11-23-2018 2:30 PM

Improved techniques for atmospheric ozone retrievals from lidar measurements using the Optimal Estimation Method and Machine Learning

Ghazal Farhani
The University of Western Ontario

Supervisor
Dr. Robert Sica
The University of Western Ontario

Graduate Program in Physics
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Recommended Citation

Farhani, Ghazal, "Improved techniques for atmospheric ozone retrievals from lidar measurements using the Optimal Estimation Method and Machine Learning" (2018). *Electronic Thesis and Dissertation Repository*. 6020.
https://ir.lib.uwo.ca/etd/6020

Abstract

A new first-principle Optimal Estimation Method (OEM) to retrieve ozone number density profiles in both the troposphere and stratosphere using Differential Absorption Lidar (DIAL) measurements obtained at the Observatoire de Haute Provence (OHP) in France is described. The method is robust and applicable to any DIAL ozone lidar. The ozone retrievals are compared to ozonesonde measurements, and these comparisons show the profiles match within the measurement uncertainties. The OEM retrieval also successfully catches much of the structure seen by the ozonesondes. The OEM retrievals are compared with the traditional analysis, and for most heights the difference between the two methods is small. One main advantage of the OEM is that all available measurements from multiple channels as well as lidars are used in the retrieval, eliminating the need to merge or perform corrections on the raw measurement. Thus, the tropospheric and stratospheric lidar measurements can be used together to generate an ozone profile which extends from 2.5 km to about 42 km. The upper troposphere and the lower stratosphere (UTLS) coincides with the measurements overlapping region. In the UTLS, even small changes in the distribution of the greenhouse gases can result in large changes in the atmospheric radiative forcing. The OEM can significantly improve the our understanding of the UTLS by providing an ozone density profile with a well-defined statistical and systematic uncertainty budget in this region.

A new state-of-the-art machine learning technique was developed to automatically classify raw (level 0) lidar measurements to remove bad scans, and to distinguish between clear sky measurements and measurements with traces of either clouds or aerosols. We have examined different supervised learning methods and found the random forest classifier, the support vector machine (SVM), and the gradient boosting trees could successfully classify our lidar data with more than 90% accuracy score with the random forest classifier recommended because of its greater computational speed.

Keywords: Optimal Estimation Method, DIAL, Ozone Retrievals, UTLS ozone, Machine Learning, random forest classifier, support vector machine (SVM), supervised learning, stratosphere, troposphere, lidar

Co-Authorship Statement

The entire thesis is written under the supervision of Dr. Robert Sica.

The work presented in Chapter 2 was done in collaboration with Dr. Sophie Godin-Beekmann and Dr. Alexander Haefele. Dr. Godin-Beekmann provided me with the OHP stratospheric measurements. Also, she performed the traditional analysis of the ozone density profiles. Dr. Alexander Haefele helped me to understand the Optimal Estimation Method (OEM). I wrote all the necessary MATLAB codes for the OEM retrievals. I also conducted the OEM analysis and provided comparisons between the methods.

The work done in Chapter 3 is a result of collaboration with Dr. Godin Beekmann and Dr. Gèrard Ancellet who provided me with the stratospheric and tropospheric measurements as well as the traditional analysis. I was responsible for providing the MATLAB codes for the OEM retrievals, and I performed the OEM analysis.

The work done in Chapter 4 is in collaboration with Dr. Alexander Haefele who provided the raw lidar measurements, and Dakota Cecil who helped me with the data preparation. Dr. Daley also helped me to understand the Machine Learning better.

Acknowlegements

I want to thank my supervisor Dr. Robert Sica. He trusted me and gave me all the freedom to do my research and to explore the different fields of physics. He was always patient with me, and he ever listened to my ideas and helped me to implement them in my work. I would also like to thanks Dr. Sophie Godin-Beekmann who helped me a lot with my research. For two summers I was lucky to go to France and work with her. I also thank Patricia Sica, she kindly edited my works, and we had terrific chats with each other. I will never forget our adventure in the Swiss Alps.

I am grateful toward Dr. Alexey Tikhomirov, Dr. Emily McCullough, Dr. Pierre Fogal, Peter McGovern and all the staff at the Environment Canada in Eureka, Nunavut. I honestly have an amazing and unique memory of Eureka because of them!

I was lucky to have the best colleagues in this department. Shayamila Mahagammulla Gamage was a fantastic colleague and a reliable friend during all these years. Robin Wing is also an incredible coworker and friend who made my stay in Paris so memorable. The staff at the Physics department are always helpful and supportive. I am genuinely appreciative toward Jodi Guthrie. From the very first day that I got to Canada, she helped me and several times she listened to me. I thank my friends Kendra, Amanda, and Moh who made my Ph.D. studies easier and made my life fabulous! Behafarid was my first roommate in London, and she became the best friend of mine. I always had her support, and I am so grateful for that.

I want to thank Sina who is always there when I need him, and who listens to me for hours. My bird Coco is not a home pet to me; he is the most fantastic friend of mine.

Finally, I would like to thank my parents and my sister. My parents are the reason that I started my Ph.D. During all these years, they always unconditionally supported me and stayed by my side. My sister is the one who always has my back, and I know no matter what I do, I still can go to her at any time!

Contents

Abstract **i**

Acknowlegements **iii**

List of Figures **vii**

List of Tables **xvi**

1 Introduction **1**

1.1 Overview 1

1.2 Atmospheric Structure 2

1.3 Ozone in the Stratosphere 3

1.3.1 Odd Nitrogen Catalytic Cycles 5

1.3.2 Odd Chlorine Catalytic Cycles 5

1.3.3 The hydroxyl radical Catalytic Cycles 6

1.4 Stratospheric Ozone Depletion 6

1.4.1 Ozone Trends 9

1.5 Ozone in the Troposphere 9

1.6 Measurements of Atmospheric Ozone 11

1.7 DIAL Measurements and Retrieval of Ozone Density 14

1.7.1 Ozone Density Retrievals 17

1.8 Corrections Applied to the Raw Counts 19

1.8.1 DeadTime Correction 19

1.8.2 Signal Induced Noise 19

1.8.3 Merging Process 20

1.8.4 Digital Filters 21
1.9 Optimal Estimation Method 21
1.10 Machine Learning Applications to Lidar Measurements 25
1.10.1 Machine Learning Methods 26

List of Appendices **1**

2 Implementing the OEM to Retrieve Stratospheric Ozone Density **34**
2.1 Overview 34
2.2 Ozone Retrievals: Traditional versus the OEM 35
2.2.1 Applying the optimal estimation method to ozone retrievals 36
2.2.2 Ozone DIAL Forward Model 37
2.3 Implementing the optimal estimation method retrieval 39
2.4 Application of the OEM to measurements from the OHP stratospheric ozone lidar 41
2.4.1 Applying the OEM to OHP measurements on 26 July 2017 43
2.4.2 Further examples of the OEM retrieval method 51
2.5 Conclusion 52

3 Improved ozone UTLS DIAL measurements **58**
3.1 Introduction 58
3.2 METHODOLOGY 61
3.2.1 Implementing the OEM for the OHP lidars 63
3.2.2 Description of the Lidars 67
3.3 OEM ozone retrieval in the free troposphere and stratosphere 68
3.4 Tropospheric Ozone Lidar Retrievals 76
3.5 Summary 79
3.6 Conclusions 80
3.6.1 Dataset Citation 81

4 Machine Learning Methods in Lidar Measurements Classification **86**
4.1 Introduction 86

4.2 Instrumentation 87

4.3 A Brief Introduction to Machine Learning 89

4.3.1 Supervised Approach 89

4.3.2 Unsupervised Approach 91

4.4 Learning Algorithms 92

4.4.1 Support Vector Machine Algorithms 93

4.4.2 Decision Trees Algorithms 93

4.4.3 Random Forests 96

4.4.4 Gradient Boosting Tree Methods 97

4.4.5 The t-distributed Stochastic Neighbour Embedding Method 98

4.5 Result for supervised and unsupervised learning in PCL and RALMO system . 100

4.5.1 Unsupervised result 102

4.6 Summary 104

4.7 Conclusion 106

5 Conclusions and Future Work **109**

Curriculum Vitae **111**

List of Figures

1.1 The mean atmospheric temperature profile is plotted for the month of July at the Observatoire de Haute-Provence 44°N, 5.8°E. The data is retrieved from the Mass Spectrometer Incoherent Scatter Radar (MSIS) database (Hedin, 1991). 4

1.2 In the DIAL technique two wavelengths (the “on-line” and the “off-line” wavelengths) are simultaneously transmitted to the atmosphere. The back scattered signals are collected by a large mirror or by multiple small mirrors. The collected signal (via optical fibers) are sent to the PMTs. 15

1.3 The flowchart is the summery of the steps we take to calculate the retrieval and its associated uncertainties. 25

2.1 To implement the OEM, the *a priori* profiles for ozone and air density, background counts, dead time values, and lidar constants are needed. Moreover, **b** parameters should be identified and proper values for them should be calculated. The covariance matrices for *a priori* profiles, measurements, and **b** parameters need to be calculated as well. 39

2.2 Average count rates for 5 hours of measurements on 26 July 2017. Left panel:“on-line” wavelength (blue curve, low altitude; red curve, high altitude). Right panel: “off-line” wavelength (blue curve, low altitude ; red curve, high altitude). 44

2.3 Averaging kernels for the ozone density for the measurements on 26 July 2017. The horizontal dashed line is a height below which the OEM retrievals is more than 80% due to the measurements. Above this horizontal cut-off as the SNR drops, the retrieval starts to fall back to the *a priori* profile. For clarity, the averaging kernels are only shown every 1500 m in altitude. The red line shows the summation of rows in the averaging kernel matrix at each altitude. The summation is of order unity below 42.7 km. 45

2.4 Residuals between the forward model and the measurements for the "on-line" and "off-line" channel (blue curves). The red line shows the uncertainty of the measurements. 46

2.5 Right panel: The statistical uncertainty of the OEM with correlation lengths (ℓ_s = 1000 and 1400 m is plotted (red curve) against the statistical uncertainty of the OEM with correlation lengths ℓ_s = 1400 m and 5500 m (black dotted curve). Additionally, the uncertainty of retrieval in the traditional method (blue curve) is plotted. The retrieval uncertainties in the OEM and the traditional method can be compared. Left panel: The vertical resolution of the OEM with correlation lengths (ℓ_s = 1000 and 1400 m (red curve) is plotted against the vertical resolution of the OEM with correlation lengths ℓ_s = 1400 and 5500 m (black dotted curve). The vertical resolution of the traditional method is shown as well (blue curve). The horizontal dashed line indicated the maximum height at which the retrieval is independent from the *a priori*. 47

2.6 At height from 20 km to 40 km, the uncertainty of retrieval for the traditional method (assuming that it has a vertical resolution similar to the OEM vertical resolution) is plotted against the OEM retrieval uncertainty (blue curve: OEM; red curve: traditional). The horizontal dashed line indicated the maximum height at which the OEM retrieval is independent from the *a priori*. 48

2.7 OEM ozone retrieval (red curve) from 20:07 UT to 00:15 UT on 26 July 2017 as well as the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve) are plotted. The dashed black line shows the OEM retrieval when the correlation length ($ł_S$) became larger. The horizontal dashed line shows the cut-off below which the effect of the *a priori* ozone profile is small less tan 10%. 49

2.8 For the night of 26 July 2017. Left panel: The percentage difference between the OEM retrieval and the ozonesonde measurements in the form of: $(\frac{OEM-sonde}{sonde} * 100)$ (blue curve); the percentage difference between the traditional retrieval and the ozonesonde measurements in the form of $(\frac{traditional-sonde}{sonde} * 100)$ (red curve). Right panel: The percentage difference between the OEM retrieval and the traditional retrieval (blue curve); the summation of the statistical uncertainty of the traditional and OEM retrievals (red curve). 50

2.9 For the night of 26 July 2017. The left panel: the OEM retrieval using the US standard model as *a priori* profile (purple curve) and the OEM retrieval using the OHP climatology as *a priori* profile (red curve) are plotted. Furthermore, the traditional method retrieval (blue curve) is plotted, thus the OEM retrievals can be compared with each other and with the traditional retrieval. The right panel: Percentage difference between the OEM retrievals using the two different *a priori* profiles (blue curve) is plotted. This difference is with in the retrieval uncertainty. At higher altitudes (above 35 km), when the SNR drops, the difference between the two methods is less than 5%, which is smaller than the retrieval uncertainty at that height. 51

2.10 For the night of 26 July 2017. The statistical uncertainty of the OEM (blue), the Rayleigh-scatter cross section uncertainty at 308 nm (red), the ozone absorption cross section at 308 nm (orange), and the ozone absorption cross section for the 355 nm channel (purple). The horizontal dashed line shows the height below which the retrieval is independent of the *a priori* profile. 52

2.11 Left panel: The retrieved air density (blue line) is plotted against the *a priori* profile (red line). Mid panel: The percentage difference between the scaled relative air density generated from the Raman channel and the OEM air density retrievals. The difference is less than 10%. Right panel: The statistical uncertainty of the OEM retrieval of air density (blue), the Rayleigh-scatter cross section uncertainty for the 308 nm channel (red), and the ozone absorption cross section in both channels (purple). 53

2.12 Left Panel: OEM ozone retrieval on the night of 14 July 2017 (red curve) compared to the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve). Right Panel: OEM ozone retrieval on the night of 20 July 2017 (red curve) compared to the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve). These cases demonstrate the high resolution of the OEM technique as evidenced by the excellent agreement around the ozone peak with the sonde measurement. 53

2.13 For the night of 14 July 2017, (a) the percentage difference between the traditional method and the OEM retrieval (blue curve) plotted within the envelope of the total statistical uncertainty of the two method (red curve). The agreement between the two lidar ozone determinations are within the statistical uncertainty above 17 km. (b): The red curve is the percentage difference between the OEM retrieval and sonde measurements. The blue curve is the percentage difference between the traditional method and sonde measurements. Figures (c) and (d) are the same format as (a) and (b) for the night of 20 July 2017. . . . 54

3.1 Averaging kernels for tropospheric-stratospheric ozone measurements. The averaging kernels are only shown every 450 m for lower altitude (from 2.5 km to 11 km height), and every 1500 m in higher altitudes for clarity. As is shown by the response function (red curve), the area of the averaging kernel matrix has a small spike at 11 km, when the stratospheric ozone measurements are included. 70

3.2 The percentage difference between the forward model and the actual measurements are shown in blue. The statistical uncertainty is plotted in red. The four plots at the left are stratospheric forward model residuals, and the four plots at the right are the tropospheric forward model residuals. For digital counting systems, the Poisson distribution is appropriate and the variance of the measurements at each altitude is the number of photons at that altitude. However, the output signals of the analog channels do not follow a Poisson distribution and to find the variance a 3-point running filter is used. As a result the red line (which indicates the noise of measurements) for the analog channels is more structured than for the digital channels. 70

3.3 The OEM retrieval (red curve) compared to the traditional calculation of ozone using the OHP lidar systems. The tropospheric lidar starts at 2.5 km and extends upward to 14.5 km (blue curve). The ozone profile measured by the stratospheric lidar system (black curve) overlaps with the profile retrieved from the tropospheric lidar system in the UTLS. In this region, the OEM retrievals smoothly transition from relying primarily on the tropospheric lidar measurements to the stratospheric measurements. 72

3.4 Left panel: The statistical uncertainty of the OEM retrieval (red curve) is plotted against the statistical uncertainty of the traditional retrievals. The uncertainty of the retrieval for the stratospheric and tropospheric lidar systems respectively are shown in black and blue. Right panel: The vertical resolution of the OEM retrieval is shown in red. The vertical resolution of the traditional calculation from the tropospheric lidar system is shown in blue, while the vertical resolution of the retrieved profile produced from the stratospheric lidar system is shown in black. 73

3.5 The percentage difference between the OEM retrieval and the ozonesonde measurements (blue curve) is plotted within the total statistical uncertainty of the OEM retrievals plus the ozonesonde measurement (red curves). 73

3.6 Uncertainty budget on the night of 12 July 2017. The statistical uncertainty of the retrieval (blue), the Rayleigh-scatter cross section uncertainty at 308 nm (dashed line red), the Rayleigh-scatter cross section uncertainty at 289 nm (dashed line yellow), the ozone absorption cross section at 308 nm (dashed line purple), the ozone absorption cross section for the 289 nm channel (dashed line green), and the overlap function for the 289 nm channel (dashed line light blue) all contribute to the budget. The horizontal dashed line shows the height below which the retrieval is independent of the *a priori* profile. 74

3.7 OEM ozone-profile retrieval on 14 July 2017 (red curve). The tropospheric traditional retrieval (blue curve) extends from 2.5 km to 15 km, while the stratospheric traditional retrieval (black curve) extends from 10 km to 43 km. At the region where the tropospheric and stratospheric lidar measurements overlap, the OEM can smoothly makes a transition from one lidar system's measurements to the other system's measurements. 75

3.8 OEM ozone-profile retrieval on 26 July 2017 (red curve). The tropospheric traditional retrieval (blue curve) extends from 2.5 km to 14 km, while the stratospheric traditional retrieval (black curve) extends from 12.5 km to 43 km. At the region where the tropospheric and stratospheric lidar measurements overlap, the OEM can smoothly makes a transition from one lidar system's measurements to the other system's measurements. 76

3.9 Averaging kernels for the ozone density for the measurements on 12 July 2017. The horizontal dashed line is the height cut-off above which the sensitivity of the retrieval to measurements is less than 90%. The averaging kernels are only shown every 450 m in altitude. The summation of rows in the averaging kernel matrix, for each specific height, is shown by the red curve. 77

3.10 Both the OEM retrieval (red curve) and the traditional retrieval (blue curve) extend from 2.5 km to 14.2 km. The ozonesonde profile is plotted in green. The black dashed line defines the cut-off altitude of the retrieval. 77

3.11 Left panel: The statistical uncertainty of the OEM retrieval (red curve) as well as the statistical uncertainty of the traditional retrieval (blue curve) for 12 July 2017. Right panel: Vertical resolution of the OEM retrieval is shown in red, while the vertical resolution of the traditional retrieval is shown in blue. The spike at 5.5 km in the OEM uncertainty is from the inclusion of the digital channels at this height. 78

3.12 Comparison of tropospheric OEM ozone DIAL retrievals for the nights of 14 July and 26 July 2017 (red curves), the traditional method (blue curves), and ozonesonde measurements (green curves). The horizontal dashed line is the altitude below which the OEM retrieval is mostly independent of the *a priori* ozone profile assumed. Left panel: 14 July 2017; right panel: 26 July 2017. . . 79

4.1 The training set is a matrix of size ($m * n$), in which each row represents one dataset (here a lidar scan), and each element in each column (except for the n-th column) represents an attribute (for lidar scans: photon counts at each altitude). The n-th column holds the label for each scan (here we have four different labels: 0, 1, 2, and 3). 90

4.2 Example of scans taken by PCL for Rayleigh and Raman channels. Panel (a) shows examples of bad scans. In this plot, the signals in blue and dark red have extremely low laser power, the purple signal has extremely high background counts, and the signal in orange has a distorted shape and high background counts. Panel (b) shows two examples of good signals: the signal in red has a high laser power compared to the blue curve. In the supervised approach, the red signal is labeled 3, and the blue signal is labeled 1. Panel (c) shows a scan in which clouds occur at lower altitudes. 91

4.3 Using a t-SNE unsupervised algorithm, the iris dataset is clustered into three different types. Setosa is shown in violet, versicolor is shown in blue and virginica is shown in yellow. 92

4.4 Example of linear SVM: the hyperplane separates sun and star logos. The subset of data points within the margin (support vector) is the most important data point as they play an important role in defining the optimum hyperplane. IN SVM a wider margin is preferable, as it divides two classes. This figure is adapted from: (Burges, 1998). 94

4.5 The schematics of a decision tree is shown. From the root (a node with no parents), a decision question is asked and answered in the form of yes (1) or no (0). The answers (yes or no) generate two children. Leaves are the nodes with no children. Ideally, each leaf is pure (all the points inside it has the same classification). 95

4.6 The schematic of a random forest: each tree casts a vote; based on majority vote, a data point is classified. 97

4.7 Boosting methods are based on the idea that combining many "weak" approximation models (a learning algorithm which is Using the t-SNE algorithm scans for the low-gain Rayleigh measurement channel on the night of May 15 2012 were clustered to four different groups. As shown in the figure, cluster number 2 has some outliers. 98

4.8 Red curve: the Gaussian distribution for data points (extending from -5 to 5 in x-axis). The peak of the distribution is at its mean in 0. Blue curve: The Student's t-distribution for the same data points. The distribution is heavy-tailed (comparing to the Gaussian distribution). 99

4.9 Using the t-SNE algorithm scans for the low-gain Rayleigh measurement channel on the night of May 15 2012 were clustered to four different groups. As shown in the figure, the cluster number 2 has some outliers. 103

4.10 Left Panel: All scans collected for the PCL system from the low-gain Rayleigh channel which were plotted for the night of May 15 2012. The green signals have extremely low power. The red line represents all signals with lower laser signal and the blue line indicates the signals which are good scans. The black lines are signals with extremely high backgrounds. Right Panel: Each line represents an average of the signals within a cluster. The black line indicates the outliers which have extremely high background counts and are outliers belonging to cluster 3 (blue curve). The red line is the average signal for scans with lower laser power. The green line is the average signal for cluster 2. The background counts in the green line start at lower latitude of 40 km, where as for the red line the background starts at 60 km and in the blue line the background starts at 85 km. 104

4.11 Figure. 4.11 (a): Using the t-SNE algorithm, scans for the Nitrogen channel on the night of May 15 2012, collected by the PCL system, were clustered into two different groups. Figure. 4.11 (b): The red line (cluster number 2) is the average of all signals within this cluster and indicates the scans in which clouds are detectable. The blue line (cluster number 2) is the average of all signals within this cluster and indicates the clear scans (non-cloudy condition). . 105

4.12 The result of applying the t-SNE algorithm to different channels at different nights is plotted. Each cluster contains lidar scans sharing similar features. . . . 105

List of Tables

2.1 Values and associated uncertainties for the retrieved and forward model parameters. 42

2.2 Dead time values which were calculated for each channel on the night of 26 July 2017. 51

3.1 Values and associated uncertainties for the retrieved and forward model parameters. 67

3.2 Measurement periods for the tropospheric and stratospheric lidars systems and launch times for the ozonesondes. 69

4.1 The success score for each of the algorithms for 2000 scans of PCL lidar. 101

4.2 Number of scans which used in each type. 101

4.3 The success score for each of the algorithms for day time scans of the RALMO lidar. 102

4.4 The success score for each of the algorithms for day time scans of the RALMO lidar. 102

4.5 Number of scans used in each type; for day time and night time we used the same number of scans. 102

Chapter 1

Introduction

1.1 Overview

Ozone is a minor constituent in the atmosphere that mostly resides in the stratosphere. Stratospheric ozone represents about 90% of the total column ozone, and about 10% of the ozone is concentrated in the troposphere. By absorbing solar ultraviolet (UV) radiation, stratospheric ozone protects the Earth's surface from receiving excessive radiation, and this heat causes the positive temperature gradient in the stratosphere (Andrews et al., 1987).

The significant decline of stratospheric ozone became a serious concern at the end of the last century. The discovery of an ozone hole in the Antarctic and the substantial reduction in the total column ozone (Farman et al., 1985; WMO, 1988) were clear examples of ozone depletion due to heterogeneous reactions involving ozone-depleting substances (ODSs). Since 1989, the implementation of the Montreal Protocol has successfully reduced the level of ODSs in both the troposphere and the stratosphere. The effect of the decline of ODSs on the recovery of the ozone layer has yet to be carefully observed and studied. Thus, continuing measurements of stratospheric ozone to characterize the rate of ozone recovery is required.

In contrast to stratospheric ozone, ozone molecules residing in the upper troposphere act as the third most abundant greenhouse gas contributing to the radiative forcing of climate change (Ramaswamy et al., 2001). Furthermore, ozone at the surface is a pollutant and has an adverse effect on air quality, human health, and the ecosystem. Continuous measurements of the stratospheric and the tropospheric ozone enable us to monitor and understand ozone changes and its trends in the atmosphere.

Various airborne and ground-based instruments have been employed to measure ozone density. This thesis focuses on Light Detecting and Ranging (LIDAR) system, which is a ground-based remote sensing instrument. Lidar's high vertical and temporal resolution make it a suitable system for ozone monitoring.

In Chapter 1, the importance of the ozone in the atmosphere is explained in detail. Section 1.2 is devoted to explaining the structure of the atmosphere. In Section 1.3, the role of the stratospheric ozone is discussed. In Section 1.4, a summary of ozone depletion processes and ozone changes before and after the implementation of the Montreal Protocol is presented. Section 1.5 discusses the role of the tropospheric ozone in the atmosphere, and Section 1.6 briefly describes the instruments that are typically used in ozone studies. Sections 1.7 and 1.8 describe the DIAL system and ozone retrievals in detail. In Section 1.9 we introduce the Optimal Estimation Method, and in Section 1.10 we briefly describe the importance of Machine Learning approach for lidar data classification.

The focus of Chapter 2, and Chapter 3 is to introduce and to implement the Optimal Estimation Method (OEM) which is used to retrieve tropospheric and stratospheric ozone profiles, using a Differential Absorption Lidar (DIAL).

Chapter 4 of this thesis focuses on implementing a state-of-the-art machine learning method for lidar data classification. In lidar data analysis, before applying any algorithm or any preprocessing (correcting photon counts), each lidar scan should be examined for quality purposes (when dealing with level-0 raw measurements). This quality check to remove background counts is mostly done manually, which is a time-consuming and inaccurate process. We propose using machine learning techniques for raw-data classification and automating this quality check.

1.2 Atmospheric Structure

The atmosphere is a layer of gas which surrounds the Earth. Based on the vertical distribution of temperature, the atmosphere is separated into four different layers (see Fig. 1.1). The lowest layer is called the troposphere.

The main characteristic of the troposphere is its negative lapse rate. The lapse rate is defined as $\Gamma = -\frac{dT}{dh}$ and is the negative of the rate of temperature change with respect to height (negative lapse rate means that the temperature decreases with height). The troposphere is statically unstable; as a result, the tropospheric gases are well mixed. Moreover, most of the weather takes place in this layer, and it contains most of the water vapor. The tropopause is the boundary line where change in the lapse rate occurs. The tropopause is defined as the lowest level at which the rate of lapse decrease is 2 °C/km or less (WMO, 1992). The tropopause has a height of about 8 km at the poles and 15 km at the equator (Harrison, 2007). The second layer is called the stratosphere which extends from the tropopause to roughly 50 km in altitude. In the stratosphere, where ozone density is at its peak, ozone molecules absorb UV radiation from the Sun. Thus, in this layer, the temperature increases with height. Compared to the troposphere, the stratosphere is generally stable, very dry, and ozone rich. The boundary that separates the stratosphere from the next layer is called the stratopause. The mesosphere is located above the stratopause where the temperature decreases again. The air in this region is extremely thin, and 99.9% of the atmospheric mass is found below the mesosphere. In the upper layer of the mesosphere, because of the passage of meteors, “dust” exists. Moreover, in this layer high concentration of sodium, potassium, and iron can be observed. The abundance of these elements is related to meteor showers. The mesosphere is bound by the mesopause, which has a height of approximately 85 km. The thermosphere is located above the mesopause. A significant temperature inversion, due to the absorption of energetic solar radiation by oxygen molecules, can be observed in this layer (Ahrens, 1998).

1.3 Ozone in the Stratosphere

The bond energy of the oxygen molecule is 498 kJ mol^{-1}, which corresponds to the energy of a photon with a wavelength of 240 nm, thus only photons with shorter wavelengths than 240 nm can photolysis oxygen molecules. As is shown in (R 1), the photolysis of an oxygen molecule produces two oxygen atoms in their ground-level triplet state O(^{3}P). The oxygen atoms are highly re-active, and they rapidly combine with O_2 to form ozone (R 2) molecules, where M can be any molecule. Ozone molecules, with a bond energy of 445 kJ mol^{-1}, weaker than

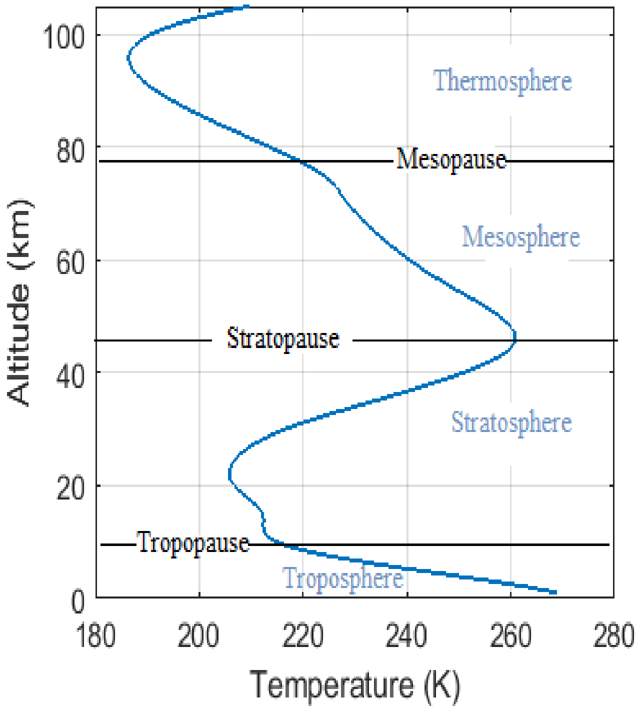

Figure 1.1: The mean atmospheric temperature profile is plotted for the month of July at the Observatoire de Haute-Provence 44°N, 5.8°E. The data is retrieved from the Mass Spectrometer Incoherent Scatter Radar (MSIS) database (Hedin, 1991).

O_2. Thus photons with lower energy (corresponding to wavelengths shorter than 270 nm) can photolyze ozone molecules (see R 3). The oxygen atoms $O(^1D)$ are in their excited singlet state, but by colliding with N_2 or O_2 molecules they will stabilize to $O(^3P)$. The rate of reaction for R 2 and R 3 is much faster than the rate of reaction for R 1 and R 4. Thus there is a rapid cycling between the atomic oxygen and ozone. It is useful to look at O and O_3 together as an odd oxygen family distinct from O_2 (even oxygen) which is a much longer-lived species. More detail about the Chapman mechanism can be found in Chapman (1930) and Brasseur and Solomon (2006).

$$O_2 + h\nu \longrightarrow O + O \tag{R 1}$$

$$O + O_2 + M \longrightarrow O_3 + M \tag{R 2}$$

$$O_3 + h\nu \longrightarrow O + O_2 \tag{R 3}$$

$$O_3 + O \longrightarrow O_2 + O_2 \tag{R 4}$$

.

The Chapman cycle is based on oxygen-only chemistry; however, the stratospheric ozone is also destroyed by hydrogen and nitrogen oxide chemistry (Bates and Nicolet, 1950; Crutzen, 1970; Johnston, 1992). Each of these species has their own odd families and can destroy odd oxygen in a catalytic process. Chlorine can also play an important role in ozone destruction as it can engage in catalytic cycles with odd oxygen (Stolarski and Cicerone, 1974).

1.3.1 Odd Nitrogen Catalytic Cycles

The nitrogen family has two cycles which destroy stratospheric ozone. The first cycle, which is dominant in the middle stratosphere (Brasseur and Solomon, 2006), involves both atomic oxygen and ozone. The second cycle does not require atomic oxygen and is more critical below 30 km. Both cycles are shown below:

$$\mathrm{NO + O_3 \longrightarrow NO_2 + O_2} \tag{R 5}$$

$$\mathrm{O + NO_2 \longrightarrow NO + O_2} \tag{R 6}$$

$$\text{Net Cycle 1: } \mathrm{O + O_3 \longrightarrow O_2.}$$

$$\mathrm{NO + O_3 \longrightarrow NO_2 + O_2} \tag{R 7}$$

$$\mathrm{NO_2 + O_3 \longrightarrow NO_3 + O_2} \tag{R 8}$$

$$\mathrm{NO_3} + \mathrm{h}\nu \longrightarrow \mathrm{NO + O_2} \tag{R 9}$$

$$\text{Net Cycle 2: } \mathrm{2\,O_3 \longrightarrow 3\,O_2.}$$

During the day, the oxidation of NO_x by OH leads to the formation of HNO_3. During the night, NO_3 and NO_2 can bond and form N_2O_5. Thus HNO_3 and N_2O_5 are reservoirs for NO_x. Details about all of these reactions can be found in Jacob et al. (1996) and Brasseur and Solomon (2006).

1.3.2 Odd Chlorine Catalytic Cycles

In the stratosphere, due to photolysis of organic chlorine species (e.g., CFCs, HCFCs and CCl_4) atomic chlorine (Cl) is produced. The free Cl reacts with chlorine monoxide radicals (ClO) as shown below:

$$\mathrm{Cl + O_3 \longrightarrow ClO + O_2} \tag{R 10}$$

$$\mathrm{ClO + O \longrightarrow Cl + O_2} \tag{R 11}$$

$$\text{Net Cycle: } \mathrm{O + O_3 \longrightarrow O_2.}$$

Atomic oxygen (which is needed for the second reaction of this cycle) is formed when UV radiation reacts with ozone and oxygen molecules. Thus, this cycle is more critical at mid-latitudes and the tropics where the UV radiation is more intense.

When Cl and ClO react with CH_4 and NO_2 they convert to HCl and $ClONO_2$ reservoirs, thus the above cycle stops:

$$Cl + CH_4 \longrightarrow HCl + CH_3 \qquad \text{(R 12)}$$

$$ClO + NO_2 + M \longrightarrow ClNO_3 + M\cdot \qquad \text{(R 13)}$$

More detail on odd chlorine catalytic cycles can be found elsewhere (Bates and Nicolet, 1950).

1.3.3 The hydroxyl radical Catalytic Cycles

Chlorine and bromine monoxides can interact hydroxyl HO_x and cause ozone destruction as follows:

$$O_3 + OH \longrightarrow HO_2 + O_2 \qquad \text{(R 14)}$$

$$XO + HO_2 + \longrightarrow HOX + O_2 \qquad \text{(R 15)}$$

$$HOX + h\nu \longrightarrow X + OH \qquad \text{(R 16)}$$

$$O_3 + X \longrightarrow XO + O_2 \qquad \text{(R 17)}$$

where X stands for either chlorine or bromine monoxides. In the lower most stratosphere, where oxygen atoms are rare, the hydroxyl chemistry dominates the gas phase loss for ozone molecules.

1.4 Stratospheric Ozone Depletion

Due to high levels of ozone-depleting substances (ODSs), chemical depletion of total ozone has been detected globally since the mid-1970s (WMO, 1999). In 1974, studies suggested that anthropogenic chlorofluorocarbons (CFCs) were major sources of stratospheric chlorine, and could play an important role in stratospheric ozone destruction (Molina and Rowland, 1974). Moreover, the gas phase reactions which were introduced earlier in this chapter could not explain the observed massive loss of the ozone.

Ozone depletion is most pronounced in Antarctica, where the stratosphere is characterized by the presence of a strong polar vortex from May to November. During the Polar Night, the stratosphere temperature drops and a low pressure system centered over the polar region, known as the polar vortex, develops. The air within the polar vortex has very low temperature, which causes a strong temperature gradient between the mid-latitudes and high-latitudes. The strong temperature gradient results in air movement from the equator to the poles, which is directed to the east by the Coriolis effect. This situation results in the formation of large horizontal pressure gradients and high jet winds at the edge of the vortex, known as the polar night jet. These winds can isolate the air inside the vortex from the warmer and ozone rich air masses at lower latitudes.

Low polar temperature conditions are key to severe ozone depletion. At low temperatures, Polar Stratospheric Clouds (PSCs) in the form of liquid or solid are formed. In polar regions, different types of liquid and solid PSCs are formed when the stratospheric temperature drops below $-78°C$. PSCs are formed between 12 km to 25 km in altitude and over large areas. Nitric acid, sulfuric acid and droplets of supercooled water (NAT) compose type IPSCs. As the temperature in the stratosphere drops below the frost point of supercooled water droplets ($-88°C$) the NATs crystallize into type II PSCs.

In Spring during Polar Sunrise ODSs chlorine compounds are activated towards species that are more detrimental for the ozone in the presence of solar radiation. PSCs play a significant role in this process, as some reservoirs such as HCl and $ClONO_2$, can react on the surface of PSCs, releasing active chlorine atoms. As a result, most of the stratospheric ozone at altitudes between 15 km to 25 km disappears (Farman et al., 1985; WMO, 2011, 2014). This phenomenon is known as the Antarctic ozone hole.

Here, one of the surface reactions is shown:

$$HCl + ClNO_3 \longrightarrow Cl_2 + HNO_3 \tag{R 18}$$

$$Cl_2 + h\nu \longrightarrow 2Cl\cdot \tag{R 19}$$

Several studies suggested that inside the polar vortex high concentration of ClO has a major effect on the ozone's destruction (Jacob et al., 1996). Due to extremely low temperatures, high amounts of ClO are sustained inside the polar vortex. During the spring because of the sun light the ClO cycle can occur:

$$Cl + O_3 \longrightarrow ClO + O_2 \tag{R 20}$$
$$Cl + O_3 \longrightarrow ClO + O_2 \tag{R 21}$$
$$ClO + ClO + M \longrightarrow Cl_2O_2 + M \tag{R 22}$$
$$Cl_2O_2 + h\nu \longrightarrow Cl + ClO_2 \tag{R 23}$$
$$ClO_2 + M \longrightarrow Cl + O_2 + M \tag{R 24}$$
$$\text{Net Cycle: } 2\,O_3 \longrightarrow 3\,O_2.$$

Furthermore, McElroy et al. (1986) and Tung et al. (1986) showed that chlorine and bromine reactions can also destroy stratospheric ozone:

$$Cl + O_3 \longrightarrow ClO + O_2 \tag{R 25}$$
$$Br + O_3 \longrightarrow BrO + O_2 \tag{R 26}$$
$$BrO + ClO \longrightarrow Br + ClO_2 \tag{R 27}$$
$$ClO_2 + M \longrightarrow Cl + O_2 + M \tag{R 28}$$
$$\text{Net Cycle: } 2\,O_3 \longrightarrow 3\,O_2.$$

As a result of these reactions, the abundance of Cl gases will significantly increase. Moreover, PSCs are the main reason for stratospheric denitrification. Below, we briefly discuss the denitrification process. Most types of PSCs form from nitric acid (HNO_3) and water (which is condensed on liquid sulfuric- acid-containing particles). As PSCs contain large particles, due to gravity, they can descend several kilometers. Thus, in the process, large amounts of HNO_3 are removed from the stratosphere which is known as the denitrification of the stratosphere. In the stratosphere, nitric acid is the source of NO_x which can convert the highly reactive Cl atoms to the reservoir molecules ($ClONO_2$). Thus, by denitrification of the stratosphere Cl remains and destroys ozone molecules.

Due to stronger planetary wave activity in the Northern Hemisphere, the polar vortex in the Arctic is less symmetric and can be very unstable; thus ozone loss in the Arctic is generally less severe. Overall stratospheric temperature in the Arctic is also higher, and there are fewer PSCs in the Arctic. Moreover, before depletion starts, as a result of stronger transport of the ozone from the tropics to the Northern Hemisphere, the abundance of ozone molecules in the Arctic is more than in the Antarctic. However, in 2011, substantial loss of the stratospheric ozone in the Northern Hemisphere was reported in the Arctic by Manney et al. (2011).

1.4.1 Ozone Trends

In the early 1990s, the total amount of global ozone was about 5% less than the average amount of ozone from 1964 through 1980 (WMO, 2014). Under the Montreal Protocol and its subsequent amendments, the emission and thus abundance of anthropogenic ODSs in the troposphere has decreased from its peak in 1994 by approximately 10% (WMO, 2014). As a result, the amount of ozone during the early 2010s lessened to 3% less than the average amount of ozone in the 1964-1985 period. Recently, the first signs of total ozone recovery over Antarctica was observed (Solomon et al., 2016). However, for non-polar regions, since 2000, no significant positive trend is detected (WMO, 2014).

Although the trends in the total column ozone are insignificant, in the upper stratosphere (around 42 km) the ozone level has significantly increased (Harris et al., 2015). This increase does not indicate that ozone in the whole stratosphere is increasing. In contrast, many studies have suggested that, at mid-latitudes and tropical latitudes, the ozone content, at the lower stratosphere has continued to decrease (Ball et al., 2018).

Trends in the ozone are on the order of few percent, for example, trends in the upper stratosphere are around 1% to 3% per decade (Harris et al., 2015). Thus, it is crucial to take ozone measurements with an instrument with high spatial and temporal resolution to detect these changes.

1.5 Ozone in the Troposphere

A small amount of ozone resides in the troposphere, where it is a greenhouse gas contributing to climate change. Moreover, tropospheric ozone near ground-level is an air pollutant damaging human health and threatening ecosystem health. The tropospheric ozone budget depends on both photochemical and physical processes. Because of the high concentration of ozone in the stratosphere, it was once assumed that ozone transportation from the stratosphere to the troposphere is the dominant source of tropospheric ozone (Junge, 1962; Danielsen, 1968). However, later studies showed that the tropospheric ozone is mostly produced from the photochemical oxidation of CO and hydrocarbons (catalyzed by HO_x and NO_x). Here, we briefly describe the two processes.

- Stratosphere-Troposphere Exchange (STE)

  In the mid-latitudes and the polar regions, air masses move along lines which have a constant potential temperature (isentropic lines), as a result, in the UTLS region, air masses from the ozone-rich stratosphere can irreversibly move down towards the upper troposphere. This process is adiabatic; thus it does not require heat. In the STE process, chemical constituents such as ozone molecules are depleted from the stratosphere (where their abundance is necessary), and their concentration increases in the troposphere (where they are greenhouse gases) (Holton et al., 1995). The rate of this exchange is between 770 ± 400 Tg/year (IPCC2007). Depending on the latitude, tropopause height, and season, the contribution of STE to tropospheric ozone concentration can vary. During spring, at high latitudes and 500 hPa, the stratospheric contribution to the tropospheric ozone is about 40%, this value drops to 25% during fall. At mid-latitudes, during spring, the contribution is between 35% and 40%, and during fall it drops to 10-15% (Cohen et al., 2018).

- Photochemistry

  Photochemistry involving CO, CH_4, and Volatile Organic Compounds (VOCs), in the presence of nitric oxides, is another source of tropospheric ozone. These molecules are known as ozone precursors. The main driver of tropospheric ozone concentration is NO_x. The primary source of NO_x in the troposphere is fossil fuel combustion (Finlayson-Pitts et al., 1999). Natural sources, including soil emissions and lightning, can contribute to NO_x formation as well (Sauvage et al., 2007). However, less than one-third of NO_x is produced from natural sources. The net flux of tropospheric ozone due to photochemical activities is 3420 ± 770 Tg/year (IPCC2007). Here, the reaction involving CO, CH_4, and Volatile Organic Compounds (VOCs) are discussed briefly. The contribution of CO to ozone production is given by the following reactions:

$$CO + OH \xrightarrow{O_2} CO_2 + HO_2 \tag{R 29}$$

$$HO_2 + NO \longrightarrow OH + NO_2 \tag{R 30}$$

$$NO_2 + h\nu \xrightarrow{O_2} NO + O_3 \tag{R 31}$$

$$\text{Net Cycle: } CO + 2\,O_2 \longrightarrow CO_2 + O_3.$$

The reactions involving CH_4 are shown below:

$$CH_4 + OH \longrightarrow CH_3 + H_2O \tag{R 32}$$

$$CH_3 + O_2 + M \longrightarrow CH_3O_2 + M \tag{R 33}$$

$$CH_3O_2 + NO \longrightarrow CH_3O + NO_2 \tag{R 34}$$

$$CH_3O + O_2 \longrightarrow CH_2O + HO_2 \tag{R 35}$$

$$CH_2O + h\nu \xrightarrow{O_2} CHO + O_2 \tag{R 36}$$

$$CHO + O_2 \longrightarrow CO + HO_2 \tag{R 37}$$

$$CO + OH \xrightarrow{O_2} CO_2 + HO_2 \tag{R 38}$$

$$4\,(HO_2 + NO \longrightarrow OH + NO_2) \tag{R 39}$$

$$5\,(NO_2 + h\nu \longrightarrow NO + O_3) \tag{R 40}$$

$$\text{Net Cycle: } CH_4 + 10\,O_2 \longrightarrow CO_2 + H_2O + 5\,O_3 + 2\,OH.$$

Finally, the VOCs contribution is as follows:

$$VOC_1 + 4O_2 + 2h\nu \longrightarrow VOC_2 + H_2O + 2O_3\cdot \tag{R 41}$$

Reactions involving HO_x are the main source for tropospheric ozone destruction:

$$HO_2 + O_3 \longrightarrow OH + 2O_2 \tag{R 42}$$

$$OH + O_3 \longrightarrow HO_2 + O_2\cdot \tag{R 43}$$

The balance between the destruction and construction of the tropospheric ozone is determined by the abundance of the precursors and HO_x molecules. A detailed discussion on the tropospheric ozone distribution and trends can be in Gebhardt et al. (2014).

1.6 Measurements of Atmospheric Ozone

Different instruments have been used to measure ozone concentration in both the troposphere and the stratosphere, each with various advantages and disadvantages. In the following section, a brief introduction to ozone measurements will be given.

Both remote sensing and *in-situ* techniques have been used to measure the concentration of ozone in the atmosphere. For *in-situ* measurements, a sample of air is taken from the atmosphere and is analyzed to determine the ozone content. In remote sensing measurements, passive and active techniques are used and the atmospheric parameter of interest is not measured directly. Instead, the radiation that is emitted, absorbed, or reflected by the atmospheric quantity of interest is measured. The remote sensing measurements can be either passive or active. Active remote sensing involves transmitting a source of electromagnetic radiation and receiving the backscattered signal. Passive remote sensing is similar to the active method; however, a natural electromagnetic radiation source (such as the Sun or the moon) is used. Ozonesondes are good examples of an *in-situ* measurement technique. Ozonesondes are attached to large weather balloons and measure the vertical ozone profile from the surface up to an altitude of approximately 35 km. These measurements have very high vertical resolutions (sample resolution of 30 m). Furthermore, They can function under severe weather conditions and in all climate regions. Ozonesondes are providing high-quality data; however, they cannot reach altitudes higher than 35 km. Usually, ozonesondes data are archived either under the World Ozone and Ultraviolet radiation Data Center (WOUDC) network or the Southern Hemisphere ADditional OZonesondes (SHADOZ) network. The Electrochemical Concentration Cell is the most common type of ozonesonde and has been widely used at different locations (Kley et al., 1996; Schulz et al., 2001; Vömel and Diaz, 2010). A chemical reaction between potassium iodide (KI) and ozone (O_3) produces iodine molecules (I_2). The concentration of I_2 molecules is proportional to the ozone concentration. This chemical reaction between potassium iodide and ozone is the basis of ozonesonde measurements and is shown below:

$$2\,\mathrm{KI} + \mathrm{O_3} + \mathrm{H_2O} \longrightarrow \mathrm{I_2} + \mathrm{O_2} + 2\,\mathrm{KOH} \qquad \text{(R 44)}$$

The iodine molecules will generate a current within the ozonesode cell. The current is directly proportional to the partial pressure of ozone in the sampled air:

$$P_{O_3} = cTt_{100}\gamma(I - I_b) \qquad (1.1)$$

where P_{O3} is the ozone partial pressure, $c = 4.309 * 10^{-4}$, T is the temperature of the sampled air, t_{100} is the time which is needed to pump 100 *ml* of air to the cell, γ is the efficiency of the pump, I is the produced current and I_b is the background current produced when there is no ozone. A detailed discussion on the topic can be found in Komhyr (1986) and Johnson et al. (2002).

The Dobson spectrometer is an example of a ground-based remote sensing instrument which is used for ozone measurements. As the Dobson spectrometer is a passive remote sensing instrument, it uses natural sources of light (like direct sunlight or diffuse light from clear or cloudy skies) for its ozone measurements. The ozone absorbs light at selected bands of the electromagnetic spectrum; thus the amount of light which is transmitted to the ground depends on the abundance of the amount of ozone along the line-of-sight, which is converted into the overhead ozone column. In this method, the ratio of sunlight intensity at two wavelengths is measured. The ozone weakly absorbs one of these wavelengths whereas the ozone mostly absorbs the other one. Thus the method is based on the differential absorption method. The instrument measures the total column ozone which is the total amount of ozone inside a vertical column extending from the ground. The Brewer spectrometer is similar to the Dobson spectrometer; however, it uses five or six wavebands. The Dobson instrument employs a selection of eight different wavelengths at UV band (from 305.5 nm to 339.8 nm). Using the Umkehr principle, the vertical ozone concentration at ten different altitude layers is retrieved. The Umkehr layers are approximately 5 km thick. Although the standard Umkehr method retrieves at ten layers, the retrieval only contains four independent pieces of information (Mateer, 1965; Mateer and DeLuisi, 1992). Thus, using the Dobson instrument, long-term stratospheric ozone (from 20 km to 40 km) measurements are produced. However, these profiles have coarse vertical resolutions (between 5 km to 10 km).

Satellites can carry passive remote sensing instruments with the ability to provide global coverage for ozone measurements. Based on different geometric viewing concept, the electromagnetic radiation which is reflected or emitted from the Earth's atmosphere can be measured in nadir, limb, or occulation modes. Nadir measurements can provide good horizontal coverage; however, their vertical resolution is poor. Limb sounders can measure vertical profiles with each measurement representing a relatively narrow layer of atmosphere, thus compare to nadir scans, the vertical resolution of limb measurements is higher. However, due to clouds, aerosols, and humidity in the troposphere, their measurement sensitivity at lower altitudes is poor. Furthermore, using solar, lunar, or stellar occultation techniques, high signal-to-noise ratio (SNR) measurements can be made. The advantage of the latter method is that it can measure troposphere ozone as well. However, solar occultation measurements can only be performed as the sun rises or sets (relative to their orbits). More detail on satellite measurements can be found elsewhere (e.g. Emery and Camps (2017)).

Lidar (LIght Detection And Ranging) is a ground-based active remote sensing technique which is similar to radar but operates using lasers. In lidar measurements, a laser beam is sent into the atmosphere, and is scatters in all directions. A portion of the light is backscattered toward the lidar. A telescope collects the backscattered photons and is detected by photomultiplier tubes (PMTs). The received signal contains information about the atmosphere (Weitkamp, 2006). In this thesis, we employed a Differential Absorption Lidar (DIAL) system to retrieve ozone density profiles. Details about the DIAL system and the ozone retrievals are presented in the next section.

1.7 DIAL Measurements and Retrieval of Ozone Density

A DIAL system is based on transmitting two wavelengths simultaneously to the atmosphere. One of the emitted wavelengths is strongly absorbed by the constituent of interest (called the "on-line" wavelength) and the other is weakly absorbed (called the "off-line" wavelength) and used as the reference wavelength. A schematic diagram of a DIAL system is shown in Fig 1.2. Here, we briefly describe how vertical ozone density profiles can be retrieved using a DIAL system. Detailed description on the topic can be found in Schotland (1974); Godin-Beekmann et al. (2003); Godin et al. (1999).

In ozone studies, selecting a wavelength pair depends on the altitude range of measurements, and wavelengths in the UV spectrum are the most efficient. A pair of wavelengths with a strong UV absorption is needed to detect the small amount of ozone which resides in the troposphere. However, for stratospheric ozone measurements, choosing a laser that can reach to higher altitudes in the stratosphere is the main concern (Megie et al., 1985; Browell, 1989; Papayannis et al., 1990). The general retrieval method for both tropospheric and stratospheric DIAL measurements is the same, and is based on the lidar equation in which the measured backscattered photocounts, $N_{obs}(z,\lambda)$, for a laser pulse at wavelength λ and at altitude z can be written as (Schotland, 1974; Fernald, 1984; Weitkamp, 2006):

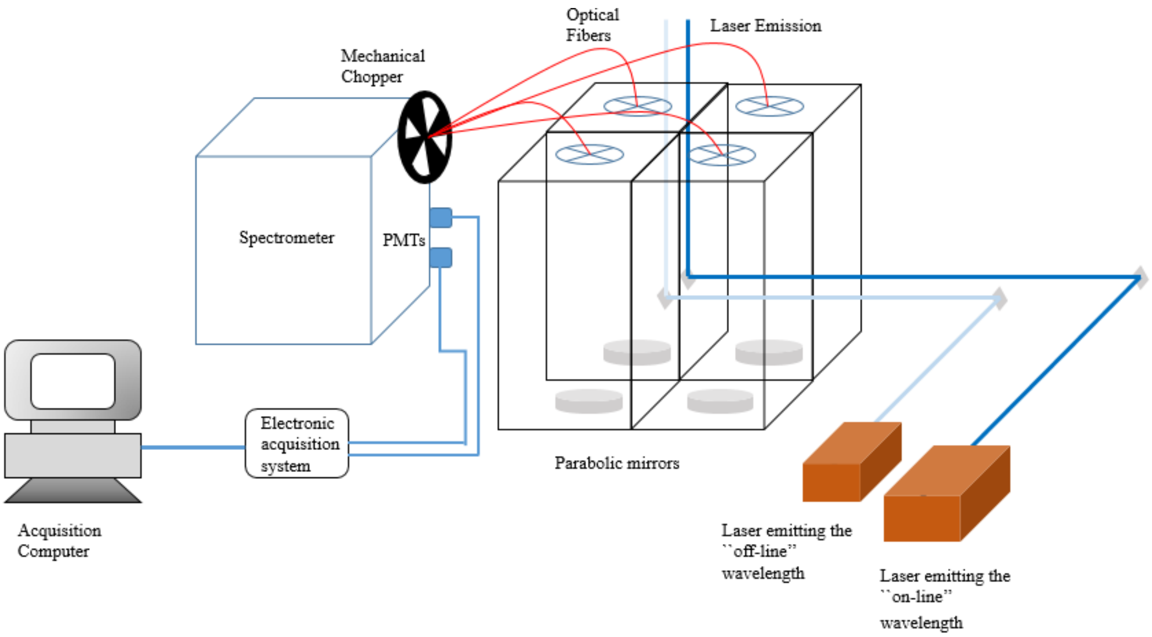
Optical Fibers
Laser Emission
Mechanical Chopper
Spectrometer
PMTs
Electronic acquisition system
Acquisition Computer
Parabolic mirrors
Laser emitting the ``off-line'' wavelength
Laser emitting the ``on-line'' wavelength

Figure 1.2: In the DIAL technique two wavelengths (the "on-line" and the "off-line" wavelengths) are simultaneously transmitted to the atmosphere. The back scattered signals are collected by a large mirror or by multiple small mirrors. The collected signal (via optical fibers) are sent to the PMTs.

$$N(z,\lambda) = \eta_{system}\exp(\tau_{emitted}(z,\lambda))\exp(\tau_{returned}(z,\lambda))O(z)\frac{P_{laser}}{\frac{\hbar c}{\lambda_{laser}}}\beta\frac{A}{4\pi z^2}\Delta t\Delta z + B(z) \tag{1.2}$$

Each of the quantities in the above equation is listed below:

η_{system}: the efficiency of the lidar system

$\tau_{emitted}(z,\lambda)$: the optical depth of the emitted photon through the atmosphere

$\tau_{returned}(z,\lambda)$: the optical depth of the backscattered photon through the atmosphere

O(z): the geometrical overlap function of the lidar

P_{laser}: the power of laser which is used in the lidar

$\hbar$: the Planck constant

c: the speed of light

λ_{laser}: the laser wavelength

β: the atmospheric backscattering coefficient

$\frac{A}{4\pi z^2}$: the effective area of the primary telescope

Δt: the temporal resolution of the lidar

Δz: the spatial resolution of the lidar

B(z): the background counts which can be function of a altitude.

In Rayleigh scattering (for which the backscattered signal has the same wavelength as the transmitted signal), the optical depths for the transmitted and returned wavelengths are the same. In Raman scattering, because the backscattered wavelength is a Raman-shifted wavelength of the transmitted signal, these two terms are different. Moreover, as the probability of multiple scattering of photons is low (even the probability of back-scattering is so low), in the above equation, we assumed that each photon only back-scattered once.

To retrieve the ozone density, in the clean (aersol free) atmospheric conditions, the Rayleigh scattering technique is used. However, in the presence of aerosols, the Raman scattering technique is added to the retrieval processes. The focus of this thesis is on retrieving ozone density for "clean" nights, thus we only explain the Rayleigh technique. In Eq.1.2, the atmospheric optical depth (considering that the molecule of interest is ozone) is given by:

$$\tau(z,\lambda) = \int_{z_0}^{z} [\sigma_{O_3}(\lambda, T(z'))n_{O_3}(z') + \alpha(\lambda, z') + \sum_{e} \sigma_e(\lambda)n_e(z')]dz' \tag{1.3}$$

where z_0 is the altitude of the lidar station, $\sigma_{O_3}(\lambda_i)$ is the ozone absorption cross section at the specific altitude and wavelength which is dependent to the atmospheric temperature $T(z')$, $n_{O_3}(z)$ is the ozone number density to be measured, $\alpha(\lambda, z)$ is the atmospheric extinction coefficient which includes both Rayleigh and Mie scattering extinction coefficients, and $\sum_e \sigma_e(\lambda)n_e(z)$ is the the extinction by other absorbers (like SO_2 and NO_2). In major volcanic eruptions the abundance of SO_2 gas in the stratosphere can significantly perturb the ozone retrievals. However, SO_2 only stays in the stratosphere for 30 to 40 days (Heath et al., 1983). In general, the amount of SO_2 mixing ratio in the stratosphere is negligible. The differential absorption cross section of NO_2 in the specified spectrum is on the order of 3×10^{-19} cm^2, thus considering the effect of NO_2 in the ozone retrievals is not essential, and the third term of Eq.1.3 is negligible (Godin-Beekmann et al., 2003).

For a specific wavelength, in Eq.1.2, the efficiency, the laser power, and the effective area of primary telescope are constant values within a lidar system, thus we can write:

$$C(\lambda) = \eta_{system} \frac{A}{4\pi z^2} \frac{P_{laser}}{\frac{\hbar c}{\lambda_{laser}}} \tag{1.4}$$

where $C(\lambda)$ is called the lidar constant.

1.7.1 Ozone Density Retrievals

By substituting Eq.1.3 and 1.4 into Eq.1.2 for "on-line" and "off-line" channels we can write:

$$\begin{aligned} N_{obs}(z, \lambda_{on}) =& C(\lambda_{on})\beta(\lambda_{on}) \exp\left(2 \int_{z0}^{z} [\sigma_{O_3}(\lambda_{on}, T(z'))n_{O_3}(z') + \alpha(\lambda_{on}, z')]dz'\right) + B(z, \lambda_{on}) \\ N_{obs}(z, \lambda_{off}) =& C(\lambda_{off})\beta(\lambda_{off}) \exp\left(2 \int_{z0}^{z} [\sigma_{O_3}(\lambda_{off}, T(z'))n_{O_3}(z') + \alpha(\lambda_{off}, z')]dz'\right) + B(z, \lambda_{off}). \end{aligned} \tag{1.5}$$

Dividing the two equations and taking the natural logarithm of the result gives:

$$\ln\left(\frac{N_{obs}(z, \lambda_{on}) - B(z, \lambda_{on})}{N_{obs}(z, \lambda_{off}) - B(z, \lambda_{off})}\right) = \ln\left(\frac{C(\lambda_{on})}{C(\lambda_{off})}\right) + \ln\left(\frac{\beta(\lambda_{on})}{\beta(\lambda_{off})}\right) + 2 \int_{z0}^{z} \Delta\delta_{o3} n_{O_3}(z')dz' + 2 \int_{z0}^{z} \Delta\alpha dz'. \tag{1.6}$$

In Eq. 1.6 and Eq.1.10 we used the term $\Delta\delta_{o3}$ which is the difference of ozone cross sections for the two different wavelengths:

$$\Delta\delta_{o3} = \sigma_{o3}(\lambda_{on}) - \sigma_{o3}(\lambda_{off}). \tag{1.7}$$

Similarly:

$$\Delta\alpha = \alpha(\lambda_{on}) - \alpha(\lambda_{off}). \tag{1.8}$$

As mentioned earlier, the atmospheric extinction term (α) includes both molecules (indicating Rayleigh scattering) and particles (indicating Mie scattering), and it can be written as:

$$\alpha(\lambda, z) = \alpha_p + \Sigma n(z)\sigma_R(\lambda) \tag{1.9}$$

where, α_p is the particulate extinction coefficient, $\sigma_R(\lambda)$ is the molecular Rayleigh cross section at a given wavelength, and $n(z)$ is the number density of molecules. The term $\beta(\lambda)$ also includes both molecular and particulate backscattering coefficients. Taking the derivative of Eq.1.6 with respect to the altitude and rearranging the equation, ozone density profile can be retrieved:

$$n_{o3(z)} = \frac{1}{2\Delta\delta_{o3}}\frac{d}{dz}\ln\left(\frac{N(\lambda_{off},z)-B_{off}(z)}{N(\lambda_{on},z)-B_{on}(z)}\right) - \frac{1}{2\Delta\delta_{o3}}\frac{d}{dz}\ln\left(\frac{\beta(\lambda_{on},z)}{\beta(\lambda_{off},z)}\right) - \frac{1}{\Delta\delta_{o3}}\Delta\alpha_p(z) - \frac{1}{\Delta\delta_{o3}}\Delta\sigma_R. \tag{1.10}$$

As the lidar constants are not functions of altitude, the derivative of them with respect to height is zero, and they do not play any role in calculating the ozone density profiles. The first term of the above equation contains the backscattered photon counts and background counts at each wavelength. The second term shows the ratio of molecular and particulate backscattering coefficients for "on-line" and "off-line" wavelengths. For "clean" nights (when the amount of aerosol in the atmosphere is insignificant) this term is negligible:

$$\frac{1}{2\Delta\delta_{o3}}\frac{d}{dz}\ln\left(\frac{\beta(\lambda_{on},z)}{\beta(\lambda_{off},z)}\right) = \frac{1}{2\Delta\delta_{o3}}\frac{d}{dz}\ln\left(\frac{\sigma_R(\lambda_{on})}{\sigma_R(\lambda_{off})}\right) = 0 \tag{1.11}$$

In the third term, the difference between the particulate extinction coefficients for the two wavelengths is shown. Similar to the second term, for "clean" nights this term is negligible. The fourth term is the difference between the Rayleigh cross sections at the two wavelengths (i.e, the difference in the molecular extinction coefficients). To calculate this term the Atmospheric density profile should be known; at altitudes below 30 km density profiles from nearby radiosondes are used. As radiosondes can not reach into higher altitudes, an atmospheric model is normally used. These models have high uncertainties; however, the ozone density uncertainty resulting from the errors on the air density above 15 km is less than 1%. The contribution of air density error below 15 km is important and at some cases can be as high as 15% (Megie et al., 1985). In summary, for ozone retrievals in a "clean" night condition (which is the focus of this thesis), the most important term of Eq.1.10 is the first term which contains the lidar measurements at the two wavelengths. However, prior to using the measurements to retrieve the ozone density profile, some corrections should be applied. In the next section, these corrections are explained in detail.

1.8 Corrections Applied to the Raw Counts

1.8.1 DeadTime Correction

For many lidar systems, at count rates below about 1 MHz, the relation between the true counts and the observed signal is linear. However, for the higher counts, the detector's response may not be linear. The nonlinearity becomes more significant as the count rate increases. Detector systems are generally limited in their useful dynamic range due typically to the detector (paralyzable) or counting system (non-paralyzable). In paralyzable systems, if the time interval between two photon strikes is shorter than the time needed to process the first photon strike (dead time γ) the detector is unable (paralyzed) to observe the second photon. This relation between the observed and true photon counts in paralyzable detectors is then:

$$N_{obs} = N_{true} \exp(-\gamma N_{true}) \tag{1.12}$$

where N_{obs} is the observed counts by the detector, N_{true} is the true counts (Donovan et al., 1995).

In non-paralyzable systems, the detector is not paralyzed but the counter is unable to record another photon in the time interval of γ after any recorded photon strike. The relation between the true and observed counts for the system shows is then:

$$N_{obs} = \frac{N_{true}}{1 + \gamma N_{true}} \tag{1.13}$$

The lidars used in this thesis have non-paralyzable counting systems. To use measurements in these systems above about 1 MHz the dead time must be specified or retrieved. In the traditional method, the lidar measurements should be corrected for the effect of deadtime. If the value of the deadtime is not known, an empirical fit can be used to estimate this value.

1.8.2 Signal Induced Noise

It is well-known that for high intensity systems, the output of the PMT can show an excess of counts some time after the signal intensity is maximum, a "tail" which is called signal-induced noise (SIN) (Hunt and Poultney, 1975). In fact, SIN is the residual signal originating from high signal intensities at low altitudes. It adds up with the background signal and is visible at altitudes where the signal-to-noise ratio (SNR) is very small (Iikura et al., 1987). Using a mechanical chopper to block high intensity light from reaching the detector is the most practical way to avoid SIN. It is important to consider the noise component from the upper altitude of lidar signals. In many lidars, the background is a constant and the effect of SIN is not detected. However, when the SIN is present in the background, the uppermost part of the signal can be fitted by an exponential function:

$$B(z) = a\exp(-bz) + c \tag{1.14}$$

where the fitting coefficients a, b, and c are empirically determined (Iikura et al., 1987). The SIN is more pronounced for the "on-line" wavelength because most of the laser power is used for the "on-line" wavelength (normally, the "on-line" wavelength has a power about 2 times of the laser power in the "off-line" wavelength). Therefore, for most nights the affect of SIN on the "off-line" wavelength is negligible, and a constant background is used.

1.8.3 Merging Process

Due to the high dynamic range of signals, usually in lidar measurements, two detecting channels are used. One channel is used for high altitude measurements while the other is optimized for the lower altitude measurements. Before applying the retrieval algorithm, the two high-altitude and low-altitude channels are merged to produce one signal. To merge the two signals an optimized height should be determined where both signals have the same SNR, and they are linear with respect to each other. A major issue appears when merging analog and digital channels is required. The most common practice is to digitize the analog signal, and then merge the two channels. The uncertainty which is introduced in the conversion process is yet to be mathematically determined. Moreover, for many lidars, the digitized analog signal does not follow the Poisson distribution, thus determining the gluing uncertainty becomes more difficult. A detailed description of the merging process can be found in Steinbercht (1994).

1.8.4 Digital Filters

In the DIAL technique, the rapid decrease of the SNR is another difficulty. Low-pass filters are used to reduce the noise of the signals. The final vertical resolution Δz_f varies by the order of filter (number of point used to make the filter) and is calculated as:

$$\Delta z_f = \nu_c \Delta z_i \tag{1.15}$$

where ν_c is the cutoff frequency of the low-pass filter, and Δz_i is the initial vertical resolution of the measurements (Godin et al., 1999; Leblanc et al., 2016). In the lower stratosphere, perturbations in the ozone profiles are well detected; however, depending on the order of filter, the perturbation can be largely attenuated and cause negative or positive biases. For higher altitudes, because of the lower SNR, the vertical resolution is decreased. Different numerical filters have been tested to optimize the ozone retrievals. In all these techniques, to overcome the SNR decrease, the number of coefficients in filters are increased with altitude (Godin et al., 1999).

In summary, for traditional DIAL analysis some corrections should be applied to raw count measurements, after which Eq.1.10 can be applied to calculate the ozone density profile. An alternative approach is to apply the Optimal Estimation Method (OEM). In this method, count correction, gluing of profiles, and pre- or post-filtering are not needed. The raw measurements from all the available channels are used as an input vector and one ozone profile is retrieved as the output. In the next section more detail on OEM is provided.

1.9 Optimal Estimation Method

Inverse modeling is a process of the transformation from data to model parameters. OEM is a matrix inverse method based on Bayes' theorem. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be the vector state of the atmosphere and $\mathbf{y} = (y_1, y_2, ..., y_n)$ be the corresponding vector measurement. The relationship between $\mathbf{x}$ and $\mathbf{y}$ can be shown with a forward model:

$$\mathbf{y} = F(\mathbf{x}, \mathbf{b}) + \epsilon \tag{1.16}$$

where $\mathbf{b}$ is the model parameter vector and ϵ is the noise in the measurements. In the absence of error from the inversion of the forward model, the exact value of $\mathbf{x}$ can be retrieved. However, all real measurements contain experimental errors. Therefore, any practical retrieval method should carry the measurement's uncertainty and the resulting uncertainty of the retrieved quantity. Bayesian statistics provides a useful way to look at this problem.

In Bayes' theorem, an *a priori* state $\mathbf{x}_a$ and its assigned probability reflecting the certainty of this state is provided. The goal is to calculate the most likely state vector $\mathbf{x}$ which is consistent with the *a priori* knowledge. Formally we can write:

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}. \tag{1.17}$$

In the above equation:

- P($\mathbf{x}$) is the *a priori* probability density function (PDF) of the state vector $\mathbf{x}$.
- $P(\mathbf{y}|\mathbf{x})$ is the PDF of the measurement vector $\mathbf{y}$ (given the true values of $\mathbf{b}$).
- $P(\mathbf{x}|\mathbf{y})$ is the PDF of $\mathbf{x}$ (given the measurement vector $\mathbf{y}$), and is called the a posteriori PDF for the state vector.

We use matrix presentation in which the uncertainties associated with the *a priori* state and measurements are shown as covariance matrices $\mathbf{S}_a$ and $\mathbf{S}_y$ where their diagonal elements are the variances of the individual elements of the *a priori* state and measurements noise. Here, for simplicity, a linear problem in which all of the PDFs follow Gaussian statistics is assumed. In most altitudes we have more than 15 photons. Hence, according to the central limit theorem, assuming a Gaussian distribution is a valid assumption. The Gaussian distribution in vector space can be written as:

$$P(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|S_{\mathbf{y}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^{\mathsf{T}}\mathbf{S}_{\mathbf{y}}^{-1}(\mathbf{y}-\bar{\mathbf{y}})\right). \tag{1.18}$$

Assuming that the forward model is linear, we define the Jacobian matrix as $\mathbf{K}_x = \frac{\partial F}{\partial \mathbf{x}}$ which indicates the sensitivity of the forward model to the state variables $\mathbf{x}$. Therefore, the PDF of the measurement vector **y** and the state vector **x**, after some rearrangements, is written as:

$$\begin{aligned} -2\ln P(\mathbf{y}|\mathbf{x}) &= (\mathbf{y}-\bar{\mathbf{y}})^{\mathsf{T}}\mathbf{S}_{\mathbf{y}}^{-1}(\mathbf{y}-\bar{\mathbf{y}}) + c_1 \\ -2\ln P(\mathbf{x}) &= (\mathbf{x}-\mathbf{x}_a)^{\mathsf{T}}\mathbf{S}_{\mathbf{a}}^{-1}(\mathbf{x}-\mathbf{x}_a) + c_2 \end{aligned} \tag{1.19}$$

where c_1 and c_2 are constants. The *posteriori* PDF is:

$$-2\ln P(\mathbf{x}|\mathbf{y}) = (\mathbf{y}-\mathbf{Kx})^T\mathbf{S}_y^{-1}(\mathbf{y}-\mathbf{Kx}) + (\mathbf{x}-\widehat{\mathbf{x}})^T\mathbf{S}_a^{-1}(\mathbf{x}-\widehat{\mathbf{x}}) \tag{1.20}$$

where $-2\ln P(x|y)$ is called the cost function. The optimal or maximum *posteriori* (MAP) solution for **x** is shown as the the maximum of $P(\mathbf{x}|\mathbf{y})$ which is the solution to $\nabla_{\mathbf{x}}(-2\ln P(\mathbf{x}|\mathbf{y})) = 0$ where ∇_x is the gradient operator in the state vector. In Eq.1.20, the first term is defining the difference between the true value (measurements) and the predicted value (the forward model) weighted by the measurement noise (error), thus this term is in fact the well-known Least Square Method which is widely used in regression problems. The second term of the equation defines the difference between an *a priori* value and the state vector weighted by *a priori* uncertainty. Depending on how much we trust the *a priori* profile (how large the *a priori* uncertainty is) the second term can play an important role. Typically, the cost is normalized to the number of measurements, and a cost of around 1 indicates a good retrieval.

The solution to $\nabla_{\mathbf{x}}(-2\ln P(\mathbf{x}|\mathbf{y})) = 0$ (the MAP solution) is:

$$\widehat{\mathbf{x}} = \mathbf{x}_a + (\mathbf{K}^{\mathsf{T}}\mathbf{S}_{\mathbf{y}}^{-1}\mathbf{K} + \mathbf{S}_a^{-1})\mathbf{K}^{\mathsf{T}}\mathbf{S}_a^{-1}(\mathbf{y}-\mathbf{Kx}_a)^{-1} \tag{1.21}$$

or equivalently:

$$\widehat{\mathbf{x}} = \mathbf{x}_a + \mathbf{G}(\mathbf{y}-\mathbf{Kx}_a) \tag{1.22}$$

where $\mathbf{G} = \frac{\partial\widehat{\mathbf{x}}}{\partial\mathbf{y}}$ is the gain matrix and represents the sensitivity of the retrieval to the observations. Moreover, the averaging kernel matrix is defined as the sensitivity of the optimal solution to the true state x:

$$\mathbf{A} = \frac{\partial\widehat{\mathbf{x}}}{\partial\mathbf{x}} = \mathbf{G}_{\mathbf{y}}\mathbf{K}_{\mathbf{x}} \tag{1.23}$$

In practice, most forward models are nonlinear and the Eq.1.20 should be solved numerically. The Newton and Gauss-Newton methods are normally used when the problem is not too non-linear, and the Levenberg-Marquart iteration is used for forward models that have higher degree of nonlinearity. Here, we use the latter method. The optimized solution for state vector **x** is found when the below iteration converges:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + [(1+\gamma_i)\mathbf{S}_a^{-1} + \mathbf{K}_i^T\mathbf{S}_y\mathbf{K}_i]^{-1}([\mathbf{K}_i^T\mathbf{S}_y^{-1}(\mathbf{y} - F(\mathbf{x}_i, \mathbf{b})] - \mathbf{S}_a^{-1}(\mathbf{x}_i - \mathbf{x}a)). \tag{1.24}$$

Here, γ_i is a damping factor, for $\gamma_i \longrightarrow 0$ the iteration becomes similar to the Gauss-Newton method, and for $\gamma_i \longrightarrow \inf$ the iteration tends to steepest descent. Detailed discussion on the choice of γ_i can be found elsewhere (Marquardt, 1963).

Finally, the uncertainty budget can be calculated as:

$$\mathbf{S}_{total} = \mathbf{S}_m + \mathbf{S}_s + \mathbf{S}_F. \tag{1.25}$$

In the latter equation, the retrieval covariance due to the measurement noise is S_m:

$$\mathbf{S}_m = \mathbf{G}_y\mathbf{S}_y\mathbf{G}_{\mathbf{y}}^T. \tag{1.26}$$

Rather than being the estimate of the true state, the retrieval is an optimal smoothed estimate (smoothed by the averaging kernel). The retrieval error resulting from smoothing should be calculated. The smoothing covariance $\mathbf{S}_s$ is:

$$\mathbf{S}_s = (\mathbf{A} - \mathbf{I}_n)\mathbf{S}_e(\mathbf{A} - \mathbf{I}_n)^T \tag{1.27}$$

where, $\mathbf{I}_n$ is the unit matrix and $\mathbf{S}_e$ is the covariance of the real ensemble of states. The error in the retrievals due to the forward model parameter uncertainties S_F is defined as:

$$\mathbf{S}_F = \mathbf{G}_\mathbf{y}\mathbf{K}_\mathbf{b}\mathbf{S}_\mathbf{b}\mathbf{K}_\mathbf{b}^T\mathbf{G}_\mathbf{y}^T. \tag{1.28}$$

In the above equation, $\mathbf{K}_b = \frac{dF}{db}$ represents the sensitivity of forward model to the **b** parameter, and $\mathbf{S}_b$ is the error on the assumed **b** parameter. A summary of the OEM procedure is shown in Fig.1.3. In Chapters 2 and 3, we describe how the OEM can be implemented to retrieve stratospheric and tropospheric ozone profiles, and we compare our results with the traditional analysis.

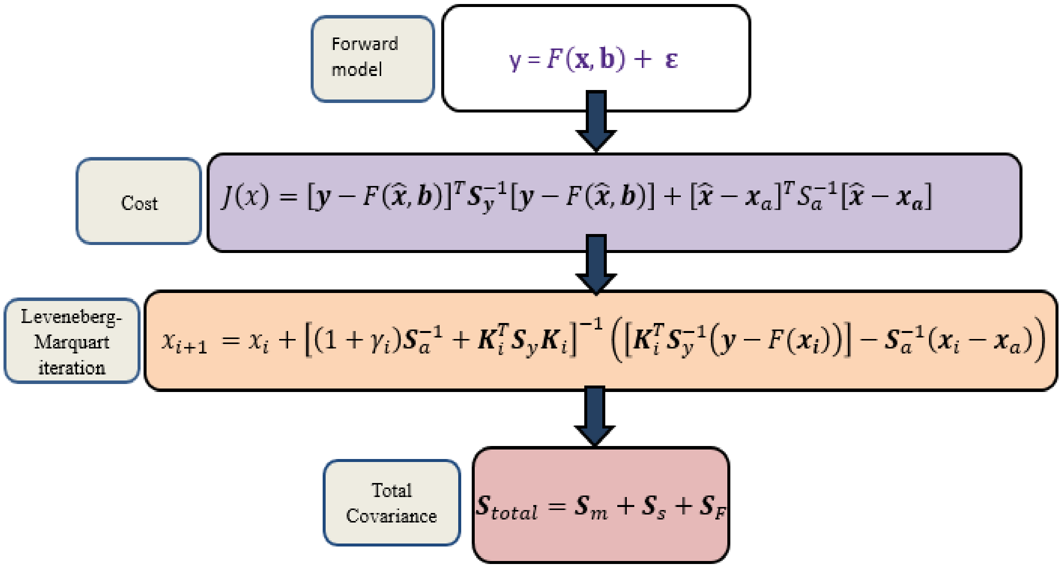
Forward model
$y = F(\mathbf{x}, \mathbf{b}) + \varepsilon$
Cost
$J(x) = [\mathbf{y} - F(\hat{\mathbf{x}}, \mathbf{b})]^T \mathbf{S}_y^{-1} [\mathbf{y} - F(\hat{\mathbf{x}}, \mathbf{b})] + [\hat{\mathbf{x}} - \mathbf{x}_a]^T S_a^{-1} [\hat{\mathbf{x}} - \mathbf{x}_a]$
Leveneberg-Marquart iteration
$x_{i+1} = x_i + \left[(1+\gamma_i)\mathbf{S}_a^{-1} + \mathbf{K}_i^T \mathbf{S}_y \mathbf{K}_i\right]^{-1} \left(\left[\mathbf{K}_i^T \mathbf{S}_y^{-1}(y - F(\mathbf{x}_i))\right] - \mathbf{S}_a^{-1}(\mathbf{x}_i - \mathbf{x}_a)\right)$
Total Covariance
$\mathbf{S}_{total} = \mathbf{S}_m + \mathbf{S}_s + \mathbf{S}_F$

Figure 1.3: The flowchart is the summery of the steps we take to calculate the retrieval and its associated uncertainties.

1.10 Machine Learning Applications to Lidar Measurements

In a lidar system, the back scattered signals are received and recorded as level 0 measurement scans. Later on, to improve the SNR, these scans are co-added in time to produce one single profile representing a period of measurement. Prior to co-adding, each individual scan should be checked to make sure if it has a good quality. Therefore, "bad scans" will be removed from the "good scans". In this process, scans with low laser signal, high background counts, or unusual shape are flagged as "bad scans". Scans with traces of clouds or aerosols are also separated from the "clean and clear" scans. These scans are not bad scans, but they may require different processing algorithms.

Scans are often classified manually as good or bad, and clean or not-clean. This method is time consuming and to some extent the classification is subject to the judgment of the observer. Some lidar groups also use simple automated routines in which a pre-defined thresh-hold for the SNR at a fixed altitude is used to classify good and bad scans. This method does not have high accuracy and bad scans may pass the thresh-hold test and incorrectly be flagged as good.

Using state-of-the-art machine learning (ML) techniques, we have developed an automated classifier. We are classifying the level 0 lidar measurements with high accuracy. ML has recently been used to distinguish between aerosols and clouds for the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) level 2 measurements (Zeng et al., 2018). Furthermore, Nicolae et al. (2018) used a Neural Network algorithm to estimate the most probable aerosol types in a set of data obtained from European Aerosol Research Lidar Network (EARLINET). We present our result for the Purple Crow Lidar (PCL) system as well as for the Meteoswiss Raman Lidar for Meteorological Observations (RALMO) system. Here, we briefly describe the ML technique.

1.10.1 Machine Learning Methods

The ML is widely used for making predictions and to recognize patterns. The ML is categorized into supervised and unsupervised methods. In the supervised approach, the aim of ML is to learn a function which can map observations (x) to correct dependent (output) values (y). In supervised learning the dataset is divided into the training set and the test set. In the traning phase, the training set (which contains x values and their corresponding y values) is used to learn the mapping function. To validate the result, the test set is used (which only has x values) to predict the output. A good supervised algorithm will produce high accuracy scores in both training and test phases. Unsupervised learning is a data driven method in which the main goal is to find similarities between data points and to cluster the data accordingly. Thus output values (y) are not needed and the algorithm is trained only by using the observations (x). It is worth noting that in inverse modelling methods (e.g., OEM), y is considered as the observable and x is the quantity of interest. However, in ML notation, x is given and y is predicted.

The ML can be used for either classification or regression. In the classification methods, targets are discrete values or categories. For example, our classifier method in which the scans are divided into good and bad scans uses a classification ML model. In regression, the target is a continuous quantity. For example, a regression model can be used to retrieve the ozone density profiles.

In Chapter 4 of this thesis, we use both supervised and unsupervised methods to classify our lidar measurements. We use different ML algorithms in our work and those which provided us with high accuracy scores are selected and discussed in detail.

Bibliography

Ahrens, C. D.: Meteorology Today: an introduction to weather, climate, and the environment, Meteorologie, 9, 74–75, 1998.

Andrews, D. G., Holton, J. R., and Leovy, C. B.: Middle atmosphere dynamics, 40, Academic press, 1987.

Ball, W. T., Alsing, J., Mortlock, D. J., Staehelin, J., Haigh, J. D., Peter, T., Tummon, F., Stübi, R., Stenke, A., Anderson, J., Bourassa, A., Davis, S. M., Degenstein, D., Frith, S., Froidevaux, L., Roth, C., Sofieva, V., Wang, R., Wild, J., Yu, P., Ziemke, J. R., and Rozanov, E. V.: Evidence for a continuous decline in lower stratospheric ozone offsetting ozone layer recovery, acp, 18, 1379–1394, 2018.

Bates, D. R. and Nicolet, M.: Atmospheric hydrogen, Publications of the Astronomical Society of the Pacific, 62, 106–110, 1950.

Brasseur, G. P. and Solomon, S.: Aeronomy of the middle atmosphere: chemistry and physics of the stratosphere and mesosphere, vol. 32, 2006.

Browell, E. V.: Differential absorption lidar sensing of ozone, IEEE, 77, 419–432, 1989.

Chapman, S.: XXXV. On ozone and atomic oxygen in the upper atmosphere, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 10, 369–383, 1930.

Cohen, Y., Petetin, H., Thouret, V., Marécal, V., Josse, B., Clark, H., Sauvage, B., Fontaine, A., Athier, G., Blot, R., et al.: Climatology and long-term evolution of ozone and carbon monoxide in the upper troposphere–lower stratosphere (UTLS) at northern midlatitudes, as seen by IAGOS from 1995 to 2013, acp, 18, 5415–5453, 2018.

Crutzen, P. J.: The influence of nitrogen oxides on the atmospheric ozone content, Quarterly Journal of the Royal Meteorological Society, 96, 320–325, 1970.

Danielsen, E. F.: Stratospheric-tropospheric exchange based on radioactivity, ozone and potential vorticity, 1968.

Donovan, D. P., Bird, J. C., Whiteway, J. A., Duck, T. J., Pal, S. R., and Carswell, A. I.: Lidar observations of stratospheric ozone and aerosol above the Canadian high arctic during the 199495 winter, Geophysical Research Letters, 22, 34893492, 1995.

Emery, W. and Camps, A.: Introduction to Satellite Remote Sensing: Atmosphere, Ocean, Land and Cryosphere Applications, Elsevier, 2017.

Farman, J. C., Gardiner, B. G., and Shanklin, J. D.: Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction, Nature, 315, 207, 1985.

Fernald, F. G.: Analysis of atmospheric lidar observations: some comments, Appl. Opt., 23, 652–653, 1984.

Finlayson-Pitts, B. J., Jr., P., and N., J.: Chemistry of the upper and lower atmosphere: theory, experiments, and applications, Elsevier, 1999.

Gebhardt, C., Rozanov, A., Hommel, R., Weber, M., Bovensmann, H., Burrows, J. P., Degenstein, D., Froidevaux, L., and Thompson, A. M.: Stratospheric ozone trends and variability as seen by SCIAMACHY from 2002 to 2012, acp, 14, 831–846, 2014.

Godin, S., Carswell, A. I., Donovan, D. P., Claude, H., Steinbrecht, W., McDermid, I. S., McGee, T. J., Gross, M. R., Nakane, H., Daan, Swart, P. J., Bergwerff, B. B., Uchino, O., von der Gathen, P., and Neuber, R.: Ozone differential absorption lidar algorithm intercomparison, Appl. Opt., 38, 6225–6236, 1999.

Godin-Beekmann, S., Porteneuve, J., and Garnier, A.: Systematic DIAL lidar monitoring of the stratospheric ozone vertical distribution at Observatoire de Haute-Provence (43.92[degree]N, 5.71[degree]E), J. Environ. Monit., 5, 57–67, 2003.

Harris, N. R. P., Hassler, B., Tummon, F., Bodeker, G. E., Hubert, D., Petropavlovskikh, I., Steinbrecht, W., Anderson, J., Bhartia, P. K., Boone, C. D., Bourassa, A., Davis, S. M., Degenstein, D., Delcloo, A., Frith, S. M., Froidevaux, L., Godin-Beekmann, S., Jones, N., Kurylo, M. J., Kyrölä, E., Laine, M., Leblanc, S. T., Lambert, J.-C., Liley, B., Mahieu, E., Maycock, A., de Mazière, M., Parrish, A., Querel, R., Rosenlof, K. H., Roth, C., Sioris, C., Staehelin, J., Stolarski, R. S., Stübi, R., Tamminen, J., Vigouroux, C., Walker, K. A., Wang, H. J., Wild, J., and Zawodny, J. M.: Past changes in the vertical distribution of ozone Part 3: Analysis and interpretation of trends, acp, 15, 9965–9982, 2015.

Harrison, R. M.: Understanding our environment: an introduction to environmental chemistry and pollution, 2007.

Heath, D. F., Schlesinger, B. M., and Park, H.: Spectral change in the ultraviolet absorption and scattering properties of the atmosphere associated with the eruption of El Chichón: Stratospheric SO2 budget and decay, Eos Trans. AGU, 64, 197, 1983.

Hedin, A. E.: Extension of the MSIS thermosphere model into the middle and lower atmosphere, J. Geophys. Res.: Space Physics, 96, 1159–1172, 1991.

Holton, J. R., Haynes, P. H., McIntyre, M. E., Douglass, A. R., Rood, R. B., and Pfister, L.: Stratosphere-troposphere exchange, Rev. Geophys., 33, 403–439, 1995.

Hunt, W. H. and Poultney, S. K.: Testing the linearity of response of gated photomultipliers in wide dynamic range laser radar systems, IEEE Trans. Nucl. Sci, 22, 116–120, 1975.

Iikura, Y., Sugimoto, N., Sasano, Y., and Shimzu, H.: Improvement on lidar data processing for stratospheric aerosol measurements, Appl. Opt., 26, 5299–5306, 1987.

IPCC2007: Climate Change 2007:Th Physical ScienceSummary for Policy makers. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,17.

Jacob, D. J., Heikes, E., Fan, S. M., Logan, J. A., Mauzerall, D. L., Bradshaw, J. D., Singh, H. B., Gregory, G. L., Talbot, R. W., Blake, D. R., et al.: Origin of ozone and NOx in the tropical troposphere: A photochemical analysis of aircraft observations over the South Atlantic basin, J. Geophys. Res.: Atmospheres, 101, 24 235–24 250, 1996.

Johnson, B. J., Oltmans, S. J., Vömel, H., Smit, H. G., Deshler, T., and Kröger, C.: Electrochemical concentration cell (ECC) ozonesonde pump efficiency measurements and tests on the sensitivity to ozone of buffered and unbuffered ECC sensor cathode solutions, Journal of Geophysical Research: Atmospheres, 107, ACH–8, 2002.

Johnston, H. S.: Atmospheric ozone, Annual review of physical chemistry, 43, 1–31, 1992.

Junge, C. E.: Global ozone budget and exchange between stratosphere and troposphere, Tellus, 14, 363–377, 1962.

Kley, D., Crutzen, P. J., Smit, H. G. J., Vömel, H., Oltmans, S. J., Grassl, H., and Ramanathan, V.: Observations of near-zero ozone concentrations over the convective Pacific: Effects on air chemistry, Science, 274, 230–233, 1996.

Komhyr, W.: Operations handbook-ozone measurements to 40-km altitude with Model 4A Electrochemical Concentration Cell (ECC) ozonesonaes (used with 1680-MHz radiosondes), Unknown, 1986.

Leblanc, T., Sica, R. J., van Gijsel, J. A. E., Godin-Beekmann, S., Haefele, A., Trickl, T., Payen, G., and Gabarrot, G.: Proposed standardized definitions for vertical resolution and uncertainty in the NDACC lidar ozone and temperature algorithms Part 1: Vertical resolution, amt, 9, 4029–4049, 2016.

Manney, G. L., Santee, M. L., Rex, M., Livesey, N. J., Pitts, M. C., Veefkind, P., Nash, E. R., Wohltmann, I., Lehmann, R., Froidevaux, L., et al.: Unprecedented Arctic ozone loss in 2011, Nature, 478, 469, 2011.

Marquardt, D. W.: An algorithm for least-squares estimation of nonlinear parameters, Journal of the society for Industrial and Applied Mathematics, 11, 431–441, 1963.

Mateer, C. L.: On the information content of Umkehr observations, J. Atmospheric Sci., 22, 370–381, 1965.

Mateer, C. L. and DeLuisi, J. J.: A new Umkehr inversion algorithm, J. Atmospheric Sol.-Terr. Phys., 54, 537–556, 1992.

McElroy, M. B., Salawitch, R. J., Wofsy, S. C., and Logan, J. A.: Reductions of Antarctic ozone due to synergistic interactions of chlorine and bromine, Nature, 321, 759, 1986.

Megie, G. J., Ancellet, G., and Pelon, J.: Lidar measurements of ozone vertical profiles, Appl. Opt, 24, 3454–3463, 1985.

Molina, M. J. and Rowland, F. S.: Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone, Nature, 249, 810, 1974.

Nicolae, D., Vasilescu, J., Talianu, C., Binietoglou, I., Nicolae, V., Andrei, S., and Antonescu, B.: A neural network aerosol-typing algorithm based on lidar data, Atmospheric Chemistry and Physics, 18, 14 511–14 537, 2018.

Papayannis, A., Ancellet, G., Pelon, J., and Megie, G.: Multiwavelength lidar for ozone measurements in the troposphere and the lower stratosphere, Appl. Opt, 29, 467–476, 1990.

Ramaswamy, V., Boucher, O., Haigh, J., Hauglustine, D., Haywood, J., Myhre, G., Nakajima, T., Shi, G. Y., and Solomon, S.: Radiative forcing of climate, Climate change, 349, 2001.

Sauvage, F., Laffont, L., Tarascon, J. M., and Baudrin, E.: Study of the insertion/deinsertion mechanism of sodium into Na0. 44MnO2, Inorganic chemistry, 46, 3289–3294, 2007.

Schotland, R.: Errors in the lidar measurement of atmospheric gases by differential absorption, Journal of Applied Meteorology (1962-1982), pp. 71–77, 1974.

Schulz, A., Rex, M., Harris, N. R. P., Braathen, G. O., Reimer, E., Alfier, R., Kilbane-Dawe, I., Eckermann, S., Allaart, M., Alpers, M., et al.: Arctic ozone loss in threshold conditions: Match observations in 1997/1998 and 1998/1999, J. Geophys. Res.: Atmospheres, 106, 7495–7503, 2001.

Solomon, S., Ivy, D. J., Kinnison, D., Mills, M. J., Neely, R. R., and Schmidt, A.: Emergence of healing in the Antarctic ozone layer, Science, p. aae0061, 2016.

Steinbercht, W.: Lidar measurements of ozone, aerosol and temperature in the stratosphere, Ph.D. thesis, York University, 1994.

Stolarski, R. S. and Cicerone, R. J.: Stratospheric chlorine: a possible sink for ozone, Can. J. Chem., 52, 1610–1615, 1974.

Tung, K., Ko, M. K., Rodriguez, J. M., and Sze, N.: Are Antarctic ozone variations a manifestation of dynamics or chemistry?, Nature, 322, 811, 1986.

Vömel, H. and Diaz, K.: Ozone sonde cell current measurements and implications for observations of near-zero ozone concentrations in the tropical upper troposphere, amt, 3, 495–505, 2010.

Weitkamp, C.: Lidar: range-resolved optical remote sensing of the atmosphere, vol. 102, 2006.

WMO: Scientic Assessment of Ozone Depletion: 1988 Global Ozone Research and Monitoring Project Report 44, World Meteorological Organization, Geneva, 1988.

WMO: Scientic Assessment of Ozone Depletion: 1992 Global Ozone Research and Monitoring Project Report 40, World Meteorological Organization, Geneva, 1992.

WMO: Scientic Assessment of Ozone Depletion: 1998 Global Ozone Research and Monitoring Project Report 44, World Meteorological Organization, Geneva, 1999.

WMO: Scientic Assessment of Ozone Depletion: 2010 Global Ozone Research and Monitoring Project Report 52, World Meteorological Organization, Geneva, 2011.

WMO: Scientic Assessment of Ozone Depletion: 2014 Global Ozone Research and Monitoring Project Report, World Meteorological Organization, Geneva, 2014.

Zeng, S., Vaughan, M., Liu, Z., Trepte, C., Kar, J., Omar, A., Winker, D., Lucker, P., Hu, Y., Getzewich, B., and Avery, M.: Application of High-Dimensional Fuzzy K-means Cluster Analysis to CALIOP/CALIPSO Version 4.1 Cloud-Aerosol Discrimination, amt Discussions, 2018, 1–40, 2018.

Chapter 2

Implementing the OEM to Retrieve Stratospheric Ozone Density

2.1 Overview

This chapter provides a detailed description of the first principle Optimal Estimation Method (OEM) which is applied to ozone retrieval analysis using Differential Absorption Lidar (DIAL) measurements. The air density, detector dead times, background coefficients, and lidar constants are simultaneously retrieved along with ozone density profiles. Using an averaging kernel, the OEM provides the vertical resolution of the retrieval as a function of altitude. A maximum acceptable height at which the *a priori* has a small contribution to the retrieval is calculated for each profile as well. Moreover, a complete uncertainty budget including both systematic and statistical uncertainties is given for each individual retrieved profile. Long term stratospheric DIAL ozone measurements have been carried out at the Observatoire de Haute-Provence (OHP) since 1985. The OEM is applied to three nights of measurements at OHP during an intensive ozone campaign in July 2017 for which coincident lidar-ozonesonde measurements are available. The retrieved ozone density profiles are in good agreement with both traditional analysis and the ozonesonde measurements. For the three nights of measurements, below 15 km the difference between the OEM and the sonde profiles is less than 25%, at altitudes between 15 km to 25 km the difference is less than 10%, and the OEM can successfully capture many variations of ozone which are detected in the sonde profiles due to its ability to adjust its vertical resolution as the signal varies. Above 25 km the difference between the OEM and the sonde profiles does not exceed 20%.

2.2 Ozone Retrievals: Traditional versus the OEM

In the traditional method, as discussed in Section 1.7.1, for retrieving ozone density from DIAL measurements, the derivative of the ratio between the "on-line" and "off-line" signals is calculated. By rewriting Eq. 1.10, the ozone number density can be retrieved as follows:

$$n_{o3(z)} = \frac{-1}{2\Delta\delta_{o3}(z)}\frac{d}{dz}\ln\left(\frac{N(\lambda_{on}, z) - B_{on}(z)}{N(\lambda_{off}, z) - B_{off}(z)}\right) + \delta n_{o_3}(z) \tag{2.1}$$

where $N(\lambda_{on}, z)$ and $N(\lambda_{off}, z)$ are, respectively, the "on-line" and "off-line" signals at altitude z, $B_{on}(z)$ and $B_{off}(z)$ are the background signals, and $\Delta\delta_{o_3}(z)$ is the differential absorption cross section between the two wavelengths. $\delta n_{o_3}(z)$ is a correction term for the effect of differential Rayleigh and Mie scattering, and the differential absorption by other absorbers (this term is equivalent to the last three terms of Eq. 1.10). More details can be found in McDermid et al. (1990), citeB205880D, and Leblanc et al. (2016b).

In the traditional ozone retrieval algorithm, several corrections are applied to the raw (level 0) counts to produce corrected photocounts, as discussed in Section 1.8. For high count rates, the dead time of the counting system is determined and a non-linearity correction is applied. Depending on the configuration of the lidar, channels with different gains may be merged (”glued”) to produce a single ozone profile. Determining the optimized height to merge the channels is typically done empirically. In the DIAL technique, the rapid decrease of sensitivity to ozone in the upper stratosphere is another important consideration. Low-pass filters are used to reduce the noise of the signals. For an ideal low-pass filter, the transfer function of all frequencies between 0 and the cut-off frequency, ν_c is 1, and the transfer function from ν_c to 1 is 0, where the reduced frequency ν is defined as $\frac{f}{f_N}$ and f_N is the Nyquist frequency. The final vertical resolution of the signal, Δz_f, varies by the order of filter, which depends on the cutoff frequency and the initial vertical resolution Δz_i:

$$\Delta z_f = \nu_c \Delta z_i. \tag{2.2}$$

A detailed discussion on the digital filtering and the vertical resolution can be found in Godin et al. (1999) and Leblanc et al. (2016a).

In the lower stratosphere, the perturbations in the ozone profiles are well detected; however, depending on the number of points in the filter (order of filter), the perturbation can be largely attenuated and cause negative or positive biases. For higher altitudes, because of the lower SNR, the vertical resolution is decreased. Different numerical filters have been tested to optimize the ozone retrievals. In all these techniques, to overcome the SNR decrease, the number of coefficients in the filters is increased with altitude (Godin et al., 1999).

2.2.1 Applying the optimal estimation method to ozone retrievals

The OEM is an inverse method in which the Bayesian theorem is used to find the probability distribution function (PDF) of the state of interest. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be the state vector, and $\mathbf{y} = (y_1, y_2, ..., y_n)$ be the vector of the measurements. The relation between the measurements and the state vector is:

$$\mathbf{y} = F(\mathbf{x}, \mathbf{b}) + \epsilon \tag{2.3}$$

where $F(\mathbf{x}, \mathbf{b})$ is called the forward model. The forward model describes our understanding of the physics of the measurements as well as the instrument's characteristics. Here, $\mathbf{b}$ is the model parameter vector which contains additional parameters needed in the forward model, and the noise in the measurements is the vector ϵ. In lidar measurements, the photon counts follow a Poisson distribution. However, for a count rate greater than 10 to 20, the PDF of the corresponding error tends toward a Gaussian distribution. Therefore, using the Bayesian approach and assuming a Gaussian PDF for all quantities, for a given measurement $\mathbf{y}$, the most likely state of $\mathbf{x}$ is found by minimizing the following cost function:

$$\mathbf{J}(x) = [\mathbf{y} - F(\widehat{\mathbf{x}}, \mathbf{b})]^T \mathbf{S}_{\mathbf{y}}^{-1} [\mathbf{y} - F(\widehat{\mathbf{x}}, \mathbf{b})] + [\widehat{\mathbf{x}} - \mathbf{x_a}]^T \mathbf{S}_a^{-1} [\widehat{\mathbf{x}} - \mathbf{x_a}] \tag{2.4}$$

where $\mathbf{S}_y$ is the covariance matrix of the measurements, $\mathbf{x_a}$ is the *a priori* profile which is an initial guess for the state vector, and $\mathbf{S}_a$ is the associated *a priori* covariance matrix. Typically, the cost is normalized to the number of measurements, and a cost of around 1 indicates a good retrieval.

As the forward model is nonlinear, the MarquardtLevenberg method is used to find the state vector. The optimized solution for the state vector **x** occurs when the following iteration converges:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + [(1+\gamma_i)\mathbf{S}_a^{-1} + \mathbf{K}_i^T\mathbf{S}_y\mathbf{K}_i^T]^{-1}([\mathbf{K}_i^T\mathbf{S}_y^{-1}(\mathbf{y} - F(\mathbf{x}_i)] - \mathbf{S}_a^{-1}(\mathbf{x}_i - \mathbf{x}a)) \tag{2.5}$$

here, $\mathbf{K} = \frac{dF}{dx}$ is the Jacobian of the forward model, and γ_i is a damping factor for the iteration. A comprehensive description on the application of the Marquardt-Levenburg method to OEM can be found in (Rodgers, 2000).

2.2.2 Ozone DIAL Forward Model

Our first-principle OEM retrieval uses the lidar equation as the forward model and the raw counts are the measurements. The lidar equation for the true counts is:

$$\begin{aligned} N_{true}(z,\lambda_{on}) =& (\frac{C_{\lambda_{on}}}{z^2})\beta(z,\lambda_{on})\Gamma_{O_3}(\lambda_{on},z)\Gamma_{atm}(\lambda_{on},z) + B_{\lambda_{on}}(z) \\ N_{true}(z,\lambda_{off}) =& (\frac{C_{\lambda_{off}}}{z^2})\beta(z,\lambda_{off})\Gamma_{O_3}(\lambda_{off},z)\Gamma_{atm}(\lambda_{off},z) + B_{\lambda_{off}(z)} \end{aligned} \tag{2.6}$$

where λ_{on} and λ_{off} represents the “on-line” and “off-line” channels, $\Gamma_{O_3}(\lambda_{on,off},z)$ and $\Gamma_{atm}(\lambda_{on,off},z)$ are respectively, the ozone and atmospheric transmissions in each wavelength, $C_{\lambda_{on}}$ and $C_{\lambda_{off}}$ are the lidar constants, and $B_{\lambda_{on}(z)}$ and $B_{\lambda_{off}(z)}$ are the background counts. For the stratospheric ozone measurements, in the altitude region of retrieval, the overlap is complete, and thus, we have not included it in our forward model. Depending on the characteristics of the data acquisition system, the true counts are related to the observed counts by either Eq. 1.13 or 1.12. In multi-channel systems, our forward model calculates the “on-line” and “off-line” wavelengths for both high altitude and low altitude channels. The transmissions are defined as:

$$\Gamma_{O_3,atm}(\lambda_i, z) = e^{-2\tau_{O_3,atm}} \tag{2.7}$$

where the optical depth $\tau_{O_3,atm}$ is previously defined in Eq. 1.3. Both atmospheric optical depth and atmospheric backscattering coefficients have contributions due to scattering from molecules and aerosols:

$$\tau_{atm} = \tau_{mol} + \tau_{aer} = \int_{z_0}^{z} [\sigma_R n_{air}(z) + \alpha(z)]dz \tag{2.8}$$

$$\beta_{atm} = \beta_{air}(z) + \beta_{aer}(z) \tag{2.9}$$

where $\beta_{air}(z)$ and $\beta_{aer}(z)$ are the corresponding air and aerosol backscattering coefficients. The "on-line" and "off-line" coefficients are related through the following equation:

$$\beta_{aer}(\lambda_{off}) = \beta_{aer}(\lambda_{on})(\frac{\lambda_{off}}{\lambda_{on}})^{-a} \tag{2.10}$$

where for aerosols the Ångstrom coefficient a equals approximately 1, and for molecular scattering the Ångstrom coefficient a equals 4. In this Chapter, we only considered the clean-night condition. Therefore, the aerosol contribution to the process is not included, but could be in the future.

Due to the presence of SIN in the "on-line" channel, the background is assumed to be a function of height in the form of Eq. 1.14, while due to a negligible presence of SIN in the "off-line" channel, a constant background is used. If necessary, it is possible and easy to assign any analytic function for the background in both channels. Therefore, if needed the background for the "off-line" channel can be assumed as a function of height as well. Using the above forward model, the ozone and air density profiles, the background coefficients, the dead time and the lidar constants for the 4 channels are simultaneously retrieved. Other parameters in the forward model are treated as model parameters. Hence, they are fixed but considered as a source of uncertainty contributing to the retrieved quantities (see Table. 2.1).

2.3 Implementing the optimal estimation method retrieval

To find the optimize solution of Eq. 3.7, *a priori* profiles for ozone and air density, as well as *a priori* values for background counts, dead time, and lidar constants are needed. Furthermore, **b** model parameter values and the covariance matrix of the measurements, *a priori* profiles, and model parameters need to be calculated. A summary of steps needed to implement the OEM for our ozone retrievals is shown in Fig. 2.1. A detailed description of these steps is provided in this section.

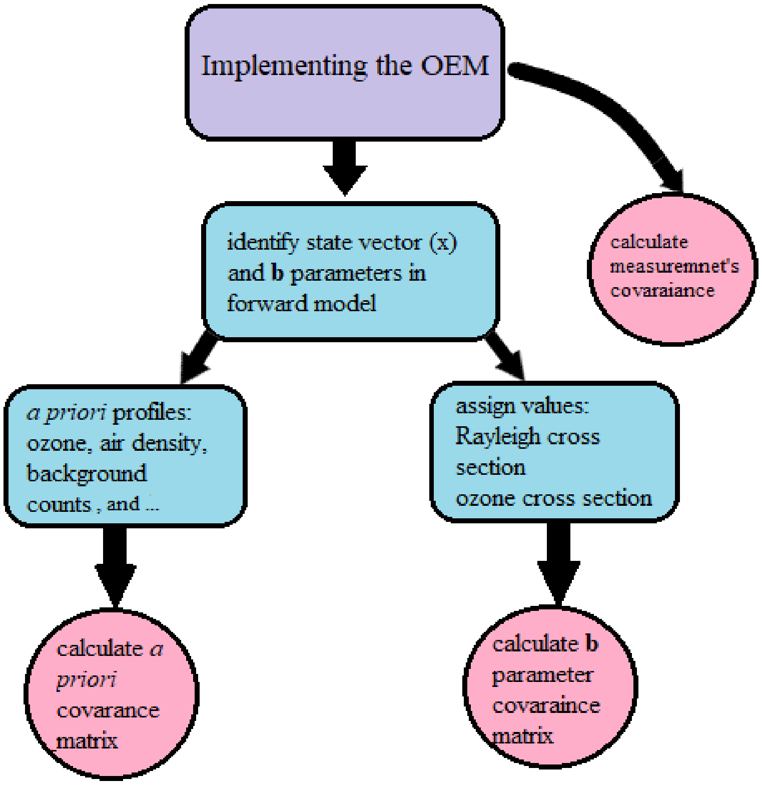
Implementing the OEM
identify state vector (x) and **b** parameters in forward model
calculate measuremnet's covaraiance
a priori profiles: ozone, air density, background counts , and ...
assign values: Rayleigh cross section ozone cross section
calculate *a priori* covarance matrix
calculate **b** parameter covaraince matrix

Figure 2.1: To implement the OEM, the *a priori* profiles for ozone and air density, background counts, dead time values, and lidar constants are needed. Moreover, **b** parameters should be identified and proper values for them should be calculated. The covariance matrices for *a priori* profiles, measurements, and **b** parameters need to be calculated as well.

The *a priori* ozone profile used for all retrievals is from an OHP ozone climatology. The climatology contains monthly-averaged ozone profile using the last 30 years of OHP DIAL and SAGE II satellite overpass measurements. The standard deviation to the 2σ level for this climatology is 50% below 25 km and 10% above 20 km altitude. Alternatively, we have used the U.S. Standard model (Krueger and Minzner, 1976) as *a priori* ozone profile, which yields similar results for our ozone retrievals.

In the traditional method, the ratio of the "on-line" to "off-line" channels is calculated. Thus, there is no need to assume an air density profile to retrieve ozone. However, in the correction term (Eq. 2.1), the air density profile is needed and an atmospheric model or a measurement is used. In the OEM, we are retrieving the air density as a state vector, and the Mass Spectrometer Incoherent Scatter Radar (MSIS) air density profile is used as *a priori* profile. The MSIS profiles are generally in good agreement with the ozonesonde measurements of air density. An uncertainty of 15% is assigned to the *a priori* of air density.

In the case of ozone and air density there is a vertical correlation between the elements of retrieval states. This corresponds to the off-diagonal elements of the *a priori* covariance matrix. Generally, it is difficult to quantify the vertical length of this correlation. We have used a correlation length ($ł_s$) of 1000 m for ozone at altitudes below 18 km and the correlation length of 1400 m at higher altitudes. The air density has a correlation length of 1400 m for all regions. A tent function is used to model the decay of correlation (**?**).

For the "off-line" channel the mean of the counts above 80 km are taken as *a priori* backgrounds, and their variances divided by the number of bins in the selected altitude region is used as *a priori* uncertainties in the background counts. For the "on-line" channel, an exponential function in the form of Eqn. 1.14 is fitted to counts above 80 km. The coefficients of the function are the *a priori* values. Depending on how good the initial fit is, uncertainties are assigned to the *a priori* coefficients, but for most nights a 20% uncertainty is chosen.

Using the forward model, the *a priori* lidar constants for both channels were estimated and an initial standard deviation of 10% for both channels is assigned. In a range in which photon counting measurements are linear (or non-linearity is correctable), Poisson statistics is applied. Thus, the measurement variances are the number of photons in each atmospheric layer located at altitude Δz, and there is no correlation between different layers (the off-diagonal elements of the matrix are zero).

The following quantities are calculated for the **b** parameters in the forward model. The Rayleigh extinction which is calculated using the Nicolet formula (Nicolet, 1984), and the temperature-dependent ozone absorption coefficients, as suggested by (Orphal et al., 2016), are calculated based on the BrionDaumontMalicet (BDM) database (Malicet et al., 1995). Uncertainties of 0.3% and 2% (Leblanc et al., 2016a) are respectively assigned to the Rayleigh and ozone cross sections. The ozone absorption cross-section is a function of temperature. The BDM database provides values for 5 different temperatures; in order to find the ozone cross section for the whole region where ozone is retrieved, the temperature is interpolated. For the interpolation, the sonde temperature profiles are used at lower altitudes (up to the altitude at which sonde measurements are available) ,and the MSIS temperature profiles are used for higher altitudes. Thus, the effect of temperature uncertainty on the ozone cross section and the final retrievals needs to be calculated as well. An uncertainty of 19 K is assigned to sonde measurements of temperature, and an uncertainty of 35 K is used for the MSIS profiles. The covariance matrix of the **b** parameters will be used later to calculate the systematic uncertainty of the retrieved quantities.

Values and associated uncertainties of the *a priori* profiles for the parameters which we are retrieving, as well as the forward model parameters which are considered as fixed parameters, (and thus, are not being retrieved) are summarized in Table. 2.1. As mentioned earlier, we are testing our model on a reasonably clear night condition from a high altitude site, therefore, we are assuming that the effects of aerosols are negligible. After calculating $\mathbf{S}_y$, $\mathbf{S}_a$, $\mathbf{S}_b$, $\mathbf{x}_a$, and **b** values, we used the Qpack software for our OEM retrieval. Qpack is a free Matlab package designed for forward and inverse modelling (Eriksson et al., 2005).

2.4 Application of the OEM to measurements from the OHP stratospheric ozone lidar

OHP is located in the south of France at (44°N, 6°E, 650 m ASL). Long term stratospheric ozone DIAL measurements have been performed since 1985. In addition, the OHP lidar is part of the international Network for the Detection of Atmospheric Composition Change (NDACC). In the OHP DIAL system, the “on-line” wavelength is provided by a XeCl excimer laser emitting at 308 nm with an emission energy of 200 mJ and a repetition rate of 100 Hz. The “off-line”

Parameter	Value	Standard Deviation
Measurements	measured	Poisson statistics
Retrieved *a priori* values		
Ozone density	OHP climatology	50% to 10%
Air density	MSIS	15%
Dead time	empirical fitting	20%
Background ("off-line")	mean above 80 km	standard deviation above 80 km
Coefficients of SIN ("on-line")	empirical fitting above 80 km	20%
Lidar constants	estimate from FM	20%
Forward model parameters		
Rayleigh-scatter cross section	Nicolet 1984	0.3%
Ozone absorption cross section	BDM 1986	2%
Temperature profile	sonde measurements	19 K
Temperature profile	MSIS	35 K

Table 2.1: Values and associated uncertainties for the retrieved and forward model parameters.

wavelength is generated from the third harmonic (355 nm) of a Continuum Nd:Yag laser, with an output energy of 40 mJ and the repetition rate of 50 Hz. In the receiving end of the DIAL system, four similar F/3 mirrors of 0.53 m diameter collect the backscattered signals. The altitude steps of measurements is 150 m. The collected signal is separated to the Rayleigh signals at the transmitted wavelengths (308 nm and 353 nm), and the corresponding 1st Stokes wavelengths in the nitrogen Raman spectrum (332.8 nm and 386.7 nm). Furthermore, to handle the high dynamic range of lidar signals in the whole altitude range, the Rayleigh signals are separated to the high and low gain channels. More details on the instrumentation can be found elsewhere (Godin-Beekmann et al., 2003).

The optical fibres transmit the receiving signals to the optical analysis device. The signals are detected by bialkali PMTs (Hamamatsu R2693P). The photon counting systems become nonlinear in the lowermost stratosphere. To correct for the saturation effect the following equation is used:

$$N_c = 1 + ((1 - x)N_r - 1)\exp(-xN_r) \tag{2.11}$$

where N_c is the observable counts, N_r is the true counts, and x is an adjustment parameter which equals the inverse of the maximum observed counts which is the definition of the dead time (Pelon and Mégie, 1982). To correct for the saturation, using Eq.2.11, the parameter x is adjusted for each wavelength in order to get a best agreement between the slopes of high and low altitude signals. The altitude at which the two profiles are combined can vary from night to night (Godin-Beekmann et al., 2003). For the two wavelengths and two different altitude channels, the dead time can differ. Therefore, we are retrieving the dead times for each altitude and at each wavelength. A dead time value which corresponds to the x parameter of each channel at each night is used as our *a priori*, and an uncertainty of ±20% is assigned to it.

Using the OEM, we retrieve the ozone density and air density profiles, as well as the dead time values for the four channels, the background counts for the "off-line" channel, and the SIN coefficients (three values) for the "on-line" channel. In total, we retrieved eight quantities along with the ozone density and air density profiles. The degree of freedom for our measurements, which is the trace of the averaging kernel, is $\approx$ 78. Below we present the ozone retrieval for 26 July 2017 in detail. In order to show that the OEM is a robust method, the results for the nights of 14, and 20 of July are presented as well. In all these nights, ozonesonde balloons were coincidentally launched, thus the OEM is validated against both the traditional method and the sonde measurements.

2.4.1 Applying the OEM to OHP measurements on 26 July 2017

Figure. 2.2 shows the averaged counts over 4 hours of measurements for two different channels at "on-line" and "off-line" wavelengths on the night of 26 July 2017. The coincident ozonesonde is launched within one hour after the start of the measurement, and takes approximately 2 hours to reach 30 km. For each retrieval, the averaging kernel matrix is calculated. The averaging kernel is a diagnostic variable which describes how the retrieval sees changes in the real atmosphere. Therefore, it contains information on the sensitivity (area of the averaging kernel function) and on the smoothing (shape of the averaging kernel function) of the retrievals. Ideally the averaging kernel is a unity matrix preserving any change in the retrieved quantity from the *a priori* state. The area is defined as the vector product $\mathbf{Au}$ where $\mathbf{u}$ is a unit vector. When the retrieval comes solely from the measurements then the area equals 1, and at altitudes where the *a priori* profile is contributing to the retrievals the area decreases, where an area equal to 0 would mean nothing is being retrieved.

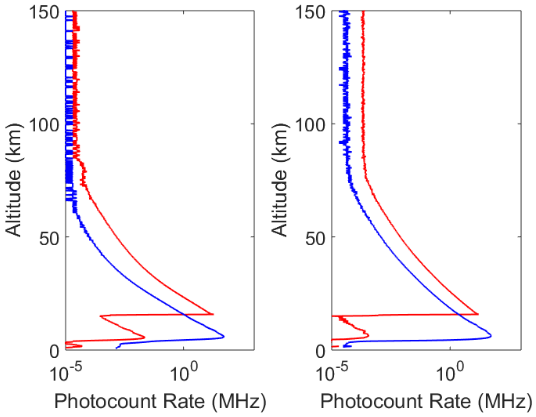

Figure 2.2: Average count rates for 5 hours of measurements on 26 July 2017. Left panel:“online” wavelength (blue curve, low altitude; red curve, high altitude). Right panel: “off-line” wavelength (blue curve, low altitude ; red curve, high altitude).

Figure. 2.3 shows the averaging kernels for the ozone density. The red line shows that the averaging kernel for ozone density equals 1 up to 42.7 km, thus below this altitude the retrieval is independent of the *a priori* profile. Ozone is a minor constituent in the atmosphere; due to the poor SNR of signals at higher altitudes, the sensitivity of the averaging kernel decreases. Here, the retrieval falls back to the *a priori* values.

In a good retrieval, the difference between the forward model and the measurements, which is called the residual, should be within the uncertainty of the measurements. Figure. 2.4 shows the residual plots, which confirm that our forward model has correctly characterized the physics of the atmosphere and is capable of retrieving the quantity of interest.

The OEM retrieval grid starts at 500 m and increases to 700 m at 18 km. The full width half maximum of the averaging kernel at each height is defined as the vertical resolution of the retrieval. At lower altitudes, the averaging kernel is broad, and the retrieval resolution is close to the spacing of the retrieval grid (for this specific retrieval around 500 m). As shown in Fig. 2.5 (right panel), by increasing the altitude, the retrieval resolution decreases consequently, such that at 40 km the resolution is 2.8 km. Traditionally, the vertical resolution decreased by height as well. Figure.2.5 (left panel) shows the vertical resolution of the retrieval in both traditional methods and the OEM. At the first 2 km of retrieval the OEM provides a better retrieval resolu-

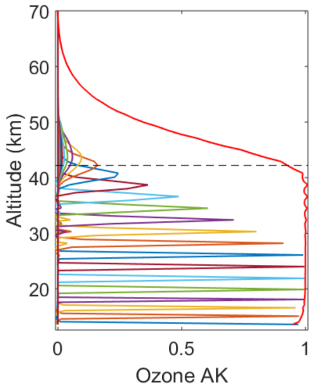

Figure 2.3: Averaging kernels for the ozone density for the measurements on 26 July 2017. The horizontal dashed line is a height below which the OEM retrievals is more than 80% due to the measurements. Above this horizontal cut-off as the SNR drops, the retrieval starts to fall back to the *a priori* profile. For clarity, the averaging kernels are only shown every 1500 m in altitude. The red line shows the summation of rows in the averaging kernel matrix at each altitude. The summation is of order unity below 42.7 km.

tion, however from 14.5 km to 17 km the traditional method has a better resolution. At around 17 km both methods show the same retrieval resolution; however, the traditional resolution decreases faster such that at 42.2 km the retrieval resolution is around 7 km. The trade-off between the retrieval resolution and the retrieval uncertainty should be considered when comparing the methods.

Having a poorer vertical resolution leads to a better (that is, smaller) retrieval uncertainty. As shown in Fig. 2.5 (right panel), the statistical uncertainty of the retrievals for the traditional method is around 12% at 15 km (where the vertical resolution is 200 m and the low altitude Rayleigh channel is used) and it decreases to less than 1% at 25 km (where the vertical resolution is around 2 km and the high altitude Rayleigh channel is used). In contrast, the statistical uncertainty of retrieval in the OEM is around 10% at 15 km (where the vertical resolution is 500 m) and decreases to 2.2% at 25 km (where the vertical resolution is 700 m).

To demonstrate the mentioned trade-off in the OEM, we increased the correlation length of the *a priori* from 1000 m to 1500 m in the lower altitudes (below 18 km) and from 1400 m to 5500 m in higher altitudes (above 18 km). As a result, the retrieval has a poorer vertical resolution and smaller retrieval uncertainties. Assuming a higher correlation length indicates that at each altitude, the retrieved ozone density is dependent on the ozone distribution above

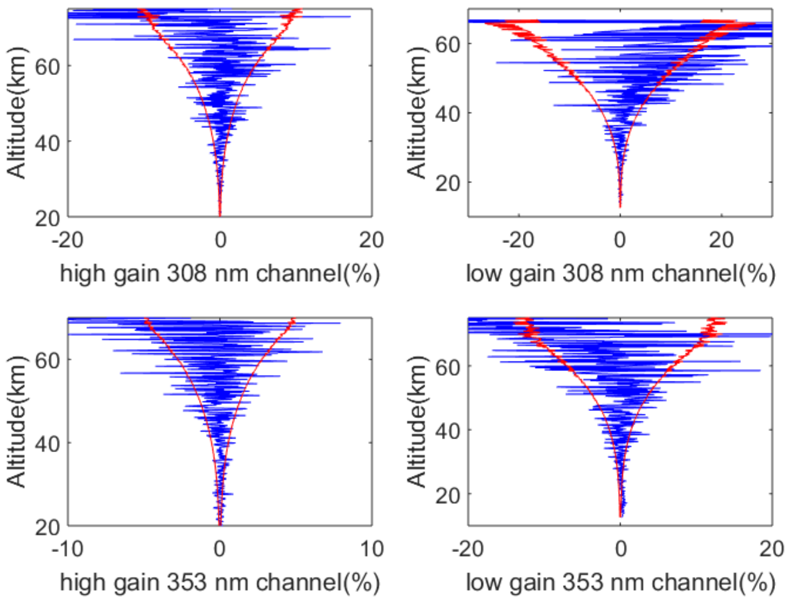

Figure 2.4: Residuals between the forward model and the measurements for the "on-line" and "off-line" channel (blue curves). The red line shows the uncertainty of the measurements.

and below the indicated altitude, thus, the retrieved ozone density looks smoother.

The vertical resolution and uncertainty for the traditional method as well as for the OEM with low and high correlation lengths are plotted in Fig.2.5.

In the traditional method, the relation between the final vertical resolution and the retrieval uncertainty is defined as follows:

$$\epsilon_s \propto (A \Delta {z_f}^3 P_0 t_a)^{-\frac{1}{2}} \tag{2.12}$$

where ϵ_s is the retrieval uncertainty, A is the area of the telescope, Δz_f is the final vertical resolution, P_0 is the emitted power, and t_a is the acquisition time (Godin et al., 1999). Assuming that the traditional method has the same vertical resolution as the OEM, using the above relation we can calculate the retrieval uncertainty which corresponds to the higher vertical resolution. Despite the difference in the vertical resolution values, at altitudes below 20 km, both the traditional method and the OEM have similar uncertainties (the difference is less than 1%). At altitudes above 20 km, assuming that the traditional method has the same vertical resolution as OEM, the retrieval uncertainty in the traditional method is calculated. Figure. 2.6 shows the comparison between OEM uncertainty and the modified traditional uncertainty for altitudes above 20 km. As is shown in the figure, from 20 km to 35 km the difference between the uncertainties is insignificant (less than 1%) and above 35 km the difference grows to 4.5%.

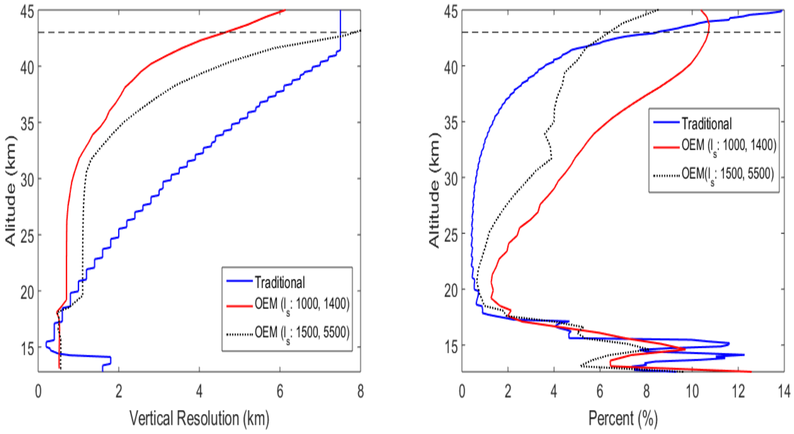

Figure 2.5: Right panel: The statistical uncertainty of the OEM with correlation lengths (ℓ_s = 1000 and 1400 m is plotted (red curve) against the statistical uncertainty of the OEM with correlation lengths ℓ_s = 1400 m and 5500 m (black dotted curve). Additionally, the uncertainty of retrieval in the traditional method (blue curve) is plotted. The retrieval uncertainties in the OEM and the traditional method can be compared. Left panel: The vertical resolution of the OEM with correlation lengths (ℓ_s = 1000 and 1400 m (red curve) is plotted against the vertical resolution of the OEM with correlation lengths ℓ_s = 1400 and 5500 m (black dotted curve). The vertical resolution of the traditional method is shown as well (blue curve). The horizontal dashed line indicated the maximum height at which the retrieval is independent from the *a priori*.

Figure 2.7 shows our retrieved ozone density compared to the sonde measurements and the traditional retrieval. Consistent with Fig. 2.5 we have plotted the OEM retrievals for two different sets of correlation lengths. The ozonesonde measurements have better vertical resolutions compared to the DIAL measurements, albeit with larger random uncertainty. Also, the sonde profiles show more vertical structure of the ozone distribution. Compared to the traditional retrieval, the OEM can successfully catch many of these variations.

As shown in Fig. 2.7 (right panel) results of a comparison between the two methods indicates that for higher altitudes (above 25 km) the difference between the two retrievals is insignificant. However, for lower altitudes (between 15 km to 21 km) the difference between the two methods becomes significant. Moreover, as shown in the left panel of Fig. 2.7, on this particular night, the difference between the two methods below 15 km can be as large as 60% relative to the ozonesonde, with the OEM retrieval in better agreement with the sonde

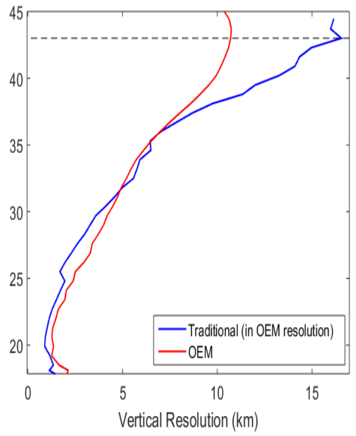

Figure 2.6: At height from 20 km to 40 km, the uncertainty of retrieval for the traditional method (assuming that it has a vertical resolution similar to the OEM vertical resolution) is plotted against the OEM retrieval uncertainty (blue curve: OEM; red curve: traditional). The horizontal dashed line indicated the maximum height at which the OEM retrieval is independent from the *a priori*.

at most heights below 21 km. For higher altitudes the two methods agree well with the sonde measurements.

To investigate the effect of *a priori* profiles on retrievals, the OHP climatology and the US standard model were used to retrieve ozone density (see Fig. 2.9). The OEM retrievals resulting from these two *a priori* profiles as well as the traditional retrieval are plotted in the left panel of Fig. 2.9. As shown in the right panel of this figure, below 35 km the difference between the two OEM retrievals is less than 0.5%. Above this altitude, the percentage difference between the two methods reaches 2.5% which is much smaller than the retrieval uncertainty at altitudes above 35 km. Thus, the choice of *a priori* has a small effect on the retrievals.

The OEM provides a complete systematic and statistical uncertainty budget. Fig. 2.10 shows the uncertainty of the ozone retrieval shown in Fig. 2.7. The forward model parameters, the Rayleigh cross sections, the ozone absorption cross section, and the temperature profiles assumed for the ozone cross section contribute to the systematic uncertainty of the retrieval. Below 20 km, these uncertainties are comparable with the statistical uncertainty; however, in the higher altitudes systematic uncertainties are less than 1%. The Rayleigh-scatter cross section uncertainty, at the bottom of the retrieval, is around 7% while at higher altitudes the uncertainty decreases to less than 1%. These values agree with the Rayleigh-scatter uncertainty of

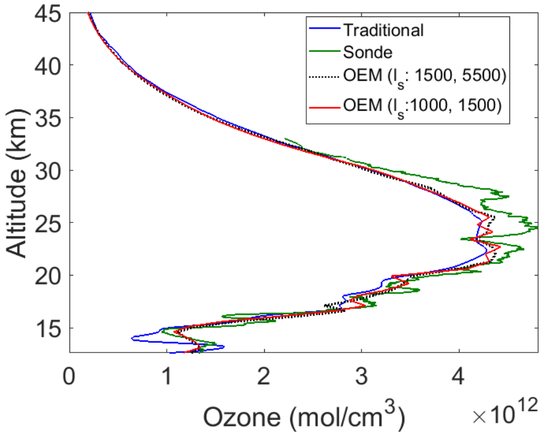

Figure 2.7: OEM ozone retrieval (red curve) from 20:07 UT to 00:15 UT on 26 July 2017 as well as the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve) are plotted. The dashed black line shows the OEM retrieval when the correlation length (ł_S) became larger. The horizontal dashed line shows the cut-off below which the effect of the *a priori* ozone profile is small less tan 10%.

8% which is calculated in the Leblanc et al. (2016b) uncertainty budget. The ozone absorption cross section for 308 nm channel reached a maximum of 4% at the bottom of the retrieval, which is higher than the calculated uncertainty of 1% in Leblanc et al. (2016b). The uncertainty due to temperature is less than 0.05%. The uncertainty due to the ozone absorption cross section at 355 nm channel is negligible as well.

The ozone retrieval extends from 12 km to 70.2 km. The averaging kernel of the air density extends much higher, as the air density contributes in both back-scattering coefficients and the extinction coefficient terms in the forward model. Therefore, in air density retrievals, the maximum height of acceptable retrieval is 70.2 km. However, we show the retrievals below 42.7 km to be consistent with the ozone density retrievals. As shown in Fig. 2.11 (left panel), the relative air density profile is retrieved as well.

To validate our result, we used the nitrogen Raman spectrum at 386.7 nm. The "off-line" wavelength is transmitted to the atmosphere at 355 nm channel, and the corresponding Raman wavelength is received at 386.7 nm channel. The Raman channel is not sensitive to the aerosol contents of the atmosphere, and the wavelength is not absorbed by ozone ("off-line" Raman channel). Thus, the atmospheric back scattering and extinction terms are mostly determined by the air density. This makes the Raman "off-line" channel a good candidate for our validation.

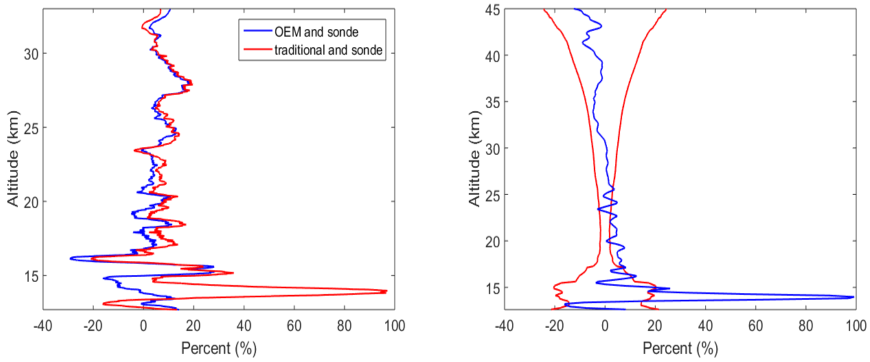

Figure 2.8: For the night of 26 July 2017. Left panel: The percentage difference between the OEM retrieval and the ozonesonde measurements in the form of: ($\frac{OEM-sonde}{sonde} * 100$) (blue curve); the percentage difference between the traditional retrieval and the ozonesonde measurements in the form of ($\frac{traditional-sonde}{sonde} * 100$) (red curve). Right panel: The percentage difference between the OEM retrieval and the traditional retrieval (blue curve); the summation of the statistical uncertainty of the traditional and OEM retrievals (red curve).

We can assume that $N(\lambda_{off}, z) \propto \frac{n_{air}}{z^2}$.

Using the above relation, the relative air density profile can be generated. The relative air density is scaled against the OEM retrieval of air density, and the percentage difference is calculated (Fig.2.11; center panel). As shown in the figure, the difference between the scaled relative air density generated from the Raman counts and the OEM relative air density is less than 10%. However, in higher altitudes (above 35 km) the difference can reach up to 50%. This difference is governed by the higher measurement noise in the Raman channel. This result provides confidence that the density retrieval is reasonable. The right panel of Fig.2.11 shows the uncertainty of the relative air density retrieval. For the air density retrieval the statistical uncertainty is small (around 0.1% at the bottom of the retrieval). The Rayleigh-scatter cross section uncertainty is small as well and the ozone absorption cross section uncertainties are negligible.

The OHP analysis employing the traditional method uses a different value of saturation correction for each wavelength. In our OEM code, we are retrieving 4 different dead times, each corresponding to one of the channels. For *a priori* values, we are using the provided x value which is discussed earlier in this section. As shown in Table 2.2, the retrieved dead time values for 26 July 2017 are similar to the provided x values. The only major difference is detected for the “on-line” low-altitude channel, where the x value is 4.6 ns and the retrieved

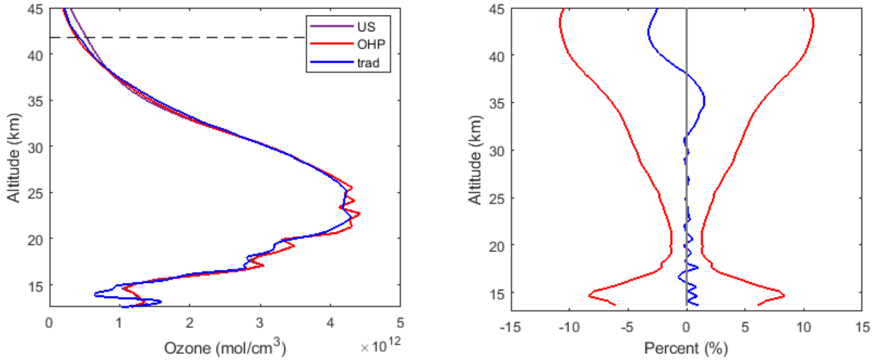

Figure 2.9: For the night of 26 July 2017. The left panel: the OEM retrieval using the US standard model as *a priori* profile (purple curve) and the OEM retrieval using the OHP climatology as *a priori* profile (red curve) are plotted. Furthermore, the traditional method retrieval (blue curve) is plotted, thus the OEM retrievals can be compared with each other and with the traditional retrieval. The right panel: Percentage difference between the OEM retrievals using the two different *a priori* profiles (blue curve) is plotted. This difference is with in the retrieval uncertainty. At higher altitudes (above 35 km), when the SNR drops, the difference between the two methods is less than 5%, which is smaller than the retrieval uncertainty at that height.

Dead time	OEM (ns)	*a priori (ns)*
“on-line” high-altitude	2.78 ± 0.55	2.80
“on-line” low-altitude	5.05 ± 0.92	4.60
“off-line” high-altitude	4.60 ± 0.92	4.60
“off-line” low-altitude	2.56 ± 0.51	2.50

Table 2.2: Dead time values which were calculated for each channel on the night of 26 July 2017.

value is 5.05 ns.

2.4.2 Further examples of the OEM retrieval method

Using the OEM, the retrieved profiles for the nights of 14 July and 20 July are plotted against the sonde measurements as well as the traditional ozone retrievals (Fig.2.12). The night of 14 July 2017 includes 4.5 h of measurements. The retrieval extends from 9.6 km to 40.2 km. Above 16 km, the difference between the two traditional methods and the OEM retrieval is within the statistical uncertainty of the measurements. Below 16 km the difference is about 15% with the OEM retrieval closer to the sonde measurements (Fig. 2.13). For 20 July 2017 the retrieval is the result of 4 hours of measurements. The ozone retrieval extends from 11 km

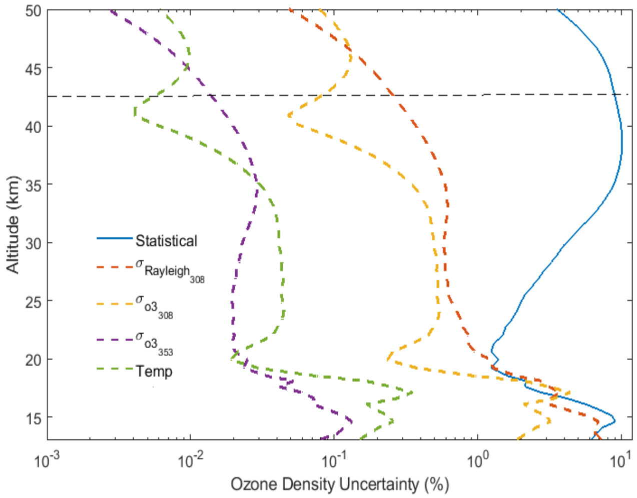

Figure 2.10: For the night of 26 July 2017. The statistical uncertainty of the OEM (blue), the Rayleigh-scatter cross section uncertainty at 308 nm (red), the ozone absorption cross section at 308 nm (orange), and the ozone absorption cross section for the 355 nm channel (purple). The horizontal dashed line shows the height below which the retrieval is independent of the *a priori* profile.

to 36.8 km. Our results indicate that the differences between the two methods are within the retrieval uncertainty. Thus, these two additional nights help to demonstrate that the OEM can produce ozone density profiles consistent with the traditional retrievals.

2.5 Conclusion

We have introduced a first-principle OEM retrieval for stratospheric ozone profiles applicable to stratospheric DIAL lidar measurements, and tested this method using measurements from the OHP stratospheric DIAL system. The discussion of the implementation of OEM for our retrievals is summarized below.

1. The forward model used in this study is capable of providing a robust estimate of the ozone profiles for clear nights.

2. Multiple measurements channels are used. The raw (uncorrected) photocounts are used for the retrieval, and no gluing process is needed. As a result, a single ozone profile consistent with all measurements is retrieved.

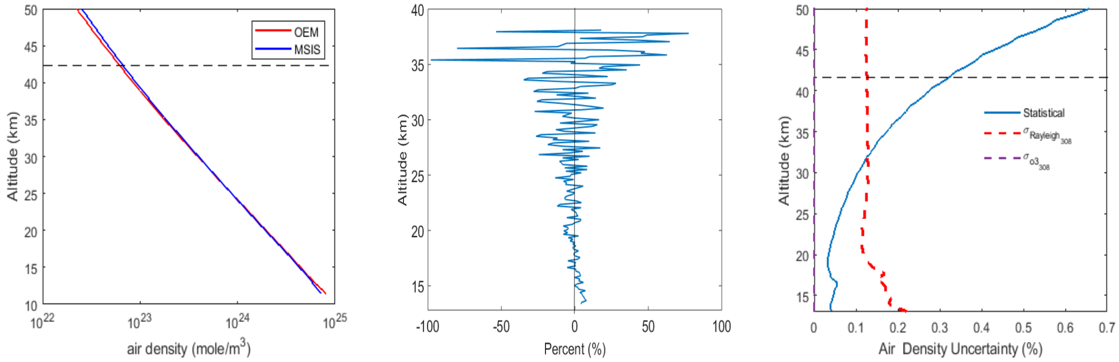

Figure 2.11: Left panel: The retrieved air density (blue line) is plotted against the *a priori* profile (red line). Mid panel: The percentage difference between the scaled relative air density generated from the Raman channel and the OEM air density retrievals. The difference is less than 10%. Right panel: The statistical uncertainty of the OEM retrieval of air density (blue), the Rayleigh-scatter cross section uncertainty for the 308 nm channel (red), and the ozone absorption cross section in both channels (purple).

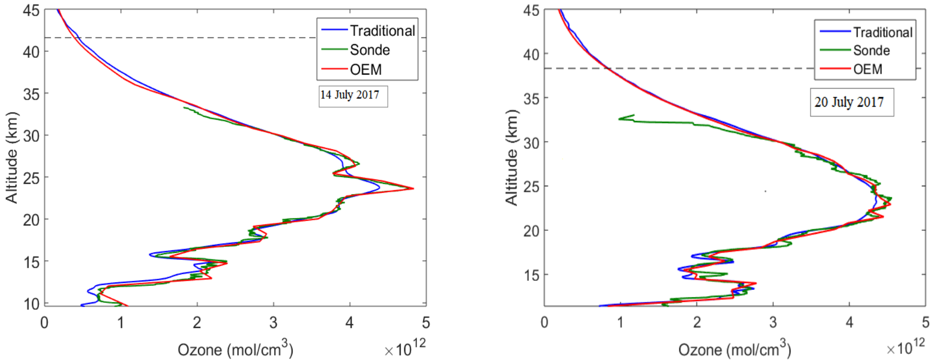

Figure 2.12: Left Panel: OEM ozone retrieval on the night of 14 July 2017 (red curve) compared to the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve). Right Panel: OEM ozone retrieval on the night of 20 July 2017 (red curve) compared to the ozonesonde profile (green curve) and the traditional ozone retrieval (blue curve). These cases demonstrate the high resolution of the OEM technique as evidenced by the excellent agreement around the ozone peak with the sonde measurement.

3. The OEM is applied to the OHP lidar measurements for three different nights in July 2017, all of which had coincident ozonesonde launches. Comparison with the radiosondes was good.

4. The OEM's averaging kernels allow the contribution of the *a priori* relative to the measurements to be accessed as a function of altitude, as well as allowing better comparison with other instrument.

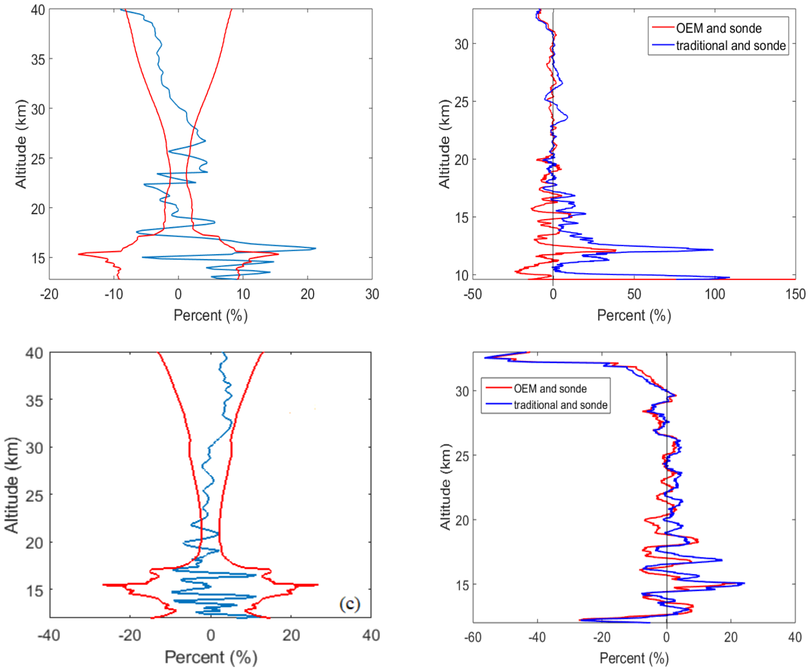

Figure 2.13: For the night of 14 July 2017, (a) the percentage difference between the traditional method and the OEM retrieval (blue curve) plotted within the envelope of the total statistical uncertainty of the two method (red curve). The agreement between the two lidar ozone determinations are within the statistical uncertainty above 17 km. (b): The red curve is the percentage difference between the OEM retrieval and sonde measurements. The blue curve is the percentage difference between the traditional method and sonde measurements. Figures (c) and (d) are the same format as (a) and (b) for the night of 20 July 2017.

5. The OEM and the traditional method are show good agreement, and for most heights their difference is small.

6. Increasing the correlation length in the retrieval allows the vertical resolution to be degraded and the statistical uncertainty decreased. Comparisons with the OEM retrievals at degraded resolution showed agreement to the traditional method to within the measurements statistical uncertainty.

7. The OEM provides a full uncertainty budget. Thus, using the OEM, for each individual retrieved profile both statistical and systematic uncertainties are calculated. The systematic uncertainties are compared with the uncertainty budget for the traditional method given by (Leblanc et al., 2016a) and are similar.

Currently we are working on a retrieval which can use measurements from both the OHP tropospheric and stratospheric lidars which will allow us to retrieve ozone profile from just above the boundary layer throughout the stratosphere. Also, we plan to include the Raman measurements into our forward model, allowing the retrieval of the ozone profiles in the presence of strong aerosol layers and thin clouds. Also, we are planning to apply our OEM retrieval to the last three decades of OHP measurements. Applying the OEM to the entire OHP lidar ozone profile database will provide an improved statistical evaluation of the differences between the traditional and the OEM methods, as well as allowing improved ozone estimates in the upper troposphere and lower stratospheric region.

Bibliography

Eriksson, P., Jimenez, C., and Buehler, S.: Qpack, a general tool for instrument simulation and retrieval work, J. Quant. Spectrosc. Radiat. Transfer, 91, 47–64, 2005.

Godin, S., Carswell, A. I., Donovan, D. P., Claude, H., Steinbrecht, W., McDermid, I. S., McGee, T. J., Gross, M. R., Nakane, H., Daan, Swart, P. J., Bergwerff, B. B., Uchino, O., von der Gathen, P., and Neuber, R.: Ozone differential absorption lidar algorithm intercomparison, Appl. Opt., 38, 6225–6236, 1999.

Godin-Beekmann, S., Porteneuve, J., and Garnier, A.: Systematic DIAL lidar monitoring of the stratospheric ozone vertical distribution at Observatoire de Haute-Provence (43.92[degree]N, 5.71[degree]E), J. Environ. Monit., 5, 57–67, 2003.

Krueger, A. J. and Minzner, R. A.: A midlatitude ozone model for the 1976 US Standard Atmosphere, Journal of Geophysical Research: Atmospheres (19842012), 81, 44774481, 1976.

Leblanc, T., Sica, R. J., van Gijsel, J. A. E., Godin-Beekmann, S., Haefele, A., Trickl, T., Payen, G., and Gabarrot, G.: Proposed standardized definitions for vertical resolution and uncertainty in the NDACC lidar ozone and temperature algorithms Part 1: Vertical resolution, amt, 9, 4029–4049, 2016a.

Leblanc, T., Sica, R. J., van Gijsel, J. A. E., Godin-Beekmann, S., Haefele, A., Trickl, T., Payen, G., and Liberti, G.: Proposed standardized definitions for vertical resolution and uncertainty in the NDACC lidar ozone and temperature algorithms Part 2: Ozone DIAL uncertainty budget, amt, 9, 4051–4078, 2016b.

Malicet, J., Daumont, D., Charbonnier, J., Parisse, C., Chakir, A., and Brion, J.: Ozone UV spectroscopy. II. Absorption cross-sections and temperature dependence, J. ATMOS. CHEM., 21, 263–273, 1995.

McDermid, I. S., Godin, S. M., and Walsh, D.: Lidar measurements of stratospheric ozone and intercomparisons and validation, Appl. Opt., 29, 4914–4923, 1990.

Nicolet, M.: On the molecular scattering in the terrestrial atmosphere: An empirical formula for its calculation in the homosphere, Planet. Space Sci., 32, 1467–1468, 1984.

Orphal, J., Staehelin, J., Tamminen, J., Braathen, G., De Backer, M., Bais, A., Balis, D., Barbe, A., Bhartia, P. K., Birk, M., et al.: Absorption cross-sections of ozone in the ultraviolet and visible spectral regions: Status report 2015, J. Mol. Spectrosc., 327, 105–121, 2016.

Pelon, J. and Mégie, G.: Ozone monitoring in the troposphere and lower stratosphere: Evaluation and operation of a ground-based lidar station, J. Geophys. Res. Oceans, 87, 4947–4955, 1982.

Rodgers, C. D.: Inverse methods for atmospheric sounding: theory and practice, vol. 2, World scientific, 2000.

Chapter 3

Improved ozone UTLS DIAL measurements

3.1 Introduction

The upper troposphere and lower stratosphere (UTLS) region extends from about 6 km to 25 km in height and plays a significant role in the atmospheric climate system. In this region of the atmosphere, even small changes in temperature and in the distribution and concentration of greenhouse gases can result in large changes in atmospheric radiative forcing, which can trigger climate change IPCP (2007); Logan (1985).

Ozone in the upper troposphere acts as the third largest greenhouse gas contributing to the radiative forcing of climate change Ramaswamy et al. (2001); IPCP (2007). The ozone distribution in the UTLS is the result of transport mechanisms and photochemical reactions. Because of stratospheric tropospheric exchange, large spatial and temporal variability can be observed in the UTLS Forster et al. (1997).

In many studies on the UTLS ozone, satellite-borne instruments are used. In limb-viewing instruments, the elevation angle of the line-of-sight varies during the measurements. As a result, limb sounders can provide good vertical resolution (about 2 km to 4 km). However, at lower altitudes (lower troposphere), the atmosphere becomes nearly opaque, and the limb-viewing instruments have difficulties measuring trace gases. Nadir-viewing instruments can provide measurements in the lower troposphere, but their vertical resolution is limited (about

6 km to 7 km). Occultation instruments use the Sun or other stars as the source of radiation, and they can obtain measurements with higher vertical resolution (about 1 km to 2 km). Solar occultation instruments are restricted by the number of sunsets and sunrises they encounter in one orbit, while stellar occultation instruments are limited by the weakness of the stellar source compared to the Sun. The combination of measurements from different geometrical-based satellite instruments has been used to measure ozone density.

The OEM methodology we describe will allow improved estimates of UTLS ozone and the associated random and systematic uncertainties, as well as provide averaging kernels for the lidar measurements. The availability of averaging kernels will improve future intercomparisons between ground-based, balloon-borne and space-based ozone measurements.

Differential Absorption Lidar (DIAL) systems provide ozone measurements with high vertical and temporal resolutions. For example, observatories such as the Canadian Network for the Detection of Atmospheric Change (CANDAC) Polar Environment Atmospheric Research Laboratory (PEARL) in Eureka, Maïdo observatory in Reunion Island, the Observatoire de Haute Provence (OHP) in France, and the NASA Table Mountain Observatory (TMO) in the United States are equipped with both tropospheric and stratospheric lidars. At the Eureka observatory, the tropospheric lidar system makes measurements from 0.5 km to about 8 km in altitude and the stratospheric lidar system operates from about 4 km to 35 km (**??**). At the Maïdo observatory, the tropopspheric DIAL makes measurements from 6 km to 16 km, and the stratospheric DIAL operates in the 13 km to 38 km region Baray et al. (2013). At the OHP observatory, the tropospheric DIAL system operates from 2.5 km to about 14.5 km, and the stratospheric DIAL operates from about 10 km to 45 km Gaudel et al. (2015); Godin-Beekmann et al. (2003). At the TMO, the tropospheric DIAL system obtains measurements from 3 km to 18 km, and the stratospheric DIAL system from 10 km to 40 km Megie et al. (1985); Mérienne et al. (2001). Although these systems can produce satisfactory ozone profiles in their overlapping region (from tropospheric lidar to stratospheric lidar), the uncertainty of merging is not well defined. Providing a single ozone profile with a full uncertainty budget using both sets of measurements can significantly improve our measurements of ozone in the UTLS Holton et al. (1995); Stohl et al. (2003); Cohen et al. (2018).

Here we apply the Optimal Estimation Method (OEM) to tropospheric and stratospheric

DIAL measurements. Measurements from these two systems are simultaneously used by the retrieval to obtain a single ozone profile. Using the OEM there is no need to "merge" or "glue" Level 0 profiles. Moreover, the input measurements can be in different units with different measurement grids (for example a mix of analog and digital measurements). Additionally, a full uncertainty budget, including both the systematic and statistical uncertainties, is calculated for each individual profile. The OEM also provides averaging kernels of the retrievals, which allows comparison of the profiles with other measurements which can account for differences in vertical resolution, such as when compared to space-based measurements. Other atmospheric and systematic parameters such as air density, the dead time of the system, and the background counts can be retrieved along with ozone profiles. The application of OEMs to aerosol lidar measurements, Rayleigh scatter temperatures, and Raman scatter water vapour retrievals has been studied and discussed in detail Povey et al. (2014); Sica and Haefele (2015, 2016). In addition, we have recently demonstrated an OEM for DIAL stratospheric ozone retrievals (Farhani et al., 2018), which we will now expand to include measurements from tropospheric ozone DIAL systems.

In this chapter, focusing on the UTLS region, we show a first principle OEM to retrieve a single ozone profile by using both tropospheric and stratospheric DIAL measurements directly from the raw (Level 0) measurements using the lidar equation as the forward model. In Section 3.2, pre-processing steps prior to applying the traditional DIAL algorithm, as well as the OEM, are discussed. Moreover, the state vectors and the **b** parameter quantities are defined and a brief overview of the lidar's specifications is given. In Section 3.3 results of the OEM retrieval, using both tropospheric and stratospheric lidar measurements, are discussed in detail. In this Section, we also show our results of comparison between the ozonesonde profiles and our retrievals. Details of how to apply our method to a standalone tropospheric DIAL are given in the 3.4. Section 3.5 is the summary of the chapter, and in Section 3.6 we discuss our future plans.

3.2 METHODOLOGY

In the DIAL system, two wavelengths are simultaneously transmitted to the atmosphere. One of the emitted wavelengths is strongly absorbed by the constituent of interest (called the “on-line” wavelength) and the other is weakly absorbed (called the “off-line” wavelength). For ozone measurements, selecting a wavelength pair depends on the altitude range of the measurements. For most studies, the ultraviolet (UV) spectrum is the most efficient spectral region. A pair of wavelengths with strong UV absorption is needed to detect the small amount of ozone which resides in the troposphere. However, for stratospheric ozone measurements, choosing a laser with a longer wavelength that can reach higher altitudes in the stratosphere is the main concern Megie et al. (1985); Browell (1989); Papayannis et al. (1990).

The traditional analysis method for ozone number density uses the derivative of the ratio between the “on-line” and “off-line” channels to calculate the ozone number density $n_{o_3(z)}$. A detailed discussion on the tropospheric and stratospheric ozone retrievals can be found in Ancellet et al. (1989); Godin et al. (1999); McDermid et al. (1990); Godin-Beekmann et al. (2003); Leblanc et al. (2016). In the traditional analysis, some corrections are applied to the raw lidar measurements, for example background counts should be removed. In many systems this requires including the effects of signal-induced-noise (SIN). Any corrections due to nonlinearity of the counting system (because of saturation) should also be applied to the raw counts. Finally, the signals from different channels need to be merged to form a single measurement profile. This corrected count profile is then used to calculate the ozone number density profiles. With the OEM, a forward model encapsulates the geophysical properties and instrumental characteristics of the system, and our OEM retrieval uses the raw (Level 0) measurements from all available channels. Unlike the traditional method, the OEM does not require corrections to the raw measurements. Furthermore, the measurements from different channels are not glued (or merged), but are input directly into the OEM routine as the measurement input vector. The dead time of the system and the background values for the “on-line channel and the “off-line channel are part of the state vector, while the overlap function is a model parameter whose contribution to the retrieved ozone profile is assessed in the uncertainty budget. The dead times, backgrounds, and lidar constants are simultaneously retrieved along with the ozone number density and air number density profiles. A comprehensive explanation of the OEM can be found in Rodgers (2000); a brief description of the OEM follows below.

The OEM is an inverse modeling technique which is based on Bayes' theorem. In the OEM a forward model is defined as the relation between the measurement vector $\mathbf{y} = (y_1, y_2, ..., y_n)$, and the state vector $\mathbf{x} = (x_1, x_2, ..., x_n)$. The forward model is:

$$\mathbf{y} = F(\mathbf{x}, \mathbf{b}) + \boldsymbol{\epsilon} \tag{3.1}$$

where $\mathbf{b}$ are the forward model parameters, which are assumed to be known, and $\boldsymbol{\epsilon}$ is the measurement noise.

We use the lidar equation as the forward model, where the raw counts are the measurements. The lidar equation for unsaturated counts, N_{true}, is:

$$N_{true}(z, \lambda_i) = \frac{C(\lambda_i)O(z)}{z^2}\beta(\lambda_i, z)\exp[-2\int_0^\infty [\sigma_{O_3}(\lambda, T(z))n_{O_3}(z)+\alpha(\lambda, z)+\sum_e \sigma_e(\lambda)n_e(z)]dz]+N_b(z, \lambda_i) \tag{3.2}$$

where $N_{true}(z, \lambda_i)$ is the number of backscattered photons. $C(\lambda_i)$ is the lidar constant, which contains the area of the receiving telescope, the total efficiency of the lidar system, and energy of the scattered photon. The geometrical overlap is $O(z)$, and $\beta(\lambda_i, z)$ are the atmospheric backscattering coefficients which includes both molecular and aerosol terms. The first term inside the integral corresponds to ozone absorption in which $\sigma_{O_3}(T(z), \lambda_i)$ is the ozone absorption cross section, which depends on atmospheric temperature, and $n_{O_3}(z)$ is the ozone number density. The second term of the integral, $\alpha(\lambda, z)$ contains the extinction coefficient which is the sum of the extinction due to molecules and particles, and the last term $\sum_e \sigma_e(\lambda)n_e(z)$ is the extinction by other absorbers. For ozone studies, the most common interfering gases are SO_2 , NO_2 and O_2. The effect of O_2 is only considered when the selected "on-line" laser wavelength is shorter than 294 nm Fally et al. (2000); Mérienne et al. (2001). In the case of heavy volcanic eruption, SO_2 and NO_2 can significantly affect ozone retrievals Heath et al. (1983). However, in most cases, for both stratospheric and tropospheric ozone studies the effect of these gases in final ozone retrievals is negligible. Thus, the last term of integration is typically neglected Godin-Beekmann et al. (2003).

The background counts are written as $N_b(z)$. In the presence of SIN, the background is fitted to an exponential function of the form:

$$N_b(z) = a\exp(-bz) + c \tag{3.3}$$

where a, b, and c are coefficients of the fit, which in the traditional method are determined analytically, but are retrieved in our OEM retrieval using the analytic values as *a priori* coefficients Hunt and Poultney (1975).

When the intensity of the backscattered signal is high, the counting system can be affected by saturation. This saturation can result in an observed count rate which is less than the true count rate (N_{true}). For a paralyzable detector, true counts are related to the observed counts N_{obs} as follows:

$$N_{obs} = N_{true}\exp(-\kappa N_{true}) \tag{3.4}$$

and, for non-paralyzable detectors, the following equation can be used:

$$N_{obs} = \frac{N_{true}}{1 + \kappa N_{true}} \tag{3.5}$$

where κ is the dead time of the detecting system. For the OEM retrieval the value of the dead time for each channel is retrieved.

3.2.1 Implementing the OEM for the OHP lidars

Knowledge of the measurement vector $\mathbf{y}$ and its covariance matrix, $\mathbf{S}_y$, along with *a priori* values of the state vector, $\mathbf{x}_a$, and its associated covariance matrix, $\mathbf{S}_a$, enables the OEM to calculate an optimal *a posteriori* state by minimizing a cost function with respect to $\mathbf{x}$ given by:

$$Cost = (\mathbf{y} - \mathbf{Kx})^T\mathbf{S}_y^{-1}(\mathbf{y} - \mathbf{Kx}) + (\mathbf{x} - \mathbf{x}_a)^T\mathbf{S}_a^{-1}(\mathbf{x} - \mathbf{x}_a) \tag{3.6}$$

where $\mathbf{K} = \frac{dF}{dx}$, the Jacobian matrix, is the linearisation term for the nonlinear forward model. In the OEM it is assumed that the measurement noise is described by a normal distribution, but lidar photocount measurements follow a Poisson distribution. Since at most heights the number of photocounts is large, the two distributions are indistinguishable. Thus, we assumed that the photon counts are distributed normally. This assumption causes the residuals to become biased at very low count rates, but this bias occurs at altitudes far above where the retrieval is valid as defined by the response function, the area of the averaging kernel matrix which is of order unity, when the retrieval depends fully on the measurement rather than the *a priori* profile.

As our forward model is nonlinear, an iterative numerical method is used. For our problem the Levenberg-Marquardt iteration is a suitable numerical method. Then, the optimized solution for the state vector **x** is given as:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + [(1+\gamma_i)\mathbf{S}_a^{-1} + \mathbf{K}_i^T\mathbf{S}_y\mathbf{K}_i]^{-1}([\mathbf{K}_i^T\mathbf{S}_y^{-1}(\mathbf{y} - F(\mathbf{x}_i, \mathbf{b})] - \mathbf{S}_a^{-1}(\mathbf{x}_i - \mathbf{x}_a)) \tag{3.7}$$

where i is the iteration term, and γ_i is a damping factor for the iteration, which is chosen at each step to minimize the cost function. As suggested by Fletcher Fletcher (2013) if the value of the cost function increases in a step, γ_i will increase by a factor of 10, and if the value of the cost function decreases in a step, γ_i will decrease by a factor of 2. The iteration stops when the cost function decreases to a value much smaller than the number of measurements. There are other criteria which result in ceasing the iteration. Further details can be found in Rodgers (2000).

To understand how measurements and *a priori* profiles contribute in the final retrievals, an averaging kernel can be used. The relation between the retrieved state and the true state is described by the averaging kernel of the retrieval. The averaging kernel is calculated as:

$$\mathbf{A} = \frac{d\widehat{\mathbf{x}}}{d\mathbf{x}} = [\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}]^{-1}\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} \tag{3.8}$$

The retrieved quantity $(\widehat{\mathbf{x}})$ can be written as follows:

$$\widehat{\mathbf{x}} = (\mathbf{I} - \mathbf{A})\mathbf{x}_a + \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon_r} \tag{3.9}$$

where $\boldsymbol{\epsilon_r}$ is the retrieval uncertainty and **I** is a unity matrix. A perfect retrieval, in the sense all the information comes from the measurement with no effect from the *a priori* state, has averaging kernels equal to one, where the first term of the above equation becomes zero. The width of the averaging kernel gives the resolution of the retrieval at each height, here defined as the Full Width Half Maximum (FWHM) of each averaging kernel.

In order to find the state vector (from Eq. 3.7) the following quantities should be known: the measurements and their covariances, the *a priori* profiles, the *a priori* profile's covariance, and the model (**b**) parameters. The **b** parameters are quantities in the forward model that are not being retrieved, because they are either well-known or retrieving them is not possible. The uncertainty associated with the retrieval due to the **b** parameters is calculated after the last iteration of the solution. We used the Qpack package to perform the retrievals. Qpack is a free software package which written in MATLAB, and is part of the ATMOS package developed for retrieval of geophysical quantities from passive instruments in the millimeter and sub-millimeter wavelength regions. Here, we have developed our own forward model, and we have also calculated the Jacobians ourselves. After assigning the *a priori* profiles, **b** parameters and their associated uncertainties, we then use the Qpack OEM solver (oem.m) to retrieve the state vector using the Marquardt-Levenberg technique. Details of the Qpack software are given by Eriksson et al. Eriksson et al. (2005).

Here we retrieve the ozone density profile, relative air density, dead time values, and background counts. Overlap functions, ozone cross sections, and Rayleigh scattering cross sections are considered as **b** parameters in the forward model. Below, we discuss our choices of *a priori* profiles and **b** parameter values. The covariance matrices associated with the measurements and *a priori* profiles are discussed as well, and these values are summarized in Table 3.1.

In photon counting mode, when the signal is linear, the measurements statistical uncertainty follows a Poisson distribution, and the number of counts at each altitude represents the measurement's variance at that height. There is no correlation between the digital counts in different layers of the atmosphere, so the off-diagonal elements of the measurement's covariance matrix are zero. However, for the OHP lidars, analog measurements do not follow Poisson distributions. Calculating the measurement variance for each measurement point requires selecting n points before and after the specified point, and then fitting a straight line to these $2n+1$ points, which is then removed. Next, the residual variance is calculated. For our measurements, we tried different values for n, and $n = 3$ provided the best fit for our measurements.

In order to determine the background counts in both the tropospheric and stratospheric measurements, the mean of the counts above a specific height is calculated and used as the *a priori* for the “off-line” channels, since SIN is negligible in these channels. The variance of the background counts divided by the number of bins in the selected region is the uncertainty for the background *a priori* value. For “on-line” channels where SIN is present, an exponential function is fitted to the signal and the coefficients of the fit are used as *a priori* values. An uncertainty of 20% is assigned to these coefficient *a priori* values. The altitude above which the background counts are determined is different for tropospheric and stratospheric lidars. Also, as the laser power in the “online” channel is about 2 times stronger than the laser power in the “off-line” channel, the effect of SIN in the “on-line” channel is more pronounced. The values we chose for the OHP lidars measurements are shown in Table 3.1.

We retrieve the logarithm of ozone density, $q(z) = \ln\{n_{o_3}(z)\}$, so in Eq. 3.2, n_{o_3} is replaced by e^q, as is commonly done for atmospheric retrievals (e.g. Deeter et al. (2007) and Sica and Haefele (2016)). Retrieving the logarithm of the number density is numerically more favourable to show small changes in large numbers. The U.S standard model is used for the *a priori* ozone profile (Krueger and Minzner, 1976). An uncertainty of 20% is assigned to this profile.

We retrieve the air density for both tropospheric and stratospheric measurements. However, below 15 km the air density profile retrieved is in fact a convolution of air density and aerosol load. Mass Spectrometer Incoherent Scatter Radar (MSIS) total density profiles are used as *a priori* profiles, and an uncertainty of 20% is assigned to it (Hedin, 1991). To generate a full length *a priori* covariance matrix for both air and ozone density profiles at altitudes below 12 km, a tent function with a correlation length of 300 m was used. At altitudes above 12 km the correlation length was increased to 900 m. This choice of correlation length is because above 12 km, the stratospheric lidar measurements have the most weight and the retrieval grid for these measurements starts at 300 m. Using the forward model, the *a priori* of the lidar constants for both tropospheric and stratospheric lidars are estimated. We assign a 10% uncertainty to the *a priori* of the lidar constants to account for changes with time of laser power, atmospheric transmission, and system efficiency.

The overlap function, Rayleigh cross sections, and ozone absorption cross sections are assumed as **b** parameters. Although these parameters are not being retrieved, the uncertainty

Parameter	Value	Standard Deviation
Stratospheric lidar measurements	measured	Poisson statistics
Tropospheric lidar measurements (digital channels)	measured	Poisson statistics
Tropospheric lidar measurements (analog channels)	measured	3-point running standard deviation
Retrieved quantities	*a priori* profiles	Standard Deviation
Ozone density	U.S standard model	20%
Air density	MSIS	20%
Deadtime	empirical fitting	20%
Background for stratospheric measurements (“off-line”)	mean above 80 km	σ above 80 km
Coefficients of SIN for stratospheric measurements (“on-line”)	empirical fitting above 80 km	20%
Background for tropospheric measurements (“off-line”) digital channel	mean above 20 km	σ above 20 km
Background for tropospheric measurements (“off-line and “on-line”) analog channel	mean above 12 km	σ above 12 km
Coefficients of SIN for tropospheric measurements (“on-line”) digital channel	empirical fitting above 12 km	20%
Lidar constants	estimate from FM	20%
Forward model parameters	Sources	Standard Deviation
Rayleigh-scatter cross section	Eberhard Eberhard (2010)	2%
Ozone absorption cross section	BDM Malicet et al. (1995)	2%
Overlap function	available at Dataset Citation 3.6.1	10%

Table 3.1: Values and associated uncertainties for the retrieved and forward model parameters.

associated to them contributes to the final uncertainty budget. The assigned values for the overlap function, Rayleigh cross sections and ozone cross section along with their standard deviations are listed in Table (3.1).

3.2.2 Description of the Lidars

The OHP Observatory (44°N, 6°E) has made routine measurements of ozone in the free troposphere and stratosphere for the last three decades using two lidar systems, a tropospheric DIAL and a stratospheric DIAL. The transmitter for the tropospheric system uses the fourth harmonic of a Continuum Nd:YAG laser (266 nm) frequency shifted by Raman Stimulated Scattering in a D2 high pressure cell Ancellet and Beekmann (1997). The DIAL measurement makes use of the 1st and 2nd Raman Stokes lines at 289 nm (the “on-line” wavelength) and 316 nm (the “off-line” wavelength). Backscattered photons are collected by a Cassegrain telescope which is equipped with a 80 cm mirror. For the spectral separation of the two backscatter signals the collected signal is passed from the telescope to a spectrometer.The signals are detected by a photomultiplier tube (PMT). The system configuration is bi-axial, and the distance between the laser and the telescope axes is 0.5 m. The overlap, $O(z)$, is significant from the surface to about 4 km Halldorsson and Langerholc (1978).

The stratospheric lidar system uses an XeCl excimer laser at 308 nm, with a repetition rate of 100 Hz. This laser has an output energy of about 200 mJ for the “on-line” channel, while the “off-line” wavelength at 355 nm is generated by the third harmonic of a continuum Nd:YAG laser with an energy of 50 mJ at 50 Hz. The backscattered signal is collected by four Newtonian telescopes, each with diameter of 0.5 m. The collected signal is sent to a spectrometer which separates the signal into four wavelengths. Two of these correspond to the emitted Rayleigh signals at 308 nm and 353 nm. The other two correspond to the Nitrogen Raman shifted spectrum at 331.8 nm and 386.7 nm, respectively. The Rayleigh signals are separated further into high- and low-gain photomultiplier channels. Hence, in total 6 photocount profile measurements are obtained. Further details on the OHP tropospheric and stratospheric lidars can be found in Ancellet and Beekmann (1997); Gaudel et al. (2015); Godin-Beekmann et al. (2003).

3.3 OEM ozone retrieval in the free troposphere and stratosphere

In this section we present the result of combining the two lidar measurements to retrieve a single ozone profile. We choose measurements from 12 July 2017 as this night has both clear skies and coincident measurements from the NDACC-LAVANDE intercomparison campaign. The results for the nights of 14 and 26 July 2017 are presented as well (see Table. 3.2). Our first example retrieval will be from the night of 12 July 2012, where the tropospheric lidar operated from 2049 to 2357, and the stratospheric lidar operated from 2106 to 0142 (all local time). An

Dates	12 July 2017	14 July 2017	26 July 2017
Tropospheric lidar	2049 – 2357	2350 – 0219	2140 – 2237
Stratospheric lidar	2106 – 0142	2151 – 0221	2007 – 0016
Ozone sonde	2153	2348	2133

Table 3.2: Measurement periods for the tropospheric and stratospheric lidars systems and launch times for the ozonesondes.

ozonesonde was launched at 2153 from the OHP station and the tropopause height was at 14.7 km.

For the tropospheric lidar system, the native resolution of measurements is 7.5 m for the analog channels and 30 m for the digital channels. For the stratospheric lidar system, the native resolution of measurements for all six channels is 150 m. Our retrieval starts at 2.6 km with a resolution grid of 150 m. At 11 km the retrieval grid changes to 500 m, and at 21 km it changes to 1500 m, and finally, at 25 km height, it becomes 1700 m. We chose these retrieval grids to be closer in vertical resolution to traditional retrieval grids.

The averaging kernels calculated for the retrieval are shown in Fig. 3.1. The shape of the averaging kernels define the sensitivity of the retrieval to the true state. As shown in Eq. 3.9, when the averaging kernel equals 1, the retrieval is sensitive only to the measurements. The vector area of the averaging kernels is defined as $\mathbf{Au}$, where $\mathbf{u}$ is a unity vector. When the area is close to 1, the retrieval is mostly independent of its *a priori* value. In our retrieval the averaging kernel has an area of about 1 up to 42.2 km, indicating the retrieval is mostly independent of the *a priori* profile. At about 11 km the stratospheric measurements are added to the retrieval. The stratospheric measurements have much lower noise at these heights than the tropospheric measurements. This rapid change in the measurement variance results in the spike seen in the response function. This behaviour of the response functions is well known in satellite OEM retrievals, and, if severe, can be mitigated by adjusting the covariances in the transition region. For our retrievals the perturbation region is narrow in height and does not cause any significant variation in the ozone profile retrieved in the UTLS, so we did not adjust the retrieval. We plan to investigate ways to smooth the response function in the hand-off region as we improve our retrieval.

The residual plots, which show the difference between the forward model and the actual

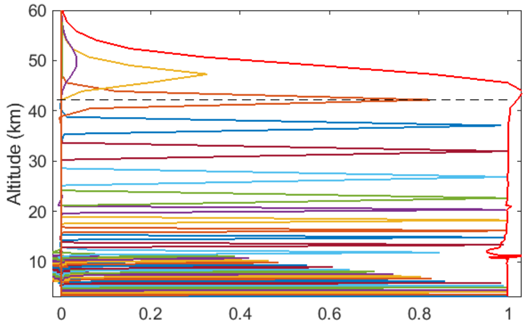

Figure 3.1: Averaging kernels for tropospheric-stratospheric ozone measurements. The averaging kernels are only shown every 450 m for lower altitude (from 2.5 km to 11 km height), and every 1500 m in higher altitudes for clarity. As is shown by the response function (red curve), the area of the averaging kernel matrix has a small spike at 11 km, when the stratospheric ozone measurements are included.

measurements, for both the tropospheric and stratospheric lidar are shown in Fig. 3.2. The four plots on the left are the residuals for the stratospheric measurements, and the four plots on the right are residuals for the tropospheric measurements. As shown in the figure, for both low-altitude and high-altitude channels the forward model can successfully encapsulate the physics of the atmosphere and the characteristics of the lidars, and, up to 50 km, the difference between the forward model and the actual measurements is less than 5%.

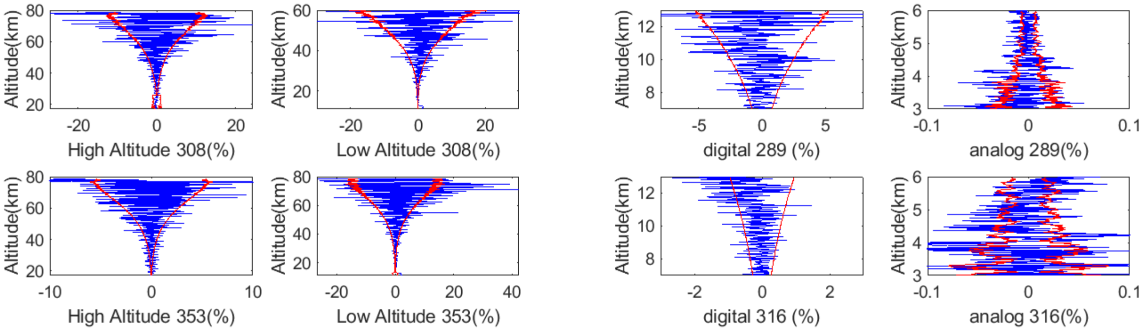

Figure 3.2: The percentage difference between the forward model and the actual measurements are shown in blue. The statistical uncertainty is plotted in red. The four plots at the left are stratospheric forward model residuals, and the four plots at the right are the tropospheric forward model residuals. For digital counting systems, the Poisson distribution is appropriate and the variance of the measurements at each altitude is the number of photons at that altitude. However, the output signals of the analog channels do not follow a Poisson distribution and to find the variance a 3-point running filter is used. As a result the red line (which indicates the noise of measurements) for the analog channels is more structured than for the digital channels.

The ozone profile for the night of 12 July 2017 is retrieved from 2.6 km to 42.2 km altitude. In Fig. 3.3 the OEM retrieval is plotted against the traditional stratospheric and tropospheric retrievals. The traditional ozone retrieval starts from 2.5 km and extends to 14.5 km, and the traditional stratospheric ozone retrieval starts at 11 km and extends to about 42.2 km. The tropopause height on this night is at 14.7 km. For comparison purposes the ozonesonde profile which starts from the ground and goes up to 33 km is shown as well. On this night of measurements, the ozonesonde balloon was released at 2153 from the station (44°N, 5.8°E). During the time of fly it drifted southeastward, such that 1.5 hours later, at the altitude of 33 km, its location was (43.6°N, 6°E).

To demonstrate how the OEM retrieval performs in the region where the tropospheric measurements are merged with the stratospheric measurements, we consider the retrievals in the region between 5 km to 18 km (Fig. 3.3). At around 6 km altitude the OEM retrieval shows a decrease in ozone not apparent in the traditional method or by the ozonesonde. However, the vertical resolution of the OEM retrieval at this altitude is half that of the traditional retrieval, with a subsequent doubling of the statistical uncertainty (Fig. 3.6). Thus, over the 5.5 km to 6.5 km altitude region the lidar ozone profiles are similar to within the measurement uncertainty, and less than the ozonesonde consistently up to about 15 km altitude. In the hand-off region between the tropospheric and stratospheric systems, the OEM retrieval is in general closer to the ozonesonde than the traditional method applied individually to both lidars, except around 11 km. Here there is a bump in the traditional method for the tropospheric system not seen in the OEM retrieval. At this altitude the traditional method has a vertical resolution of about 800 m and like the OEM a statistical uncertainty of about 5%. As was true for the feature around 6 km, the different in lidar ozone values is within the statistical uncertainty when the ozone values in the region of the bump are averaged over 800 m.

The vertical resolution of the OEM retrieval is calculated from the full width at half maximum (FWHM) of the averaging kernel at each altitude. The vertical resolutions as well as the statistical uncertainties of the OEM and the traditional retrievals are plotted in Fig. 3.4. In the free troposphere, at an altitude of 2.5 km, the vertical resolution for the OEM retrieval is 150 m, increasing to 300 m at 11 km. The traditional vertical resolution starts at 150 m as well, but designed to grow faster such that at 11 km the vertical resolution is 1000 m (Fig. 3.4,

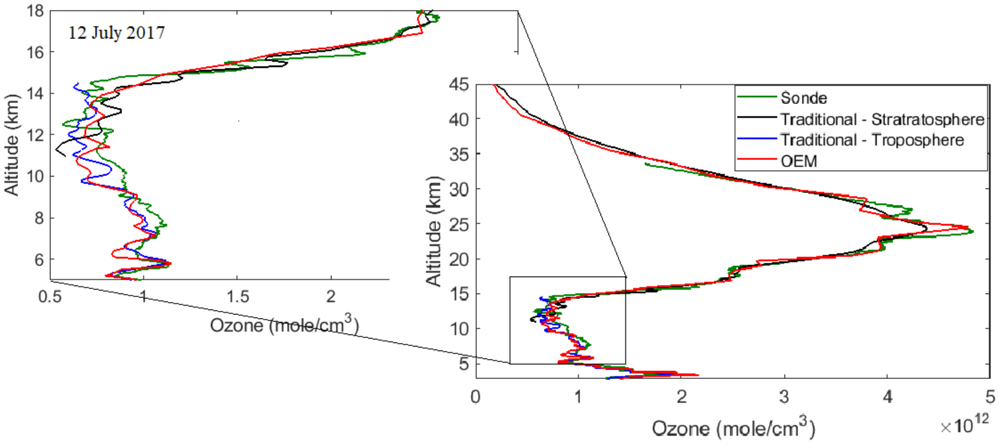

Figure 3.3: The OEM retrieval (red curve) compared to the traditional calculation of ozone using the OHP lidar systems. The tropospheric lidar starts at 2.5 km and extends upward to 14.5 km (blue curve). The ozone profile measured by the stratospheric lidar system (black curve) overlaps with the profile retrieved from the tropospheric lidar system in the UTLS. In this region, the OEM retrievals smoothly transition from relying primarily on the tropospheric lidar measurements to the stratospheric measurements.

right panel). The trade-off is that the uncertainty of the retrieval in the traditional method is smaller, so that at 11 km it is 4.5% as opposed to the OEM retrieval which has a larger uncertainty of 7.5%. At 11 km where the stratospheric measurements are added, the OEM vertical resolution is 300 m and gradually increases up to 600 m at 14.5 km, whereas, in the traditional tropospheric method, the retrieval resolution increases to 1900 m at the same height. At 21 km altitude, the vertical resolution is 1500 m, while at 25 km it is 1700 m. The vertical resolution does not change until 40 km, where due to the rapid drop in SNR it increases to 2000 m.

The percentage difference between the OEM retrieval and the ozonesonde measurements is shown in Fig. 3.5. For most heights the difference between the OEM retrieval and the sonde measurements in within the uncertainty of the two profiles. At 10 km, the difference between the sonde and the OEM retrieval is almost 30%. Above this altitude, and in the UTLS, the difference between the two profiles is less than 10%

The systematic and statistical uncertainties for the retrieved ozone profile for this night are shown in Fig. 3.6. The systematic uncertainties due to Rayleigh cross sections, the ozone absorption cross sections, and the overlap function are the **b** parameters in the forward model which contribute to the uncertainty of the ozone retrieval.

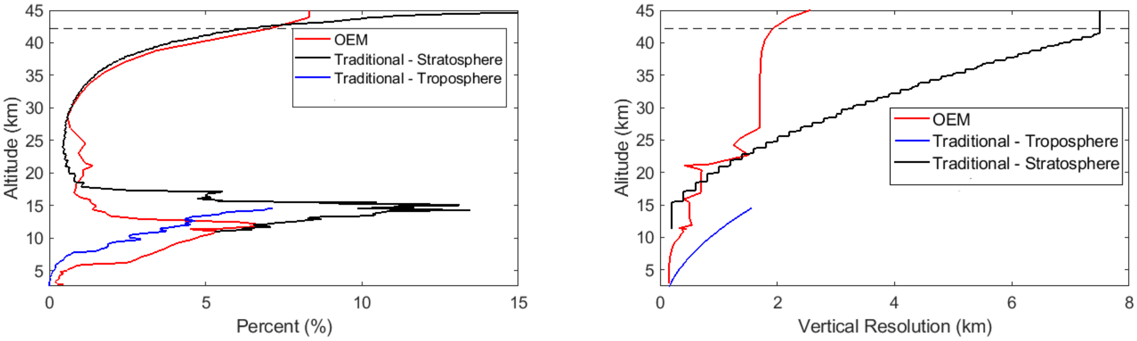

Figure 3.4: Left panel: The statistical uncertainty of the OEM retrieval (red curve) is plotted against the statistical uncertainty of the traditional retrievals. The uncertainty of the retrieval for the stratospheric and tropospheric lidar systems respectively are shown in black and blue. Right panel: The vertical resolution of the OEM retrieval is shown in red. The vertical resolution of the traditional calculation from the tropospheric lidar system is shown in blue, while the vertical resolution of the retrieved profile produced from the stratospheric lidar system is shown in black.

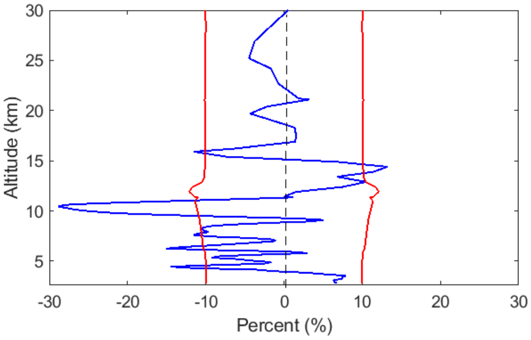

Figure 3.5: The percentage difference between the OEM retrieval and the ozonesonde measurements (blue curve) is plotted within the total statistical uncertainty of the OEM retrievals plus the ozonesonde measurement (red curves).

The contribution of the Rayleigh-scatter cross section uncertainty to the ozone retrievals for both the tropospheric and stratospheric lidars is less than 1%, consistent with the calculations given by Leblanc et al. Leblanc et al. (2018). The ozone absorption cross section for the 289 nm channel is about 5% which is close to the 7% uncertainty calculated by Leblanc et al. (2016). For stratospheric measurements, the ozone uncertainty has its maximum of 4% at the bottom of retrievals, which is higher than the calculated uncertainty of 2% in uncertainty budget of Leblanc et al. (2016). This difference is due to the OHP tropospheric DIAL having a larger wavelength separation than that in the NDACC LWG calculation.

The uncertainties due to the overlap function is 5% at the bottom of the retrieval and, at

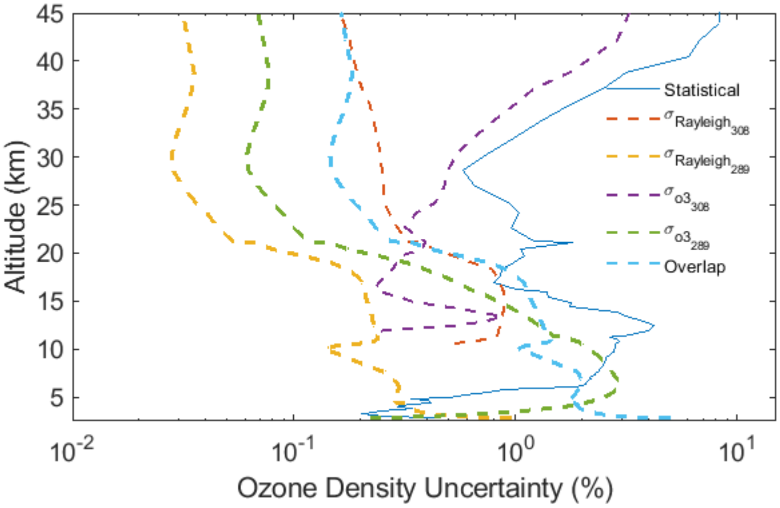

Figure 3.6: Uncertainty budget on the night of 12 July 2017. The statistical uncertainty of the retrieval (blue), the Rayleigh-scatter cross section uncertainty at 308 nm (dashed line red), the Rayleigh-scatter cross section uncertainty at 289 nm (dashed line yellow), the ozone absorption cross section at 308 nm (dashed line purple), the ozone absorption cross section for the 289 nm channel (dashed line green), and the overlap function for the 289 nm channel (dashed line light blue) all contribute to the budget. The horizontal dashed line shows the height below which the retrieval is independent of the *a priori* profile.

10 km it drops to about 1%. The uncertainty due to the ozone cross sections at 316 nm and 353 nm are negligible and are not shown in this plot. Also, the uncertainty on the retrieved ozone profile due to the temperature uncertainty of the ozone cross section is negligible as well.

Ozone density profiles are also retrieved for the nights of 14 July 2017, and 26 July 2017, each of which also had coincident ozonesonde measurements. On July 14 2017 the tropopause height is 12.5 km. The traditional retrievals of tropospheric and stratospheric lidars, at altitudes between 10 km to 15 km, are not consistent with each other (Fig. 3.7). The percentage difference between the traditional tropospheric and stratospheric ozone profiles in this region reaches its maximum of 33% at a height of 11.8 km. The OEM retrieval, similar to the night of 12th July, smoothly hands off from one lidar's measurements to the other one, and in this case is closer to the traditional stratospheric DIAL measurement. Also, the statistical uncertainty of the OEM retrieval reaches its maximum at about 12 km. The figure also shows the OEM retrieval compared with the sonde measurements. The ozonesonde was released from the station (44°N, 5.8°E) and it flew to the southeast. At its maximum height, the ozonesonde was located at (43.5°N, 6.3°E), and was not more than 50 km away from the OHP station. Similar to 12 July 2017, the percentage difference between the OEM and the sonde measurements is within the two profiles uncertainty.

On the night of 26 July 2017, the two ozone profiles calculated by the traditional method are inconsistent in the region from 12 km to 14 km are inconsistent with each other in the region of the tropopause (13.3 km). The OEM retrieval can smoothly transition from the tropospheric measurements to the stratospheric measurements (Fig. 3.8). Although, the OEM and the traditional analysis in the lower troposphere are match well (their difference is about 2%), the sonde profile is far from the two retrievals, and the difference between the sonde measurements and the lidar measurements (both the OEM and the traditional analysis) is greater than 60%. The ozonesonde, similar to the other nights, was released from the station, but comparing to the other nights it moved slightly farther toward the south, such that at its maximum height its location was (43.1°N, 5.8°E). Thus, the sonde was within approximately 100 km of its launch point.

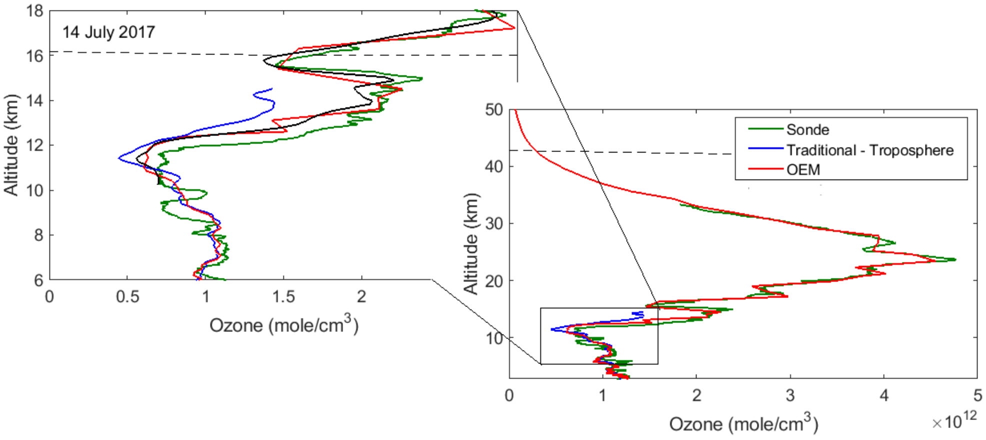

Figure 3.7: OEM ozone-profile retrieval on 14 July 2017 (red curve). The tropospheric traditional retrieval (blue curve) extends from 2.5 km to 15 km, while the stratospheric traditional retrieval (black curve) extends from 10 km to 43 km. At the region where the tropospheric and stratospheric lidar measurements overlap, the OEM can smoothly makes a transition from one lidar system's measurements to the other system's measurements.

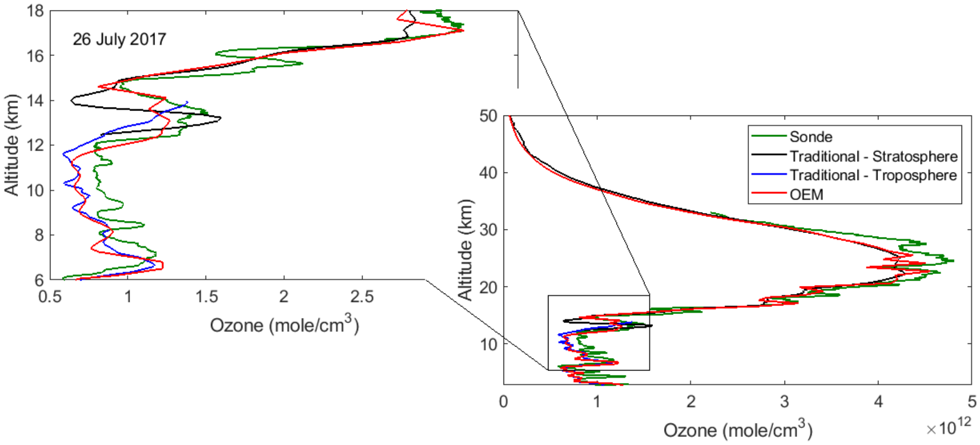

Figure 3.8: OEM ozone-profile retrieval on 26 July 2017 (red curve). The tropospheric traditional retrieval (blue curve) extends from 2.5 km to 14 km, while the stratospheric traditional retrieval (black curve) extends from 12.5 km to 43 km. At the region where the tropospheric and stratospheric lidar measurements overlap, the OEM can smoothly makes a transition from one lidar system's measurements to the other system's measurements.

3.4 Tropospheric Ozone Lidar Retrievals

We have demonstrated a retrieval for stratospheric ozone profiles using the OEM and measurements from a stratospheric DIAL lidar Farhani et al. (2018). In addition to the combined retrieval discussed here, the OEM retrieval presented in this work can be applied separately to tropospheric lidar systems. For tropospheric DIAL lidars, the overlap function must be added to the forward model, and this is the main difference between the two systems aside from including the different parameters associated with the different choice of wavelengths in a tropospheric DIAL system. Another difference is that for the OHP tropospheric lidar, the analog channel counts (as discussed in details in 3.2.1) do not follow a Poisson distribution.

The averaging kernels for the tropospheric retrieval is shown in Fig. 3.9. Below 14.2 km (where the horizontal dashed line is plotted), at least 90% of the ozone profile is retrieved from the measurements. Although, the retrieval extends to 16.5 km, we only consider the retrieved profile up to 14.2 km in height (above this altitude, the ozone profile starts falling into the *a priori* profile). In higher altitudes when the SNR drops, the averaging kernel becomes smaller and the retrieval falls back to its *a priori* value. The residual plots are similar to one shown in Fig. 3.2.

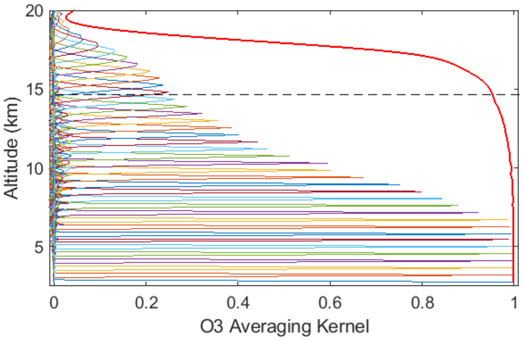

Figure 3.9: Averaging kernels for the ozone density for the measurements on 12 July 2017. The horizontal dashed line is the height cut-off above which the sensitivity of the retrieval to measurements is less than 90%. The averaging kernels are only shown every 450 m in altitude. The summation of rows in the averaging kernel matrix, for each specific height, is shown by the red curve.

The retrieval starts from 2.5 km, with 150 m steps, and extends to 14.2 km. The ozone retrieval resulting from the OEM code is plotted against the sonde profile and the traditional profiles (see Fig. 3.10). The OEM retrieval and the traditional method are within good agreement for most heights. Above 12 km the difference between the OEM and the traditional profile reaches to its maximum of 25%.

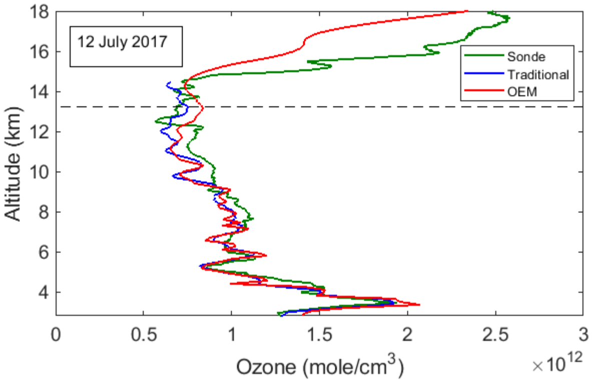

Figure 3.10: Both the OEM retrieval (red curve) and the traditional retrieval (blue curve) extend from 2.5 km to 14.2 km. The ozonesonde profile is plotted in green. The black dashed line defines the cut-off altitude of the retrieval.

The statistical uncertainties and vertical resolution of the traditional and OEM retrievals are shown in Fig. 3.11. The OEM retrieval has a better vertical resolution but higher uncertainty. The OEM retrieval resolution at 2.5 km is 150 m and at 14.2 km it becomes 600 m. As shown in the right panel of Fig. 3.11, the vertical resolution in the OEM retrieval is 200 m until about

10 km altitude. At an altitude of 5.5 km, where the photon counting signals are added, a small spike is observed. The traditional vertical resolution starts at 150 m and reaches to 1500 m at 14.2 km. The uncertainty of the OEM and the traditional method are similar for the first few kilometers. At 5.5 km, where the digital measurements begin, both methods have an uncertainty smaller than 1%. Above 5.5 km, the uncertainty in the OEM retrieval grows larger, and at 14.2 km it becomes 10.2%. The uncertainty of the traditional retrieval becomes larger as well, however at 14.2 km it is 7%. As shown in our stratospheric retrieval, the data and retrieval grids, as well as the correlation lengths, in the OEM method can be chosen to trade off larger vertical resolution for lower statistical uncertainty, like in the traditional method.

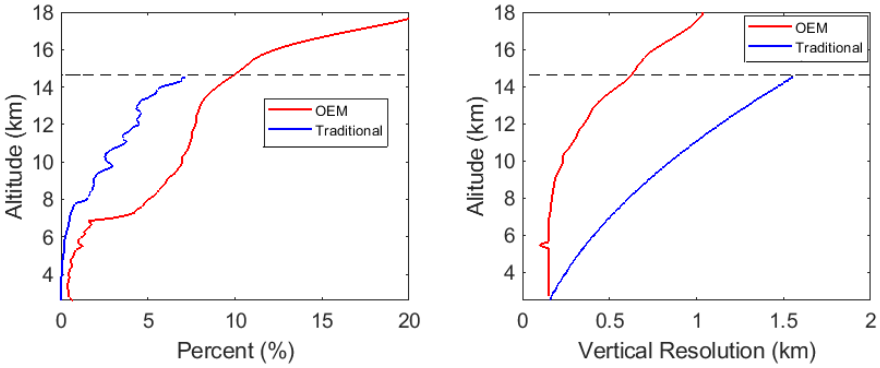

Figure 3.11: Left panel: The statistical uncertainty of the OEM retrieval (red curve) as well as the statistical uncertainty of the traditional retrieval (blue curve) for 12 July 2017. Right panel: Vertical resolution of the OEM retrieval is shown in red, while the vertical resolution of the traditional retrieval is shown in blue. The spike at 5.5 km in the OEM uncertainty is from the inclusion of the digital channels at this height.

OEM tropospheric ozone profile have also been retrieved for measurements on the nights of 14 July 2017 and 26 July 2017. The OEM retrieval for the night of 14th July is in good agreement with the traditional method, and for most heights, the difference between the two methods is small. At 11.5 km, the OEM retrieval has a better agreement with the sonde profile, but the difference between the two methods is only 2.5%, which is within their statistical uncertainty (see Fig. 3.12). On 26 July, at all altitudes above 3 km, the difference between the traditional and the OEM retrievals is less than 2%.

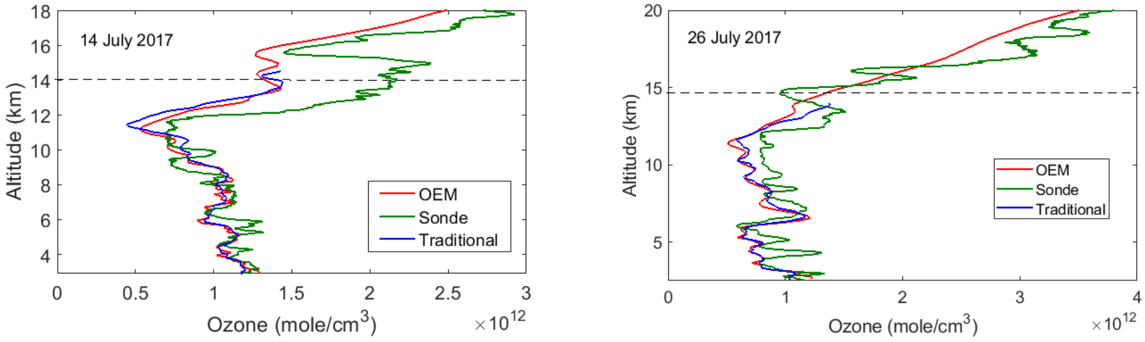

Figure 3.12: Comparison of tropospheric OEM ozone DIAL retrievals for the nights of 14 July and 26 July 2017 (red curves), the traditional method (blue curves), and ozonesonde measurements (green curves). The horizontal dashed line is the altitude below which the OEM retrieval is mostly independent of the *a priori* ozone profile assumed. Left panel: 14 July 2017; right panel: 26 July 2017.

3.5 Summary

We have introduced a first-principle OEM retrieval for tropospheric ozone profiles, as well as for a combination of tropospheric and stratospheric ozone profiles. Using the DIAL lidar measurements, we retrieved ozone profiles starting in the free troposphere and extending to the upper stratosphere. The results from our implementation of the OEM are summarized below.

1. The forward model uses the lidar equation and works directly with the raw measurements. The forward model provides a robust estimate of ozone profiles for clear nights.

2. The combined stratospheric-tropospheric DIAL OEM retrieval calculate a single ozone profile consistent with all the measurements.

3. A new retrieval method for tropospheric DIAL ozone lidars is given in the Appendix.

4. We used four different channels for tropospheric ozone retrievals, and eight different channels for the stratospheric-tropospheric ozone retrievals. The OEM has the advantage of using all these measurements at the raw (Level 0) stage; thus, no gluing or merging of profiles is needed.

5. For the tropospheric retrievals, the traditional method and our OEM retrieval produce similar results.

6. In the UTLS, the OEM retrieval smoothly transitions from one lidar system to the other system. The vertical resolution of the OEM retrievals in this region is about 600 m, and the retrieval uncertainty due to measurement noise does not exceed 7%.

7. Both tropospheric and tropospheric-stratospheric retrievals provide a full uncertainty budget which includes both statistical and systematic uncertainties.

3.6 Conclusions

Our OEM implementation brings benefits to the analysis of DIAL ozone measurements. Our retrieval has no need for "gluing" or "merging" the tropospheric and stratospheric measurements, as all measurements are simultaneously considered while retrieving a single ozone profile from multiple analog and digital channels measured by the two lidars. It provides a single ozone profile consistent with the measurements from both lidar systems, and includes the vertical resolution as a function of height, a detailed uncertainty budget, and averaging kernels to facilitate comparisons with other instruments.

While our initial implementation of our retrieval for ozone in the free troposphere and stratosphere has advantages, it also has limitations. Our forward model has been tested under clear sky conditions. However, in the UTLS region, clouds and significant aerosol loads can occur. We are planning to augment our forward model to allow for inclusion of aerosols, as well as other trace gases. We have taken steps in this direction by including the retrieval of overlap and particle extinction in our forward model for rotational-Raman temperature retrievals Mahagammulla Gamage et al. (2018). Another limitation of our current forward model is a difference in how the measurements must be handled during highly variable sky conditions. In the traditional method, individual scans can be corrected for dead time effects and then added in time. Our current OEM implementation assumes that scans added in times have roughly similar count rates. For a situation with high variability in count rates, such as the presence of patchy clouds, we recommend using shorter temporal integrations, retrieving the associated ozone profile, and then averaging the resulting ozone profiles. As the computational requirements for our retrievals are minimal, processing measurements in short time blocks (e.g. minutes) is practical even on a modest laptop computer.

Future work, in addition to improvements to our forward model, includes comparing our retrievals with satellite measurements in the UTLS. We also plan to re-process the OHP DIAL lidar measurements using the OEM technique.

3.6.1 Dataset Citation

G. Ancellet, ftp://ftp.cpc.ncep.noaa.gov/ndacc/meta/lidar/ga-ohp-tropo-ldr.txt

Bibliography

Ancellet, G. and Beekmann, M.: Evidence for changes in the ozone concentrations in the free troposphere over southern France from 1976 to 1995, Atmos. Env., 31, 2835–2851, 1997.

Ancellet, G., Papayannis, A., Pelon, J., and Megie, G.: DIAL tropospheric ozone measurement using a Nd: YAG laser and the Raman shifting technique, JTECH, 6, 832–839, 1989.

Baray, J. L., Courcoux, Y., Keckhut, P., Portafaix, T., Tulet, P., Cammas, J. P., Hauchecorne, A., Godin Beekmann, S., Mazière, M. D., Hermans, C., et al.: Maïdo observatory: a new high-altitude station facility at Reunion Island (21 S, 55 E) for long-term atmospheric remote sensing and in situ measurements, amt, 6, 2865–2877, 2013.

Browell, E. V.: Differential absorption lidar sensing of ozone, IEEE, 77, 419–432, 1989.

Cohen, Y., Petetin, H., Thouret, V., Marécal, V., Josse, B., Clark, H., Sauvage, B., Fontaine, A., Athier, G., Blot, R., et al.: Climatology and long-term evolution of ozone and carbon monoxide in the upper troposphere–lower stratosphere (UTLS) at northern midlatitudes, as seen by IAGOS from 1995 to 2013, acp, 18, 5415–5453, 2018.

Deeter, M., Edwards, D., and Gille, J.: Retrievals of carbon monoxide profiles from MOPITT observations using lognormal a priori statistics, Journal of Geophysical Research: Atmospheres, 112, 2007.

Eberhard, W. L.: Correct equations and common approximations for calculating Rayleigh scatter in pure gases and mixtures and evaluation of differences, Applied optics, 49, 1116–1130, 2010.

Eriksson, P., Jimenez, C., and Buehler, S.: Qpack, a general tool for instrument simulation and retrieval work, J. Quant. Spectrosc. Radiat. Transfer, 91, 47–64, 2005.

Fally, S., Vandaele, A. C., Carleer, M., Hermans, C., Jenouvrier, A., Mérienne, M. F., Coquart, B., and Colin, R.: Fourier transform spectroscopy of the O2 Herzberg bands. III. Absorption cross sections of the collision-induced bands and of the Herzberg continuum, J. Mol. Spectrosc., 204, 10–20, 2000.

Farhani, G., Sica, R. J., Godin-Beekmann, S., and Haefele, A.: Optimal Estimation Method Retrievals of Stratospheric Ozone Profiles from a DIAL Lidar, AMTD, 2018, 1–22, https://doi.org/10.5194/amt-2018-310, URL https://www.atmos-meas-tech-discuss.net/amt-2018-310/, 2018.

Fletcher, R.: Practical methods of optimization, John Wiley & Sons, 2013.

Forster, F., Piers, M., and Shine, K. P.: Radiative forcing and temperature trends from stratospheric ozone changes, Journal of Geophysical Research: Atmospheres, 102, 10 841–10 855, 1997.

Gaudel, A., Ancellet, G., and Godin-Beekmann, S.: Analysis of 20 years of tropospheric ozone vertical profiles by lidar and ECC at Observatoire de Haute Provence (OHP) at 44 N, 6.7 E, Atmospheric Environment, 113, 78–89, 2015.

Godin, S., Carswell, A. I., Donovan, D. P., Claude, H., Steinbrecht, W., McDermid, I. S., McGee, T. J., Gross, M. R., Nakane, H., Daan, Swart, P. J., Bergwerff, B. B., Uchino, O., von der Gathen, P., and Neuber, R.: Ozone differential absorption lidar algorithm intercomparison, Appl. Opt., 38, 6225–6236, 1999.

Godin-Beekmann, S., Porteneuve, J., and Garnier, A.: Systematic DIAL lidar monitoring of the stratospheric ozone vertical distribution at Observatoire de Haute-Provence (43.92[degree]N, 5.71[degree]E), J. Environ. Monit., 5, 57–67, 2003.

Halldorsson, T. and Langerholc, J.: Geometrical form factors for the lidar function, Appl. Opt., 17, 240–244, 1978.

Heath, D. F., Schlesinger, B. M., and Park, H.: Spectral change in the ultraviolet absorption and scattering properties of the atmosphere associated with the eruption of El Chichón: Stratospheric SO2 budget and decay, Eos Trans. AGU, 64, 197, 1983.

Hedin, A. E.: Extension of the MSIS thermosphere model into the middle and lower atmosphere, J. Geophys. Res.: Space Physics, 96, 1159–1172, 1991.

Holton, J. R., Haynes, P. H., McIntyre, M. E., Douglass, A. R., Rood, R. B., and Pfister, L.: Stratosphere-troposphere exchange, Rev. Geophys., 33, 403–439, 1995.

Hunt, W. H. and Poultney, S. K.: Testing the linearity of response of gated photomultipliers in wide dynamic range laser radar systems, IEEE Trans. Nucl. Sci, 22, 116–120, 1975.

IPCP: Climate change 2007: The physical science basis, Agenda, 6, 333, 2007.

Krueger, A. J. and Minzner, R. A.: A midlatitude ozone model for the 1976 US Standard Atmosphere, Journal of Geophysical Research: Atmospheres (19842012), 81, 44774481, 1976.

Leblanc, T., Sica, R. J., van Gijsel, J. A. E., Godin-Beekmann, S., Haefele, A., Trickl, T., Payen, G., and Liberti, G.: Proposed standardized definitions for vertical resolution and uncertainty in the NDACC lidar ozone and temperature algorithms Part 2: Ozone DIAL uncertainty budget, amt, 9, 4051–4078, 2016.

Leblanc, T., Brewer, M. A., Wang, P. S., Granados-Muñoz, M. J., Strawbridge, K. B., Travis, M., Firanski, B., Sullivan, J. T., McGee, T. J., Sumnicht, G. K., Twigg, L. W., Berkoff, T. A., Carrion, W., Gronoff, G., Aknan, A., Chen, G., Alvarez, R. J., Langford, A. O., Senff, C. J., Kirgis, G., Johnson, M. S., Kuang, S., and Newchurch, M. J.: Validation of the TOLNet lidars: the Southern California Ozone Observation Project (SCOOP), Atmos. Meas. Tech., 11, 6137–6162, https://doi.org/10.5194/amt-11-6137-2018, URL https://www.atmos-meas-tech.net/11/6137/2018/, 2018.

Logan, J. A.: Tropospheric ozone: Seasonal behavior, trends, and anthropogenic influence, Journal of Geophysical Research: Atmospheres, 90, 10 463–10 482, 1985.

Mahagammulla Gamage, S., Sica, R. J., Haefele, A., and Martucci, G.: Retrieval of Temperature From a Multiple Channel Pure Rotational Raman-Scatter Lidar Using the Optimal Estimation Method, Appl. Opt., in press, 2018.

Malicet, J., Daumont, D., Charbonnier, J., Parisse, C., Chakir, A., and Brion, J.: Ozone UV spectroscopy. II. Absorption cross-sections and temperature dependence, J. ATMOS. CHEM., 21, 263–273, 1995.

McDermid, I. S., Godin, S. M., and Walsh, D.: Lidar measurements of stratospheric ozone and intercomparisons and validation, Appl. Opt., 29, 4914–4923, 1990.

Megie, G. J., Ancellet, G., and Pelon, J.: Lidar measurements of ozone vertical profiles, Appl. Opt, 24, 3454–3463, 1985.

Mérienne, M. F., Jenouvrier, A., Coquart, B., Carleer, M., Fally, S., Colin, R., Vandaele, A. C., and Hermans, C.: Improved Data Set for the Herzberg Band Systems of 16O 2, J. Mol. Spectrosc., 207, 120–120, 2001.

Papayannis, A., Ancellet, G., Pelon, J., and Megie, G.: Multiwavelength lidar for ozone measurements in the troposphere and the lower stratosphere, Appl. Opt, 29, 467–476, 1990.

Povey, A. C., Grainger, R. G., Peters, D. M., and Agnew, J. L.: Retrieval of aerosol backscatter, extinction, and lidar ratio from Raman lidar with optimal estimation, amt, 7, 757–776, 2014.

Ramaswamy, V., Boucher, O., Haigh, J., Hauglustine, D., Haywood, J., Myhre, G., Nakajima, T., Shi, G. Y., and Solomon, S.: Radiative forcing of climate, Climate change, 349, 2001.

Rodgers, C. D.: Inverse methods for atmospheric sounding: theory and practice, vol. 2, World scientific, 2000.

Sica, R. J. and Haefele, A.: Retrieval of temperature from a multiple-channel Rayleigh-scatter lidar using an optimal estimation method, Appl. Opt, 54, 1872–1889, 2015.

Sica, R. J. and Haefele, A.: Retrieval of water vapor mixing ratio from a multiple channel Raman-scatter lidar using an optimal estimation method, Appl. Opt, 55, 763–777, 2016.

Stohl, A., Bonasoni, P., Cristofanelli, P., Collins, W., Feichter, J., Frank, A., Forster, C., Gerasopoulos, E., Gäggeler, H., James, P., et al.: Stratosphere-troposphere exchange: A review, and what we have learned from STACCATO, J. Geophys. Res.: Atmospheres, 108, 2003.

Chapter 4

Machine Learning Methods in Lidar Measurements Classification

4.1 Introduction

Lidar equipment consists of a transmitter unit (laser and beam formation optics), a receiver (telescope), and a signal detector unit (photomultiplier tubes and some optics). The recorded backscattered measurements (also known as level-0 measurement scans) are co-added for a period. Before co-adding, all scans should be checked for quality purposes to remove the "bad scans" from "good scans". Recorded measurements with extremely low power laser signals, high background counts, outliers, and scans with distorted or unusual shapes are bad scans. Over many nights, we may collect a series of scans that are not identified as bad but have lower laser power, which means they may need to be separated from high power scans before processing for geophysical quantities. Furthermore, depending on the lidar system and the purpose of the measurements, scans with traces of clouds or aerosol might be classified separately.

Over a night of measurement, atmospheric changes and laser fluctuations can cause both the background counts and the signal power to change abruptly. These changes make it difficult to identify bad scans quantitatively. Furthermore, identifying outlier signals is a complex task. During an observation night, signal quality can change for different reasons including any change in light pollution, the appearance of thin clouds, and drops in laser power.

Scans are often analyzed daily and classified manually as good or bad. A more common way of classifying scans is to define a threshold for the signal-to-noise ratio at some altitude: any scan which does not meet the pre-defined threshold value is flagged bad. In this method, bad scans may be incorrectly flagged as good, as they might pass the threshold criteria, but have the wrong shape at other altitudes. Recently, Wing et al. (2018) suggested that a Mann-Whitney-Wilcoxon rank-sum metric could be used to identify bad scans. In the Mann-Whitney-Wilcoxon test, the null hypothesis that the two populations are the same is tested against the alternate hypothesis that there is a significant difference between the two populations. The main advantage of this method is that it can be conducted when the data distribution is not Gaussian. However, defining a local median is a subjective task.

Using state-of-the-art machine learning techniques, we have introduced an automated and robust classification method for lidar scans. We present our result for two different lidar systems, the Purple Crow Lidar (PCL) system and the Meteoswiss Raman Lidar for Meteorological Observations (RALMO) system. In Section 4.2, we briefly describe the experimental setup for both PCL and RALMO. In Section 4.3, we explain how machine learning (ML) works. In Section 4.4, the algorithms used in this chapter are explained in detail. In Section 4.5, we show our results, and in Sections 4.6 and 4.7, the summary of the ML approach is provided, and is compared with other methods.

4.2 Instrumentation

The PCL is a Rayleigh-Raman lidar which has been operational since 1992. From 1992 to 2010, the lidar was located at the Delaware Observatory (42.5°N, 81.2°W) near London, Ontario, Canada. In 2012, the lidar was moved to the Environmental Science Western Field Station (43.1°N, 81.3°W). The PCL uses the second harmonic of an Nd: YAG solid-state laser. The laser operates at 532 nm and has a repetition rate of 30 Hz at 1000 mJ. The primary receiving optics is a liquid mercury mirror with the diameter of 2.2 m. Currently, the PCL has four detection channels:

1. A high gain Rayleigh channel which detects the backscattered counts from 30 km to 110 km altitude.

2. A low gain Rayleigh channel which detects the backscattered counts from 30 km to 110 km altitude. This channel is optimized to detect counts at lower altitudes where the high intensity backscattered counts can saturate the detector, and cause nonlinearity in the observed signal. Thus, using the low gain channel, at lower altitudes, the signal remains linear.

3. A Nitrogen Raman channel which detects the Raman shifted backscattered counts from 0.5 km to 20 km altitude.

4. A Water Vapour Raman channel which detects the Raman shifted backscattered counts from 0.5 km to 20 km altitude.

The Rayleigh channels are used for atmospheric temperature retrievals, and the water vapour and Nitrogen channels are used together to retrieve relative humidity profiles. Details about PCL instrumentation can be found in Sica et al. (1995).

The RALMO system is located at the MeteoSwiss meteorological station in Payerne, Switzerland (46.8°N, 6.9°E), and has been fully operational since 2008.For RALMO Nd: YAG laser produces a beam at 354.7 nm. The laser has a maximum energy of 400 mJ per pulse, and has a repetition rate of 30 Hz. To collect the backscattered photon counts, four f/3.33 parabolic mirrors, each with a diameter of 30 cm, are used. The RALMO system has the following measurement channels:

1. A Rayleigh channel which detects the backscattered counts from 0.3 km to about 30 km in altitude; this signal is used to retrieve the temperature profile.

2. A water vapour channel which is optimized to make measurements from 0.3 km to 20 km altitude during night time and from 0.3 km to 15 km altitude during day time.

3. A Nitrogen channel which is optimized to make measurements from from 0.3 km to 20 km altitude during night time and from 0.3 km to 15 km altitude during day time.

Raman scattering from the water vapour and Nitrogen channels are used to retrieve water vapour mixing ratio profiles (Cooney, 1970; Melfi, 1972). More specifications about the lidar can be found in Dinoev et al. (2012).

4.3 A Brief Introduction to Machine Learning

Machine learning (ML) is a branch of artificial intelligence that uses algorithms to predict outcomes. Machine learning observations, called *instances*, are described by number of variables called *attributes*. Supervised machine learning is a branch of machine learning in which instances are labeled. Thus, each instance has a set of independent attributes x and a dependent attribute y which is the label for each instance. The purpose of ML is to compute a function to map x to y. Formally, we are trying to learn a prediction function $f(x) : x \rightarrow y$ which minimizes the expectation of some loss function $L(y, f) = \Sigma(y_{true} - y_{predicted})$, where y_{true} is the actual value (label) and $y_{predicted}$ is the prediction generated from the prediction function.

The category of modeling can be either classification or regression. In classification, the target is a discrete value, and in regression, the target is a continuous quantity. In unsupervised learning, instances are not labeled, and the purpose of the method is to find similarities within the data. Hence, clustering instances are the primary goal in the unsupervised approach. Unsupervised ML can be used to detect outliers in a dataset. Here, in regards to their application in lidar data classification, both supervised and unsupervised techniques are explained.

4.3.1 Supervised Approach

In supervised learning, the dataset is divided into a training set and a test set. The training set contains both x (attributes) and y (output values). The test set only contains x values. For our lidar scan classification, we have a training set in which, for each scan, photon counts at each altitude are considered as an attribute. The classification of the scan is the output value. Thus the training set is a matrix of size $(m * n)$, and each row of the matrix presents a lidar scan in the training set. The columns of this matrix (except the last column), are photon counts at each altitude. The last column of the matrix shows the classifications of each scan as shown in 4.1.

Here, all scans with a distorted shape, low laser power, or high background counts are labeled as 0. The scans with lower than normal laser power (but still in an acceptable range) are labeled as 1. The scans with indications of clouds are labeled as 2. All other scans which are scans with high laser power and low background counts are labeled as 3. Examples of each of these scans are shown in Fig. 4.2.

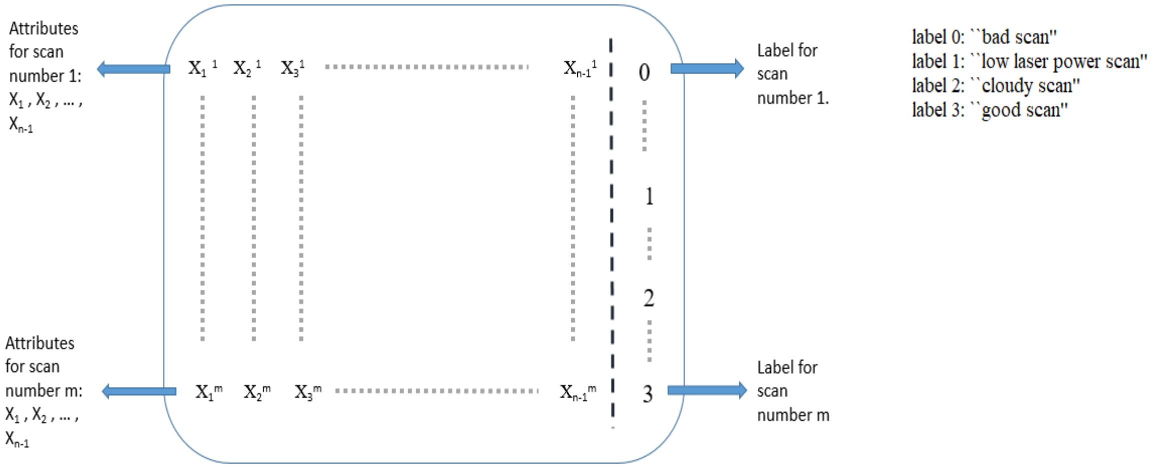

Figure 4.1: The training set is a matrix of size $(m*n)$, in which each row represents one dataset (here a lidar scan), and each element in each column (except for the n-th column) represents an attribute (for lidar scans: photon counts at each altitude). The n-th column holds the label for each scan (here we have four different labels: 0, 1, 2, and 3).

The goal of a training set in supervised learning is to train the machine (i.e., have it find a mapping function between x and y) such that it can predict the output of an unlabeled dataset correctly. Following the training phase, the test set can be used to check the accuracy of the ML model. There is no defined fraction of data that should be used for the training phase, but it is essential to select an unbiased set of training data which can represent the whole dataset.

In the training phase, a common issue arises when the developed algorithm becomes too complex. Although, the algorithm provides highly accurate results for the training set (often approaching 100% accuracy) it fails to perform well in the test phase. This is known as the overfitting problem. To overcome this problem, a simpler algorithm can be used. Compared to complex algorithms, simple algorithms have a lower accuracy rate in the training set; in return, in the test phase, they can perform better. Thus, highly complex models perform well in the training phase, but they have a poor generalization. It is said that they have high variance and low bias. Although simple models cannot perform perfectly in the training phase, they have a better generalization ability. Thus, they have low variance and high bias. Hence, there is a “trade-off” between the variance and bias, such that a model cannot be complicated and perform perfectly for a set of data, and at the same time, have an excellent generalization. The best practice is to take a model with a level of complexity that shows an acceptable performance

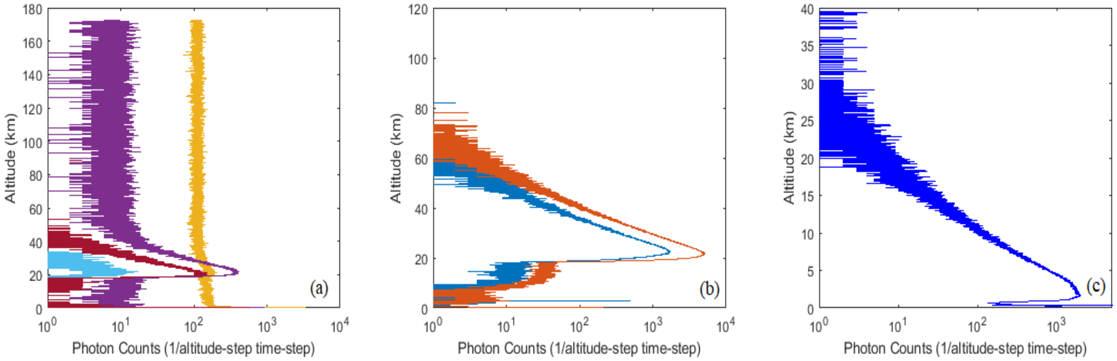

Figure 4.2: Example of scans taken by PCL for Rayleigh and Raman channels. Panel (a) shows examples of bad scans. In this plot, the signals in blue and dark red have extremely low laser power, the purple signal has extremely high background counts, and the signal in orange has a distorted shape and high background counts. Panel (b) shows two examples of good signals: the signal in red has a high laser power compared to the blue curve. In the supervised approach, the red signal is labeled 3, and the blue signal is labeled 1. Panel (c) shows a scan in which clouds occur at lower altitudes.

for both the training and test sets. Furthermore, selecting a larger dataset (if possible) can help to avoid overfitting and make a better model.

Finally, it is worth noting that for inverse modeling methods (like OEM) discussed in previous chapters, y is considered as the observable and x is the quantity of interest. However, in machine learning, x is given and y is predicted.

4.3.2 Unsupervised Approach

In unsupervised learning, all of the given data are unlabeled. This method is mostly used to find the underlying structure of the data by clustering, then looking for similarities. In unsupervised algorithms, some similarity measures are considered for clustering. After the algorithm creates clusters from the data, the user should look at each determined cluster to confirm that data in each independent cluster are similar to each other and different from data in other clusters. Moreover, by comparing data points from different clusters, the differences between the clusters should be identified. Here, we present a simple example to demonstrate how an unsupervised algorithm can make clusters. We use the iris flowers dataset (Dheeru and Karra Taniskidou, 2017) which includes three different types of irises (Setosa, Versicolour, and Virginica). Each individual iris flower has four attributes (features); this means that the dataset is in a four dimensional (4D) space. In many unsupervised learning methods, the focus is to reduce the dimension of datasets while preserving the most important characteristics of it. Here, we reduce the dimension from 4D to a 2D space, as a 2D space is possible for us to visualize. As shown in Fig. 4.3 with the help of an unsupervised learning algorithm (in this example we used the t-distributed stochastic neighbor embedding (t-SNE) method) we can make three clusters. Each cluster represents a type of iris. For our lidar scan classification, we are interested in seeing if the lidar scans based on their similarities (similar features) can be clustered. For our clustering task, a good ML method will distinguish between high background counts, low laser power scans, clouds, and high laser scans, and put each of these in a different cluster. A detailed description of unsupervised learning can be found in (Hastie et al., 2009).

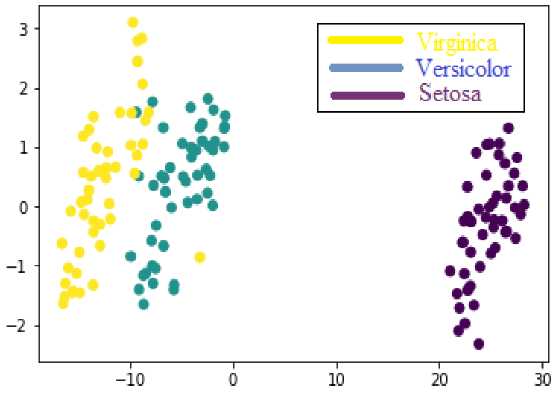

Figure 4.3: Using a t-SNE unsupervised algorithm, the iris dataset is clustered into three different types. Setosa is shown in violet, versicolor is shown in blue and virginica is shown in yellow.

4.4 Learning Algorithms

Many algorithms have been developed for both supervised and unsupervised learning. Here, we introduce those which are used in this chapter.

4.4.1 Support Vector Machine Algorithms

The Support Vector Machine (SVM) algorithms are popular in the remote sensing community because they can successfully handle small training sets, while producing highly accurate predictions (Mantero et al., 2005; Foody and Mathur, 2004). Moreover, unlike some statistical methods such as Maximum Likelihood Estimation, which assume data is distributed normally, the SVM does not require this assumption. This property makes it suitable for datasets with unknown distributions. Here, we briefly, describe how SVM works. More details on the topic can be found in (Vapnik, 2013; Burges, 1998).

The SVM algorithm finds an optimal hyperplane which separates the dataset into a distinct predefined number of classes (Vapnik, 2013). A hyperplane is a subspace with a dimension which is one less than its ambient configuration space. The optimal hyperplane is a decision boundary that minimizes misclassifications and is obtained from an iterative process in the training phase. The simplest form of SVM is a linear binary classifier in which a two-dimensional input space will be divided into two-class classification (see Fig. 4.4). As shown in the figure, an optimal hyperplane provides a maximum margin (separation line between the two different classes) between the two classes. Here, the hyperplane is used to separate the star logos from the sun logos. In the figure, few objects are misclassified. It is possible to obtain more complex hyperplanes; however, it can lead to the overfitting problem. To define the hyperplane, there is no need to use all the data in the training set; and only a subset of data points that lie on the margin (called support vector) is used.

To use SVM as a multi-class classifier, some adjustments need to be made to the simple SVM binary model. Methods like a directed acyclic graph, one-against-all, and one-against-others are among the most successful techniques for multi-class classification. Details about these methods can be found in Knerr et al. (1990).

4.4.2 Decision Trees Algorithms

Decision trees are nonparametric algorithms. Decision trees allow complex relations between inputs and outputs, to be modeled. Moreover, decision trees are the foundation of both random forest and boosting methods. A comprehensive introduction to the topic can be found in

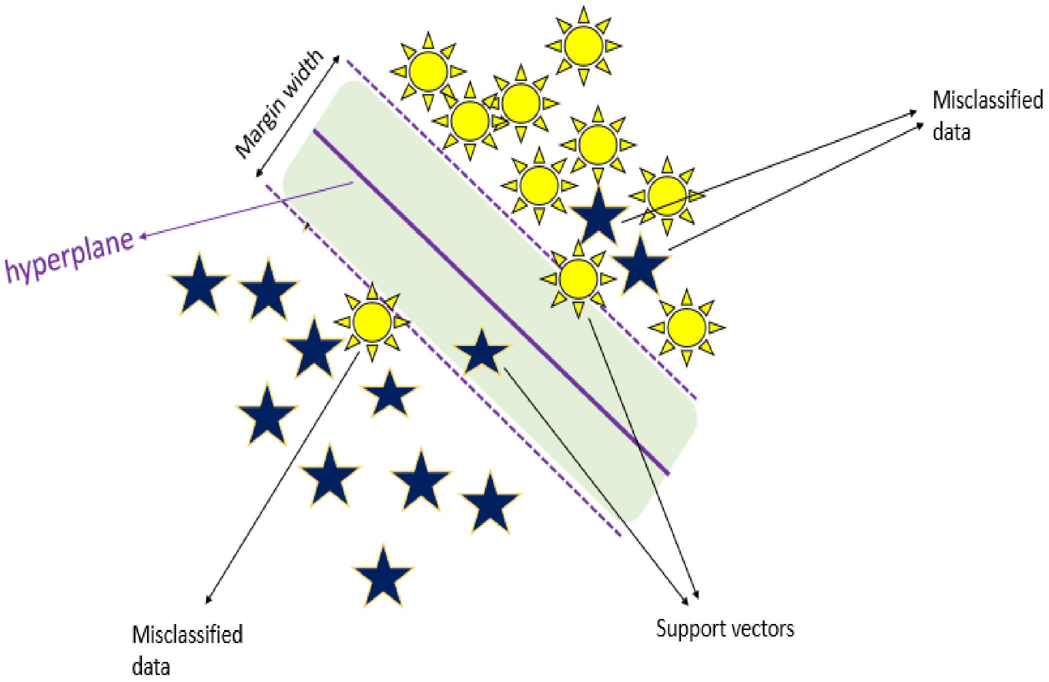

Figure 4.4: Example of linear SVM: the hyperplane separates sun and star logos. The subset of data points within the margin (support vector) is the most important data point as they play an important role in defining the optimum hyperplane. IN SVM a wider margin is preferable, as it divides two classes. This figure is adapted from: (Burges, 1998).

Quinlan (1986), here, we briefly describe how a decision tree is built.

A decision tree is a set of (binary) decisions represented by an acyclic graph directed outward from a root node to each leaf (see Fig. 4.6). Each node has one parent (except the root), and can have two children. A node with no children is called a leaf. Decision trees can be complex and this depends largely on the dataset. Later, the tree is simplified by pruning (from leaves to upper parts of the tree). To grow a decision tree, the following steps should be taken.

- Defining a set of candidate splits: We should answer a question about the value of a selected input feature to split the dataset into two groups.

- Evaluating the splits: Using a score measure, at each node, we can decide what the best question is to be asked and what the best feature is to be used. As the goal of splitting is to find the purest learning subset (in each leaf, we want the output labels to be the same), the purity improves of each split candidate should yield to the score measure.Shannon

Entropy (see below) is used to evaluate the purity of each subgroup. Thus, a split that reduces the entropy from one node to its descendent is favorable

- Deciding to stop splitting: We should set some rules to define when the splitting should be stopped, and a node becomes a leaf. This decision can be data-driven. For example, we can stop splitting when all objects in a node have the same label (pure node). The decision can be defined by a user as well. For example, we can limit the maximum depth of the tree (length of the path between root and a leaf).

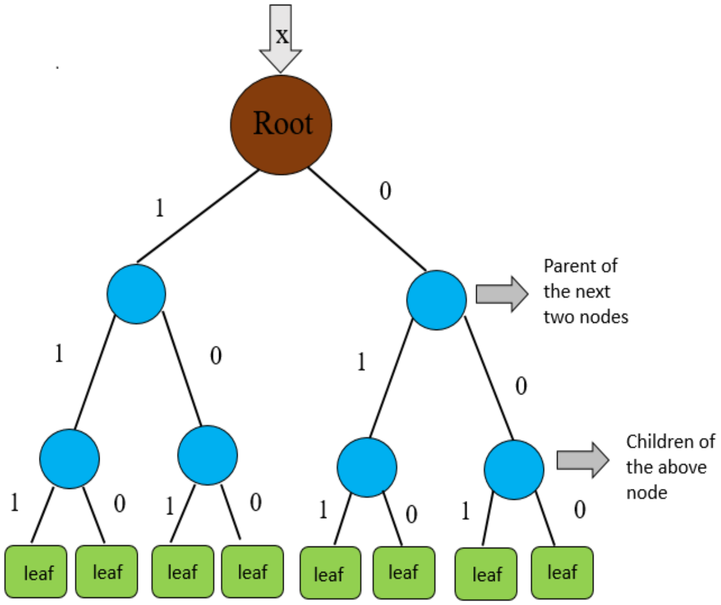

Figure 4.5: The schematics of a decision tree is shown. From the root (a node with no parents), a decision question is asked and answered in the form of yes (1) or no (0). The answers (yes or no) generate two children. Leaves are the nodes with no children. Ideally, each leaf is pure (all the points inside it has the same classification).

In a decision tree, by performing a full scan of attribute space the optimal split (at each local node) is selected, and irrelevant attributes are discarded. This method allows us to identify the attributes that are most important in our decision-making process. In summary, the simplicity of decision trees makes them suitable algorithms for both classification and regression processes.

The metric used to judge the quality of the tree splitting is Shannon entropy (Shannon, 1948). Shannon Entropy describes the amount of information gained with each event and is calculated as follows:

$$H(x) = -\Sigma p_i \log p_i \tag{4.1}$$

where p_i represents a set of probabilities. For example, if we toss a coin and are interested in the Shannon entropy of the output, we would calculate it as follows. The probabilities can be written as p (probability of the outcome being heads) and $q = 1 - p$ (probability of the outcome being tails). The Shannon entropy is then:

$$H(x) = -(p_i \log p_i + q_i \log q_i). \tag{4.2}$$

where $H(x)$ defines how much information is gained from the observation. To understand this result consider an unfair coin, where both sides of the coin are heads, and before doing any measurements, we know that flipping the coin will give us no new information (that is $p = 1$, $q = 0$ and $H(x) = 0$). In comparison, in the process of tossing a fair coin (where $p = 0.5$ and $q = 0.5$) we do not know the outcome of any individual flip, so a measurements gives us a maximum amount of information, that is $H = 1$.

4.4.3 Random Forests

A critical issue with trees is their high variance. This high variance can occur when a tree gets too complex in order to make pure leaves. To overcome this issue, we can grow an ensemble of trees which is known as the random forest (RF) method. If the number of trees is large, overfitting is not a problem, (Breiman, 2001). In an RF algorithm, to decide the class type all trees vote, and the vote of the majority is selected as the output. In the RF algorithm, to grow each tree, the bootstrap aggregating (bagging) method is used to sample from the training set. In bagging, a dataset is iteratively resampled with replacements. The bagging allows a data point to be included in a sample more than once.

Parameters which can significantly influence RFs are the number of trees and the tree depth. As mentioned earlier, increasing the number of trees can help with the overfitting problem, and growing more trees in a forest yield a smaller prediction error. Finding the optimal depth of each tree is a crucial task. While leaves in a short tree may contain heterogeneous data (the leaves are not pure), tall trees can suffer from poor generalization. Thus, the optimal depth provides a tree with pure leaves and great generalization.

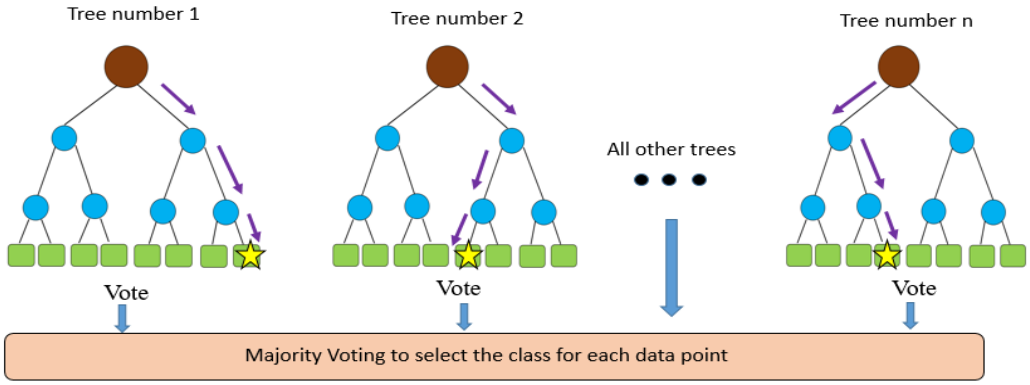

Figure 4.6: The schematic of a random forest: each tree casts a vote; based on majority vote, a data point is classified.

4.4.4 Gradient Boosting Tree Methods

Boosting methods are based on the idea that combining many “weak” approximation models (a learning algorithm which is slightly more accurate than 50%) will eventually boost the predictive performance (Knerr et al., 1990; Schapire, 1990). Thus, many “local rules” are combined to produce highly accurate models.

In the gradient boosting method, simple parametrized models (base models) are sequentially fitted to current residuals (known as pseudo-residual) at each iteration. The residuals are the gradients of the loss function (they show the difference between the predicted value and the true value) which we are trying to minimize. The Gradient Boosting Trees (GBT) algorithm is a sequence of simple trees generated such that each successive tree is grown based on the prediction residual of the preceding tree with the goal of reducing the new residual (see Fig.4.7. This “additive weighted expansion” of trees will eventually become a strong classifier (Knerr et al., 1990). This method can be successfully used even when the relation between the instances and output values are complex. Compared to the RF model, which is based on building many independent models and combining them (using some averaging techniques), the gradient boosting method is based on building sequential models.

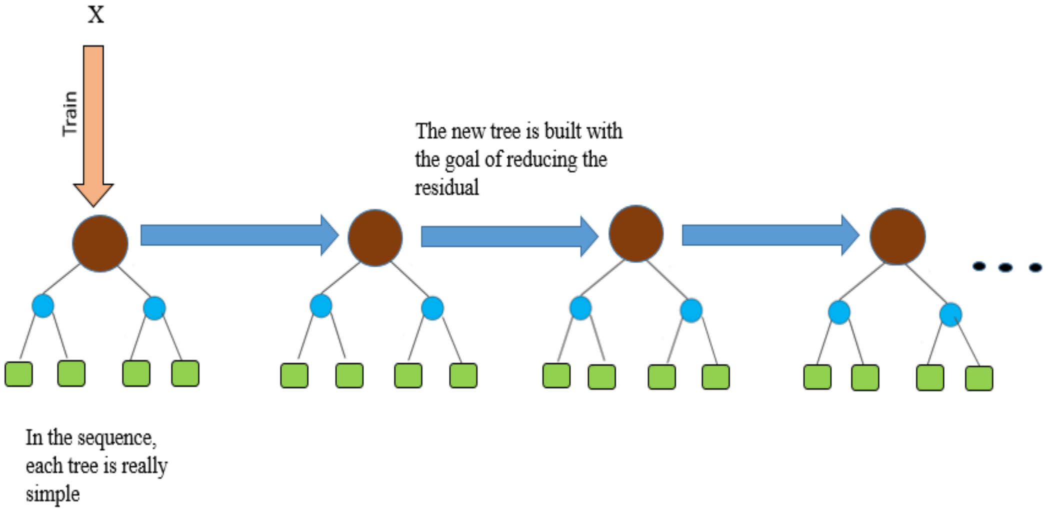
X
Train
The new tree is built with the goal of reducing the residual
In the sequence, each tree is really simple

Figure 4.7: Boosting methods are based on the idea that combining many "weak" approximation models (a learning algorithm which is Using the t-SNE algorithm scans for the low-gain Rayleigh measurement channel on the night of May 15 2012 were clustered to four different groups. As shown in the figure, cluster number 2 has some outliers.

4.4.5 The t-distributed Stochastic Neighbour Embedding Method

The t-SNE method is an unsupervised ML algorithm which is based on Stochastic Neighbor Embedding (SNE). In the SNE, the data points are placed into a low-dimensional space such that the neighborhood identity of each data point is preserved. The SNE is based on finding the probability that data point (i) has data point (j) as its neighbor, which can formally be written as:

$$P_{i,j} = \frac{exp(-d_{i,j}^2)}{\sum_{k \neq i} exp(-d_{i,k}^2)} \tag{4.3}$$

where $P_{i,j}$ is the probability of i selecting j as its neighbour and $d_{i,j}^2$ is the squared Euclidean distance between two points in the high dimensional space, and can be written as:

$$d_{i,j}^2 = \frac{\| (x_i - x_j) \|^2}{2\sigma_i^2} \tag{4.4}$$

where σ_i is defined so that the entropy of the distribution becomes $\log \kappa$, and κ is the "perplexity" which is set by the user, and its value determines how many neighbors will be around a selected point.

The SNE tries to model each data point x_i, at the higher dimension, by a point y_i at a lower dimension such that the similarities in $P_{i,j}$ are conserved. In this low dimensional map, we assume that the points follow a Gaussian distribution. Thus, the SNE tries to make the best match between the original distribution ($p_{i,j}$) and the induced probability distribution ($q_{i,j}$). This match is determined by minimizing the error between the two distributions, and the best match is developed. The induced probability is defined as:

$$q_{i,j} = \frac{exp(-\parallel (y_i - y_j) \parallel^2)}{\sum_{k \neq i} exp(-\parallel (y_i - y_k) \parallel^2)} \tag{4.5}$$

The t-SNE uses a similar approach but assumes a lower dimensional space which instead of being a Gaussian distribution follows Student's t-distribution with a single degree of freedom. Thus, since a heavy-tailed distribution is used to measure similarities between the points in the lower dimension, the data points which are less similar will be located further from each other. The above approach gives t-SNE an excellent capability for visualizing data. For our unsupervised approach, we use this method.

To demonstrate the difference between the Student's t-distribution and the Gaussian distribution, we plot the two distributions in Fig. 4.8. Here, the x values are within 5 and -5. The Gaussian distribution with the mean at 0 and the Student's t-distribution with the degree of freedom of 1 are generated. As is shown in the figure, the t-distribution peaks at a lower value and has a more pronounced tail.

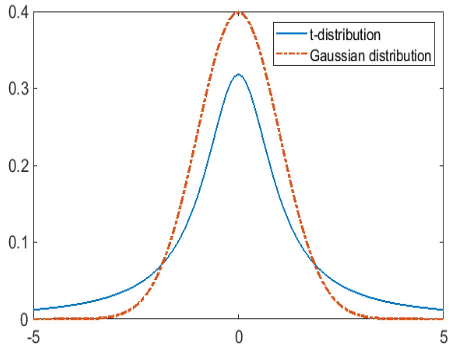

Figure 4.8: Red curve: the Gaussian distribution for data points (extending from -5 to 5 in x-axis). The peak of the distribution is at its mean in 0. Blue curve: The Student's t-distribution for the same data points. The distribution is heavy-tailed (comparing to the Gaussian distribution).

4.5 Result for supervised and unsupervised learning in PCL and RALMO system

To apply supervised learning to the PCL system, we randomly chose about 2000 scans from the low-gain and high-gain Rayleigh channels and the Nitrogen vibrational Raman channel (see Table.4.2). These measurement scans were taken from different nights, to represent different atmospheric conditions. For the low-gain and high-gain Rayleigh channels, the scans were labeled as “bad scans”, “scans with low laser power”, and “good scans”. For the Nitrogen channel we added one more label which represents scans with traces of cloud, called “cloudy” scans, where we must keep in mind by cloud we mean a substantial increase in scattering relative to a clean atmosphere, which could be caused by clouds or aerosol layers. Also, for scans of the Nitrogen channel, we did not find any “scan with low laser power”; thus for the Nitrogen channel this classification was not needed. Furthermore, in the Rayleigh channels because measurements start at 30 km, there is no need for a cloud classification. Furthermore, labeling the water vapour channel is not possible as there is no specific and well-defined pattern between different scans in this channel, so for most of the scans at this channel, we can not easily distinguish between “bad” and “good”.

To overcome the overfitting issue we used the k-fold cross-validation technique in which the dataset is divided into k equal subsets. The “training” set contains k-1 of the subsets and one subset is used as a “test” set. The test set is not used in the training phase, and the accuracy result is reported based on the accuracy of the unseen test data. This process is then iteratively repeated k times, and each time a different test set is chosen. The accuracy score is the ratio of correct predictions to the total number of predictions. Thus an accuracy score of 0.85 means that our predictions are 85% accurate. Using the k-fold cross-validation method (with k = 5), we calculated the accuracy score of SVM, random forest, and gradient boosting algorithms. The results are shown in Table. 4.1. The number of scans that we used for the training at each category (“good”, ”bad”, “lower laser power”, and “cloudy”) and for each channel are shown in Table.4.2.

As shown in Table. 4.1 all these algorithms perform excellently for all the channels. The only exception is the multi-class SVM for the Nitrogen channel, in which although the accuracy

Channel	multi class SVM	RF	GBT
High-gain Rayleigh	99%	98%	99%
Low-gain Rayleigh	98%	98%	98%
Nitrogen	99%	91%	98%

Table 4.1: The success score for each of the algorithms for 2000 scans of PCL lidar.

Channel	number of “good scan”	number of “scans with lower laser power”	number of “bad scan”	number of “cloud”
High-gain Rayleigh	172	40	131	not-applicable
Low-gain Rayleigh	134	76	163	not-applicable
Nitrogen	173	not-applicable	134	94

Table 4.2: Number of scans which used in each type.

score is still high (91%), the good scans and cloud scans can get labeled wrongly. In the Nitrogen channel, the RF and the GBT can classify the lidar scans with an accuracy above 98%, and nearly all scans are labeled correctly; thus the Nitrogen channel classification is still performing well.

For the RALMO system, similar to the PCL system, random scans from the Rayleigh channel and from the Nitrogen channel were selected. The labeling is binary as the scans are either labeled as cloud or clean. The Rayleigh channel in the RALMO system takes measurements from lower altitudes (as low as 50 m above the ground), so the cloudy scan label was used for the Rayleigh channel as well. Furthermore, as RALMO is operational during both days and nights, two different sets of data (one representing the day time scans and one representing the night time scans) were provided (see Table.4.5). The reason for this division is that the background counts during the day time, due to the solar radiation, are much higher than at night. Thus, a scan which is labeled as a good scan during day, would be considered a bad scan at night. Using the k-fold Cross Validation technique, accuracy scores for training around 2500 scans in day time and night time are calculated and respectively shown in Table. 4.3, and Table. 4.4. Similar to the PCL system result, the accuracy scores are high.

Channel	multi class SVM	RF	GBT
Rayleigh Night Time	99%	97%	99%
Nitrogen Night Time	99%	96%	98%

Table 4.3: The success score for each of the algorithms for day time scans of the RALMO lidar.

Channel	multi class SVM	RF	GBT
Rayleigh Day Time	98%	97%	97%
Nitrogen Day Time	96%	95%	98%

Table 4.4: The success score for each of the algorithms for day time scans of the RALMO lidar.

Channel	number of “good scan”	number of “cloud”
Rayleigh day time	513	316
Rayleigh night time	513	316
Nitrogen day time	257	200
Nitrogen night time	257	200

Table 4.5: Number of scans used in each type; for day time and night time we used the same number of scans.

4.5.1 Unsupervised result

The TSNE algorithm allows scans for each night to be clustered. The clustering can differ from night to night (and day to day for RALMO). On nights/days where most scans look similar, fewer clusters are seen, and on other nights/days when the atmospheric or the instrument conditions were not stable, more clusters are generated. To demonstrate how clustering works, we present our result for May 15 2012 (a cloudy night) for data collected with the PCL low-gain Rayleigh channel, as well as, the clustering result for the Nitrogen channel. As shown in Fig. 4.9, the t-SNE algorithm generates three distinct clusters for the low-gain Rayleigh channel. Figure. 4.10 (left panel) shows all the signals for each of the clusters. The maximum number of photon counts and the value and the height of the background counts are the identifiers between different clusters. Thus, cluster 3 with low background counts and high maximum counts (this indicates the power of the laser) represents a group of scans which are labeled as good scans in our supervised algorithms. Cluster 1 represents the scans with lower than normal laser powers, and clusters 2 shows scans with extremely low laser powers. To better understand the difference between these clusters, Fig.4.10 (right panel) shows the average signal within each cluster which shows the difference clearly. Furthermore, the outliers of cluster 2 (shown in black) identify the scans with extremely high background counts. This result is consistent with our supervised method, in which we had good scans (here is cluster 3), scans with lower laser power (here is cluster 1), and extremely low laser power scans (here cluster 2).

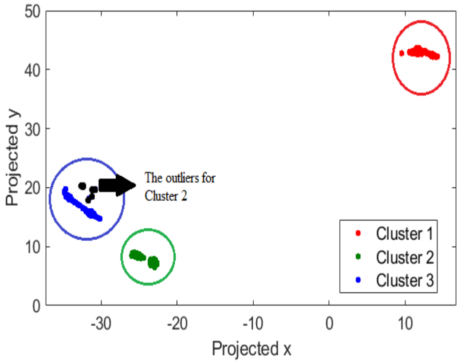

Figure 4.9: Using the t-SNE algorithm scans for the low-gain Rayleigh measurement channel on the night of May 15 2012 were clustered to four different groups. As shown in the figure, the cluster number 2 has some outliers.

Using the t-SNE, we have clustered scans for the Nitrogen channel with the data collected on 26 May 2012 at PCL system. This night was selected because the sky conditions changed from clear to cloudy and this was detected by the lidar system. The data from this night allows us to test our algorithm and determine how well it can distinguish cloudy scans from the non-cloudy scans. The result of clustering is shown in Fig.4.11 (left panel) in which two well-distinguished clusters are generated, where one cluster represents, the cloudy and the other represents the non-cloudy scans. The averaged signal for each cluster is plotted in Fig.4.11 (right panel).

Figure.4.12 shows results of clustering for few nights of measurements for different channels. By looking at scans within each cluster we can find what are the similarities among scans of a cluster, and decide how to label each cluster. Thus, the t-SNE is a powerful tool which can be used to visualize our data (by reducing its dimensionality) for each night of measurements, and to cluster them.

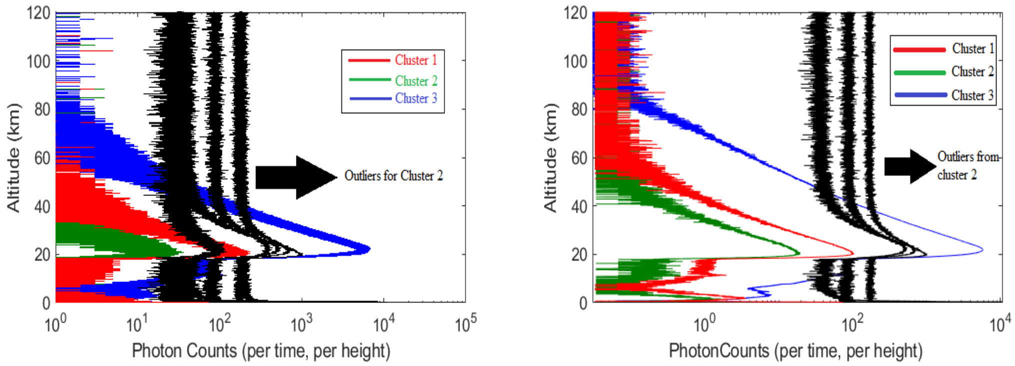

Figure 4.10: Left Panel: All scans collected for the PCL system from the low-gain Rayleigh channel which were plotted for the night of May 15 2012. The green signals have extremely low power. The red line represents all signals with lower laser signal and the blue line indicates the signals which are good scans. The black lines are signals with extremely high backgrounds. Right Panel: Each line represents an average of the signals within a cluster. The black line indicates the outliers which have extremely high background counts and are outliers belonging to cluster 3 (blue curve). The red line is the average signal for scans with lower laser power. The green line is the average signal for cluster 2. The background counts in the green line start at lower latitude of 40 km, where as for the red line the background starts at 60 km and in the blue line the background starts at 85 km.

4.6 Summary

We introduce a machine learning method to classify raw lidar measurements. We used different ML methods on elastic and inelastic measurements from both the PCL and RALMO lidar systems. The ML methods we used and our results are summarized as follows.

1. We tested different supervised ML algorithms, among which the SVM, RF, and the GBT performed better, with a success rate above 90% for both PCL and RALMO systems.

2. The t-SNE unsupervised algorithm can successfully cluster scans on nights with both consistent and varying lidar scans due to both atmospheric conditions and system alignment/performance. For example, if during the measurements the laser power dropped or clouds became present, the t-SNE showed different clusters representing these conditions.

3. Unlike the traditional method of defining a fixed threshold for the background counts, in supervised ML approach the machine can distinguish high background counts by looking at the labels of the training set. In the unsupervised ML approach, by looking at the

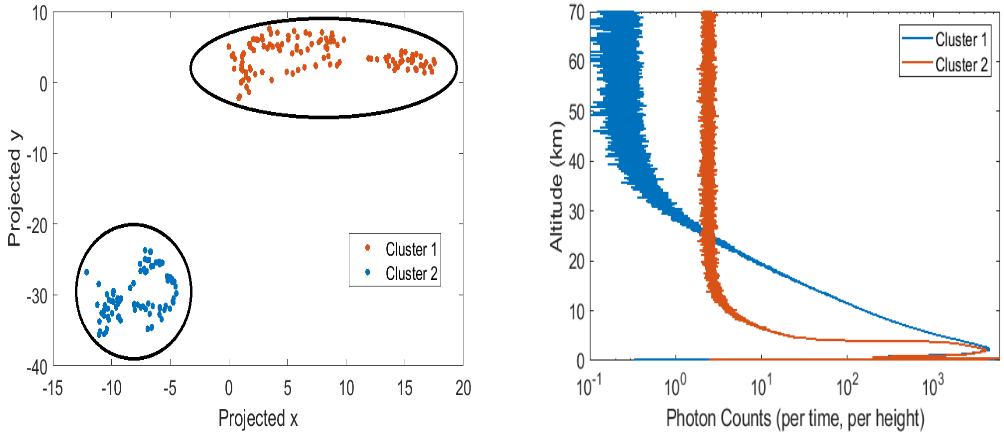

Figure 4.11: Figure. 4.11 (a): Using the t-SNE algorithm, scans for the Nitrogen channel on the night of May 15 2012, collected by the PCL system, were clustered into two different groups. Figure. 4.11 (b): The red line (cluster number 2) is the average of all signals within this cluster and indicates the scans in which clouds are detectable. The blue line (cluster number 2) is the average of all signals within this cluster and indicates the clear scans (non-cloudy condition).

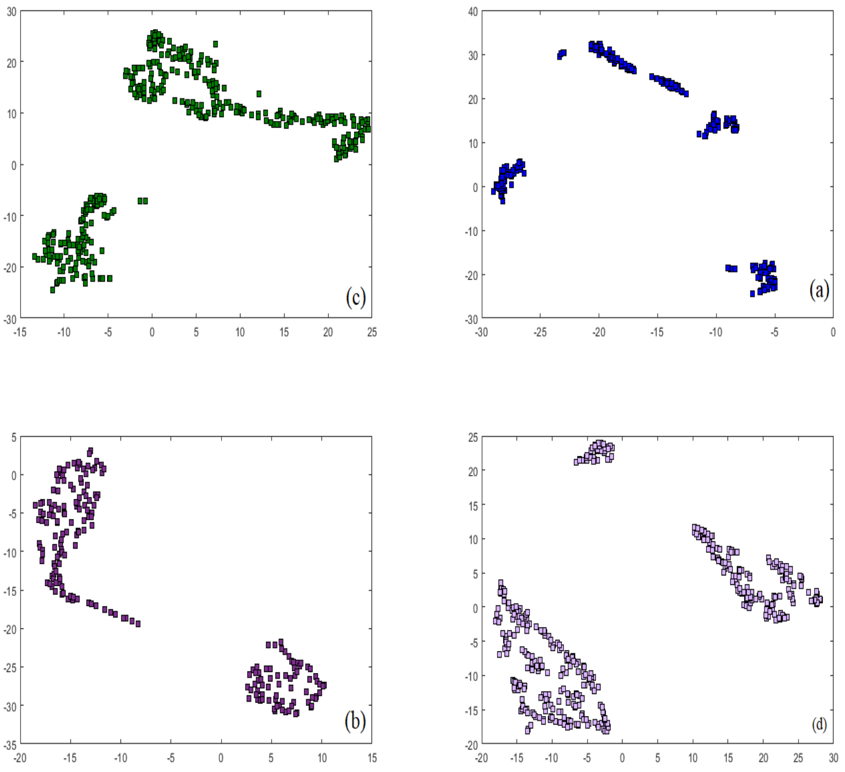

Figure 4.12: The result of applying the t-SNE algorithm to different channels at different nights is plotted. Each cluster contains lidar scans sharing similar features.

similarities between the two scans and defining a distance scale high background counts can be grouped in one cluster.

4.7 Conclusion

We successfully implemented supervised and unsupervised ML algorithms to classify lidar measurement scans. The ML is a robust method with high accuracy which enables us to precisely classify thousands of lidar scans within a short period of time. Thus, with accuracy of higher than 90% this method has a significant advantage over previous methods of classifying. For example, in the supervised ML, we train the machine by showing (labeling) different scans in different conditions. When the machine has seen enough examples of each class (which is a really small fraction of the entire data base), it can classify the un-labeled scans with no need to pre-define any condition for the system. Furthermore, in the unsupervised learning method, no labeling is needed, and the whole classification is free from subjective biases of the individual marking the scans (which for large atmospheric datasets ranging over decades is important). For example, for lidars in which observers, without using any pre-defined code identify each scan by eye, a common issue arises when an observer takes a scan as a good scan while another might classify it as a bad- scan, especially for edge cases. However, as mentioned earlier, these kind of mistakes are minimized by using an unsupervised ML approach.

Our results, in this chapter, indicate that ML is a powerful technique which can be used in lidar classifications. We encourage our colleagues in the lidar community to use both supervised and unsupervised ML algorithms for their lidar scans. To investigate rare atmospheric events, we are planning to use unsupervised ML to simultaneously analyze several nights of measurements.

As mentioned in the previous chapters, using OEM requires a well-defined forward model; in contrast the ML's focus is to learn the function which maps the state vector ($\mathbf{x}$) to the measurements ($\mathbf{y}$). Thus, we are planning to apply ML for our ozone retrievals.

Bibliography

Breiman, L.: Random forests, Machine learning, 45, 2001.

Burges, C. J.: A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2, 121–167, 1998.

Cooney, J.: Remote measurements of atmospheric water vapor profiles using the Raman component of laser backscatter, Journal of Applied Meteorology, 9, 182–184, 1970.

Dheeru, D. and Karra Taniskidou, E.: UCI Machine Learning Repository, URL `http://archive.ics.uci.edu/ml`, 2017.

Dinoev, T. S., Simeonov, V. B., Arshinov, Y. F., Bobrovnikov, S. M., Ristori, P., Calpini, B., Parlange, M. B., and van den Bergh, H.: Raman Lidar for Meteorological Observations, RALMO-Part I: Instrument description, amtd, 5, 6867–6914, 2012.

Foody, G. M. and Mathur, A.: A relative evaluation of multiclass image classification by support vector machines, IEEE Transactions on geoscience and remote sensing, 42, 2004.

Hastie, T., Tibshirani, R., and Friedman, J.: Unsupervised learning, in: The elements of statistical learning, pp. 485–585, 2009.

Knerr, S., Lé, P., and Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network, in: Neurocomputing, pp. 41–50, Springer, 1990.

Mantero, P., Moser, G., and Serpico, S. B.: Partially supervised classification of remote sensing images through SVM-based probability density estimation, IEEE Transactions on Geoscience and Remote Sensing, 43, 2005.

Melfi, S.: Remote measurements of the atmosphere using Raman scattering, Applied Optics, 11, 1605–1610, 1972.

Quinlan, J. R.: Induction of decision trees, Machine learning, 1, 81–106, 1986.

Schapire, R. E.: The strength of weak learnability, Machine learning, 5, 197–227, 1990.

Shannon, C. E.: A mathematical theory of communication, Bell system technical journal, 27, 379–423, 1948.

Sica, R., Sargoytchev, S., Argall, P. S., Borra, E. F., Girard, L., Sparrow, C. T., and Flatt, S.: Lidar measurements taken with a large-aperture liquid mirror. 1. Rayleigh-scatter system, Applied optics, 34, 1995.

Vapnik, V.: The nature of statistical learning theory, Springer science & business media, 2013.

Wing, R., Hauchecorne, A., Keckhut, P., Godin-Beekmann, S., Khaykin, S., McCullough, E. M., Mariscal, J.-F., and d'Almeida, E.: Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a 20 year cross-validation of two co-located French lidars, Atmospheric Measurement Techniques Discussions, 2018, 1–29, 2018.

Chapter 5

Conclusions and Future Work

In this thesis, a first-principle Optimal Estimation Method (OEM) is used to retrieve stratospheric and tropospheric ozone density profiles measured by a Differential Absorption Lidar (DIAL) system located at the Observatoire de Haute Provence (France). Although the stratospheric and tropospheric ozone profiles are retrieved individually, the OEM is capable of using all available channels from different measurements to retrieve a single optimum ozone profile extending from the troposphere to the upper stratosphere, without any need to merge results obtained from different analysis routines. For tropospheric and stratospheric lidar measurements, this feature of OEM offers a significant improvement over traditional methods, as the measurements' region of overlap is in the upper troposphere and the lower stratosphere (UTLS) region, and the traditional method in which the two profiles were glued together had no well-defined uncertainty. Understanding the UTLS is of vital importance because even small changes in the distribution of greenhouse gases in the UTLS can significantly affect the climate.

The OEM retrieval I developed in this thesis is valid for clear conditions with low to moderate aerosol loading. A future plan is to add the Raman channels to our retrievals to be able to provide retrievals in the presence of strong aerosol loads, as well as thin clouds. Another future direction would be to update the OHP ozone climatology using the OEM analysis. Analyzing the entire OHP database will significantly contribute to our understanding of the ozone distribution, particularly in the UTLS region. The OEM-retrieved ozone profiles have lower statistical uncertainty compared to previous lidar measurements and satellites, which in the UTLS in particular will provides better constraints on ozone models. Also, improved instrument intercomparison and satellite validation will be possible in both the troposphere and stratosphere by our calculations of lidar averaging kernels.

I introduced a machine learning routine in which the level 0 lidar measurements are automatically classified in categories corresponding to bad scans, clear sky scans, and scans with cloud or aerosol layers present. This scheme is a significant improvement for the lidar community, as typically groups use either simple methods based on defining a fixed threshold at a specific height or do the classification manually, which is time-consuming. The classifier is computationally fast and is suitable even for the lidars which analyze their measurements in near "real-time."

Five methods were tested, and the random forest classifiers, the support vector machines, and the gradient decision trees methods had the highest success scores. Currently, our classifying techniques cannot distinguish between clouds and aerosols loads, but they can tell if clouds and/or layers are present in a measurement. We are planning to extend our machine learning routine by adding more features to our training data to be able to distinguish between clouds and aerosol layers. I also employed the t-distributed Stochastic Neighbour Embedding (t-SNE) method which is an unsupervised algorithm. The t-SNE can successfully divide our lidar data into clusters, such that scans in each cluster share similar features. The t-SNE has potential to be in important tool, as it should be able to detect unusual events (like high loads of volcanic aerosols, fires, and meteor shower traces), in addition to evaluating scan quality without requiring training data. We are planning to employ the t-SNE to classify the entire Purple Crow Lidar system at Western to search for these unusual events.

Another future direction is using machine learning technique directly for ozone retrievals, eliminating the need for a pre-defined forward model. The goal in machine learning is to find a function which maps our measurements to the state of interest, allowing retrieval of the quantity of interest. We plan to extend our machine learning system to retrieve both temperature and composition, including ozone density and possibly water vapour mixing ratio.

Curriculum Vitae

Name:	Ghazal Farhani
Post-Secondary Education and Degrees:	The University of Western Ontario Astronomy 2011 - 2013 M.Sc.
	The University of Western Ontario Atmospheric Physics 2014 - 2018 Ph.D.
Honours and Awards:	Faculty of Science scholarship 2014-2016
	Northern Scientific Training Program 2015-2018
Related Work Experience:	Teaching Assistant The University of Western Ontario 2011 - 2018

Publications:

G. Farhani, R. J. Sica, S. Godin-Beekmann, A. Haefele, *"Stratospheric Ozone Density Retrieval Using the Optimal Estimation Method (OEM)"*, 2018, EDP science

G. Farhani, R. J. Sica, S. Godin-Beekmann, A. Haefele, *"Optimal Estimation Method Retrievals of Stratospheric Ozone Profiles from a DIAL Lidar"*, 2018, AMTD

G. Farhani, R. J. Sica, S. Godin-Beekmann, G. Ancellet, A. Haefele, *"Improved ozone UTLS DIAL measurements, using an Optimal Estimation Method"*, 2018, (to be submitted at applied optics)

G. Farhani, R. J. Sica, S. Godin-Beekmann, A. Haefele, *"The Classification of Lidar Measurement Scans, Using Machine Learning Methods"*, 2018, (to be submitted at amt)

Selected Conference talks and posters:

The European Geosciences Union meeting, 2018, Austria *“Stratospheric Ozone Density Retrieval Using the Optimal Estimation Method (OEM)”*

The International Laser-Radar Conference, 2017, Romania *“The application of the OEM on ozone retrievals of the Eureka DIAL system”*