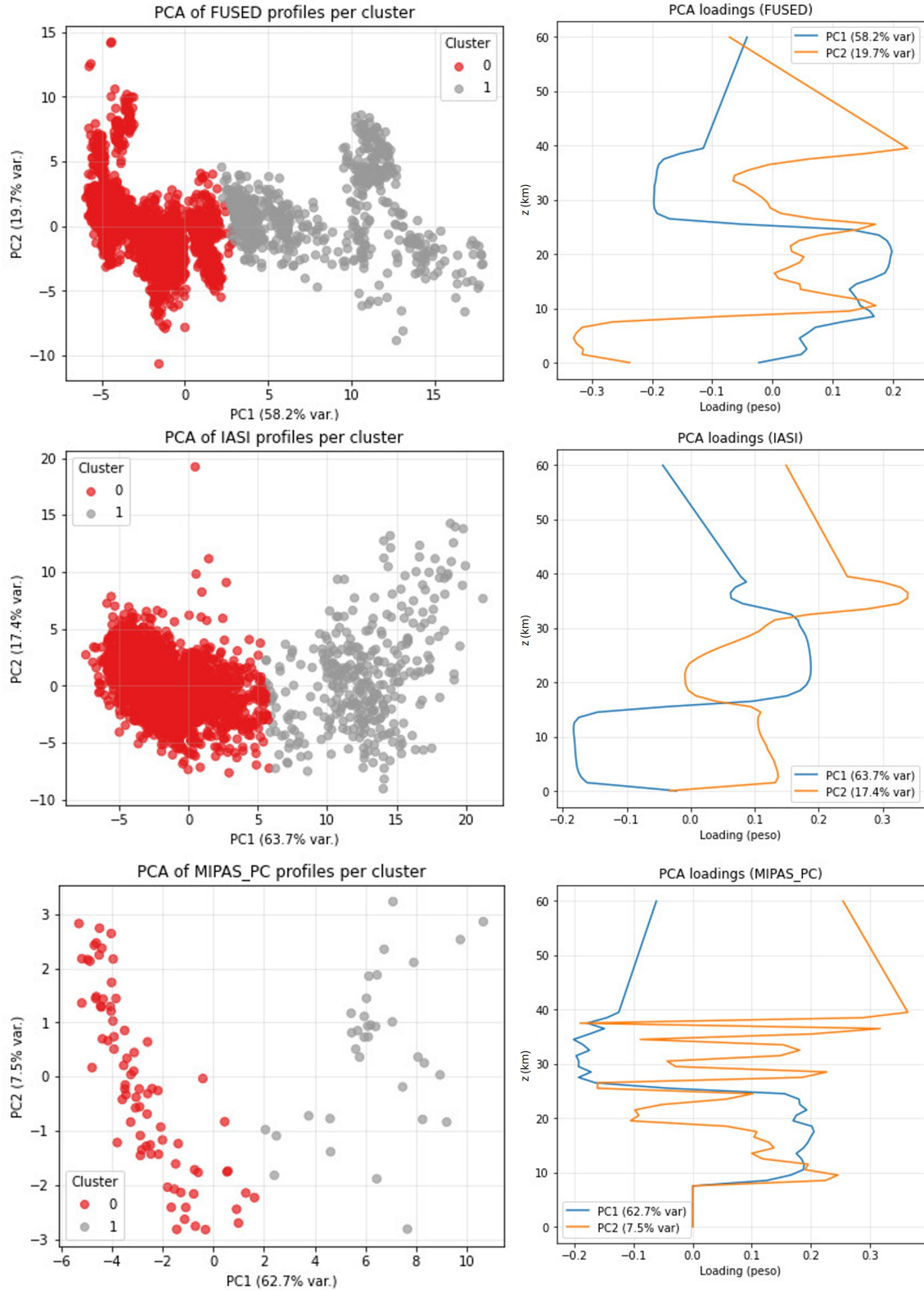


Figure 3.25: PCA of ozone profiles from the fused, IASI, and MIPAS datasets of the case study. Left panels show the projection of the profiles onto the first two principal components (PC1 and PC2), coloured according the two clusters. Right panels display the corresponding PCA loadings or weights as a function of altitude.

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

negative in the troposphere - representing the dominant contrast between stratospheric and tropospheric ozone. PC2 (19.7%) highlights variability concentrated in the UTLS region (10–20 km), likely related to the tropopause height and ozone intrusions.

The IASI dataset exhibits loading profiles with a similar vertical structure but opposite sign compared to the fused and MIPAS data, for both components. It indicates that both IASI and the fused product capture the same dominant vertical mode. The fused loadings appear oscillatory compared to IASI, suggesting that the fusion procedure effectively reduces noise while preserving the key vertical features. In contrast, the MIPAS loadings display more complex and oscillatory patterns above 25 km, reflecting higher vertical resolution but also less stable variance partitioning between components.

Overall, PCA analysis shows that the two clusters are well separated in the PC space, and that the first two components explain the majority of the variance of the profile space.

Figure 3.26 shows the mean of the profiles within the two clusters and provide the classification performances.

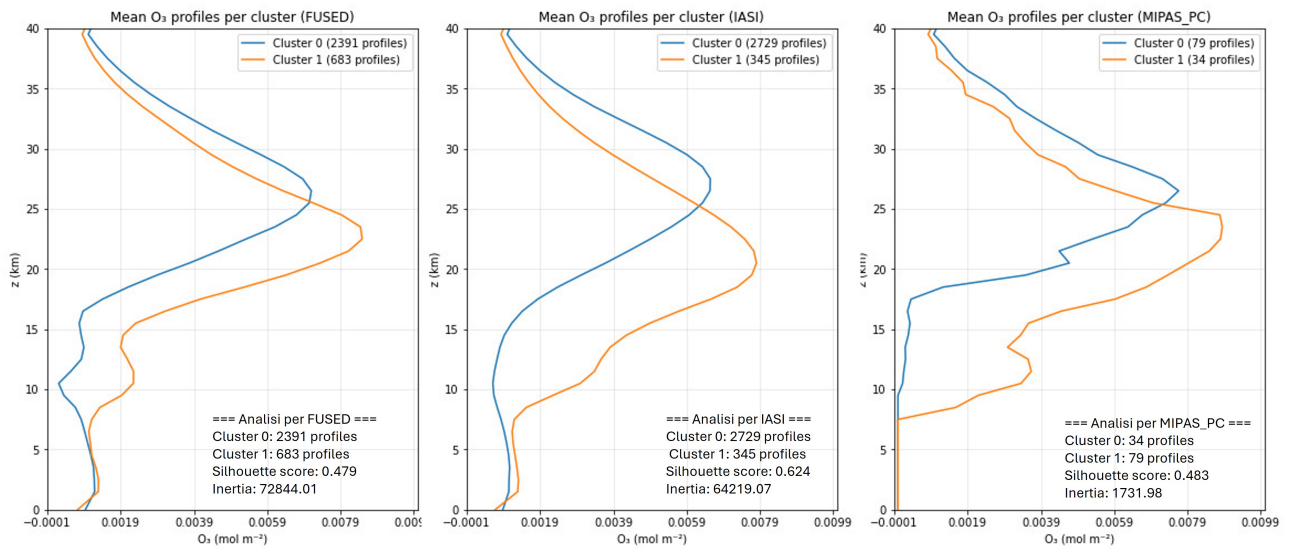


Figure 3.26: Mean of the profiles per clusters and classification performances.

All products, fused and L2, clearly allow to distinguish two clusters, as expected by the PCA analysis. These two clusters are very similar to the mean and median profiles observed in Figure 3.23, so that we could conclude that Cluster 1 represents event profiles in Figure 3.23 and Cluster 0 non-events profiles. The reader might wonder about the differing numbers of profiles. It should be noted, however, that whereas previously two sub-regions over six days were analyzed, here we are considering the entire study region over the same six-day period. The lower number of MIPAS profiles is due to the fact that the fusion was centered on IASI: multiple IASI profiles share the same MIPAS profile, which is therefore not counted more than once in the classification.

The performance of the classification is represented by the silhouette score. In general, the classification is good for all cases. IASI provides the best performance (0.624), followed by MIPAS (0.483), and the fused product (0.479). The reason why the fused product has a lower silhouette score is still under analysis.

The classification presented up to this point provides limited insights into the underlying physics, as it is not based on physical parameters. Thus, in this first analysis, we decided to re-conduct the clustering and the PCA analysis based on few physical parameters that were shown to be relevant in the previous section. As a first step, we reproduce the PCA using the variables pair ozone maximum and its altitude, leaving for feature developments the inclusion of the ratio between stratospheric and tropospheric column.

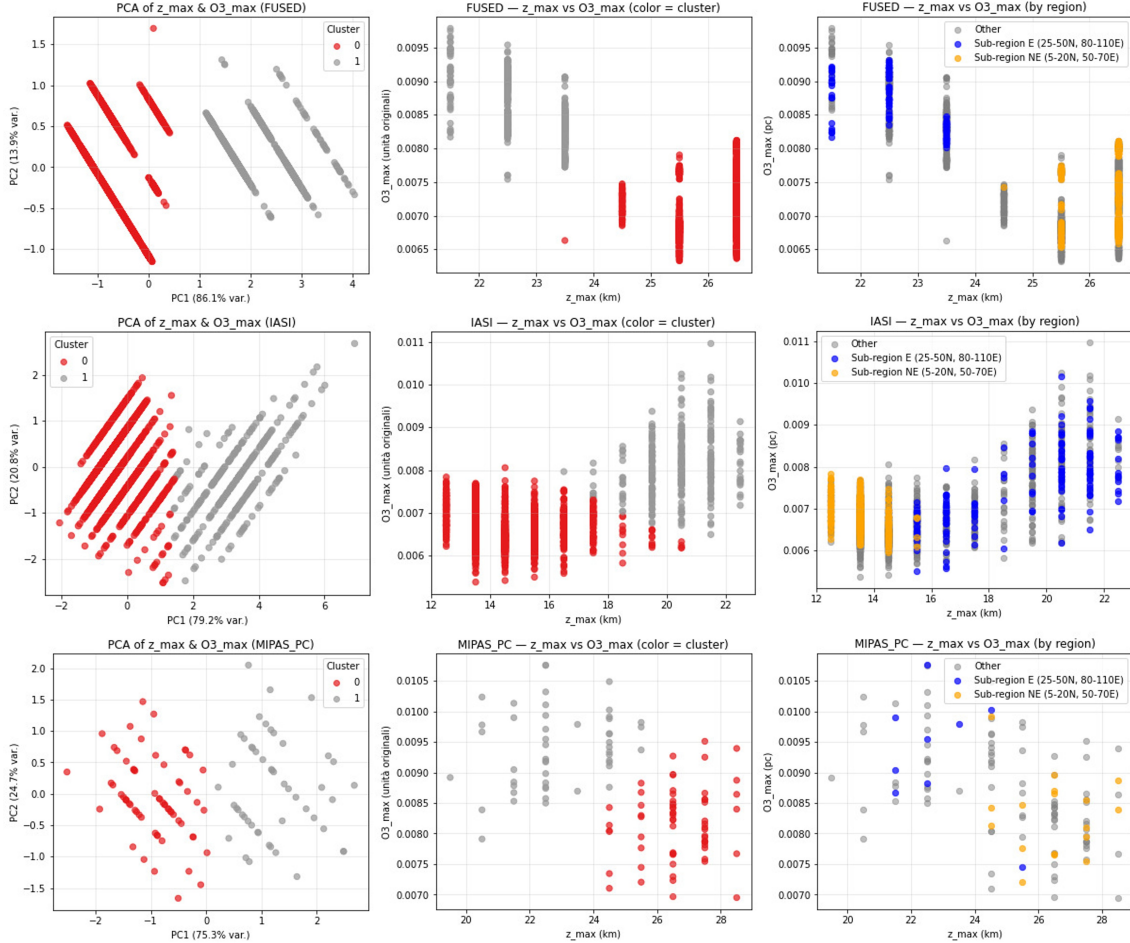


Figure 3.27: PCA scatterplots for fused (first row), IASI (second row), MIPAS (third row) products. First column shows scatterplots in the PC-plane. Second column shows scatterplots in the z_{max} , O_{3max} plane. Third column shows which of the profiles represented in the scatterplots lies in the two sub-regions selected in the Section 3.4.3 as sub-regions with event and non-event profile.

Figure 3.27 shows the results of the PCA through the scatterplots in the PC-space (first column) and in the variable space (second column). These columns indicate that, for all datasets, the first principal component (PC1) effectively captures the dominant mode of variability, while the second component (PC2) represents a smaller but non-negligible portion of the variance. Specifically, the FUSED dataset exhibits a very strong uni-dimensional structure, with PC1 explaining about 86% of the total variance and PC2 contributing only 13%. This suggests that the variability in O_{3max} and z_{max} in the fused product is largely governed by a single linear combination of the two variables, likely reflecting their strong correlation and the increased internal consistency introduced

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

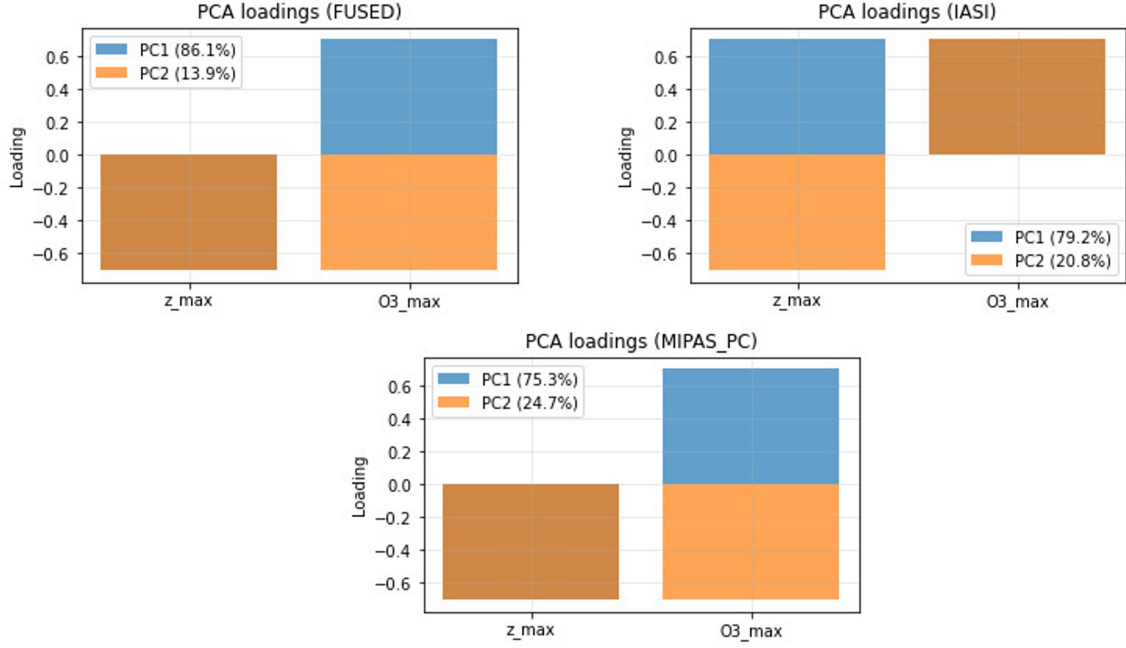


Figure 3.28: PCA loadings for fused, MIPAS and IASI products.

by the fusion process. In contrast, the IASI and MIPAS datasets display a more balanced variance distribution between PC1 and PC2 (79–21% and 75–25%, respectively), implying a less rigid coupling between O_3max and z_{max} , and a higher degree of independent variability.

Across all three cases, the two clusters (indicated in red and grey) are clearly separated along PC1, revealing that this axis captures the contrast between two distinct regimes of ozone distribution. The FUSED dataset shows well-separated clusters even if not well compacted (see the silhouette score lower than IASI and MIPAS), confirming that the fusion enhances coherence and reduces internal noise relative to the individual IASI and MIPAS datasets. The banded patterns visible in the scatter plots of IASI and MIPAS, particularly the vertical alignment of points around central z_{max} values, likely stem from the regular 1 km vertical discretization of the datasets, rather than from physical variability, suggesting that some sampling-related structure is retained in the original observations. Overall, the PCA results highlight that the fused dataset provides a cleaner and more coherent representation of the relationship between maximum ozone concentration and its altitude, while retaining the key physical distinction between the two observed regimes.

The PCA loadings (Figure 3.28) clarify the contribution of each variable to the principal components derived from the three datasets (FUSED, IASI, and MIPAS) and better explain the correlation between maximum ozone concentration and its altitude. For the FUSED and MIPAS datasets, PC1 displays opposite signs for the two variables (positive for O_3max and negative for z_{max}), indicating a clear anti-correlation between ozone concentration and its altitude. This means that higher ozone maxima tend to occur at lower altitudes, consistent with typical stratosphere–troposphere exchange dynamics and with what observed in previous analysis. Conversely, the IASI dataset shows concordant (positive) loadings for both variables on PC1, suggesting that in the IASI data the dominant variability reflects joint increases or decreases in O_3max and z_{max} rather than an inverse relationship. This difference may arise from the distinct vertical sensitivity and retrieval

characteristics of IASI, which can smooth vertical gradients and yield more coherent but less vertically contrasted ozone profiles.

To briefly reconnect to the reference truth — namely, the distinction between profiles classified as event or non-event in the previous section — we attempted to distinguish in the scatterplots within the variable space (second column of Figure 3.27) the points corresponding to the two subregions of event and non-event (third column of Figure 3.27). This link to the reference truth shows that the fused dataset is able to clearly separate the two clusters, corresponding to events / non-events, performing better than both MIPAS and IASI. Although still limited to this case study and not yet fully developed, this type of analysis appears to be a promising approach.

3.5 Chapter Conclusions

In this chapter, the study moved closer to addressing the scientific problem of stratospheric ozone intrusions into the troposphere over the Himalayan region, with the aim of investigating the usefulness of the fused dataset for studying such phenomena and of exploiting the information contained within ozone profiles. To initially reduce the complexity of the analysis, a case study was selected by combining in situ measurements from the NCO-P station with the outputs of STEFLUX, a Lagrangian transport model. The analysis of the dataset derived from these sources allowed the identification of a specific six-day event, occurring from 28 January to 2 February 2008.

As a first step, after defining an extended region around the in situ Himalayan station, we examined the characteristics of the event through the visualization of ERA5 and CAMS reanalyses and the EMAC chemistry–climate model, comparing their respective representations. This comparison was particularly relevant given the substantial differences between these datasets, both in terms of their formulation and in their ability to represent dynamical and chemical processes, as well as their different horizontal and vertical resolutions. Both the reanalyses and the model were able to capture the intrusion, which appeared in the typical form of a tropopause folding. The occurrence of the intrusion was confirmed not only through ozone concentrations but also via its correlation with other proxy variables such as potential vorticity, relative humidity, and tropospheric ozone (in the case of CAMS and EMAC). Their results were compared and analyzed with respect to the selected case study.

To focus the analysis on the ozone profiles, we then investigated the differences within the case study region—between event and non-event profiles. Based on the analysis of ERA5, CAMS, and EMAC - which are at this step used as a pseudo-truth - we selected two subregions within the study area where all three datasets either consistently detected or did not detect the intrusion signature. Within these two subregions, we computed and compared mean and median profiles from both the reanalyses and the model (known to represent the intrusion) and from the observational datasets (fused product, MIPAS, and IASI). This fundamental step allowed us to characterize profiles classified as events versus non-events, providing a reference “truth” basis for subsequent analyses.

In both the reanalyses and the model, we observed that converting to partial-column representations preserves the fundamental features of the vertical profiles. Moreover, in the partial-column representation—especially in the observational datasets—distinct physical

Chapter 3. Investigations of Stratospheric Ozone Intrusions over the Himalayas Using the New Dataset

characteristics such as the height and magnitude of the ozone maximum, as well as the presence of secondary peaks, clearly differentiated event profiles from non-event ones.

Since the fused product is by design a standalone dataset—not necessarily tied to the physical problem of intrusions—and given the lack of a robust truth dataset with sufficient statistics, we developed a fully unsupervised classification analysis based on k-means clustering and Principal Component Analysis (PCA). The goal was to test whether, without any reference to the ground truth, it was still possible to identify two distinct groups corresponding to event and non-event profiles, using only the profile information itself.

This first classification attempt yielded promising results. The analysis revealed two well-separated groups that, when compared with the previous profile-based analysis of the two subregions, could indeed be associated with event and non-event cases. Both the fused product and the Level 2 datasets demonstrated good ability to distinguish between the two clusters. Although the compactness of the clusters (as indicated by the silhouette score) suggested that IASI performed slightly better overall, the fused product appeared to provide a clearer separation between the two groups and to capture finer physical details, such as the ozone maximum dynamics and the altitude of the peak. Moreover, when compared with the reference truth of profiles within events / non-events sub-regions, the fused dataset appeared more effective in distinguishing event from non-event cases, reinforcing its added value over the individual IASI and MIPAS datasets.

Although this analysis is still at an early stage, it demonstrates the potential of the fused product as a consistent observational basis for investigating ozone intrusions in complex orographic regions such as the Himalayas. It establishes a solid foundation for future developments, which will first aim to refine the case study by incorporating more robust truth or pseudo-truth references, and subsequently extend the same framework to additional case studies, eventually generalizing the analysis to the full dataset and to other regions.

Conclusions and Future Developments

This thesis addressed the study of stratospheric ozone intrusion as a part of stratosphere–troposphere exchange (STE) processes, with a particular focus on the Himalayan region — one of the global hotspots for stratospheric intrusions due to its complex orography and dynamic meteorological conditions. Through the synergistic use of satellite observations, data fusion techniques, and statistical classification methods, the work aimed to improve both the observational basis and the analytical framework for investigating stratospheric ozone intrusions and their influence on tropospheric ozone variability.

The first part of the thesis focused on the development and validation of a new ozone dataset for the period 2008–2011, obtained through the Complete Data Fusion (CDF) of limb observations from MIPAS and nadir observations from IASI. This fusion approach successfully combined the high vertical resolution of MIPAS with the broad horizontal resolution, and lower-tropospheric sensitivity of IASI, resulting in enhanced vertical information and reduced retrieval uncertainties, particularly in the Upper Troposphere–Lower Stratosphere (UTLS) region. Validation against ozonesonde data from the WOUDC network confirmed the robustness of the fused product, demonstrating improved degrees of freedom for signal and reduced errors compared with the individual input datasets. Overall, this was a dominant focus of this work. This emphasis is motivated by the intrinsic complexity of handling and harmonizing heterogeneous satellite datasets, each characterized by distinct vertical sensitivities, spatial coverages, and uncertainty structures. Developing a consistent and validated fused product required extensive effort in data preprocessing, bias correction, and intercomparison across multiple observing systems and validation sources. Addressing these methodological challenges was a necessary step to ensure that the resulting dataset could serve as a reliable foundation for the subsequent analysis of stratosphere-to-troposphere exchange events. Beyond the technical validation and next steps, this work contributed a novel, observation-based dataset that bridges the gap between limb and nadir perspectives, offering a valuable resource for both

regional and global studies of ozone dynamics. The dataset developed in this thesis will be made publicly available in a standardized format as part of next work steps, ensuring transparency and facilitating its use by the atmospheric composition community

Building on this foundation, the second part of the thesis explored the application of the fused dataset to the detection and characterization of stratospheric intrusions. A six-day case study (28 January – 2 February, 2008) was selected as a proof of concept to test the capability of the dataset to capture the vertical and spatial structure of an intrusion event. By combining in situ, reanalysis, and model data (ERA5, CAMS, EMAC) with the fused and original satellite retrievals, the analysis provided an integrated view of the event’s dynamics. The work demonstrated that the fused product could reproduce the main physical features associated with the intrusion, such as the tropopause folding structure and the ozone peak descent, while maintaining consistency with independent datasets.

As a first attempt to move beyond qualitative assessments, a profile-based, unsupervised classification approach was applied using Principal Component Analysis and k-means clustering, relying solely on vertical ozone profiles. This exploratory analysis showed that, even without a direct link to a reference truth dataset, it was possible to separate ozone profiles into two distinct classes corresponding to intrusion and non-intrusion conditions. The fused dataset performed comparably to, and in some aspects better than, the individual MIPAS and IASI datasets, particularly in representing the vertical structure of the ozone maximum and its variability. The a posteriori comparison with the classification based on reanalysis and models, confirm that the fused dataset can represent an added value with respect to IASI and MIPAS alone. These findings highlight the added value of the fusion approach not only for improving observational quality but also for enabling new analytical frameworks applicable to large-scale satellite datasets.

Overall, this thesis demonstrated the feasibility and scientific value of combining complementary satellite observations to enhance the representation of atmospheric processes that are otherwise difficult to observe, such as stratosphere-to-troposphere exchange. The new fused dataset and the classification framework developed here constitute two complementary outcomes: the first provides a physically consistent and validated observational product, while the second establishes the methodological groundwork for automated STE detection in satellite data.

The future directions of this work span multiple fronts. Although the first part is currently under publication, it offers several opportunities for further investigation. The CDF framework presents notable flexibility, particularly in modifying the constraint strength and vertical grid of a given product. Building on the current study, future work will include a more detailed analysis of these aspects, as well as a deeper exploration of the additional errors identified during the tuning phase. This flexibility opens up a wide range of possibilities for further research. Moreover, the datasets used in this study, benefiting from two distinct observation geometries, are not the only ones available in the remote sensing landscape. Applying the CDF to new real observations would be of great interest, and the fusion scheme adopted here is not the only viable approach. More optimized fusion strategies could be developed to better serve scientific applications.

In the second part of this work, still in its early stages, the analysis could be extended to multiple case studies and longer temporal scales, enabling the construction of regional and global climatologies of ozone intrusion events. Further refinement of the classification approach - potentially integrating machine learning algorithms and additional atmospheric variables derived from ozone profiles —could enhance both the accuracy and the gener-

alizability of the framework. Indeed, the unsupervised algorithm used here (k-means) is not the only option, and expanding the dataset could allow for the creation of a robust ground truth dataset, which could then be used to train supervised algorithms. Finally, integrating the fused product into data assimilation systems and atmospheric models would represent a significant step forward in improving the representation of ozone variability in both the troposphere and stratosphere, thereby advancing our understanding of its role in climate and air quality.

In conclusion, this work contributes to the broader effort of bridging observational and modeling approaches in atmospheric science. By leveraging data fusion, it enhances our ability to observe, interpret, and ultimately predict the dynamics of ozone and stratosphere–troposphere exchange key processes in the Earth’s climate system. Finally, it is worth noting that the long-term potential of CDF-based limb–nadir fusion depends critically on the availability of limb-sounding satellite measurements. Since no new infrared limb-sounding missions are currently planned, future datasets comparable to MIPAS will not be available. This limits the possibility of applying similar data fusion approaches and reduces access to high vertical resolution ozone observations in dynamically important regions such as the UTLS, essential for STE studies. This underlines the importance of preserving, exploiting, and extending existing limb datasets, while also advocating for future missions capable of providing high-vertical-resolution ozone profiles to support both scientific and climate monitoring needs.



Basic Statistics

A.1 Mean, Variance, Covariance and the Gaussian Distribution

For a set $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ of M vectors of dimension N the *mean value* $\langle \mathbf{x} \rangle$ is:

$$\langle \mathbf{x} \rangle = \frac{\sum_{j=1}^M \mathbf{x}_j}{M}, \quad (\text{A.1})$$

where the mean for the i – *th* component of each vector is:

$$\langle \mathbf{x}_i \rangle = \frac{\sum_{j=1}^M \mathbf{x}_{j,i}}{M}. \quad (\text{A.2})$$

We can also define the spread of these values around the mean value, represented by the *variance* σ^2 , which for a set of N variables is:

$$\langle x^2 \rangle = \frac{\sum_{i=1}^N x_i^2}{N} - \left(\frac{\sum_{i=1}^N x_i}{N} \right)^2 = \langle x^2 \rangle - \langle x \rangle^2. \quad (\text{A.3})$$

Considering vectors instead of numbers, we have to define the *covariance*, which between two variables of the vector $\mathbf{x}$ is:

$$\text{Cov}(x_a, x_b) = \frac{1}{M} \sum_{j=1}^M (x_{a,j} - \langle x_a \rangle)(x_{b,j} - \langle x_b \rangle) = \langle x_a x_b \rangle - \langle x_a \rangle \langle x_b \rangle. \quad (\text{A.4})$$

At this point we can define the *variance-covariance matrix* $\mathbf{S}$, as the symmetric matrix having on its diagonal elements the variance of the vector elements, and on its off-diagonal elements are the covariance between vector elements, which represent the correlation between them.

A Gaussian distribution is a probability density function used to describe and approximate

random real variable which tend to close up to a single value called mean value, with a spread represented by the variance. It is defined as:

$$P(x) = N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (\text{A.5})$$

such as $P(x)dx$ is the probability that x lies in the interval $(x - dx, x + dx)$. It is often called *Normal* distribution, and has several advantages in its properties, reason why treating problems with Gaussian distribution is convenient. The mean value and the variance of a continuous Normal pdf will be:

$$\mu = \int xP(x)dx, \quad \sigma^2 = \int (x - \mu)^2 P(x)dx. \quad (\text{A.6})$$

A useful consequence is that the product of two Gaussian pdfs is again a Gaussian pdf with mean value that is the weighted mean of the mean values of each pdf on each variance, and variance being the sum of the inverse of variances.

In N dimensions, the Normal distribution becomes:

$$P(\mathbf{x}) = N(<\mathbf{x}>, \mathbf{S}) = \frac{1}{(2\pi)^{\frac{1}{N}} |\mathbf{S}|^{\frac{1}{2}}} e^{-\frac{1}{2}[\mathbf{x} - <\mathbf{x}>]^T \mathbf{S}^{-1} (\mathbf{x} - <\mathbf{x}>)}, \quad (\text{A.7})$$

where:

$$<\mathbf{x}> = \int \mathbf{x}P(\mathbf{x})d\mathbf{x}. \quad (\text{A.8})$$

A.2 Errors and Propagation

The process of measurement is affected by errors due to several reasons, and each measurement can be seen as the result of several trials within a infinite set of possible measurement. The error can be considered the sum of two components: a systematic error component and a random error component. While the former can be estimate and corrected, the second not. However its variation can be described by a Gaussian pdf (central limit theorem: *the sum of n random variables tends to a Gaussian as n tend to ∞ .*) having null expected value of the random error. A set of random errors $\epsilon_1, \epsilon_2, \dots, \epsilon_N$ has a mean value of zero by definition, hence the standard deviation, which is the *root mean square (RMS)*, is calculated as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N \epsilon_i^2} = \sqrt{<\epsilon^2>}. \quad (\text{A.9})$$

Being $\mathbf{y}$ an m dimensional measurement vector, each element will be affected by an error, hence we will have an m dimensional error vector $\boldsymbol{\epsilon}$. The uncertainty on the measurement will be represented by the variance-covariance matrix $\mathbf{S}_y$. Recalling the formula of the covariance A.4 and that $<\epsilon_i> = 0$, we have that the variance-covariance matrix is:

$$\begin{bmatrix} Cov_{11} & Cov_{12} & \dots & Cov_{1m} \\ Cov_{21} & Cov_{22} & \dots & Cov_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ Cov_{m1} & Cov_{m2} & \dots & Cov_{mm} \end{bmatrix} = \begin{bmatrix} <\epsilon_1\epsilon_1> & <\epsilon_1\epsilon_2> & \dots & <\epsilon_1\epsilon_m> \\ <\epsilon_2\epsilon_1> & <\epsilon_2\epsilon_2> & \dots & <\epsilon_2\epsilon_m> \\ \vdots & \vdots & \vdots & \vdots & \\ <\epsilon_m\epsilon_1> & <\epsilon_m\epsilon_2> & \dots & <\epsilon_m\epsilon_m> \end{bmatrix} = <\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}>.$$

If a measure is indirect, then errors propagate within calculations. If we suppose to have a measurement described by a function of other parameters, such as $y = f(x)$. To estimate the uncertainty in the measurement we assume:

$$\langle y \rangle = f(\langle x \rangle). \quad (\text{A.10})$$

The spread of y_i values is found from a Taylor expansion about the mean:

$$y_i - \langle y \rangle = (x_i - \langle x \rangle) \frac{\partial \langle y \rangle}{\partial \langle x \rangle}, \quad (\text{A.11})$$

and the variance will be:

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \langle y \rangle)^2 = \frac{1}{m} \sum_{i=1}^m \left[(x_i - \langle x \rangle) \frac{\partial \langle y \rangle}{\partial \langle x \rangle} \right]^2. \quad (\text{A.12})$$

If y is function of two variables, x_1 and x_2 , the equation A.12 becomes:

$$\begin{aligned} \sigma^2 &= \frac{1}{m} \sum_{i=1}^m \left[(x_{1i} - \langle x_1 \rangle) \frac{\partial \langle y \rangle}{\partial \langle x_1 \rangle} + (x_{2i} - \langle x_2 \rangle) \frac{\partial \langle y \rangle}{\partial \langle x_2 \rangle} \right]^2 = \\ &= \sigma_{x_1}^2 \left(\frac{\partial \langle y \rangle}{\partial \langle x_1 \rangle} \right)^2 + \sigma_{x_2}^2 \left(\frac{\partial \langle y \rangle}{\partial \langle x_2 \rangle} \right)^2 + 2Cov(x_1, x_2) \frac{\partial \langle y \rangle}{\partial \langle x_1 \rangle} \frac{\partial \langle y \rangle}{\partial \langle x_2 \rangle}. \end{aligned} \quad (\text{A.13})$$

This formulation has to be generalized in the case in which measurements, parameters and function are the vectors $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_m]$, $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, and $\mathbf{f} = [\mathbf{f}_1, \dots, \mathbf{f}_m]$. We define the variance covariance matrix $\mathbf{S}_x$ and we will have that the variance become the matrix $\mathbf{S}_y$, given by:

$$\mathbf{S}_y = \mathbf{K} \mathbf{S}_x^{-1} \mathbf{K}^T. \quad (\text{A.14})$$

where $\mathbf{K}$ is the $n \times m$ *Jacobian* matrix of the set of $\mathbf{f}$ with respect to $\mathbf{x}$:

$$\begin{aligned} \mathbf{K} &= \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}. \end{aligned} \quad (\text{A.15})$$

B

Complete Data Fusion Background

B.1 Equivalence between CDF and Synergistic Retrieval

Proof. In the linear case, the forward model $\mathcal{F}$ is such that:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}. \quad (\text{B.1})$$

If we suppose a Gaussian distribution of errors, we have that the χ^2 is:

$$\chi^2(\mathbf{x}) = (\mathbf{y} - \mathcal{F}(\mathbf{x}))^T \mathbf{S}_{\text{ny}}^{-1} (\mathbf{y} - \mathcal{F}(\mathbf{x})). \quad (\text{B.2})$$

Substituting B.1 in B.2, and exploiting the linearity, we have that:

$$\chi^2(\mathbf{x}) = (\mathbf{K}\mathbf{x}_{\text{true}} + \boldsymbol{\epsilon} - \mathbf{K}\mathbf{x})^T \mathbf{S}_{\text{ny}}^{-1} (\mathbf{K}\mathbf{x}_{\text{true}} + \boldsymbol{\epsilon} - \mathbf{K}\mathbf{x}). \quad (\text{B.3})$$

Developing B.3:

$$\begin{aligned} \chi^2(\mathbf{x}) &= (\mathbf{K}(\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\epsilon})^T \mathbf{S}_{\text{ny}}^{-1} (\mathbf{K}(\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\epsilon}) = \\ &= (\mathbf{x}_{\text{true}} - \mathbf{x})^T \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} (\mathbf{x}_{\text{true}} - \mathbf{x}) + 2\boldsymbol{\epsilon}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} (\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\epsilon}^T \mathbf{S}_{\text{ny}}^{-1} \boldsymbol{\epsilon}. \end{aligned} \quad (\text{B.4})$$

The equation B.4 can be expressed in terms of the retrieved profile error ($\boldsymbol{\sigma}_{\text{n}}$), its related covariance matrix ($\mathbf{S}_{\text{n}}$), and the averaging kernel ($\mathbf{A}$), which we recall for convenience:

$$\boldsymbol{\sigma}_{\text{n}} = \mathbf{G}\boldsymbol{\epsilon} = (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_{\text{a}})^T \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \boldsymbol{\epsilon}; \quad (\text{B.5})$$

$$\mathbf{S}_{\text{n}} = (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_{\text{a}})^{-1} \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_{\text{a}})^{-1}; \quad (\text{B.6})$$

$$\mathbf{A} = (\mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} + \mathbf{S}_{\text{a}})^{-1} \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K}. \quad (\text{B.7})$$

From eq.s B.5, B.6, and B.7, we obtain the following two equations:

$$\begin{aligned} \boldsymbol{\epsilon}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} &= \boldsymbol{\sigma}_{\text{n}}^T \mathbf{S}_{\text{n}}^{-1} \mathbf{A}; \\ \mathbf{K}^T \mathbf{S}_{\text{ny}}^{-1} \mathbf{K} &= \mathbf{A}^T \mathbf{S}_{\text{n}}^{-1} \mathbf{A}. \end{aligned} \quad (\text{B.8})$$

Substituting B.8 in B.4, we find that:

$$\chi^2(\mathbf{x}) = (\mathbf{x}_{\text{true}} - \mathbf{x})^T \mathbf{A}^T \mathbf{S}_{\mathbf{n}}^{-1} \mathbf{A} (\mathbf{x}_{\text{true}} - \mathbf{x}) + 2\boldsymbol{\sigma}_{\mathbf{n}}^T \mathbf{S}_{\mathbf{n}}^{-1} \mathbf{A} (\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\epsilon}^T \mathbf{S}_{\mathbf{ny}}^{-1} \boldsymbol{\epsilon}. \quad (\text{B.9})$$

According to the CDF method:

$$\boldsymbol{\alpha} = \mathbf{A} \mathbf{x}_{\text{true}} + \boldsymbol{\sigma}_{\mathbf{n}}, \quad (\text{B.10})$$

that can be written as:

$$\boldsymbol{\alpha} - \mathbf{A} \mathbf{x} = \mathbf{A} (\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\sigma}_{\mathbf{n}}. \quad (\text{B.11})$$

Therefore, we have that:

$$\mathbf{A} (\mathbf{x}_{\text{true}} - \mathbf{x}) = \boldsymbol{\alpha} - \mathbf{A} \mathbf{x} - \boldsymbol{\sigma}_{\mathbf{n}}. \quad (\text{B.12})$$

Substituting in B.9:

$$\chi^2(\mathbf{x}) = (\boldsymbol{\alpha} - \mathbf{A} \mathbf{x})^T \mathbf{S}_{\mathbf{n}}^{-1} (\boldsymbol{\alpha} - \mathbf{A} \mathbf{x}) + \boldsymbol{\sigma}_{\mathbf{n}}^T \mathbf{S}_{\mathbf{n}}^{-1} \boldsymbol{\sigma}_{\mathbf{n}} + \boldsymbol{\epsilon}^T \mathbf{S}_{\mathbf{ny}}^{-1} \boldsymbol{\epsilon}. \quad (\text{B.13})$$

If we suppose to have N -independent measurements $\mathbf{y}_i$, for each of these we will have a chi-function χ_i . The cost function becomes:

$$\begin{aligned} c(\mathbf{x}) = & \sum_{i=1}^N (\boldsymbol{\alpha}_i - \mathbf{A}_i \mathbf{x})^T \mathbf{S}_{\mathbf{n}i}^{-1} (\boldsymbol{\alpha}_i - \mathbf{A}_i \mathbf{x}) + (\mathbf{x} - \mathbf{x}_{\mathbf{a}})^T \mathbf{S}_{\mathbf{a}}^{-1} (\mathbf{x} - \mathbf{x}_{\mathbf{a}}) + \\ & + \sum_{i=1}^N \boldsymbol{\sigma}_{\mathbf{n}i}^T \mathbf{S}_{\mathbf{n}i}^{-1} \boldsymbol{\sigma}_{\mathbf{n}i} + \sum_{i=1}^N \boldsymbol{\epsilon}_i^T \mathbf{S}_{\mathbf{ny}_i}^{-1} \boldsymbol{\epsilon}_i. \end{aligned} \quad (\text{B.14})$$

If we minimize the cost function, we have that the last two terms do not depend on x and the expression in (B.14) becomes equal to the 2.12 one, with the same solution for $\mathbf{x}_{\mathbf{f}}$.

The same result can be rigorously reached considering the case of the linearised forward model. Indeed, If we consider the linearization around a reference state $\mathbf{x}_0$, we have:

$$\mathcal{F}(\mathbf{x}) = \mathcal{F}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x} - \mathbf{x}_0); \quad (\text{B.15})$$

evaluating it in $\mathbf{x}_{\text{true}}$ we have:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x}_{\text{true}} - \mathbf{x}_0) + \boldsymbol{\epsilon}. \quad (\text{B.16})$$

By subtracting B.15 to B.16, we obtain:

$$\mathbf{y} - \mathcal{F}(\mathbf{x}) = \mathbf{K}(\mathbf{x}_{\text{true}} - \mathbf{x}) + \boldsymbol{\epsilon}. \quad (\text{B.17})$$

from which using equation B.2, we obtain:

$$\chi^2(\mathbf{x}) = (\mathbf{K} \mathbf{x}_{\text{true}} + \boldsymbol{\epsilon} - \mathbf{K} \mathbf{x})^T \mathbf{S}_{\mathbf{ny}}^{-1} (\mathbf{K} \mathbf{x}_{\text{true}} + \boldsymbol{\epsilon} - \mathbf{K} \mathbf{x}). \quad (\text{B.18})$$

the same equation in B.3. $\square$



STE Events Background

C.1 Conserved Properties

The two important conserved quantities in the atmosphere are the potential temperature and potential vorticity. These two quantities are invariant for a particular air parcel even as the parcel moves about. Because of its fixed value, it allows the individual air parcel to be traced, and hence the conserved property acts as a tracer.

C.1.1 Potential Temperature

Potential Temperature θ is the temperature an air parcel would have if it were adiabatically compressed or expanded from the original pressure p , to the standard pressure p_0 , usually 1000 hPa. Lines of constant θ are called *isentropes*, and the flow along surfaces of constant potential temperature is called *isentropic flow* Maahn et al., 2020. Starting from the State Equation for an ideal gas within a system in thermodynamic equilibrium, it is possible obtain the definition of θ , also called *Poisson's Equation*.

In dry air case, we have:

$$p\alpha = RT, \quad (\text{C.1})$$

where p is the pressure, α is the specific volume equal to the inverse of the density ρ , T is the absolute temperature, and R is the specific gas constant (287 J kg⁻¹ K).

Because of the first law of thermodynamics, we have that:

$$\frac{DQ}{Dt} = C_v \frac{DT}{Dt} + p \frac{D\alpha}{Dt}, \quad (\text{C.2})$$

where the left-hand side is the heating rate, and C_v is the specific heat at constant volume. Differentiating the equation of state (eq. C.1) with respect to time, knowing that $C_p = C_v + R$, and exploiting eq C.2, we can obtain:

$$C_p \frac{D \ln T}{Dt} - R \frac{D \ln p}{Dt} = 0. \quad (\text{C.3})$$

Integrating it from a given pressure p and temperature T , to a reference pressure p_0 and to the potential temperature θ . we obtain:

$$\theta = T \left[\frac{p_0}{p} \right]^{\frac{R}{C_p}}. \quad (\text{C.4})$$

Potential temperature is conserved when there are no diabatic effects. An air parcel will tend to move on an isentropic surface, or surface of constant potential temperature. θ is related to the *entropy* of the system, reason why constant potential temperature surfaces are also at constant entropy. Therefore, in adiabatic process, if an air parcel moves on these surfaces, θ is a conserved quantity following the parcel trajectory. This is usually a fairly good approximation for periods of 5–10 days. The vertical gradient of the potential temperature, such as of temperature, is strictly related to the atmospheric air stability. In general when it is positive it means stable air, negative it means unstable air, and null is a neutral case. The stratosphere is the layer of the atmosphere where the potential temperature increases with altitude; hence, it is a stably stratified region.

In presence of water vapor and in a diabatic case it is necessary to take into account different processes such as condensation Wallace and Hobbs, 2006. In those cases, to take into account of the water vapor presence, the (virtual temperature) $T_v = T(1 + 0.61w)$ and substituted to the temperature in eq C.5, where w is the mixing ratio, i.e the ratio of the mass of water vapor to the mass of dry air. In case of saturation, an *equivalent potential temperature* θ_e is defined:

$$\theta_e = \theta e^{\frac{L_v w_s}{C_p T}}, \quad (\text{C.5})$$

where L_v is the latent heat ($2.25 \times 10^6 \text{ J kg}^{-1}$).

C.1.2 Circulation, Vorticity and Potential Vorticity

Circulation and vorticity are the two primary measures of rotation in a fluid Mohanakumar, 2008, the atmosphere in our case.

Circulation is the flow or motion of fluid through a given area or volume, used to describe (in atmospheric science) air motion around pressure systems. It is a macroscopic measure of the rotation of the fluid and is a scalar quantity. Circulation about a closed contour in a fluid is defined as the line integral of the velocity vector that is locally tangent to the contour C :

$$C = \oint \vec{V} \cdot d\vec{l}, \quad (\text{C.6})$$

where $\vec{V}$ is the velocity vector and $d\vec{l}$ is the displacement vector along the contour.

Vorticity is a microscopic measure of the rotation of a fluid considered at a point, hence is a vector field, and it is defined as the curl of the velocity. Since in atmospheric science we are interested in rotational motion around axis perpendicular to the Earth surface,

we usually consider the vertical component of the vorticity. If we consider the relative horizontal velocity vector $\vec{V} = (u, v)$, the associated relative vorticity will be:

$$\zeta = \hat{k} \cdot (\vec{\nabla} \times \vec{V}) = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}. \quad (\text{C.7})$$

Considering that atmospheric air masses experience also vorticity due to the Earth rotation, called planetary vorticity f , we define *absolute vorticity*, the following quantity:

$$\eta = \zeta + f. \quad (\text{C.8})$$

Potential vorticity is a quantity proportional to the product of vorticity and stratification that, following an air parcel and can only be changed by diabatic or frictional processes. Baroclinic instability requires the presence of a potential vorticity gradient along which waves amplify during cyclogenesis. In a homogeneous incompressible fluid, it is defined as:

$$PV = \frac{\eta}{\delta z}, \quad (\text{C.9})$$

where δz is the depth of air parcel. It is possible to relate this to potential temperature and pressure, obtaining:

$$PV = \eta \left[-g \frac{\partial \theta}{\partial p} \right], \quad (\text{C.10})$$

also called *Ertel's PV*, which consists of both vorticity and a component due to vertical changes in potential temperature, a related atmospheric static stability. It follows that where the static stability is higher, also the PV will be higher. Hence, stratosphere, being a stable region, will be characterized by higher potential vorticity values than those in the troposphere. The Ertel's potential vorticity is a conservative quantity during adiabatic evolution. It is usually measured in PVU, where a PVU is $10^{-6} \text{ m}^2 \text{Ks}^{-1} \text{kg}^{-1}$. The PV is relatively low in the troposphere (smaller than 1 pvu) and the PV gradient is flat. There is, however, a tendency to have a slightly higher PV at high latitudes than near the equator because of the planetary component of the absolute vorticity. At the tropopause, the PV gradient becomes very strong because of the very high values of the static stability in the stratosphere. Above the surface of 1.5 or 2 PVU, the PV increases very quickly in the stratospheric air. The isopotential vorticity surfaces of 1.5 or 2 PVU define the *dynamical tropopause*, used as limit between troposphere and stratosphere.



MIPAS Validation

D.1 Validation Technique

To validate MIPAS, the year 2008 was selected, using WOUDC ozonesondes as reference measurements and applying the same methodology later employed for the validation of the fused (Section 2.3.2). Here, only the main aspects are summarized.

MIPAS reprocessed data, as described in Section 2.2.1, are provided as volume mixing ratios. In this work, the validation was carried out using both volume mixing ratios and partial column data, in order to better assess the role of the unit conversion algorithm and to evaluate the impact of the applied approximations.

Another key aspect of the validation concerns the treatment of the interpolation grid and, in particular, the handling of edge points during the interpolation process. Indeed, it has to be taken into account that MIPAS does not have fixed vertical grid.

Two interpolation approaches were employed in this work.

The first approach used the binning method as described by Cortesi et al., 2007. In this case, the ozonesonde profiles, together with their associated uncertainties, were first interpolated onto the MIPAS vertical grid and subsequently smoothed by applying the MIPAS averaging kernels. The differences between MIPAS and the smoothed ozonesonde profiles were then computed, and the resulting datasets were organized into vertical bins in pressure, in such a way that each bin contained at most one point from each profile. Within each bin, statistical quantities such as mean values, counts, or sums were derived and used for the graphical representation of the comparison. This methodology should reduce the impact of the different vertical resolutions and sampling strategies of the two datasets, and ensures that the comparison is not biased by irregularly distributed measurement points.

The second approach adopted for the validation was the interpolation onto a fixed grid, the IASI grid. In this case, the ozonesonde profiles, together with their associated un-

certainties, were first interpolated onto the MIPAS vertical grid. The ozonesonde data were then smoothed using the MIPAS averaging kernels, and the differences between MIPAS and the smoothed ozonesonde profiles were calculated. These differences were subsequently interpolated onto a common fixed grid, allowing for a homogeneous vertical representation across all coincidences. In parallel, both the smoothed ozonesonde profiles and their associated variances, originally defined on the MIPAS grid (which varies for each coincidence), were interpolated onto the same fixed grid. Finally, the statistical quantities of interest—such as mean values, counts, or sums—were computed on this fixed grid, providing a consistent framework for graphical representation and comparison.

An additional source of systematic uncertainty arises from the fact that the MIPAS vertical grid and the ozonesonde grid are not perfectly aligned. This is particularly critical when working in terms of partial columns, as the definition of the common vertical support affects the calculation of the differences. In order to address this issue, particular care was taken in the treatment of the extreme points (both upper and lower) of the profiles.

For the upper extreme, the procedure starts by identifying the minimum of the maximum altitudes sampled by the two grids, hereafter denoted as Z_M . If this minimum belongs to the coarser (MIPAS) grid, the interpolation is considered reliable up to the highest level of the MIPAS grid. Conversely, if Z_M belongs to the finer ozonesonde grid, the closest upper level of the MIPAS grid, $Z[i]$, is identified. In this case, if the distance between $Z[i]$ and Z_M is smaller than 5% of the vertical spacing of the MIPAS grid (i.e. $Z[i] - Z_M < 0.05 (Z[i] - Z[i - 1])$), the profile is masked from $Z[i]$ upwards. Otherwise, the masking is applied starting from the previous MIPAS level, $Z[i - 1]$, upwards.

The lower extreme is treated in a dual way, although in practice this case is less critical, since ozonesonde profiles always start from the ground and therefore extend to lower altitudes than satellite measurements. Nevertheless, the same criterion is adopted to ensure that the interpolation at the boundaries does not introduce spurious contributions to the calculated differences.

D.2 Validation Results

As shown in Figure D.1, the comparison between MIPAS and ozonesonde profiles highlights a consistent bias pattern that is largely independent of the interpolation method. In both cases, the bias is close to zero between about 15 and 25 km, while larger deviations are evident in the lower (10–15 km, underestimation) and upper layers (above 25–30 km, overestimation). The fixed-grid interpolation (left panel) introduces stronger variability, especially near the profile extremes, where the mismatch between the MIPAS and ozonesonde vertical grids leads to additional systematic errors. This results in noisier bias estimates and enhanced spread of the standard deviation. In contrast, the binned approach (right panel) yields smoother bias profiles and reduced variability, particularly at the top and bottom of the vertical domain, thus improving the robustness of the validation. Overall, while the general features of the MIPAS bias are consistent across both methods, the binned approach is more effective in mitigating edge effects and providing statistically stable results.

Figure D.2 shows that volume mixing ratio to partial columns conversion affects the validation results shifting the bias in positive. However, the MIPAS bias remains non

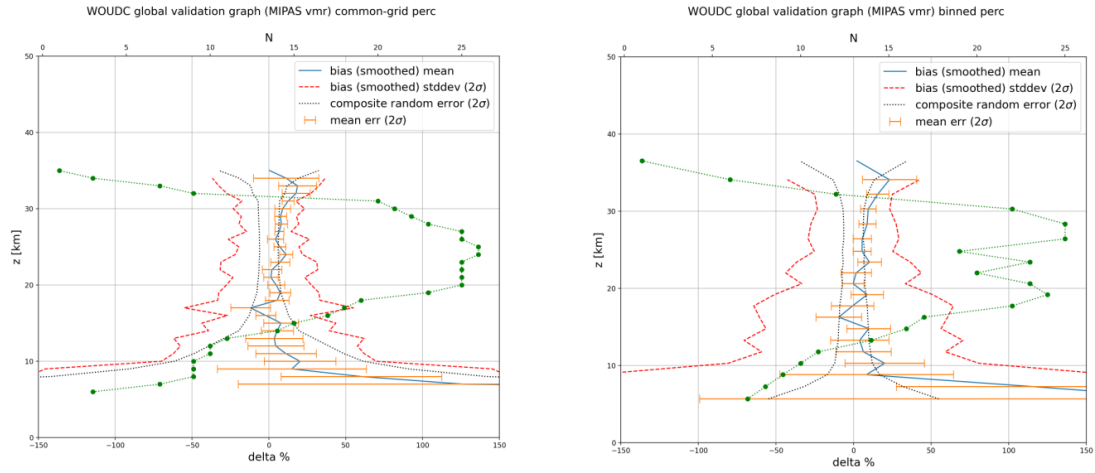


Figure D.1: MIPAS validation against WOUDC ozononesondes in 2008. Left panel: validation using fixed-grid interpolation approach. Right panel: validation using binning approach.

significant, except for the points in the upper and lower part of the grid, where the bias diverges reaching higher bias values. This is probably due to the problem mentioned before.

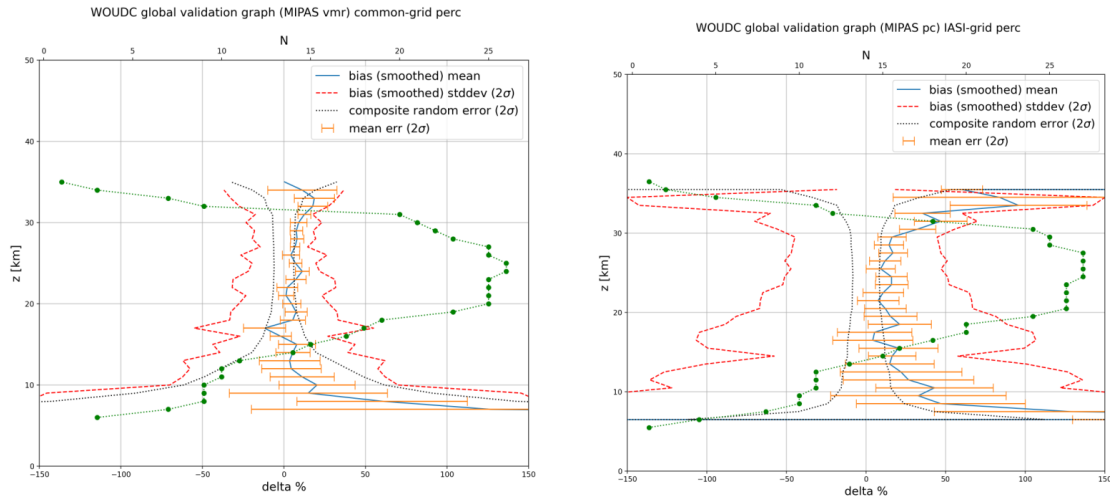


Figure D.2: MIPAS validation against WOUDC ozononesondes in 2008. Left panel: validation using fixed-grid interpolation approach and volume mixing ratio as unit measurements. Right panel: validation using fixed-grid interpolation approach and partial columns as unit measurements.



In Situ NCO-P Data Analysis

This Appendix presents the results of the analysis performed on the NCO-P dataset.

First of all we worked on values of ozone and relative humidity measured by the station at the surface. To analyze the daily variability of ozone (O_3) and relative humidity (RH) relative to their monthly mean, separate boxplots were generated for each type of methodology (SIO and STEFLUX). Specifically, for each day, the difference between the observed value during an event and the corresponding monthly mean was calculated, providing a normalized measure of the daily deviation.

These daily differences were then visualized using boxplots (Figure E.1), grouped by month, to highlight the distribution of daily values within each month of the year. This approach allows for a straightforward comparison of intra-monthly variability between different event types and facilitates the identification of potential outliers or seasonal trends. Each boxplot displays the median, first and third quartiles, and extreme values (“outliers”), while a line at zero was included to indicate reference difference with the monthly mean, making it easier to interpret deviations from the monthly norm. This method therefore provides a clear and concise characterization of the daily variability of ozone and relative humidity associated with each event type.

The analysis of the boxplots shows that during the intrusion events, O_3 values generally have a median above zero, while RH values have a median below zero. This indicates that, for most days of the month, the events are associated with an increase in ozone and a decrease in relative humidity compared to the monthly average. Both STEFLUX and SIO agree on that, and with previous studies mentioned in this thesis. Another prominent feature is the seasonality: summer months exhibit fewer events compared to autumn and winter months, which also show larger deviations from the monthly mean but with lower dispersion. This pattern is particularly evident in STEFLUX, with almost no events occurring in June, July, and August.

An additional analysis conducted in this work focuses on the combined frequency and

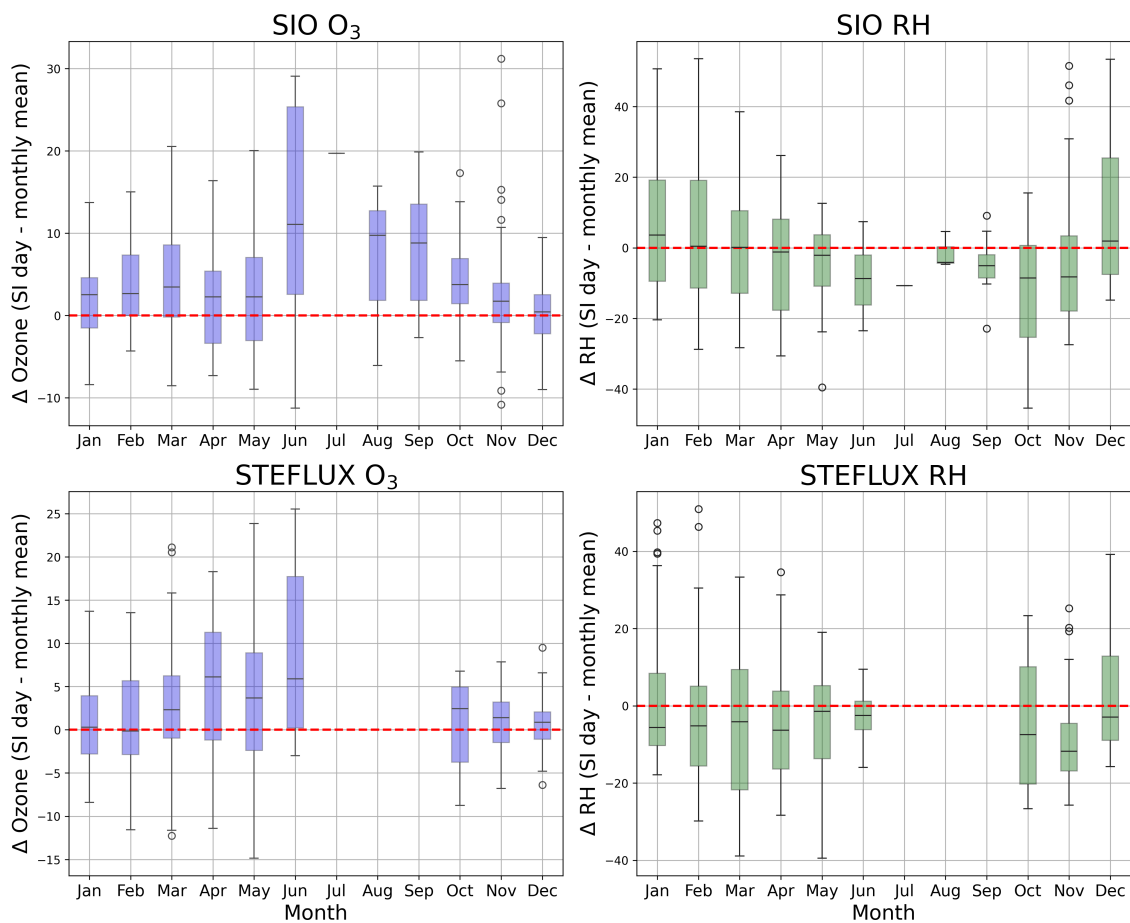
Distribution of $\Delta(\text{SI Day Values} - \text{Monthly Mean})$ for O_3 and RH

Figure E.1: Distribution of daily event deviations of ozone (O_3) and relative humidity (RH) from the monthly mean for SIO and STEFLUX events. Each boxplot shows the median (black line in the box), interquartile range (the box), and outliers (the whiskers) for each month, with a reference red line at zero.

duration of SI events (Figure E.2). Here, an event is defined as a period of one or more consecutive days during which the considered methodology registers a value of 1, indicating the presence of the phenomenon. For each year, events are identified and classified according to their duration, allowing the computation of annual statistics such as the total number of events, the total number of days affected, and the average duration per event. These statistics are visualized in the form of annual histograms, where each bar represents the total number of days with events, and overlaid labels indicate both the number of events and their mean duration. This approach provides a clear overview of not only how frequently events occur, but also how persistent they are, enabling a direct comparison between the STEFLUX and SIO methodologies.

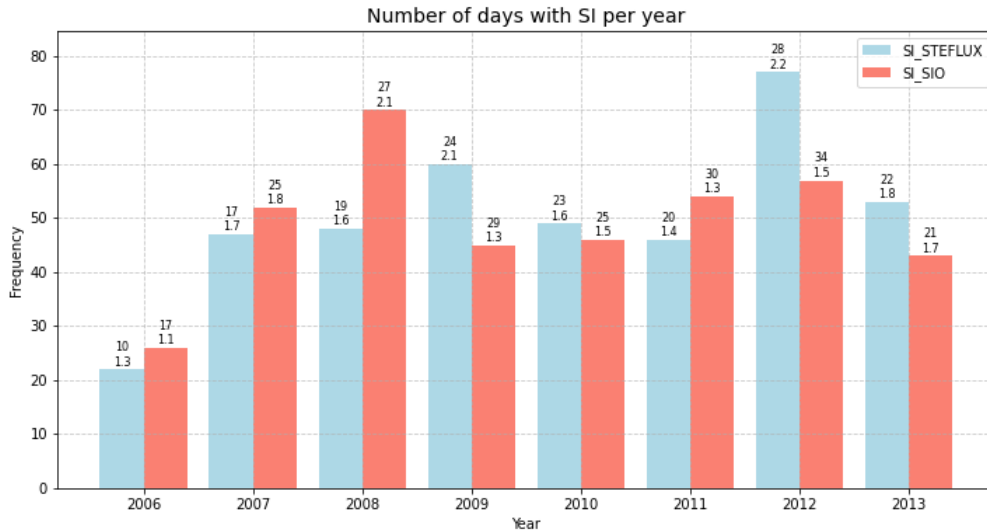


Figure E.2: Annual histogram of SI events detected by the STEFLUX (light blue) and SIO (light red) methodologies. Each bar represents the total number of days per year in which events occurred, while overlaid labels indicate the number of distinct events and their average duration (in days). Events are defined as one or more consecutive days with a methodology value of 1, highlighting both the frequency and persistence of the phenomena.

A visual inspection of the annual histogram suggests a possible pattern of alternating higher and lower event frequencies approximately every two years, observable in both STEFLUX and SIO datasets. However, this analysis alone is not sufficient to confirm a biennial cycle, and further statistical investigation would be required to assess the significance and persistence of such a pattern. Anyway, the combined analysis of frequency and duration provides insight into both how often events occur and how persistent they are over time. In general, event duration ranges from isolated single days to about one/two weeks, with higher frequency of short events, and longer-lasting events being less frequent. Years with high events number and the highest mean duration are 2008 and 2012.

To better investigate the biennial cycle a trend analysis has been performed. The dataset was analyzed using multiple complementary trend detection methods applied to event frequency, duration, and values of ozone and relative humidity. These methods combine

parametric, non-parametric, and robust approaches to provide a comprehensive understanding of long-term and seasonal variations while minimizing the influence of outliers and autocorrelation.

Linear regression and Pearson correlation were first applied to assess average temporal changes and the strength of linear relationships. However, as intrusion events are episodic and may not follow a strict linear pattern, Mann-Kendall (MK) tests were employed to detect monotonic trends without assuming linearity. To account for seasonality inherent in monthly observations, the Seasonal MK test was applied, preventing seasonal cycles from being misinterpreted as long-term trends. Robust trend estimators, such as the Theil-Sen slope and the Yue-Pilon method, were used to reduce sensitivity to extreme events and provide confidence intervals, ensuring reliable estimates even in the presence of outliers.

Finally, series were deseasonalized using STL decomposition, enabling robust trend estimation on the residuals, while spectral analysis helped identify hidden cyclicities in ozone and RH frequency and intensity values associated with events.

Figure E.3 shows the results of the trend analysis.

Variable	OLS regression		Seasonal Kendall		Deseasonalized Y-P	
	p-value	mean change /yr	p-value	mean change /yr	p-value	mean change /yr
Freq_SI_STEFLUX	0.040	−0.005	0.043	−0.033	0.004	−0.173
Duration_mean_SI_STEFLUX	0.252	−0.002	0.171	−0.012	0.068	−0.074
Duration_max_SI_STEFLUX	0.049	−0.005	0.066	−0.039	0.006	−0.173
intensity_o3_SI_STEFLUX	0.485	−0.029	0.547	−	0.226	−0.955
intensity_rh_SI_STEFLUX	0.739	0.015	1	−	0.905	0.051
Freq_SI_SIO	0.920	0.0003	0.939	−	0.834	0.016
Duration_mean_SI_SIO	0.887	0.0001	0.758	−	0.390	−0.023
Duration_max_SI_SIO	0.368	−0.002	0.875	−	0.817	−0.01
intensity_o3_SI_SIO	0.278	0.035	0.579	−0.029	0.635	0.305
intensity_rh_SI_SIO	0.002	0.136	0.002	4.97	0.0003	5.07

Figure E.3: Results of OLS regression, Seasonal Kendall, and Deseasonalized Y-P tests. Filled cells indicate statistical significance.

The analysis indicates a statistically significant negative trend in the frequency of STEFLUX SI events, while the relative humidity intensity of SIO SI events shows a positive trend. In contrast, ozone intensity does not exhibit statistically significant changes, despite some apparent increases through SIO. These results suggest that different aspects of SI events may be evolving in distinct ways, potentially reflecting site-specific dynamics. However, given the relatively short length of the available time series and the variability across methods, the robustness of these trends as indicators of long-term changes remains uncertain.

Figures E.4 and E.6 show the STL decomposition of the time series of stratospheric intrusion event frequency and the mean tropospheric ozone values associated with these events recorded by STEFLUX methodology at the NCO-P station over the period 2006–2013. In both decompositions, the seasonal component exhibits a clear annual cycle, with recurring peaks and troughs occurring during the same months each year. This indicates a stable and regular seasonal behavior.

The trend component shows a slower and irregular evolution over time, characterized by two main maxima around 2007–2008 and 2010–2011, followed by a gradual decline towards the end of the record. This alternation between phases of increasing and decreasing frequency and intensity suggests the presence of a quasi-biennial variability, possibly linked

to the Quasi-Biennial Oscillation (QBO). The QBO modulates tropical and subtropical stratospheric circulation, influencing planetary wave propagation and, consequently, the likelihood of stratosphere–troposphere exchange at mid-latitudes. Depending on the QBO phase (easterly or westerly), changes in the position and strength of the subtropical jet can modify the dynamical conditions favorable to STE events, leading to observable modulations in both event frequency and associated ozone intensity, even on a regional scale such as the Himalayas.

Finally, the residual component is larger for the ozone intensity series than for event frequency, reflecting higher episodic variability likely related to individual strong events or local meteorological influences. Overall, the STL decomposition reveals a marked annual seasonality superimposed on a quasi-biennial interannual variability, consistent with possible dynamical modulations by the QBO or similar hemispheric-scale oscillations. However, the available time series is too short to allow a more detailed investigation of this relationship.

Figure E.5 and E.7 show the same results obtained by SIO methodology. Despite some differences in the overall pattern and in the contributions of the different signal components, the agreement between the two decompositions suggests that despite methodological differences, both approaches identify a robust annual seasonality and a quasi-biennial modulation possibly linked to the QBO. However, as said before, given the limited temporal extent of the dataset, the interpretation of this interannual variability remains tentative and would require longer time series for confirmation.

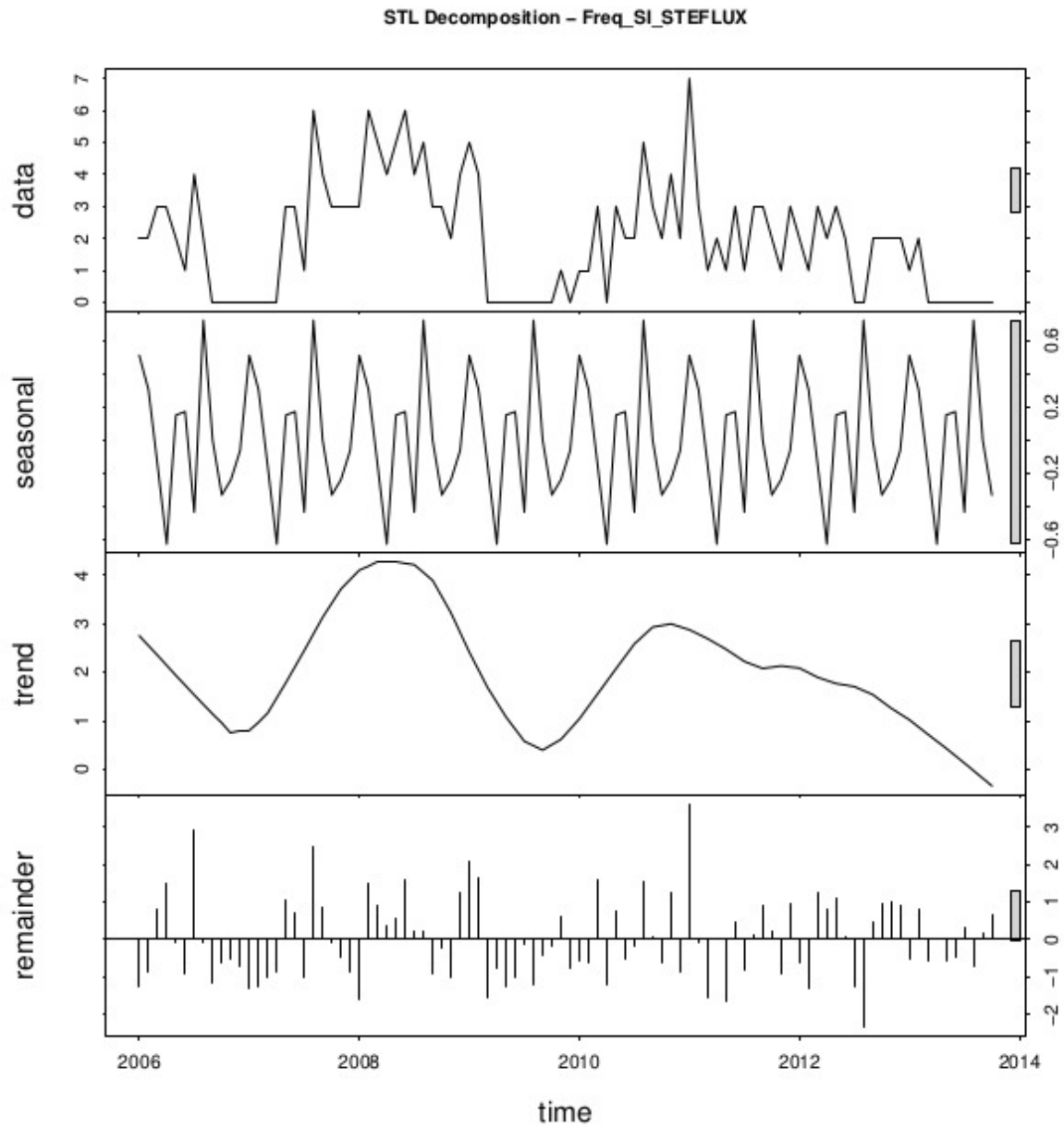


Figure E.4: STL decomposition of the time series of stratospheric intrusion events frequency identified by the STEFLUX methodology at NCO-P for the period 2006–2013.

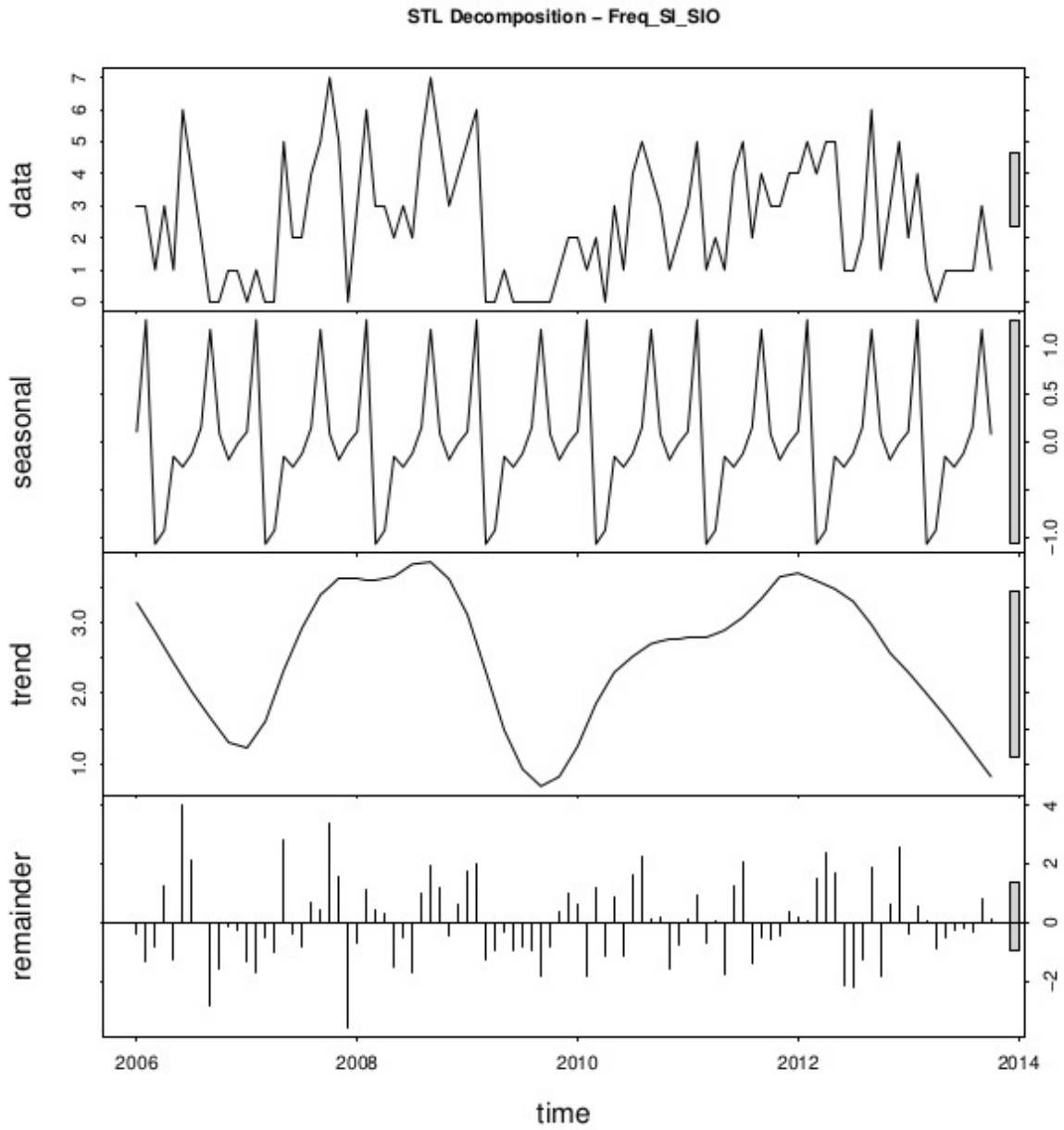


Figure E.5: STL decomposition of the time series of stratospheric intrusion events frequency identified by the SIO methodology at NCO-P for the period 2006–2013.

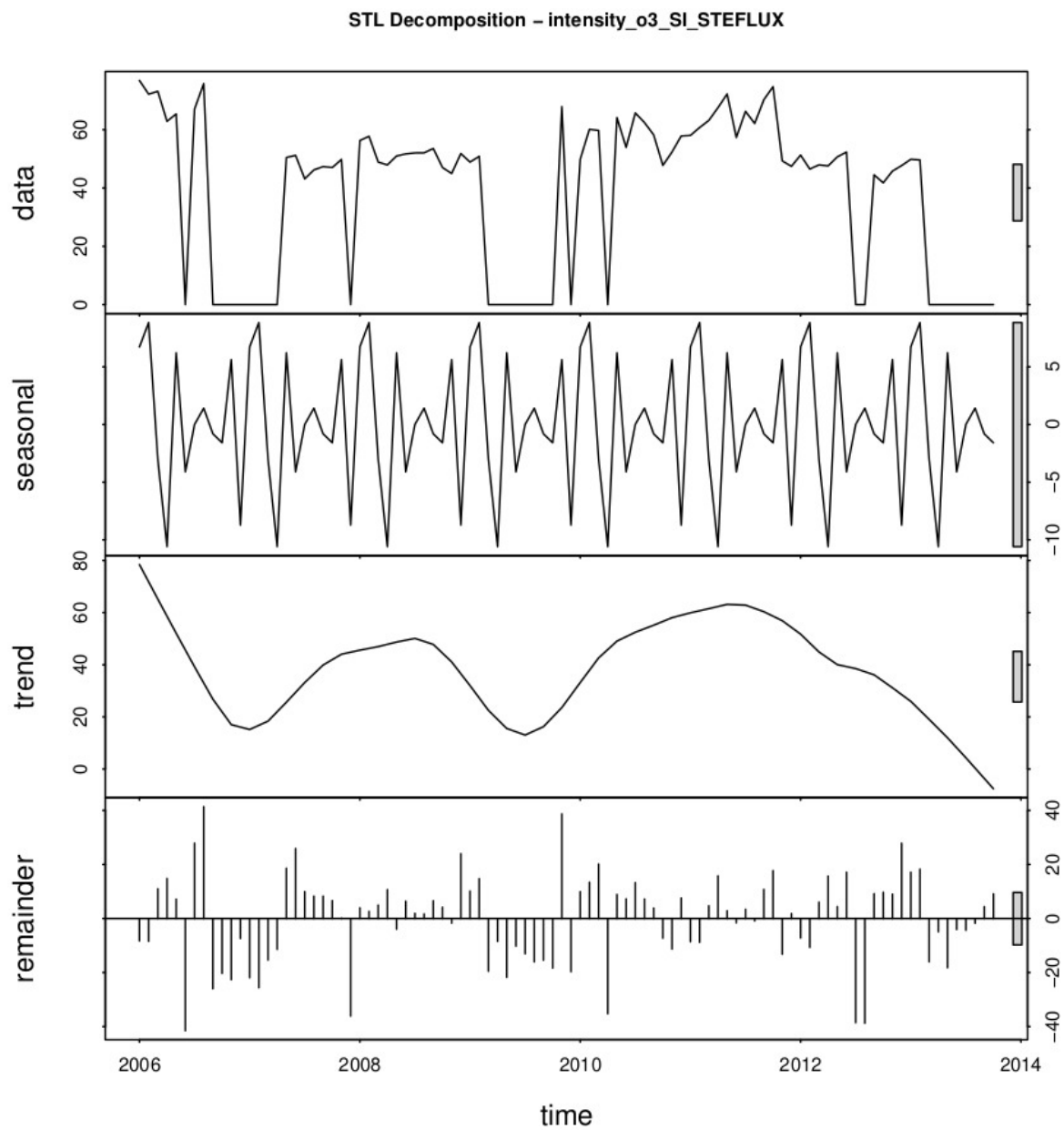


Figure E.6: STL decomposition of the time series of surface ozone values during stratospheric intrusion events identified by the STEFLUX methodology at NCO-P for the period 2006–2013.

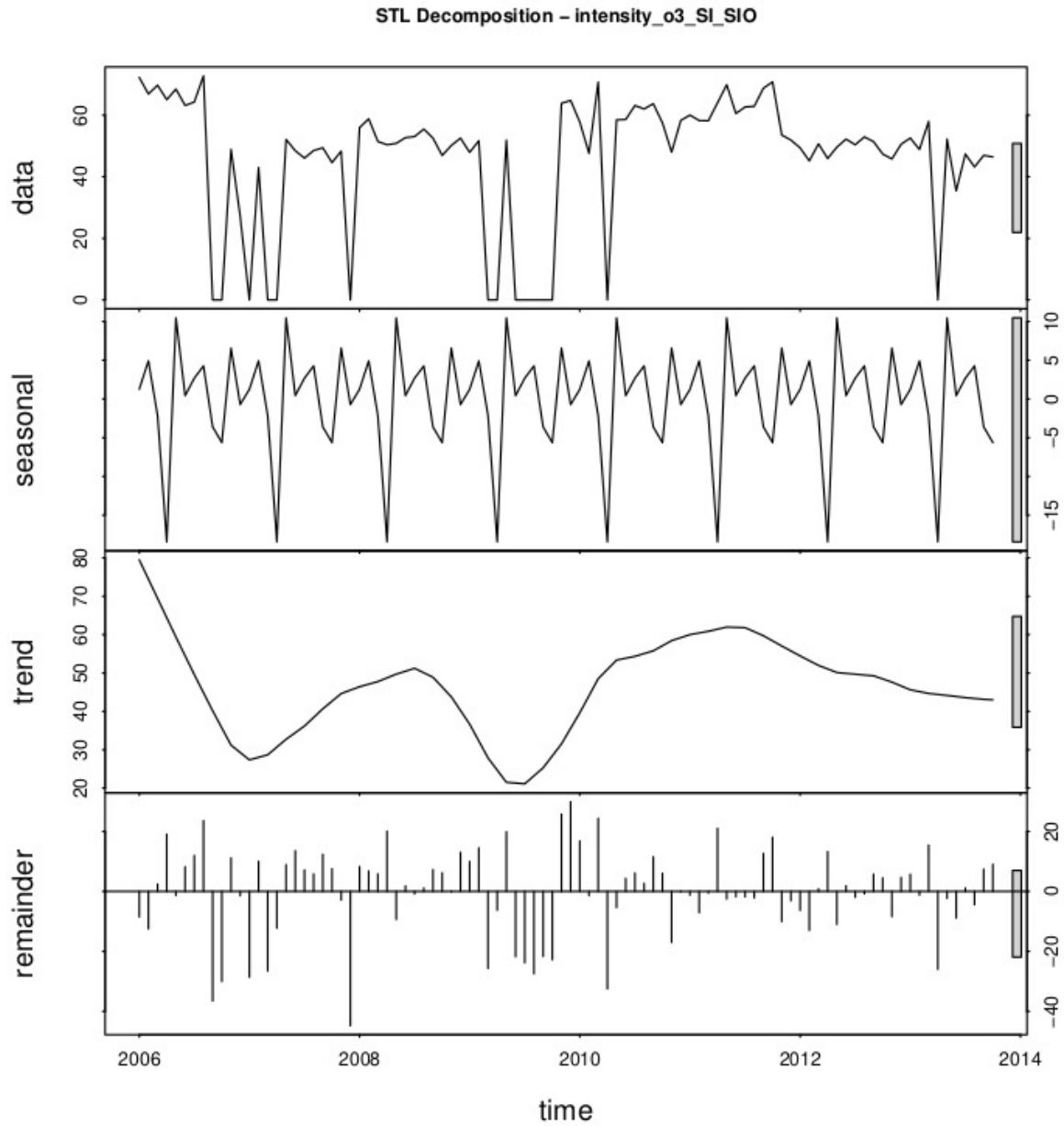


Figure E.7: STL decomposition of the time series of surface ozone values during stratospheric intrusion events identified by the SIO methodology at NCO-P for the period 2006–2013.

Bibliography

- Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.
- Ajtić, J. V., Sarvan, D., Djurdjevic, V. S., Hernández-Ceballos, M. A., & Brattich, E. (2017). Beryllium-7 surface concentration extremes in europe. *Facta Universitatis, Series: Physics, Chemistry and Technology*, 15(1), 045–055.
- Akritidis, D., Katragkou, E., Zanis, P., Pytharoulis, I., Melas, D., Flemming, J., Inness, A., Clark, H., Plu, M., & Eskes, H. (2018). A deep stratosphere-to-troposphere ozone transport event over europe simulated in cams global and regional forecast systems: Analysis and evaluation. *Atmospheric Chemistry and Physics*, 18(20), 15515–15534.
- Akritidis, D., Pozzer, A., Zanis, P., Tyrllis, E., Škerlak, B., Sprenger, M., & Lelieveld, J. (2016). On the role of tropopause folds in summertime tropospheric ozone over the eastern mediterranean and the middle east. *Atmospheric Chemistry and Physics*, 16(21), 14025–14039.
- Baldwin, M. P., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W., Holton, J., Alexander, M., Hirota, I., Horinouchi, T., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, 39(2), 179–229.
- Banerjee, A., Maycock, A. C., Archibald, A. T., Abraham, N. L., Telford, P., Braesicke, P., & Pyle, J. A. (2016). Drivers of changes in stratospheric and tropospheric ozone between year 2000 and 2100. *Atmospheric Chemistry and Physics*, 16(5), 2727–2746.
- Beekmann, M., Ancellet, G., Blonsky, S., De Muer, D., Ebel, A., Elbern, H., Hendricks, J., Kowol, J., Mancier, C., Sladkovic, R., et al. (1997). Regional and global tropopause fold occurrence and related ozone flux across the tropopause. *Journal of Atmospheric Chemistry*, 28(1), 29–44.
- Blume, C., Matthes, K., & Horenko, I. (2012). Supervised learning approaches to classify sudden stratospheric warming events. *Journal of the atmospheric sciences*, 69(6), 1824–1840.
- Blumstein, D., Chalon, G., Carlier, T., Buil, C., Hebert, P., Maciaszek, T., Ponce, G., Phulpin, T., Tournier, B., Simeoni, D., et al. (2004). Iasi instrument: Technical overview and measured performances. *Infrared Spaceborne Remote Sensing XII*, 5543, 196–207.
- Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, 8(4), 341–357.
- Boynard, A., Hurtmans, D., Koukouli, M. E., Goutail, F., Bureau, J., Safieddine, S., Lerot, C., Hadji-Lazaro, J., Wespes, C., Pommereau, J.-P., et al. (2016). Seven years of iasi ozone retrievals from forli: Validation with independent total column

-
- and vertical profile measurements. *Atmospheric Measurement Techniques*, 9(9), 4327–4353.
- Brattich, E., Liu, H., Zhang, B., Hernández-Ceballos, M. Á., Paatero, J., Sarvan, D., Djurdjevic, V., Tositti, L., & Ajtić, J. (2021). Observation and modeling of high-7 be concentration events at the surface in northern europe associated with the instability of the arctic polar vortex in early 2003. *Atmospheric Chemistry and Physics*, 21(23), 17927–17951.
- Brattich, E., Orza, J., Cristofanelli, P., Bonasoni, P., & Tositti, L. (2017). Influence of stratospheric air masses on radiotracers and ozone over the central mediterranean. *Journal of Geophysical Research: Atmospheres*, 122(13), 7164–7182.
- Brenninkmeijer, C. A., & Cai, D. S. (2016). Earth system chemistry integrated modelling (escimo) with the modular earth submodel system (messy) version 2.51. *Geoscientific Model Development*, 9(3), 1153.
- Butchart, N. (2014). The brewer-dobson circulation. *Reviews of geophysics*, 52(2), 157–184.
- Ceccherini, S., Cortesi, U., Del Bianco, S., Raspollini, P., & Carli, B. (2010). Iasi-metop and mipas-envisat data fusion. *Atmospheric Chemistry and Physics*, 10(10), 4689–4698. <https://doi.org/10.5194/acp-10-4689-2010>
- Ceccherini, S. (2016). Equivalence of measurement space solution data fusion and complete fusion. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 182, 71–74.
- Ceccherini, S., Carli, B., & Raspollini, P. (2015). Equivalence of data fusion and simultaneous retrieval. *Optics Express*, 23(7), 8476–8488.
- Ceccherini, S., Carli, B., Tirelli, C., Zoppetti, N., Del Bianco, S., Cortesi, U., Kujanpää, J., & Dragani, R. (2018). Importance of interpolation and coincidence errors in data fusion. *Atmospheric Measurement Techniques*, 11(2), 1009–1017.
- Ceccherini, S., Raspollini, P., & Carli, B. (2009). Optimal use of the information provided by indirect measurements of atmospheric vertical profiles. *Optics Express*, 17(7), 4944–4958.
- Ceccherini, S., Zoppetti, N., & Carli, B. (2022). An improved formula for the complete data fusion. *Atmospheric Measurement Techniques*, 15(23), 7039–7048.
- Chang, A., Tang, T. M., Zikry, T. M., & Allen, G. I. (2025). Unsupervised machine learning for scientific discovery: Workflow and best practices. *arXiv preprint arXiv:2506.04553*.
- Chauhan, S., Höpfner, M., Stiller, G., von Clarmann, T., Funke, B., Glatthor, N., Grabowski, U., Linden, A., Kellmann, S., Milz, M., et al. (2009). Mipas reduced spectral resolution utls-1 mode measurements of temperature, o₃, hno₃, n₂o, h₂o and relative humidity over ice: Retrievals and comparison to mls. *Atmospheric Measurement Techniques*, 2(2), 337–353.
- Chen, D., Zhou, T., Guo, D., & Ge, S. (2021). Simulation of the multi-timescale stratospheric intrusion processes in a typical cut-off low over northeast asia. *Atmosphere*, 13(1), 68.
- Clerbaux, C., Boynard, A., Clarisse, L., George, M., Hadji-Lazaro, J., Herbin, H., Hurtmans, D., Pommier, M., Razavi, A., Turquety, S., et al. (2009). Monitoring of atmospheric composition using the thermal infrared iasi/metop sounder. *Atmospheric Chemistry and Physics*, 9(16), 6041–6054.
- Coheur, P.-F., Clarisse, L., Turquety, S., Hurtmans, D., & Clerbaux, C. (2009). Iasi measurements of reactive trace species in biomass burning plumes. *Atmospheric Chemistry and Physics*, 9(15), 5655–5667.

-
- Cortesi, U., Lambert, J., De Clercq, C., Bianchini, G., Blumenstock, T., Bracher, A., Castelli, E., Catoire, V., Chance, K., De Maziere, M., et al. (2007). Geophysical validation of mipas-envisat operational ozone data. *Atmospheric Chemistry and Physics*, 7(18), 4807–4867.
- Cristofanelli, P., Bonasoni, P., Bonafé, U., Calzolari, F., Duchi, R., Marinoni, A., Roccato, F., Vuillermoz, E., & Sprenger, M. (2009). Influence of lower stratosphere/upper troposphere transport events on surface ozone at the everest-pyramid gaw station (nepal): First year of analysis. *International Journal of Remote Sensing*, 30(15-16), 4083–4097.
- Cristofanelli, P., Bracci, A., Sprenger, M., Marinoni, A., Bonafè, U., Calzolari, F., Duchi, R., Laj, P., Pichon, J. M., Roccato, F., et al. (2010). Tropospheric ozone variations at the nepal climate observatory-pyramid (himalayas, 5079 m asl) and influence of deep stratospheric intrusion events. *Atmospheric Chemistry and Physics*, 10(14), 6537–6549.
- Cristofanelli, P., Scheel, H.-E., Steinbacher, M., Saliba, M., Azzopardi, F., Ellul, R., Fröhlich, M., Tositti, L., Brattich, E., Maione, M., et al. (2015). Long-term surface ozone variability at mt. cimone wmo/gaw global station (2165 m asl, italy). *Atmospheric Environment*, 101, 23–33.
- Crutzen, P. J., & Lelieveld, J. (2001). Human impacts on atmospheric chemistry. *Annual review of earth and planetary sciences*, 29(1), 17–45.
- Cui, J., Sprenger, M., Staehelin, J., Siegrist, A., Kunz, M., Henne, S., & Steinbacher, M. (2009). Impact of stratospheric intrusions and intercontinental transport on ozone at jungfraujoch in 2005: Comparison and validation of two lagrangian approaches. *Atmospheric Chemistry and Physics*, 9(10), 3371–3383.
- Dhomse, S. S., Arosio, C., Feng, W., Rozanov, A., Weber, M., & Chipperfield, M. P. (2021). Ml-tomcat: Machine-learning-based satellite-corrected global stratospheric ozone profile data set from a chemical transport model. *Earth System Science Data*, 13(12), 5711–5729.
- Di Girolamo, P., Summa, D., & Ferretti, R. (2009). Multiparameter raman lidar measurements for the characterization of a dry stratospheric intrusion event. *Journal of Atmospheric and Oceanic Technology*, 26(9), 1742–1762.
- Dinelli, B. M., Raspollini, P., Gai, M., Sgheri, L., Ridolfi, M., Ceccherini, S., Barbara, F., Zoppetti, N., Castelli, E., Papandrea, E., et al. (2021). The esa mipas/envisat level2-v8 dataset: 10 years of measurements retrieved with orm v8. 22. *Atmospheric Measurement Techniques*, 14(12), 7975–7998.
- Ding, A., & Wang, T. (2006). Influence of stratosphere-to-troposphere exchange on the seasonal cycle of surface ozone at mount waliguan in western china. *Geophysical Research Letters*, 33(3).
- Donzelli, G., & Suarez-Varela, M. M. (2024). Tropospheric ozone: A critical review of the literature on emissions, exposure, and health effects. *Atmosphere*, 15(7), 779.
- Eremenko, M., Dufour, G., Foret, G., Keim, C., Orphal, J., Beekmann, M., Bergametti, G., & Flaud, J.-M. (2008). Tropospheric ozone distributions over europe during the heat wave in july 2007 observed from infrared nadir spectra recorded by iasi. *Geophysical Research Letters*, 35(18).
- Fadnavis, S., Chakraborty, T., & Beig, G. (2010). Seasonal stratospheric intrusion of ozone in the upper troposphere over india. *Annales Geophysicae*, 28(11), 2149–2159. <https://doi.org/10.5194/angeo-28-2149-2010>

-
- Fahrin, F., Jones, D. C., Wu, Y., Keeble, J., & Archibald, A. T. (2022). Unsupervised classification of ozone profiles in ukesm1. *Atmospheric Chemistry and Physics Discussions, 2022*, 1–27.
- Fischer, H., Birk, M., Blom, C., Carli, B., Carlotti, M., Von Clarmann, T., Delbouille, L., Dudhia, A., Ehhalt, D., Endemann, M., et al. (2008). Mipas: An instrument for atmospheric and climate research. *Atmospheric Chemistry and Physics, 8*(8), 2151–2188.
- Fischer, H., & Oelhaf, H. (1996). Remote sensing of vertical profiles of atmospheric trace constituents with mipas limb-emission spectrometers. *Applied Optics, 35*(16), 2787–2796.
- Fleming, Z. L., Doherty, R. M., Von Schneidmesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., et al. (2018). Tropospheric ozone assessment report: Present-day ozone distribution and trends relevant to human health. *Elem Sci Anth, 6*, 12.
- Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., & Zhang, H. (2021). Chapter 7: The earth’s energy budget, climate feedbacks, and climate sensitivity. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 923–1054). Cambridge University Press. <https://doi.org/10.1017/9781009157896.009>
- Fueglistaler, S., Dessler, A., Dunkerton, T., Folkins, I., Fu, Q., & Mote, P. W. (2009). Tropical tropopause layer. *Reviews of Geophysics, 47*(1).
- Fuhrer, J., Skärby, L., & Ashmore, M. R. (1997). Critical levels for ozone effects on vegetation in europe. *Environmental pollution, 97*(1-2), 91–106.
- Gadgil, S. (2003). The indian monsoon and its variability. *Annual Review of Earth and Planetary Sciences, 31*(1), 429–467.
- Garcia, R. R., & Randel, W. J. (2008). Acceleration of the brewer–dobson circulation due to increases in greenhouse gases. *Journal of the Atmospheric Sciences, 65*(8), 2731–2739.
- Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P.-F., Cuesta, J., Cuevas, E., et al. (2018). Tropospheric ozone assessment report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation. *Elem Sci Anth, 6*, 39.
- Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., et al. (2017). The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate, 30*(14), 5419–5454.
- Gettelman, A., Hoor, P., Pan, L., Randel, W., Hegglin, M. I., & Birner, T. (2011). The extratropical upper troposphere and lower stratosphere. *Reviews of Geophysics, 49*(3).
- Greenslade, J. W., Alexander, S. P., Schofield, R., Fisher, J. A., & Klekociuk, A. K. (2017). Stratospheric ozone intrusion events and their impacts on tropospheric ozone in the southern hemisphere. *Atmospheric Chemistry and Physics, 17*(17), 10269–10290.

-
- Gronoff, G., Berkoff, T., Knowland, K., Lei, L., Shook, M., Fabbri, B., Carrion, W., & Langford, A. (2021). Case study of stratospheric intrusion above hampton, virginia: Lidar-observation and modeling analysis. *Atmospheric Environment*, 259, 118498.
- Hegglin, Gettelman, A., Hoor, P., Krichovsky, R., Manney, G., Pan, L., Son, S.-W., Stiller, G., Tilmes, S., Walker, K., et al. (2010). Multimodel assessment of the upper troposphere and lower stratosphere: Extratropics. *Journal of Geophysical Research: Atmospheres*, 115(D3).
- Hegglin, M. I. L. A., Fahey, D. W., McFarland, M., Montzka, S. A., & Nash, E. R. (2015). *Twenty questions and answers about the ozone layer: 2014 update* (tech. rep. No. Report No. 57) (Accessed: September 2025). World Meteorological Organization. https://ozonewatch.gsfc.nasa.gov/facts/hole_SH.html
- Hegglin & Shepherd, T. G. (2009). Large climate-induced changes in ultraviolet index and stratosphere-to-troposphere ozone flux. *Nature Geoscience*, 2(10), 687–691.
- Henne, S., Klausen, J., Junkermann, W., Kariuki, J. M., Aseyo, J. O., & Buchmann, B. (2008). Representativeness and climatology of carbon monoxide and ozone at the global gaw station mt. kenya in equatorial africa. *Atmospheric Chemistry and Physics*, 8(12), 3119–3139.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730), 1999–2049.
- Hess, P., & Zbinden, R. (2013). Stratospheric impact on tropospheric ozone variability and trends: 1990–2009. *Atmospheric Chemistry and Physics*, 13(2), 649–674.
- Hitchman, M. H., & Rowe, S. M. (2021). On the formation of tropopause folds and constituent gradient enhancement near westerly jets. *Journal of the Atmospheric Sciences*, 78(7), 2057–2074.
- Hofmann, C., Kerkweg, A., Hoor, P., & Jöckel, P. (2016). Stratosphere-troposphere exchange in the vicinity of a tropopause fold. *Atmospheric Chemistry and Physics Discussions*, 2016, 1–26.
- Holton, J. R., Haynes, P. H., McIntyre, M. E., Douglass, A. R., Rood, R. B., & Pfister, L. (1995). Stratosphere-troposphere exchange. *Reviews of geophysics*, 33(4), 403–439.
- Holton, J. R., & Lelieveld, J. (1996). Stratosphere-troposphere exchange and its role in the budget of tropospheric ozone. In *Clouds, chemistry and climate* (pp. 173–190). Springer.
- Höpfner, M., Boone, C., Funke, B., Glatthor, N., Grabowski, U., Günther, A., Kellmann, S., Kiefer, M., Linden, A., Lossow, S., et al. (2015). Sulfur dioxide (so 2) from mipas in the upper troposphere and lower stratosphere 2002–2012. *Atmospheric Chemistry and Physics*, 15(12), 7017–7037.
- Hua, J., Zhang, Y., Chen, J., Shang, Y., Hu, X., de Foy, B., Sulaymon, I. D., Zhang, Y., & Wu, Z. (2024). Unravelling the impacts of stratospheric intrusions on near-surface ozone during the springtime ozone pollution episodes in lhasa, china. *Atmospheric Research*, 311, 107687.
- Hurtmans, D., Coheur, P.-F., Wespes, C., Clarisse, L., Scharf, O., Clerbaux, C., Hadji-Lazaro, J., George, M., & Turquety, S. (2012). Forli radiative transfer and retrieval code for iasi. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 113(11), 1391–1408.
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., et al. (2019). The

-
- cams reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6), 3515–3556.
- Jöckel, P., Brinkop, S., Graf, P., Eichinger, R., Garny, H., Mertens, M., Nützel, M., & Pozzer, A. (2024). Rd1sd: Emac ccmi-2022 hindcast simulations with specified dynamics, era-5, 1979-2019 (additional data).
- Joiner, J., Schoeberl, M., Vasilkov, A., Oreopoulos, L., Platnick, S., Livesey, N., & Levelt, P. (2009). Accurate satellite-derived estimates of the tropospheric ozone impact on the global radiation budget. *Atmospheric Chemistry and Physics*, 9(13), 4447–4465.
- Jordan, C., Dibb, J. E., & Finkel, R. (2003). 10be/7be tracer of atmospheric transport and stratosphere-troposphere exchange. *Journal of Geophysical Research: Atmospheres*, 108(D8).
- Keppens, A., Di Pede, S., Hubert, D., Lambert, J.-C., Veefkind, P., Snee, M., De Haan, J., Ter Linden, M., Leblanc, T., Compennolle, S., et al. (2024). 5 years of sentinel-5p tropomi operational ozone profiling and geophysical validation using ozonesonde and lidar ground-based networks. *Atmospheric Measurement Techniques*, 17(13), 3969–3993.
- Kerkweg, A., Sander, R., Tost, H., & Jöckel, P. (2006). Implementation of prescribed (offlem), calculated (onlem), and pseudo-emissions (tnudge) of chemical species in the modular earth submodel system (messy). *Atmospheric Chemistry and Physics*, 6(11), 3603–3609.
- Koch, D., & Rind, D. (1998). Beryllium 10/beryllium 7 as a tracer of stratospheric transport. *Journal of Geophysical Research: Atmospheres*, 103(D4), 3907–3917.
- Kunz, A., Sprenger, M., & Wernli, H. (2015). Climatology of potential vorticity streamers and associated isentropic transport pathways across pv gradient barriers. *Journal of Geophysical Research: Atmospheres*, 120(9), 3802–3821.
- Langford, A., Aikin, K., Eubank, C., & Williams, E. (2009). Stratospheric contribution to high surface ozone in colorado during springtime. *Geophysical Research Letters*, 36(12).
- Larchevêque, G., Balin, I., Nessler, R., Quaglia, P., Simeonov, V., van den Bergh, H., & Calpini, B. (2002). Development of a multiwavelength aerosol and water-vapor lidar at the jungfrau-joch alpine station (3580 m above sea level) in switzerland. *Applied optics*, 41(15), 2781–2790.
- Li, D., Vogel, B., Müller, R., Bian, J., Günther, G., Li, Q., Zhang, J., Bai, Z., Vömel, H., & Riese, M. (2018). High tropospheric ozone in lhasa within the asian summer monsoon anticyclone in 2013: Influence of convective transport and stratospheric intrusions. *Atmospheric chemistry and physics*, 18(24), 17979–17994.
- Li, K., Tan, R., Qiao, W., Lee, T., Wang, Y., Zhang, D., Tang, M., Zhao, W., Gu, Y., Fan, S., et al. (2025). Surface and tropospheric ozone over east asia and southeast asia from observations: Distributions, trends, and variability. *Atmospheric Chemistry and Physics*, 25(19), 11575–11596.
- Lin, M., Fiore, A. M., Cooper, O. R., Horowitz, L. W., Langford, A. O., Levy, H., Johnson, B. J., Naik, V., Oltmans, S. J., & Senff, C. J. (2012). Springtime high surface ozone events over the western united states: Quantifying the role of stratospheric intrusions. *Journal of Geophysical Research: Atmospheres*, 117(D21).
- Liu, J., Rodriguez, J. M., Oman, L. D., Douglass, A. R., Olsen, M. A., & Hu, L. (2020). Stratospheric impact on the northern hemisphere winter and spring ozone interan-

-
- nual variability in the troposphere. *Atmospheric Chemistry and Physics*, 20(11), 6417–6433.
- Liu, X., Fu, Y., Zhang, L., Li, H., Burr, G. S., Bi, Y., & Zhao, G. (2024). Deep stratospheric intrusion events in china revealed on the ground by cosmogenic 10be/7be. *Communications Earth & Environment*, 5(1), 553.
- Logan, J. A. (1999). An analysis of ozonesonde data for the troposphere: Recommendations for testing 3-d models and development of a gridded climatology for tropospheric ozone. *Journal of Geophysical Research: Atmospheres*, 104(D13), 16115–16149.
- Luo, J., Liang, W., Xu, P., Xue, H., Zhang, M., Shang, L., & Tian, H. (2019). Seasonal features and a case study of tropopause folds over the tibetan plateau. *Advances in Meteorology*, 2019(1), 4375123.
- Maahn, M., Turner, D. D., Löhnert, U., Posselt, D. J., Ebell, K., Mace, G. G., & Comstock, J. M. (2020). Optimal estimation retrievals and their uncertainties: What every atmospheric scientist should know. *Bulletin of the American Meteorological Society*, 101(9), E1512–E1523.
- Martius, O., Schwierz, C., & Davies, H. (2010). Tropopause-level waveguides. *Journal of the Atmospheric Sciences*, 67(3), 866–879.
- McKenna, D. S., Konopka, P., Groöf, J.-U., Günther, G., Müller, R., Spang, R., Offermann, D., & Orsolini, Y. (2002). A new chemical lagrangian model of the stratosphere (clams) 1. formulation of advection and mixing. *Journal of Geophysical Research: Atmospheres*, 107(D16), ACH–15.
- McPeters, R. D., Labow, G. J., & Logan, J. A. (2007). Ozone climatological profiles for satellite retrieval algorithms. *Journal of Geophysical Research: Atmospheres*, 112(D5).
- Mkololo, T., Mbatha, N., Sivakumar, V., Bègue, N., Coetzee, G., & Labuschagne, C. (2020). Stratosphere–troposphere exchange and o₃ variability in the lower stratosphere and upper troposphere over the irene shadoz site, south africa. *Atmosphere*, 11(6), 586.
- Mohanakumar, K. (2008). Stratosphere–troposphere exchange. *Stratosphere Troposphere Interactions: An Introduction*, 331–355.
- Monks, P. S., Archibald, A., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G., et al. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric chemistry and physics*, 15(15), 8889–8973.
- Moore, G., & Semple, J. L. (2004). High himalayan meteorology: Weather at the south col of mount everest. *Geophysical Research Letters*, 31(18).
- Morgenstern, O., Hegglin, M. I., Rozanov, E., O’Connor, F. M., Abraham, N. L., Akiyoshi, H., Archibald, A. T., Bekki, S., Butchart, N., Chipperfield, M. P., et al. (2017). Review of the global models used within phase 1 of the chemistry–climate model initiative (ccmi). *Geoscientific Model Development*, 10(2), 639–671.
- Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.
- NASA. (2023a). Ozone facts [Accessed: September 2025].
- NASA. (2023b). Ozone facts [Accessed: September 2025].

-
- Nepal Climate Observatory - Pyramid. (2011). Nepal climate observatory - pyramid (himalaya - nepal) [Accessed: 2025-09-05].
- Nieto, R., Sprenger, M., Wernli, H., Trigo, R., & Gimeno, L. (2008). Identification and climatology of cut-off lows near the tropopause. *Annals of the New York Academy of Sciences*, 1146(1), 256–290.
- NOAA. (2014). Q2: How is ozone formed in the atmosphere? [Accessed: September 2025].
- NOAA. (2018). Q3: How is total ozone distributed over the globe? [Accessed: September 2025].
- NOAA. (2022). Scientific assessment of ozone depletion: 2022 [Accessed: September 2025].
- NOAA. (2023). Ozone and water vapor [Accessed: September 2025].
- Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of k-means clustering algorithms. *criterion*, 12(08), 22–23.
- Philippopoulos, K., Deligiorgi, D., & Kouroupetroglou, G. (2014). Performance comparison of self-organizing maps and k-means clustering techniques for atmospheric circulation classification. *methods*, 13, 14.
- Prather, M. J., & Zhu, X. (2024). Lifetimes and timescales of tropospheric ozone: Global metrics for climate change, human health, and crop/ecosystem research. *Elementa: Science of the Anthropocene*, 12(1).
- Putero, D., Cristofanelli, P., Sprenger, M., Škerlak, B., Tositti, L., & Bonasoni, P. (2016). Steflux, a tool for investigating stratospheric intrusions: Application to two wmo/gaw global stations. *Atmospheric Chemistry and Physics*, 16(22), 14203–14217.
- Ramanathan, V., & Carmichael, G. (2008). Global and regional climate changes due to black carbon. *Nature geoscience*, 1(4), 221–227.
- Randel, W. J., & Park, M. (2006). Deep convective influence on the asian summer monsoon anticyclone and associated tracer variability observed with atmospheric infrared sounder (airs). *Journal of Geophysical Research: Atmospheres*, 111(D12).
- Raspollini, P., Arnone, E., Barbara, F., Bianchini, M., Carli, B., Ceccherini, S., Chipperfield, M. P., Dehn, A., Della Fera, S., Dinelli, B. M., et al. (2022). Level 2 processor and auxiliary data for esa version 8 final full mission analysis of mipas measurements on envisat. *Atmospheric Measurement Techniques*, 15(6), 1871–1901.
- Razavi, A., Clerbaux, C., Wespes, C., Clarisse, L., Hurtmans, D., Payan, S., Camy-Peyret, C., & Coheur, P.-F. (2009). Characterization of methane retrievals from the iasi space-borne sounder. *Atmospheric Chemistry and Physics*, 9(20), 7889–7899.
- Remedios, J. J., Leigh, R. J., Waterfall, A. M., Moore, D. P., Sembhi, H., Parkes, I., Greenhough, J., Chipperfield, M. P., & Hauglustaine, D. (2007). Mipas reference atmospheres and comparisons to v4.61/v4.62 mipas level 2 geophysical data sets. *Atmospheric Chemistry and Physics Discussions*, 7, 9973–10017. <https://doi.org/10.5194/acpd-7-9973-2007>
- Richards, J. A. (2022). Clustering and unsupervised classification. In *Remote sensing digital image analysis* (pp. 369–401). Springer.
- Rodgers, C. D. (2000). *Inverse methods for atmospheric sounding: Theory and practice* (Vol. 2). World scientific.
- Roelofs, G.-J., & Lelieveld, J. (1997). Model study of the influence of cross-tropopause o₃ transports on tropospheric o₃ levels. *Tellus B: Chemical and Physical Meteorology*, 49(1), 38–55.
- Roy, C., Ravishankara, A., Newman, P. A., David, L. M., Fadnavis, S., Rathod, S. D., Lait, L., Krishnan, R., Clark, H., & Sauvage, B. (2023). Estimation of stratospheric

-
- intrusions during indian cyclones. *Journal of Geophysical Research: Atmospheres*, 128(3), e2022JD037519.
- Runde, T., Dameris, M., Garny, H., & Kinnison, D. (2016). Classification of stratospheric extreme events according to their downward propagation to the troposphere. *Geophysical Research Letters*, 43(12), 6665–6672.
- Sander, R., Baumgaertner, A., Gromov, S., Harder, H., Jöckel, P., Kerkweg, A., Kubistin, D., Regelin, E., Riede, H., Sandu, A., et al. (2011). The atmospheric chemistry box model caaba/mecca-3.0. *Geoscientific Model Development*, 4(2), 373–380.
- Sander, R., Jöckel, P., Kirner, O., Kunert, A., Landgraf, J., & Pozzer, A. (2014). The photolysis module jval-14, compatible with the messy standard, and the jval pre-processor (jvpp). *Geoscientific Model Development*, 7(6), 2653–2662.
- Schoeberl, M., Jensen, E., Pfister, L., Ueyama, R., Wang, T., Selkirk, H., Avery, M., Thornberry, T., & Dessler, A. (2019). Water vapor, clouds, and saturation in the tropical tropopause layer. *Journal of Geophysical Research: Atmospheres*, 124(7), 3984–4003.
- Seguel, R. J., Opazo, C., Cohen, Y., Cooper, O. R., Gallardo, L., Sinnhuber, B.-M., Obersteiner, F., Zahn, A., Hoor, P., Rohs, S., et al. (2025). Hemispheric differences in ozone across the stratosphere–troposphere exchange region. *Atmospheric Chemistry and Physics*, 25(15), 8553–8573.
- Sembhi, H., Remedios, J., Trent, T., Moore, D., Spang, R., Massie, S., & Vernier, J.-P. (2012). Mipas detection of cloud and aerosol particle occurrence in the utls with comparison to hirdls and caliop. *Atmospheric Measurement Techniques*, 5(10), 2537–2553.
- Serra, A., & Tagliaferri, R. (2019). Unsupervised learning: Clustering.
- Singh, H. B., Viezee, W., Johnson, W. B., & Ludwig, F. L. (1980). The impact of stratospheric ozone on tropospheric air quality. *Journal of the air pollution control association*, 30(9), 1009–1017.
- Škerlak, B., Sprenger, M., & Wernli, H. (2014). A global climatology of stratosphere–troposphere exchange using the era-interim data set from 1979 to 2011. *Atmospheric Chemistry and Physics*, 14(2), 913–937.
- Sprenger, M., & Wernli, H. (2003). A northern hemispheric climatology of cross-tropopause exchange for the era15 time period (1979–1993). *Journal of Geophysical Research: Atmospheres*, 108(D12).
- Sprenger, M., & Wernli, H. (2015). The lagranto lagrangian analysis tool–version 2.0. *Geoscientific Model Development*, 8(8), 2569–2586.
- Srivastava, A., RAY, K. S., & Kulkarni, R. (2002). Seasonal variations, trend in the tropopause height/temperature over indian stations and its modulations by sst anomalies of east pacific ocean. *Mausam*, 53(4), 439–446.
- Stauffer, R. M., Morris, G. A., Thompson, A. M., Joseph, E., Coetzee, G. J. R., & Nalli, N. R. (2014). Propagation of radiosonde pressure sensor errors to ozonesonde measurements. *Atmospheric Measurement Techniques*, 7(1), 65–79. <https://doi.org/10.5194/amt-7-65-2014>
- Sterling, C. W., Johnson, B. J., Oltmans, S. J., Smit, H. G., Jordan, A. F., Cullis, P. D., Hall, E. G., Thompson, A. M., & Witte, J. C. (2018). Homogenizing and estimating the uncertainty in noaa’s long-term vertical ozone profile records measured with the electrochemical concentration cell ozonesonde. *Atmospheric Measurement Techniques*, 11(6), 3661–3687.

-
- Stevenson, D., Dentener, F., Schultz, M., Ellingsen, K., Van Noije, T., Wild, O., Zeng, G., Amann, M., Atherton, C., Bell, N., et al. (2006). Multimodel ensemble simulations of present-day and near-future tropospheric ozone. *Journal of Geophysical Research: Atmospheres*, 111(D8).
- Stohl, A., Bonasoni, P., Cristofanelli, P., Collins, W., Feichter, J., Frank, A., Forster, C., Gerasopoulos, E., Gäggeler, H., James, P., et al. (2003). Stratosphere-troposphere exchange: A review, and what we have learned from staccato. *Journal of Geophysical Research: Atmospheres*, 108(D12).
- Tarasick, D., Carey-Smith, T., Hocking, W., Moeini, O., He, H., Liu, J., Osman, M., Thompson, A., Johnson, B., Oltmans, S., et al. (2019). Quantifying stratosphere-troposphere transport of ozone using balloon-borne ozonesondes, radar windprofilers and trajectory models. *Atmospheric Environment*, 198, 496–509.
- Tarasick, D., Jin, J., Fioletov, V., Liu, G., Thompson, A., Oltmans, S., Liu, J., Sioris, C., Liu, X., Cooper, O., et al. (2010). High-resolution tropospheric ozone fields for inter and arctic from ion ozonesondes. *Journal of Geophysical Research: Atmospheres*, 115(D20).
- Thompson, A. M., Witte, J. C., Oltmans, S. J., Schmidlin, F. J., Logan, J. A., Fujiwara, M., Kirchhoff, V. W., Posny, F., Coetzee, G. J., Hoegger, B., et al. (2003). Southern hemisphere additional ozonesondes (shadoz) 1998–2000 tropical ozone climatology 2. tropospheric variability and the zonal wave-one. *Journal of Geophysical Research: Atmospheres*, 108(D2).
- Tirelli, C., Ceccherini, S., Cortesi, U., Del Bianco, S., Errera, Q., Funke, B., Höpfner, M., Kujanpää, J., Poli, G., Preusse, P., et al. (2023). Potential advantages coming from the synergy between products by limb and nadir imaging fourier transform spectrometers: 2d data fusion of cairt,iasi-ng (and sentinel 5) simulated data. *Fourier Transform Spectroscopy*, FTu5B–5.
- Tirelli, C., Ceccherini, S., Del Bianco, S., Funke, B., Höpfner, M., Cortesi, U., & Raspollini, P. (2025). Extension of the complete data fusion algorithm to tomographic retrieval products. *EGU sphere*, 2025, 1–24.
- Tirelli, C., Ceccherini, S., Zoppetti, N., Del Bianco, S., & Cortesi, U. (2021). Generalization of the complete data fusion to multi-target retrieval of atmospheric parameters and application to forum andiasi-ng simulated measurements. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 276, 107925.
- Tost, H., Jöckel, P., Kerkweg, A., Sander, R., & Lelieveld, J. (2006). A new comprehensive scavenging submodel for global atmospheric chemistry modelling. *Atmospheric Chemistry and Physics*, 6(3), 565–574.
- Tost, H., Jöckel, P., & Lelieveld, J. (2007). Lightning and convection parameterisations—uncertainties in global modelling. *Atmospheric Chemistry and Physics*, 7(17), 4553–4568.
- US EPA. (2023). What is ozone? [Accessed: September 2025].
- US EPA. (2025). Ground-level ozone pollution [Accessed: September 2025].
- van Peet, J. (2018). Averaging kernel conversion [Version 1.0, 08 March 2018. Technical document from the AURORA project. Not publicly available].
- Wallace, J. M., & Hobbs, P. V. (2006). *Atmospheric science: An introductory survey* (Vol. 92). Elsevier.

-
- Wang, M., & Fu, Q. (2021). Stratosphere-troposphere exchange of air masses and ozone concentrations based on reanalyses and observations. *Journal of Geophysical Research: Atmospheres*, 126(18), e2021JD035159.
- Wang, T., Wong, H., Tang, J., Ding, A., Wu, W., & Zhang, X. (2006). On the origin of surface ozone and reactive nitrogen observed at a remote mountain site in the northeastern qinghai-tibetan plateau, western china. *Journal of geophysical research: atmospheres*, 111(D8).
- Wernli, H., & Bourqui, M. (2002). A lagrangian “1-year climatology” of (deep) cross-tropopause exchange in the extratropical northern hemisphere. *Journal of Geophysical Research: Atmospheres*, 107(D2), ACL–13.
- Wernli, H., & Sprenger, M. (2007). Identification and era-15 climatology of potential vorticity streamers and cutoffs near the extratropical tropopause. *Journal of the atmospheric sciences*, 64(5), 1569–1586.
- Wild, O. (2007). Modelling the global tropospheric ozone budget: Exploring the variability in current models. *Atmospheric Chemistry and Physics*, 7(10), 2643–2660.
- Williams, R. S., Hegglin, M. I., Jöckel, P., Garny, H., & Shine, K. P. (2024). Air quality and radiative impacts of downward-propagating sudden stratospheric warmings (ssws). *Atmospheric Chemistry and Physics*, 24(2), 1389–1413.
- Williams, R. S., Hegglin, M. I., Kerridge, B. J., Jöckel, P., Latter, B. G., & Plummer, D. A. (2019). Characterising the seasonal and geographical variability in tropospheric ozone, stratospheric influence and recent changes. *Atmospheric Chemistry and Physics*, 19(6), 3589–3620.
- WMO, W. M. O. (2018). *Scientific assessment of ozone depletion: 2018*.
- Wu, X., Liu, X., & Zhou, Y. (2021). Review of unsupervised learning techniques. *Proceedings of 2021 Chinese Intelligent Systems Conference: Volume II*, 576–590.
- Xiao, E. (2024). Comprehensive k-means clustering. *Journal of Computer and Communications*, 12(3), 146–159.
- Yin, X., Rupakheti, D., Zhang, G., Luo, J., Kang, S., de Foy, B., Yang, J., Ji, Z., Cong, Z., Rupakheti, M., Li, P., Hu, Y., & Zhang, Q. (2023). Surface ozone over the tibetan plateau controlled by stratospheric intrusion. *Atmospheric Chemistry and Physics*, 23(17), 10137–10143. <https://doi.org/10.5194/acp-23-10137-2023>
- Zang, Z., Liu, J., Tarasick, D., Moeini, O., Bian, J., Zhang, J., Thompson, A. M., Van Malderen, R., Smit, H. G., Stauffer, R. M., et al. (2024). The improved trajectory-mapped ozonesonde dataset for the stratosphere and troposphere (tost): Update, validation and applications. *Atmospheric Chemistry and Physics*, 24(24), 13889–13912.
- Zhang, J., Wei, Y., & Fang, Z. (2019). Ozone pollution: A major health hazard worldwide. *Frontiers in immunology*, 10, 2518.
- Zhang, Y., Huang, Q., Guo, K., Wang, M., Liao, H., Chou, Y., & He, X. (2024). Tropopause folds over the tibetan plateau and their impact on water vapor in the upper troposphere-lower stratosphere. *Climate Dynamics*, 62(2), 1423–1437.
- Zhu, T., Lin, W., Song, Y., Cai, X., Zou, H., Kang, L., Zhou, L., & Akimoto, H. (2006). Downward transport of ozone-rich air near mt. everest. *Geophysical Research Letters*, 33(23).
- Zoppetti, N., Ceccherini, S., Carli, B., Del Bianco, S., Gai, M., Tirelli, C., Barbara, F., Dragani, R., Arola, A., Kujanpää, J., et al. (2021). Application of the complete data fusion algorithm to the ozone profiles measured by geostationary and low-

earth-orbit satellites: A feasibility study. *Atmospheric Measurement Techniques*, 14(3), 2041–2053.

Acknowledgements

At this point, I'd like to take time to thank all the people I worked profitably together: Dr. Erika Brattich from UNIBO, Dr. Ugo Cortesi, Dr. Piera Raspollini, Dr. Simone Ceccherini, Dr. Nicola Zoppetti, and Dr. Cecilia Tirelli, as well as all the group "Sondaggio dell'Atmosfera" of the "Istituto di Fisica Applicata Nello Carrara" in Florence, and young CNR people I met during my PhD: Stefano, Cristina, Chiara, Francesco, Simone, Antonio, and Elisa. I am particularly grateful to Nicola Zoppetti for his valuable contribution, and for believing in this work, always supporting me through the process. To all, thank you for giving me the opportunity to broaden my knowledge by working on a new, complex, and challenging topic. I am especially grateful for having learned to push beyond limits, approaching obstacles with enthusiasm and embracing the many dimensions that a research project can reveal.

A special thanks also goes to Prof. Michaela Imelda Hegglin from the Forschungszentrum Jülich GmbH, who hosted part of my work and with whom I had the pleasure of engaging in scientific exchange. I am also grateful to Mia and Vitor, who enriched my time in Germany both professionally and personally, as colleagues - Mia - and friends - both.

I am deeply grateful to my family, who have always guided and supported me like a lighthouse, helping me grow and standing by my choices—including throughout this journey. Thank you, Nonno Mario, because this path began with you, even though it could not end the same way. And finally, thank you, Luca, for being part of My family, for never letting me lose heart, and for making everything feel lighter when shared with you.