Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

A Study on the performances of Different Metrics of a Machine Learning Cloud Classificator

Supervisor:

Prof. Tiziano Maestri

Submitted by:
Federico Donat

Academic Year 2022/2023

Abstract

Detecting clouds and clouds properties is essential for meteorological research, climate modeling and weather forecasting.

CIC (Cloud Identification and Classification) is a recently proposed, innovative machine learning algorithm adopted as the cloud identification code in the ESA Far-infrared Outgoing Radiation Understanding and Monitoring (FORUM, [15]) End2End simulator (FE2ES). CIC performs a classification by defining an eigenvectors-based Similarity Index that measures the information content brought into a Training Set when it is concatenated with a new observation.

In this thesis work, a new metric called the eigenvalues-based Similarity Index is proposed and studied to quantify the change in information content. Additionally, a novel methodology is developed within the CIC algorithm framework, allowing the simultaneous utilization of multiple Similarity Indices.

A cross-validation procedure is conducted using downwelling radiance spectra collected on the Antarctic Plateau to test and compare three configurations: classical eigenvectors-CIC, eigenvalues-CIC, and double-CIC. The double-Similarity-Index CIC demonstrates superior performance and is selected for the classification of a comprehensive dataset of spectra obtained by the REFIR-PAD instrument on the Antarctic Plateau from 2013 to 2020. The analysis yields statistically consistent results with previous studies ([6], [5]).

Finally, the double-CIC has been used for the challenging analysis of a large dataset, containing the simulated radiance of the Thermal Infra-Red Spectrometer (TIRS) of the PREFIRE mission ([10]). TIRS is characterized by only 64 wavenumber channels, and the synthetic data have global coverage. Very good cloud detection performances (more than 93% of clear and cloudy sky spectra correctly classified) are obtained in this study. This constitutes an important result that witnesses the classification power of CIC-like algorithms and their suitability for satellite missions.

Contents

Chapter 1

Introduction

Clouds play a crucial role in the Earth's climate system and have a significant impact on weather and on the planet's energy balance. Detecting clouds and clouds properties is thus essential for meteorological research, climate modelling and weather forecasting. Many different kinds of cloud properties detection techniques exist.

A first type of technique is represented by in situ measurements, that involve collecting data directly within the clouds using instruments deployed on aircrafts, balloons, or ground-based platforms. This techniques provide detailed information on cloud microphysical properties, such as cloud droplet size distribution, ice particle characteristics and cloud water content, but cannot be extensively used due to practical issues, remoteness of the sites and extreme climatic conditions.

Another type of technique is represented by active instruments such as Lidar (Light Detection and Ranging) and radar. Such devices emit signals towards the target and measure the reflected or scattered signals to gather valuable information. They have different sensitivities to cloud droplet size and concentration; lidar signal is attenuated by thick clouds, while radars have less attenuation. However, the use of such active measurements is not widespread and the majority of satellites are equipped with passive devices that measure the radiation emitted or reflected by the observed scene.

A large variety of different techniques based on the analysis of collected radiances exist. Many of them rely on the part of the spectrum that spans from infrared to shortwave, which limits their applicability to daytime hours, while other methods rely on outgoing longwave radiation only ([11]).

Thresholding techniques involve setting a threshold value for specific parameters (e.g. the radiance at some wavelength, temperature) to classify scenes as cloudy or cloud-free. For instance, the brightness temperature difference between bands at 11 and 12 μm can be used as an indicator for cirrus clouds ([4]). A paradigmatic example is constituted by the MODerate resolution Imaging Spectroradiometer (MODIS), that measures bands from 0.4 to 14.5 µm, from the middle infrared part of the spectrum to the visible. MODIS' measurements are analyzed using 14 channels and several steps to consider all the possible variability and providing an efficient cloud mask. Different thresholds are applied depending on the specific surface type and solar illumination. Ackerman et al. ([17]) describe four groups of tests: IR threshold test, brightness temperature difference, solar reflectance threshold, near-infrared (NIR) thin cirrus, IR thin cirrus. The final cloud mask is then determined from the results of each group.

A large class of cloud detection and classification methods is then based on statistical analysis techniques. This methods can exploit various parts of the spectrum and are built to minimize some kind of distance from the new observation to some previously-defined Training Sets; since these methods rely on example spectra provided by the user they can be seen as supervised machine learning algorithms. Examples are the cumulative discriminant analysis by Amato et al. ([1]) and the minimization of the Mahalanobis distance described by Clarisse et al. ([4]). Often, given the high dimensionality and resolution of the spectra, the data need to undergo a dimensionality reduction. A widely used tool for dimensionality reduction is Principal Component Analysis (PCA), a technique that looks for those linear combinations of the original variables that have maximal variance and zero correlation; only the variables carrying most of the variance are then retained and used for classification. Malinowski ([12]) and Turner et al. ([18]) provided a way for identifying the exact number of

maximal variance variables associated with the physical signal and therefore discarding the noise. Such method is used by Maestri et al. ([11]) in the context of Cloud Identification and Classification algorithm, that performs a classification by measuring the information content brought to the Training Sets when they are concatenated with a new observation.
A theoretical limitation of PCA is the fact that it can detect only linear relations among the variables. Neural Networks are often used to exploit linear and, eventually, nonlinear relations. Examples of NNs used for cloud detection are given by Mastro et al. ([14]), that implemented a multilayer feed-forward neural network for the classification of Infrared Atmospheric Sounding Interferometer (IASI) spectral radiances, and by Bertossa et al. ([2]) in the context of satellite radiances simulated for the TIRS (Thermal InfraRed Spectrometer) instrument. Important issues regarding Neural Networks are represented by the huge number of training spectra needed for the learning of the optimal parameters (neural weights) and the challenging interpretation of them.
The present thesis is developed in the context of Cloud Identification and Classification algorithm.

1.1 Thesis objectives and overview

Cloud Identification and Classification (CIC) is an innovative machine learning algorithm which aims at classifying infrared radiance spectra. It is partially based on other works ([12], [18]) and has the advantages of being easy to implement, user-friendly, fast and particularly suitable for situations where few training spectra are available. The classification of a spectrum whose true class is unknown is performed adding the spectrum to previously defined Training Sets, that represent different scenes, and quantifying the change it causes to the eigenvectors of the Training Sets covariance matrices. These eigenvectors represent the directions along which the variance of the data is maximal and are referred to as Information BEaring principal Components (IBECs) or Principal Components of the data (PCs). A small eigenvectors change is interpreted as high similarity between new observation and Training Set while great changes are interpreted as information not previously contained in the Training Set, and suggest low similarity and thus a probable scene diversity. A training phase is also previously performed in order to find a numerical threshold for unambiguously assigning the observation to a class. The algorithm has been successfully used for the classification of synthetic datasets ([11]) and downwelling measured radiances ([5]).

One of the aims of this thesis work is understanding if different similarity indices, not necessarily based on the Principal Components, can be used for measuring the relatedness of a new observation with the pre-defined Training Sets. As an example, a Training Set feature that could be exploited for enhancing the classification power of the algorithm is the set of the eigenvalues of the covariance matrices. In fact, these eigenvalues represent the variance of the data along the Principal Components. Not suprisingly, some examples exist of data analysis methods involving the eigenvalues of the covariance matrix. The Mahalanobis distance, a classical measure of the distance between a data point and a multivariate distribution, can be seen as a Euclidean distance performed along the Principal Components and rescaled using the associated eigenvalue ([3]).

Another more recent example is EigenClass, a machine learning algorithm that was successfully tested on different datasets against more widespread classification algorithms ([7]).
The aims of the present work are not limited to the introduction of new similarity indices in the context of CIC algorithm, but intend to investigate the possibility of new methodologies within the CIC algorithm conceptual framework itself. In addition, these proposals and developments are going to be tested and compared using downwelling Antarctic infrared radiance spectra as well as simulated upwelling radiances relative to the whole globe.
The thesis is organized as follows:

- Chapter 2: the first two sections provide an overview of Principal Component Analysis and of the logic of CIC algorithm; in section 3 an eigenvalues-based Similarity Index is proposed and its properties are studied analytically; in section 4 a methodology for exploiting more than one Similarity Index at one time is developped.
- Chapter 3: the classical CIC algorithm, the eigenvalues-based CIC and a double-Similarity-Index CIC undergo a cross-validation procedure in order to understand their behaviour with respect to cloud detection, cloud classification and dependance on Training Set sizes.
- Chapter 4: the double-Similarity-Index CIC is used for the analysis of about one million simulated upwelling spectra relative to the whole globe.

Chapter 2

CIC: Cloud Identification and Classification Algorithm

Cloud Identification and Classification (CIC) is a machine learning algorithm based on Principal Component Analysis that has been successfully used for the classification of clear sky, ice cloud and mixed-phase cloud scenes. CIC performs a classification by defining an eigenvectors-based Similarity Index that measures the information content brought to a Training Set when it is concatenated with a new observation. Its characteristics make it particularly suitable for situations where few training spectra are available. In this chapter, after introducing the concepts of Principal Components Analysis (PCA) on which CIC is based, we describe the logic of the algorithm; later, we propose and study analytically a new metric for quantifying the relatedness of the new observation with the Training Sets; finally, we propose a methodology for exploiting more than one metric at a time aimed at enhancing the classification power of the algorithm.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction technique. It is aimed at reducing the number of variables that characterize a set of data limiting the information loss. It is to be noted that the new variables are not necessarily a subset of the original ones, but are in general a linear combination of them. This fact often implies the loss of physical meaning of the variables, but it permits to separate the physical signal from noise, to have lower computational costs and, to a certain extent, to visualize multidimensional data.

Let X be a $N_{VAR} \times N_{OBS}$ matrix of data whose N_{VAR} rows represent variables and whose N_{OBS} columns represent observations:

$$X = \left(\vec{x}_1, \ldots, \vec{x}_{N_{OBS}}\right), \tag{2.1}$$

where the $\vec{x}_i \in \mathbb{R}^{N_{VAR}}$ are column vectors. The data X have a covariance matrix that we shall indicate with S_X.

In the framework of PCA the data are expressed in a new orthonormal basis:

$$X \longrightarrow Y = PX, \tag{2.2}$$

where P is the $N_{VAR} \times N_{VAR}$ matrix whose rows are the N_{VAR} orthonormal eigenvectors of S_X:

$$P = \begin{pmatrix} \vec{p}_1^T \\ . \\ . \\ . \\ \vec{p}_{N_{VAR}}^T \end{pmatrix} \tag{2.3}$$

The eigenvectors $\vec{p}_j$ are often called the Principal Components (PCs) of the data X.

The generical observation $\vec{x}_i$ thus becomes

$$P\vec{x}_i = \begin{pmatrix} \vec{p}_1^T \\ . \\ . \\ . \\ \vec{p}_{N_{VAR}}^T \end{pmatrix} \vec{x}_i = \begin{pmatrix} \vec{x}_i \cdot \vec{p}_1 \\ . \\ . \\ . \\ \vec{x}_i \cdot \vec{p}_{N_{VAR}} \end{pmatrix} \tag{2.4}$$

The eigenvalues λ_j associated with the PCs $\vec{p}_j$ are of great importance because they represent the variance of the original data along the associated PC:

$$\lambda_j = \frac{1}{N_{VAR} - 1} \sum_{i=1}^{N_{VAR}} |(\vec{x_i} - \vec{\mu}_X) \cdot \vec{p_j}|^2, \tag{2.5}$$

where $\vec{\mu}_X$ denotes the mean of the data X. Usually (and always in the present work) the PCs are meant sorted from the one associated with the largest variance to the one associated with the lowest variance. The number of PCs of the data X is equal to the number of variables N_{VAR}; however, when a few PCs explain the great part of the variance of the data, taking only a subset of PCs preserves the physical signal and discards the noise. Malinowski ([12]) and Turner et al. ([18]) showed how to extract the number P_0 of PCs associated with the physical signal. The P_0 elements are extracted by minimising the factor indicator function (IND) defined as follows:

$$IND(j) = \frac{RE(j)}{(N_{VAR} - j)^2}, \; j \in \{1, ..., N_{VAR} - 1\} \tag{2.6}$$

where RE(i) is defined, in [18], as the real error

$$RE(j) = \sqrt{\frac{\sum_{k=j+1}^{N_{VAR}} \lambda_k}{N_{OSS}(N_{VAR} - j)}}. \tag{2.7}$$

The natural number P_0, obtained through this minimisation process, is the number of eigenvectors associated with the physical signal.

2.2 Training Phase and Classification: Elementary Approach

The first step is the definition of a training set (TR) for each class:

$$TR_i(\nu, t),\ i \in \{1, 2\},\ \nu \in \{1, ..., N_{VAR}\},\ t \in \{1, ..., T_i\} \tag{2.8}$$

where the couple (ν,t) denotes the ν-th element of the t-th spectrum and T_i represents the total number of spectra in TR_i. N_{VAR} denotes the total number of variables; in the present work the total number of variables is the number of wavenumber channels considered. In order to classify a spectrum $\vec{x}$ whose true class is unknown, $\vec{x}$ is added to the Training Sets TR_i creating the Extended Training Sets ETR_i:

$$ETR_i(\nu, t),\ i \in \{1, 2\},\ \nu \in \{1, ..., N_{VAR}\},\ t \in \{1, ..., T_i + 1\}. \tag{2.9}$$

For each TR the principal components are computed:

$$\vec{p}_{j,TR_i} \in \mathbb{R}^{N_{VAR}},\ j \in \{1, ..., N_{VAR}\}. \tag{2.10}$$

The same is done for each ETR:

$$\vec{p}_{j,ETR_i} \in \mathbb{R}^{N_{VAR}},\ j \in \{1, ..., N_{VAR}\}. \tag{2.11}$$

We will denote with $p_{j,TR_i}(k)$ (or $p_{j,ETR_i}(k)$) the k-th element of the j-th principal component of TR_i (or ETR_i). In the present work the PCs are meant sorted from the most important (i.e. associated with the largest eigenvalue) to the least important (i.e. associated with the smallest eigenvalue). The classification of $\vec{x}$ is performed through a parameter called Similarity Index (SI) that evaluates the variation of the PCs of ETR_i with respect to the PCs of TR_i. The Similarity Index of the new observation $\vec{x}$ with the generical Training Set TR$_i$ is defined as follows:

$$SI(\vec{x}, TR_i) = 1 - \frac{1}{2P_0} \sum_{j=1}^{P_0} \sum_{k=1}^{N_{VAR}} |p_{j,TR_i}(k)^2 - p_{j,ETR_i}(k)^2|, \tag{2.12}$$

where P_0 is the number of principal components that are associated to the physical signal characterising the spectrum. The SI is normalized, in fact the quantity

$$\sum_{k=1}^{N_{VAR}} |p_{j,TR_i}(k)^2 - p_{j,ETR_i}(k)^2| \tag{2.13}$$

is at most equal to 2. When it is summed over the index j it can reach a value of $2P_0$ and therefore

$$SI \in [0, 1]. \tag{2.14}$$

A Similarity Index equal to 1 means that TR_i and ETR_i have exactly the same eigenvectors; a Similarity Index close to 0 represents two very different sets of PCs, denoting that $\vec{x}$ brings new information to TR_i.
The Similarity Index is computed for both classes and the so called Similarity Index Difference (SID) is computed:

$$SID(\vec{x}) = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_2) \in [-1, 1]. \tag{2.15}$$

Finally,

$$\vec{x} \text{ is assigned to} \begin{cases} \text{class 1, if } SID(\vec{x}) = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_2) > 0 \\ \text{class 2, otherwise.} \end{cases} \tag{2.16}$$

The method can be generalized to three or more classes. If there are three classes three Training Sets must be defined: TR_1, TR_2 and TR_3, each representative of the variability within that class. In order to classify a new observation $\vec{x}$ it is necessary to compute three SIDs, one for each couple of Training Sets:

$$\begin{cases} SID(\vec{x})_{1,2} = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_2) \\ SID(\vec{x})_{1,3} = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_3) \\ SID(\vec{x})_{2,3} = SI(\vec{x}, TR_2) - SI(\vec{x}, TR_3) \end{cases} \tag{2.17}$$

The classification output is then obtained following the logical scheme in *Figure* 2.1. The white boxes represent the partial results and the green ones are the final classification outcome.

If the classes are more than three, the same procedure is adopted and a spectrum is assigned to a class if and only if a class prevails over all the others, otherwise the spectrum is considered unclassified.

It is to be noted that the Elementary method has been introduced only to clarify the process of classification but, in practice, more advanced approaches are required. In particular, the weakness of the elementary method relies on the fact that it assumes that the PCs of different Training Sets are equally inclined to undergo a change when a new spectrum is introduced. This is not necessarily true. For example, it is possible that if TR_1 has much more spectra than TR_2, the introduction of a new observation might cause a greater eigenvectors change to TR_2 than TR_1 independently from the true class of the new observation. A similar phenomenon is plausible if TR_1 is populated by spectra well representing the true variability of class 1 while TR_2 is not. In order to account for the different characteristics of the Training Sets the distributional method is assumed.

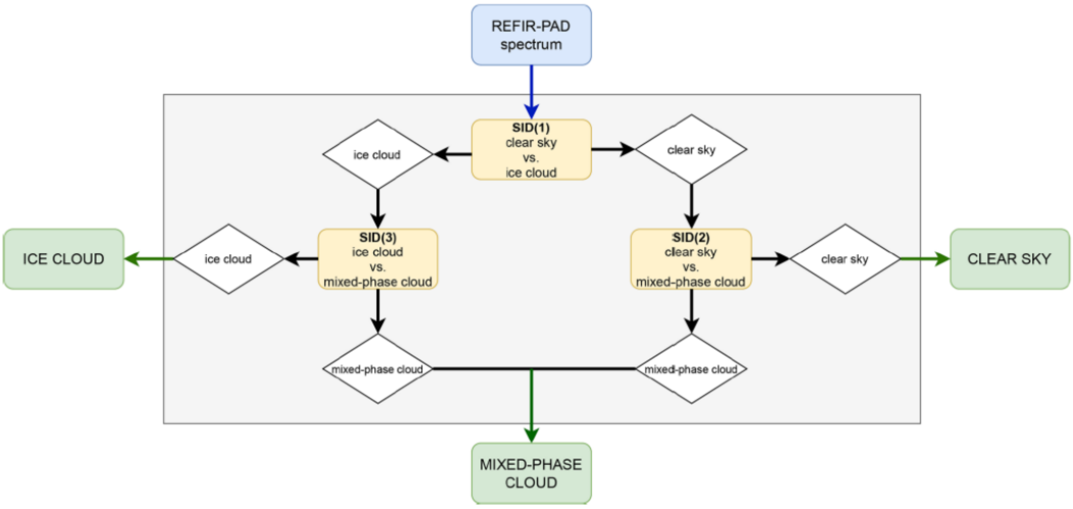


Figure 2.1: Logical scheme of the classification process for three classes

2.3 Training Phase and Classification: Distributional Method

The distributional method requires the learning of one more parameter in the Training Phase. Instead of assigning the observation $\vec{x}$ to class 1 rather class 2 based on the sign of the

$$SID(\vec{x}) = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_2) \in [-1, 1], \tag{2.18}$$

a new quantity is introduced to replace the SID. We define the Corrected Similarity Index Difference (CSID) as

$$CSID(\vec{x}) = SI(\vec{x}, TR_1) - SI(\vec{x}, TR_2) - shift, \tag{2.19}$$

where $shift$ is a number that can be learnt over the Training Sets spectra. For each spectrum of TR_1 and TR_2 the SID is computed. This operation creates two distributions: one for the spectra of TR_1 and one for the spectra of TR_2. The two distributions are used to define the most suitable delimeter for the two classes, selecting a point that maximizes the CIC performances on the Training Sets. The shift optimal value can be obtained using different functions that potentially forecast the performance of the algorithm. In the present work we define the shift as

$$shift = \text{argmax}\left[\frac{1}{2}\left(\frac{TP_1}{TP_1 + FN_1} + \frac{TP_2}{TP_2 + FN_2}\right)\right], \tag{2.20}$$

where TP and FN are the number of spectra respectively correctly classified and incorrectly classified for each class. The quantity $\frac{TP_i}{TP_i+FN_i}$ is called Hit Rate and represent the probability that a spectrum belonging to class i is correctly classified. The mean Hit Rate thus represent the probability that a TR spectrum would be correctly classified if it was a new observation. The mean HR value is close to 1 if and only if both class 1 and class 2 False Negatives are rare.

The Distributional Approach is exemplified and compared to the Elementary Approach in $Figure$ 2.2.

Previous studies of Maestri et al. (2019b) and Magurno et al. (2020)

proved that the distributional approach increases the performance of the classification algorithm, thus it will be adopted for the present work.

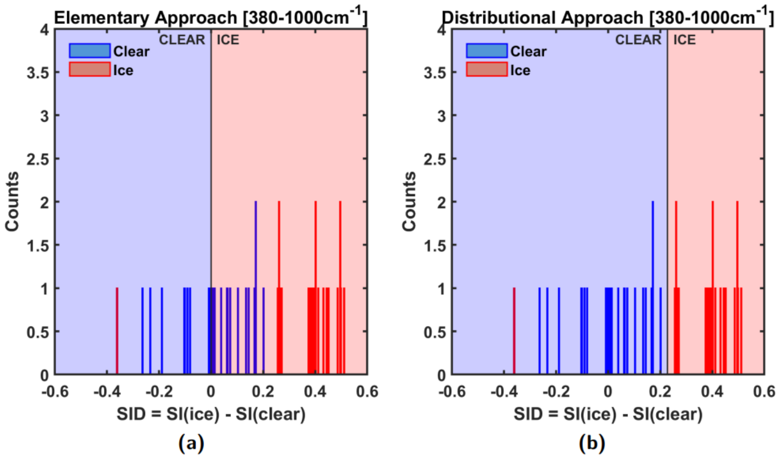


Figure 2.2: Application of the Elementary (a) and Distributional Approach (b) to Training Sets of clear sky scenes and ice cloud scenes

2.4 CIC: introduction of an eigenvalues-based Similarity Index

As seen in the previous section, CIC algorithm classifies a new observation quantifying the eigenvectors change it causes when it is added to a Training Set. In this work we propose the eigenvalues change as a valuable metric. A new Similarity Index is thus introduced:

$$SI_{eigVal}(\vec{x}, TR_i) = -\sum_{j=1}^{P_0} \frac{|\lambda_{j,TR_i} - \lambda_{j,ETR_i}|}{\lambda_{j,TR_i}}, \tag{2.21}$$

where λ_{j,TR_i} represents the eigenvalue associated with the j-th Principal Component of TR_i and λ_{j,ETR_i} represents the eigenvalue associated with the j-th Principal Component of ETR_i. $SI_{eigVal}(\vec{x}, TR_i)$ represents the total principal eigenvalues percentage change. For example, if P_0=5 and SI = -5, the new observation caused an average 100% change to the 5 principal eigenvalues. Since the eigenvalues associated to the Principal Components of some data are positive by construction, we have that

$$SI_{eigVal}(\vec{x}, TR_i) \in (-\infty, 0], \tag{2.22}$$

where 0 means maximal similarity and $-\infty$ means lowest similarity. In the context of PCA the eigenvalue associated to a PC represents the variance of the data along that PC:

$$\lambda_{j,TR_i} = \frac{1}{T_i - 1} \sum_{t=1}^{T_i} |(\vec{x}_t - \vec{\mu}_{TR_i}) \cdot \vec{p}_{j,TR_i}|^2, \tag{2.23}$$

where $\vec{x}_t$ is the t-th TR_i spectrum, T_i is the total number of spectra in TR_i and $\vec{\mu}_{TR_i}$ is the mean of TR_i. The introduction of a new spectrum $\vec{x}$, here labelled as the $(T_i + 1)$-th spectrum of ETR_i, causes the variation of the mean of TR_i. In fact

$$\vec{\mu}_{ETR_i} = \frac{1}{T_i + 1} \sum_{t=1}^{T_i+1} \vec{x}_t = \frac{1}{T_i + 1} \left(\sum_{t=1}^{T_i} \vec{x}_t + \vec{x}_{T_i+1} \right) = \frac{T_i}{T_i + 1} \vec{\mu}_{TR_i} + \frac{\vec{x}_{T_i+1}}{T_i + 1}, \tag{2.24}$$

where the last quantity can be obtained noticing that $\sum_{t=1}^{T_i} \vec{x}_t = T_i \vec{\mu}_{TR_i}$. On the other hand, even the j-th eigenvector $\vec{p}_{j,TR_i}$ is caused to change by the introduction of $\vec{x}_{T_i+1}$. The j-th eigenvalue of ETR_i is thus

$$\lambda_{j,ETR_i} = \frac{1}{T_i} \sum_{t=1}^{T_i+1} |(\vec{x}_t - \vec{\mu}_{ETR_i}) \cdot \vec{p}_{j,ETR_i}|^2. \qquad (2.25)$$

It is expected that if the new spectrum belongs to the class represented by TR_i, the new spectrum does not bring new variance to TR_i and the eigenvalues of TR_i stay approximately the same. Instead, if the new spectrum belongs to another class it should bring new variance to TR_i and the eigenvalues should vary considerably.

It can be shown that the eigenvalues-Based Similarity Index reaches its minimum value ($-\infty$) if and only if the following is true:

$$|\vec{x}_{T_i+1} - \vec{\mu}_{TR_i}| \longrightarrow +\infty. \qquad (2.26)$$

This is a very reasonable property since $|\vec{x}_{T_i+1} - \vec{\mu}_{TR_i}|$ is the Euclidean distance of the new observation $\vec{x}_{T_i+1}$ from $\vec{\mu}_{TR_i}$. A minimal Similarity Index value is thus equivalent to having an observation geometrically infinitely distant from the center of mass of the considered Training Set.

On the other hand one has that

$$|\vec{x}_{T_i+1} - \vec{\mu}_{TR_i}| = 0 \implies SI_{eigVal}(\vec{x}_{T_i+1}, TR_i) = -\frac{P_0}{T_i}. \qquad (2.27)$$

Even this property is reasonable since usually the number of Training Set spectra T_i is much greater than the number of PCs considered. In all the cases considered in this work one has that

$$|\vec{x}_{T_i+1} - \vec{\mu}_{TR_i}| = 0 \implies SI_{eigVal}(\vec{x}_{T_i+1}, TR_i) \simeq 0. \qquad (2.28)$$

A small geometrical distance of $\vec{x}_{T_i+1}$ from the center of mass of TR_i thus implies a Similarity Index close to the maximal value. The results 2.26 and 2.27 are proved in the following part of the section.

Result 1 $|\vec{x} - \vec{\mu}_{TR_i}| = \infty \iff SI_{eigVal}(\vec{x}, TR_i) = -\infty.$

Let's first prove the right implication. $|\vec{x} - \vec{\mu}_{TR_i}| = \infty$ implies that $\vec{x}$ is infinitely distant from $\vec{\mu}_{TR_i}$ along at least one direction of $\mathbb{R}^{N_{VAR}}$ that we denote with the directional vector $\vec{u} \in \mathbb{R}^{N_{VAR}}$. The eigenvectors of the Extended Training Set, $\vec{p}_{j,ETR_i}$ with $j \in \{1, ..., N_{VAR}\}$, form an orthonormal basis of $\mathbb{R}^{N_{VAR}}$. Hence $\vec{u} \in \mathbb{R}^{N_{VAR}}$ can be written as a linear combination of them. For this reason, not all of them can be orthogonal to $\vec{u}$:

$$\exists\ j \in \{1, ..., N_{VAR}\}\ s.t.\ \vec{p}_{j,ETR_i} \cdot \vec{u} \neq 0. \tag{2.29}$$

Since $\vec{x}$ is infinitely distant from $\vec{\mu}_{TR_i}$ along $\vec{u}$, it follows that $\vec{x}$ must be infinitely distant from $\vec{\mu}_{TR_i}$ along $\vec{p}_{j,ETR_i}$ too, because $\vec{p}_{j,ETR_i}$ and $\vec{u}$ are not orthogonal. It follows that the variance along $\vec{p}_{j,ETR_i}$, λ_{j,ETR_i}, is infinite, which implies a $SI_{eigVal} = -\infty$.
The proof of the left implication follows the same logic. $SI_{eigVal} = -\infty$ implies $\lambda_{j,ETR_i} = \infty$, which is equivalent to saying that $\vec{x}$ is infinitely distant from $\vec{\mu}_{TR_i}$ along at least one direction, hence $|\vec{x} - \vec{\mu}_{TR_i}| = \infty$.

Result 2 $|\vec{x} - \vec{\mu}_{TR_i}| = 0 \implies SI_{eigVal}(\vec{x}, TR_i) = -\frac{P_0}{T_i}.$

$|\vec{x} - \vec{\mu}_{TR_i}| = 0$ implies $\vec{x} = \vec{\mu}_{TR_i}$. Since $\vec{x}$ coincides with $\vec{\mu}_{TR_i}$, it must result

$$\vec{\mu}_{ETR_i} = \vec{\mu}_{TR_i}. \tag{2.30}$$

In addition, the Principal Components of TR_i stay the same:

$$\forall j \in \{1, ..., P_0\},\ \vec{p}_{j,TR_i} = \vec{p}_{j,ETR_i}. \tag{2.31}$$

Hence, the generical eigenvalue of ETR_i becomes

$$\lambda_{j,ETR_i} = \frac{1}{T_i} \sum_{t=1}^{T_i} |(\vec{x}_t - \vec{\mu}_{TR_i}) \cdot \vec{p}_{j,TR_i}|^2 = \frac{T_i - 1}{T_i} \lambda_{j,TR_i}. \tag{2.32}$$

Finally, the eigenvalues-based Similarity Index is

$$SI_{eigVal} = -\sum_{j=1}^{P_0} \frac{|\lambda_{j,TR_i} - \lambda_{j,ETR_i}|}{\lambda_{j,TR_i}} = -\sum_{j=1}^{P_0} |1 - \frac{T_i - 1}{T_i}| = -\frac{P_0}{T_i}. \tag{2.33}$$

2.4.1 Training and classification phase

With the adoption of the Eigenvalues-based Similarity Index the algorithm stays conceptually the same. A spectrum $\vec{x}$ is classified as belonging to

$$\begin{cases} \text{class 1, if } SI_{eigVal}(\vec{x}, TR_1) - SI_{eigVal}(\vec{x}, TR_2) - shift > 0 \\ \text{class 2, otherwise.} \end{cases} \tag{2.34}$$

The shift is obtained similarly to what shown before for the distributional method applied to eigenvectors.

2.5 CIC: development of a double-Similarity-Index approach

CIC algorithm defines a single Similarity Index between the new observation and the Training Sets, and assigns the new observation to the class that has the highest Corrected Similarity Index Difference (CSID) with it. In this section we develop a version of CIC algorithm that can exploit two different Similarity Indices in order to enhance the classification performances. In the present section, with SI_{eigVec} we will refer to the Similarity Index explained in 2.2.

2.5.1 Training and classification phase

The Training Sets are defined just like in the single-Similarity-Index CIC. Then, just like in the Distributional Approach, CIC is applied to the spectra of TR_1 and TR_2, whose true class is known. For each of the T_i spectra $\vec{x}_{t,TR_i}$ belonging to TR_i, $i \in \{1, 2\}$, two SIDs are computed:

$$\begin{cases} SID_{eigVec}(\vec{x}_{t,TR_i}) = SI_{eigVec}(\vec{x}_{t,TR_i}, TR_1) - SI_{eigVec}(\vec{x}_{t,TR_i}, TR_2) \\ SID_{eigVal}(\vec{x}_{t,TR_i}) = SI_{eigVal}(\vec{x}_{t,TR_i}, TR_1) - SI_{eigVal}(\vec{x}_{t,TR_i}, TR_2) \end{cases} . \tag{2.35}$$

We have thus a set of $T_1 + T_2$ couples $(SID_{eigVec}, SID_{eigVal}) \in \mathbb{R}^2$. These couples of numbers are represented as points on a plane. In order to separate the two classes a separation line in defined:

$$y = ax + b, \tag{2.36}$$

where a and b are such that the mean Hit Rate is maximal, in analogy with the definition of the $shift$ for the single-Similarity-Index CIC. The line divides the plane in two half-planes. One of the two half-planes is characterized by the presence of TR_1 spectra, while the other is characterized by the presence of TR_s spectra ($Figure$ 2.3). As an example, if TR_1 spectra lie in the half-plane above the line, the new observation $\vec{x}$ must be classified as belonging to class 1 if it falls above the line and as

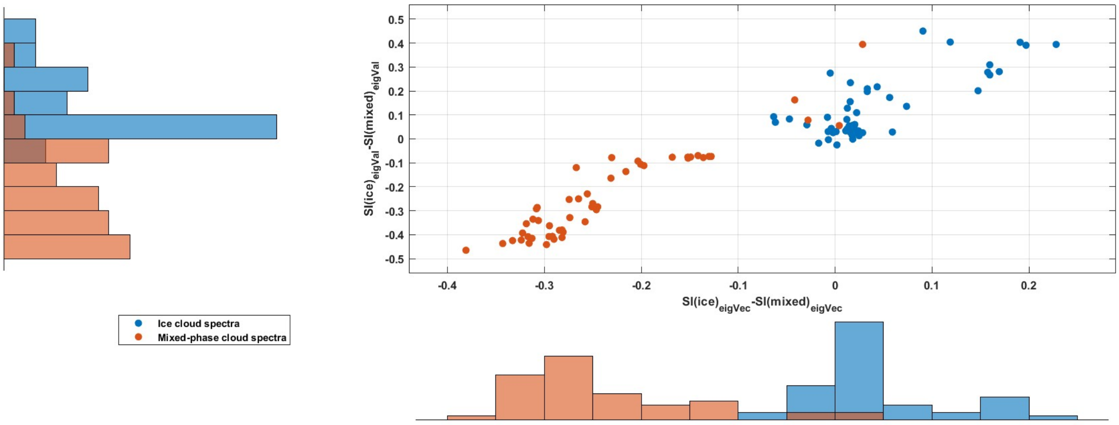


Figure 2.3: Example of application of the double-Similarity-Index approach

belonging to class 2 otherwise:

$$\vec{x} \text{ is assigned to } \begin{cases} \text{class 1, if } SID_{eigVal}(\vec{x}) > a \cdot SID_{eigVec}(\vec{x}) + b \\ \text{class 2, otherwise.} \end{cases} \tag{2.37}$$

It is to be noted that it is not necessary to compute a shift for each Similarity Index as in the single-Similarity-Index CIC since the shifts are here replaced by the class-separating line.

Chapter 3

Testing and Comparing CIC Algorithms using Downwelling Antarctic Radiance Spectra

The principal aim of this chapter is understanding if the adoption of new metrics and methodologies within the CIC algorithm framework can lead to better scene identification and understanding. In order to do so high spectral resolution downwelling radiances collected at Dome-C on the Antarctic Plateau (figure 3.1) between 2013 and 2020 are exploited. The measurements are performed by the REFIR-PAD Fourier transform spectroradiometer, in the context of the projects PRANA (Radiative Properties of Water Vapor and Clouds in Antarctica) and CoMPASs (Concordia MultiProcess Atmospheric Studies), within the Italian National Program for Research in Antarctica (PNRA) ([15]) .
In the first section of this chapter equipment and measurements are described. In section 3.2 the methods exposed in the previous chapter are tested and compared using REFIR-PAD spectra; particular importance is given to the behaviour of the various methods with respect to the number of spectra they are trained on. In section 3.4 the CIC algorithm is applied to the analysis of the entire REFIR-PAD dataset.

3.1 Instrumentation and measurements

The REFIR-PAD instrument (figure 3.3) measures downwelling radiance in the zenith direction within the range $100 - 1500\ cm^{-1}$, with a resolution of $0.4\ cm^{-1}$. It has the capability to detect atmospheric emission in both the FIR and MIR regions of the spectrum. In order to acquire a complete spectrum, four calibration acquisitions and four sky observations are performed. Each acquisition takes approximately 80 seconds, resulting in a total sequence duration of 14 minutes. This includes 5.5 minutes for calibrations, 5.5 minutes for sky observations, and additional time for detector settling (Palchetti et al., 2020). The instrument operates continuously, performing a cycle of 5-6 hours of measurements followed by 1-3 hours of analysis. REFIR-PAD is installed inside an insulated shelter (figure 3.2); within the same shelter a LiDAR (Light Detection and Ranging) instrument is present. This active remote sensing device emits radiation in the visible band at a wavelength of 532 nm. The LiDAR measures vertical profiles of backscattering and depolarization up to 7 km above the surface. By interpreting the signals obtained, valuable information about cloud layers can be derived. In clear sky conditions, the LiDAR's backscatter signal diminishes with increasing altitude. Conversely, when clouds are present, the detected radiation intensifies due to the backscatter from the cloud layer. Additionally, depolarization can indicate the phase of cloud particles. Liquid water droplets maintain the polarization state of the incident beam, whereas non-spherical ice particles exhibit partial depolarization due to internal reflections. Theoretical studies indicate that liquid water droplets cause a polarization change of 2-4%, while non-spherical ice particles cause a more pronounced depolarization of 30-40%. The determination of the cloud's thermodynamic phase threshold relies on atmospheric conditions and cloud microphysical parameters. Furthermore, clouds can consist of multiple layers, each exhibiting a distinct depolarization value. Mixed-phase clouds, for example, often comprise an upper layer of ice particles and a lower layer of water particles, where the temperature is higher. In this study, a depolarization threshold of 15% is utilized to differentiate ice clouds (signal > 0.15) from mixed-phase clouds (signal < 0.15).

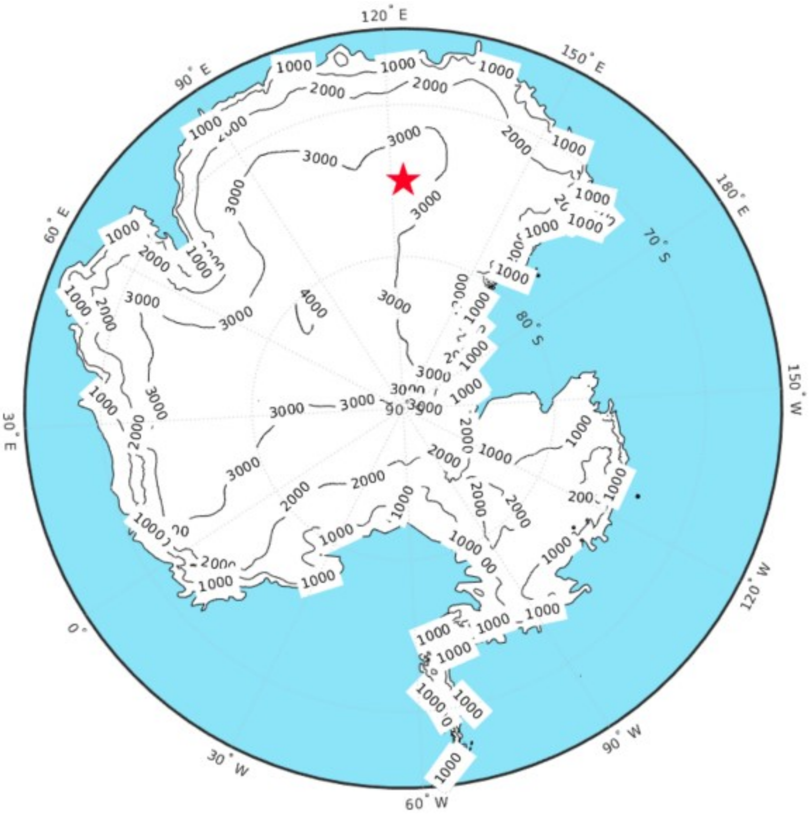


Figure 3.1: Antarctica elevation map, with Concordia Station indicated by a red star.

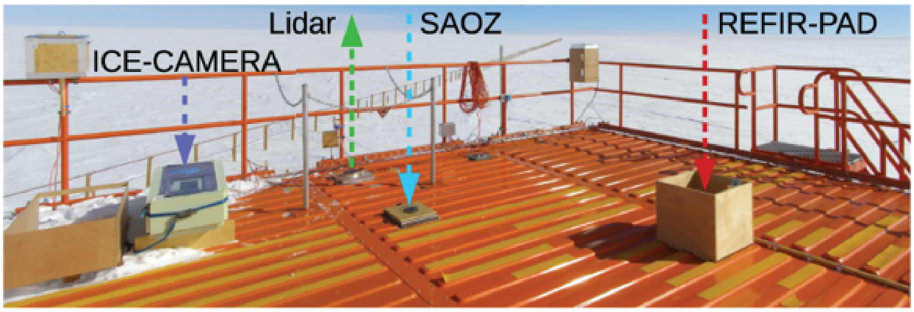


Figure 3.2: The Physical Shelter

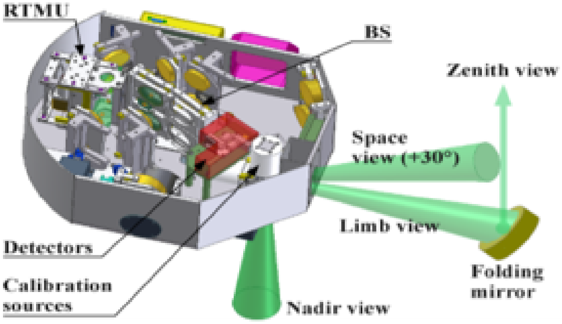


Figure 3.3: The REFIR-PAD instrument

3.2 Validation phase

In this section a methodology for testing and comparing different classification methods on downwelling radiance spectra is developped. First, the spectra collection methodology is described with particular regards to the choice of the classes for each season; later, criteria for performance evaluation are discussed; finally, after describing the methodology of analysis, the results are reported and discussed.

3.2.1 Spectra collection

Spectra used for training and testing are selected using the LiDAR instrument. The profiles obtained from backscatter and depolarization LiDAR were visually examined to identify the presence of clouds and determine their thermodynamic phase. In situations where the sky is clear, the LiDAR signal decreases as altitude increases. However, a sudden increase indicates the existence of a scattering layer, such as a cloud. To differentiate the cloud phase, the depolarization profiles are analyzed. Values above 0.15 indicate the presence of ice clouds, while mixed-phase clouds are identified in layers with depolarization below 0.15. Considering the significant variations in environmental conditions, the classification is conducted separately for two main seasons: a warm season from November to March and a cold season from April to October. Mixed-phase clouds are only observed during the warm season, with a higher occurrence in December and January. In the cold months, the extremely low temperatures prevent the formation of mixed-phase clouds. As a result, three classes are defined for the warm season (clear sky, ice cloud, mixed-phase cloud) and only two classes are defined for the cold season (clear sky and ice cloud).
All the spectra visually classified with the help of LiDAR images and used for the Validation Phase are summarized in the following table.

Class	No. of spectra
Summer clear sky	105
Summer ice cloud	147
Summer mixed-phase cloud	103
Winter clear sky	311
Winter ice cloud	485

The figures 3.4 and 3.5 represent all the collected spectra divided by class and season. The most notable discrepancies arise in two atmospheric windows: one in the far infrared range, specifically between 400 and 600 cm^{-1}, and the other in the middle infrared range, between 800 and 1000 cm^{1}.

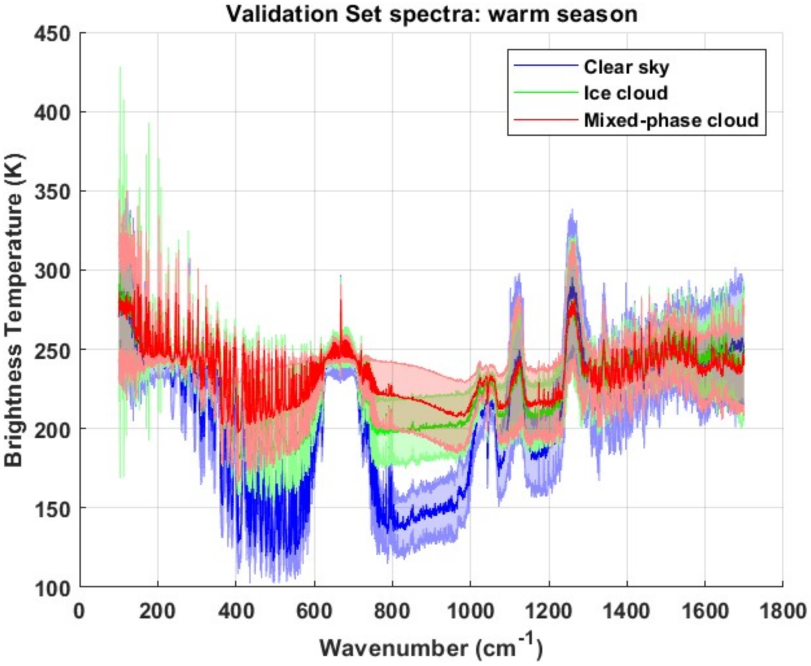


Figure 3.4:

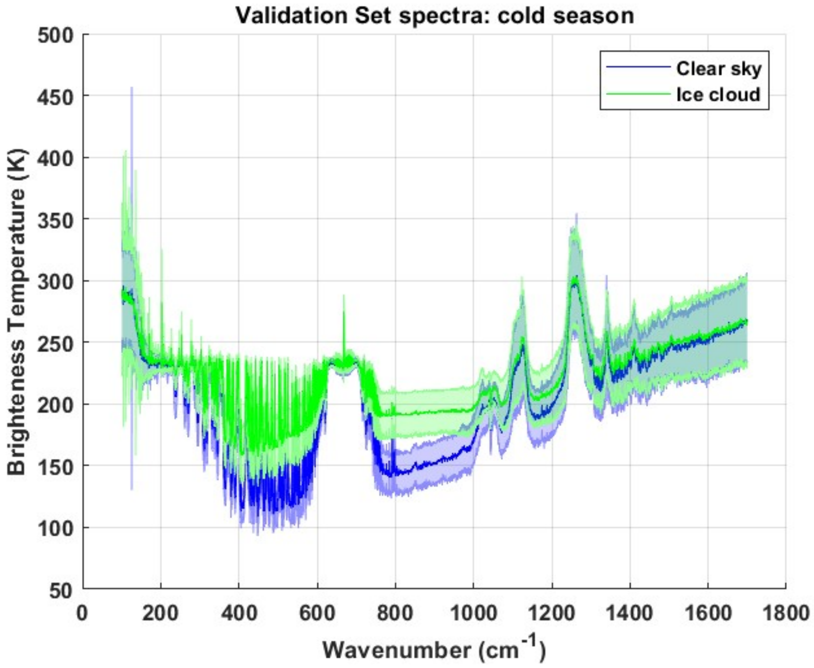


Figure 3.5:

3.2.2 Criteria for performance evaluation

We want to exploit the collected spectra in order to test and compare the different methods for various numbers of spectra in the training sets (the same number for each class). We are mainly interested in two classification features:

- Separation between clear sky scenes and cloudy sky scenes (identification)
- Separation between ice cloud scenes and mixed-phase cloud scenes (classification)

The most important feature is clearly the first one. In fact the most important thing is detecting a cloud if it is present. Furthermore, we are interested in separation between ice cloud scenes and mixed-phase cloud scenes only if we can properly separate a cloudy scene from a clear sky scene.

In order to evaluate the performance of the various methods with respect to the above features, a criterion must be chosen. It is useful to introduce these quantities:

- True Positive, TP_i. The number of elements belonging to class i and correctly classified.
- False Positive, FP_i. The number of elements not belonging to class i but classified as belonging to class i.
- True Negative, TN_i. The number of elements not belonging to class i and correctly classified as not belonging to class i.
- False Negative, FN_i. The number of elements belonging to class i but classified as not belonging to class i.
- $\mathrm{T}_i = \mathrm{TP}_i + \mathrm{FN}_i$. The number of elements belonging to class i.

Based on these quantities, different measures of the goodness of the classification are possible.

- Hit Rate, $\mathrm{HR}_i = \frac{TP_i}{TP_i+FN_i}$. The probability that a spectrum belonging to class i is correctly classified.
- Threat Score, $\mathrm{ThS}_i = \frac{TP_i}{TP_i+FN_i+FP_i}$.
- Positive Predictive Value, $\mathrm{PPV}_i = \frac{TP_i}{TP_i+FP_i}$. The probability that a spectrum classified as belonging to class i actually belongs to class i.
- Misclassification from class i to class j, $\mathrm{m}_{i \longrightarrow j} = \frac{FN_{i \longrightarrow j}}{TP_i+FN_i}$, where $\mathrm{FN}_{i \longrightarrow j}$ represents the number of elements belonging to class i but classified as belonging to class j. The probability that a spectrum belonging to class i is classified as belonging to class j.

It should be noted that the Hit Rate HR_i does not depend on the T_is, while the Threat Score and the Positive Predictive Value do since they include the number of false positives. For example, if we pass from the situation in which $T_1 = T_2$ to the situation in which $T_2 = 10T_1$, HR_1

doesn't change but ThS_1 and PPV_1 can vary greatly. In the present work the HR is preferred for its clear probabilistic interpretation and because it is less dependent on experimental conditions. In order to account for cloud identification we will introduce the identification Hit Rate:

$$HR_{ID} = \frac{1}{4}(HR_{sumCle} + HR_{sumClo} + HR_{winCle} + HR_{winClo}), \qquad (3.1)$$

that can be interpreted as the probability that a spectrum is correctly classified as being relative to a clear sky scene or to a cloudy sky scene. This interpretation can be justified by noticing that the Hit Rate is the probability that a spectrum belonging to a class is correctly classified in that class. All considered classes are given the same weight $\frac{1}{4}$ thus we are making no assumptions over the a-priori probabilities of the single scenes. HR_{ID} is equal to the quantity

$$\frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)}, \; i \in \{sumCle, sumClo, winCle, winClo\} \qquad (3.2)$$

when $T_{sumCle} = T_{sumClo} = T_{winCle} = T_{winClo}$.

In order to account for cloud classification we will introduce the classification Hit Rate:

$$HR_{CLASS} = \frac{1}{2}(\frac{TP_{sumIce}}{TP_{sumIce} + FN_{sumIce \rightarrow sumMix}} + \frac{TP_{sumMix}}{TP_{sumMix} + FN_{sumMix \rightarrow sumIce}}), \qquad (3.3)$$

that can be interpreted as the probability that a warm season spectrum correctly classified as being relative to a cloudy scene is also classified in the correct cloudy class. This interpretation can be justified by noticing that the numerator TP_{sumIce} is the number of ice cloud spectra correctly classified in the ice cloud class, while the denominator $TP_{sumIce} + FN_{sumIce \rightarrow sumMix}$ is the number of ice cloud spectra correctly classified in the cloudy class. The same reasoning can be made for the second term of HR_{CLASS}. The factor $\frac{1}{2}$ ensures that we are making no assumptions over the a-priori probabilities of the single scenes (summer ice cloud and summer mixed-phase cloud). As explained before, the importance of this quantity is conditioned by a high HR_{ID} score.

3.2.3 Methodology of analysis

In order to test and compare the methods for various numbers of spectra in the training sets, a random Monte Carlo subsampling technique has been adopted following this logic:

1. For each class, create a Training Set by randomly choosing N spectra;
2. For each method train the algorithm on the Training Sets;
3. For each method test the algorithm using all the spectra not included in the Training Sets;
4. Save the classification results
5. Repeat for 10 times

The above process is made for $N \in \{10, 20, ..., 100\}$. Point 5 is necessary for finding a mean classification performance and for assigning a standard deviation to it. It is to be noted that the spectra chosen at point 1 are not used for the Testing Phase of point 3, thus preserving the independency of the test set.
The analysis has been run using only the spectral interval $380-1000$ cm^{-1}, which was found by Cossich et al. (2021) to be the most performing for all the three classes.

3.2.4 Results

The overall results regarding the cloud identification are reported in figure 3.6. It is to be noted how for all three methods there is not a great difference among the various Training Sets sizes, that span from 10 to 100. The scores are always higher than 0.94 confirming that CIC and its variants need only a few spectra in order to properly separate clear sky and cloudy sky scenes. The second notable thing is that both the eigenvalues-based CIC and the double-SI CIC have better scores than the classical one and show more stability. After $N = 80$ a decrease is observed. This decrease is probably caused by the small number of spectra available for test when

$N = 90$ and $N = 100$; in fact for $N = 90$ only 15 summer clear sky spectra and 13 summer mixed-phase cloud spectra are available and for $N = 100$ the available spectra of those classes become respectively 5 and 3; the small number of warm season spectra for $N = 90$ and $N = 100$ determines high fluctuations that result in a worse average score.

For what concerns the capacity of the methods to classify ice cloud and mixed-phase cloud scenes in the correct cloudy class, figure 3.7 shows that, contrarily to the case of cloud identification, the performances get better when the algorithms are trained on a large number of spectra. For all methods the performances gets better and better until $N = 90$; at $N = 100$ a little decrease is observed. This decrease is totally analogous to the one observed in the previous graph.

Notable is the fact that for $N = 10$ the double-SI CIC shows a much higher score than the other methods: 0.85 instead of about 0.75. This shows that the double metric approach, being able to exploit the information coming from both eigenvectors and eigenvalues change, is particularly suited for a small number of Training Set spectra. When $N > 10$ the scores of the classical CIC are slightly better than the scores of the double-SI CIC and more decisely better than the eigenvalues-based CIC.

For what concerns the separation of clear sky and cloudy sky scenes, we can conclude that the eigenvalues-based Similarity Index permits better scores than the classical, eigenvectors-based Similarity Index (3.6). On the contrary, for what concerns the separation of ice cloud scenes and mixed-phase cloud scenes, we can conclude that the classical eigenvectors Similarity Index permits better scores than the eigenvalues-based one (3.7). The introduction of a double-Similarity Index, that maps the spectra in the 2-dimensional space generated by both the eigenvectors change and the eigenvalues change, permits scores that are always better than those of the least performing Similarity Index and close or better to those of the best performing Similarity Index. Considering all this observations and the fact that the HR_{ID} score is the most important score, the double-SI CIC will be used in this chapter to analyze the entire REFIR-PAD dataset.

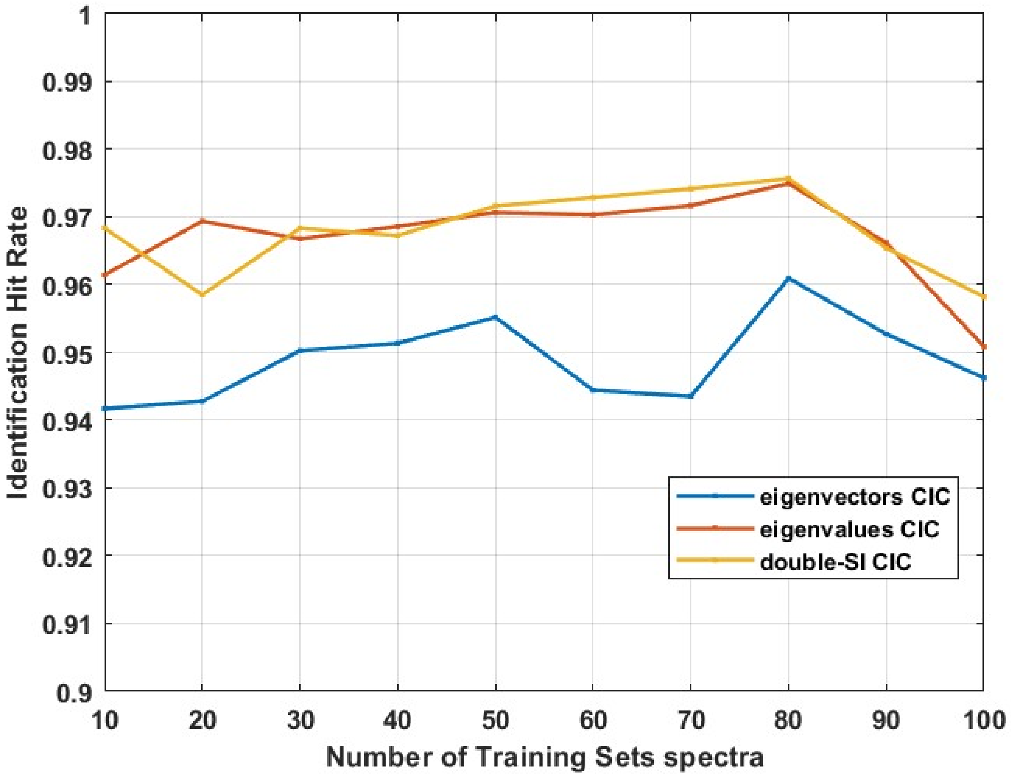


Figure 3.6: Identification Hit Rate for the classical CIC (blue), eigenvalues-based CIC (red) and double-Similarity-Index CIC (yellow) as a function of the number of spectra in the Training Sets. The reported values are averaged over 10 observations

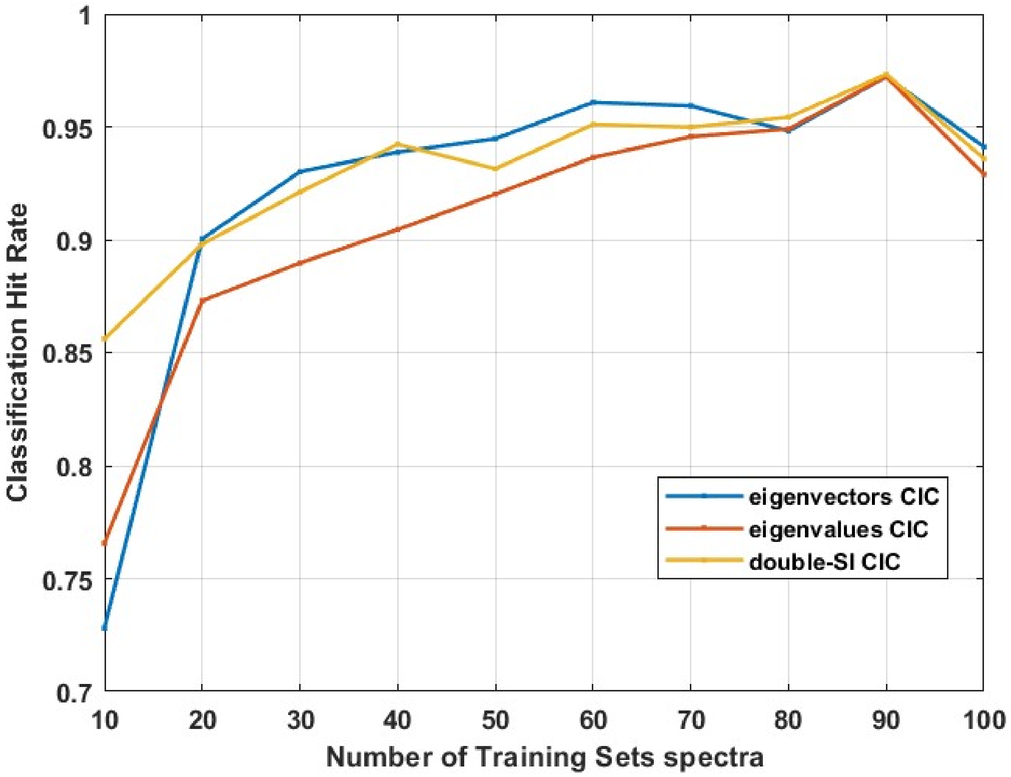


Figure 3.7: Classification Hit Rate for the classical CIC (blue), eigenvalues-based CIC (red) and double-Similarity-Index CIC (yellow) as a function of the number of spectra in the Training Sets. The reported values are averaged over 10 observations

3.3 Analysis of the entire REFIR-PAD dataset

In this section both the double-SI CIC and the classical CIC are trained over Training Sets populated with 50 spectra chosen from the Validation Set of section 3.2 and tested over the remaining spectra. The number 50 was chosen in order to have a balance between a representative Training Set and a statistically significant Test Set. Then, the results of the two methods are discussed with regards to the Test Set as well as to LiDAR images. Finally, the double-SI CIC is used to analyze the entire (2013-2020) REFIR-PAD dataset.

3.3.1 Training and Test Phase

The Training Sets spectra are chosen from the Validation Set collected as described in section 3.2. The double-SI CIC and the classical (eigenvectors-based) CIC is trained on them learning the following parameters:

	class-separating line (double-CIC)	shift (classical CIC)	P_0
sumCle/sumIce	y=-1.6669x+0.0819	0.0143	4
sumCle/sumMix	y=-1.1934x+0.4477	0.2083	8
sumIce/sumMix	y=-2.1251x-0.1973	-0.0947	8
winCle/winIce	y=-0.5255x-0.0063	-0.0043	3

The double-SI CIC is then tested on the Test Set giving the following results:

Field	No. of spectra	Hit Rate	Misclassification
Summer clear sky	55	96.4%	3.6% ice cloud
Summer ice cloud	97	96.9%	3.1% clear sky
Summer mixed-phase cloud	53	94.3%	5.7% ice cloud
Winter clear sky	261	98.1%	1.9% ice cloud
Winter ice cloud	435	95.4%	4.6% clear sky

The same is done with the classical CIC, that gives the following results:

Class	No. of spectra	Hit Rate	Misclassification
Summer clear sky	55	89.1%	5.5% ice cloud 5.4% mixed-phase cloud
Summer ice cloud	97	94.8%	3.1% clear sky 2.1% mixed-phase cloud
Summer mixed-phase cloud	53	96.2%	3.8% ice cloud
Winter clear sky	261	97.7%	2.3% ice
Winter ice cloud	435	92.2%	7.8% clear

The following table shows the overall classification performances of the two methods:

Method	HR_{ID}	HR_{CLASS}
Classical CIC	94.4%	97.0%
Double-Similarity-Index	97.0%	97.15%

The identification Hit Rate is higher for the double-SI CIC as was expected from the analysis of section 3.2. For what concerns the classification Hit Rates, a lower score for the double-SI CIC was expected. Anyway this result is totally consistent considering that the values reported in section 3.2 are average values.

The Hit Rates of the single classes are always better with the double-SI CIC except in the case of summer mixed-phase cloud spectra, where the classical CIC has a higher score. Anyway, the fact that both summer clear sky spectra and summer ice cloud spectra are sometimes misclassified as mixed-phase clouds by the classical CIC, leads to think that the latter has a bias towards mixed-phase clouds.

Figure 3.8 represents summer clear sky spectra and summer mixed-phase cloud spectra separated by the double-SI CIC (skew line) and by the classical CIC (vertical line). The left image represents the spectra on which the algorithms are trained. The double-SI CIC is able to achieve a perfect separation exploiting both the eigenvalues-based Similarity Index (on the vertical axis) and the eigenvectors-based Similarity Index (horizontal axis). Instead, the classical CIC permits a worse separation since it can only use the information coming from the eigenvectors-based Similarity Index. The right image represents the spectra belonging to the Test Sets: the skew line of the double-SI CIC permits the higher Hit Rates reported in the previous tables.

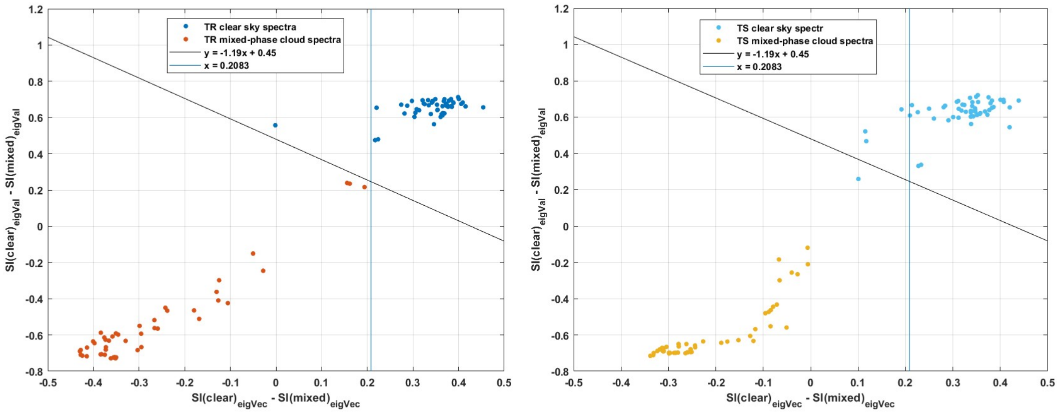


Figure 3.8: On the left spectra belonging to clear sky and mixed-phase cloud Training Sets separated by the double-SI CIC (skew line) and by the classical CIC shift (vertical line); on the right, the spectra belonging to the Test Sets of the same classes

It is useful to inspect some LiDAR images in order to have examples of different classifications performed by the two methods. On the 29th of October 2013 (3.9) clouds are present from about 12 until the end of the day, with brief clear sky interruptions. The clouds cause a depolarization ratio of at least 15, sign that their phase is not liquid. Between 12:30 and 23, 12 spectra have been measured by REFIR-PAD. 12 of this 14 spectra were classified as ice cloud spectra by the double-SI CIC (in the time intervals 13:16-13:30 and 14:38-14:52 the scene was classified as a clear sky scene). All of the 12 spectra were classified as relative to a clear sky scene by the classical CIC. After 23 the scene is characterized by a higher and multi-layered depolarizing cloud detected by both methods.

On the 27th of December 2013 (3.10) from about 15 both algorithms detect cloud presence. After 22, however, the classical CIC start detecting

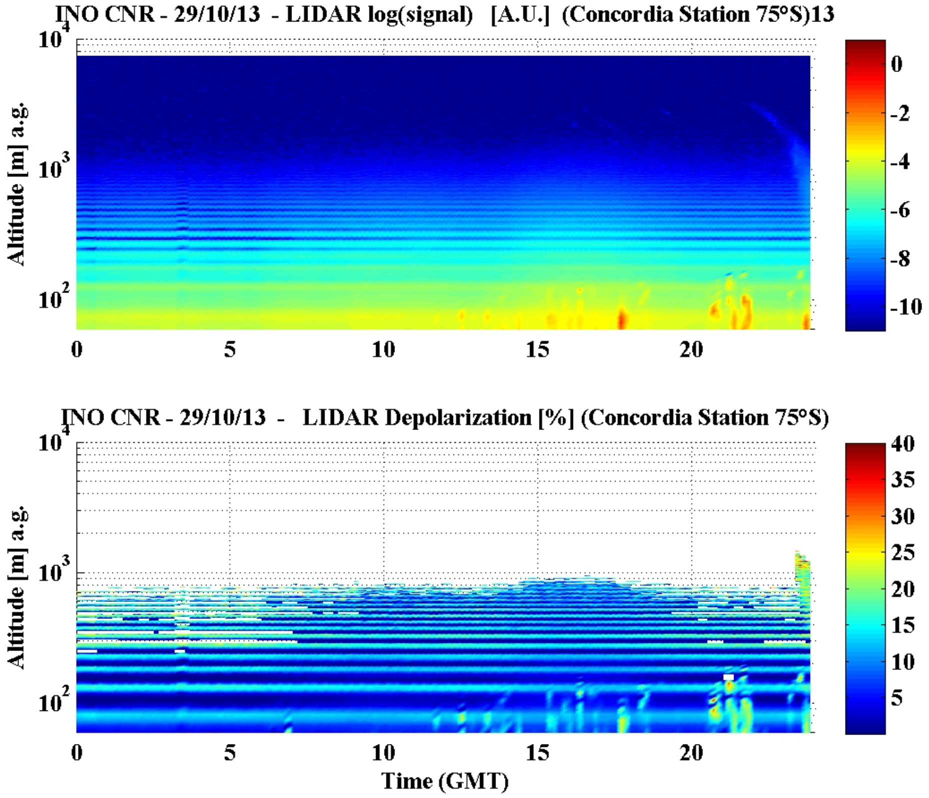


Figure 3.9: Lidar backscattering and depolarization ratio for 29 October 2013.

the presence of mixed-phase clouds while the visual inspection suggests only the presence of strongly depolarizing clouds (depolarization > 15).

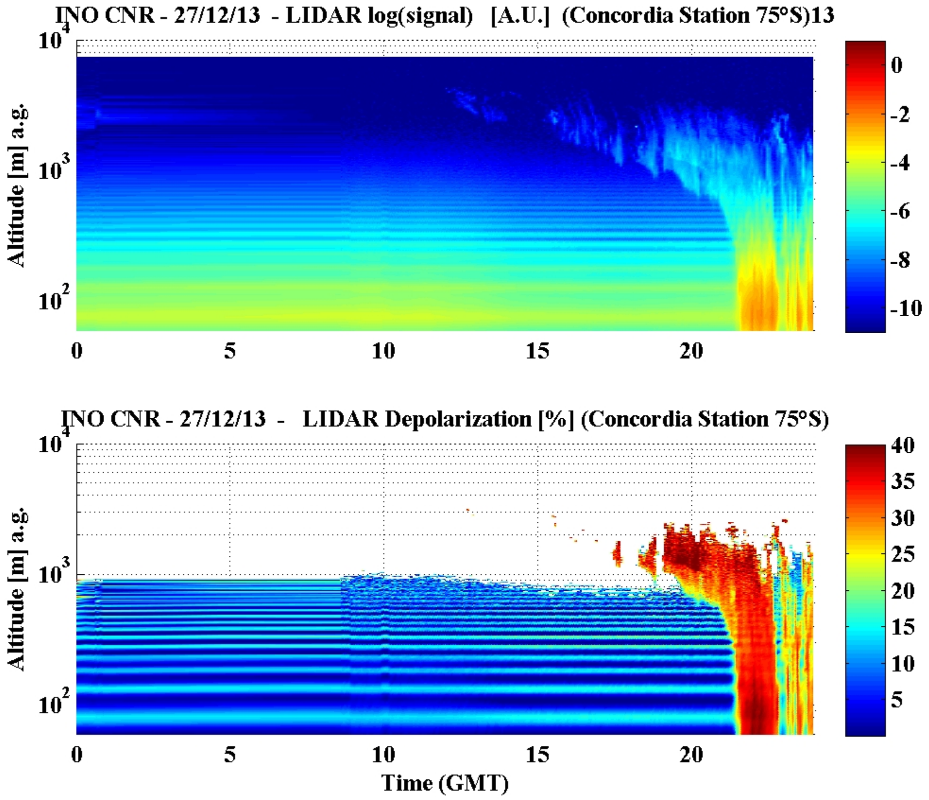


Figure 3.10: Lidar backscattering and depolarization ratio for 27 December 2013.

3.4 Results over the entire dataset

The double-SI CIC is finally run over the entire dataset using the previously defined spectral interval of $380 - 1000\ cm^{-1}$ in order to deduce clear sky and cloud statistics over the year 2013-2020. In order to report results that take into account both False Negatives and False Positives, a method exploiting the Test Set Hit Rates and misclassifications has been used. We will suppose that the Hit Rates found on the Test Set are representative of the whole dataset. By the definition of False Positives and False Negatives we can write

$$\begin{cases} N^{CIC}_{clear} = N_{clear}HR_{clear} + N_{ice}m_{ice \to clear} + N_{mixed}m_{mixed \to clear} \\ N^{CIC}_{ice} = N_{clear}m_{clear \to ice} + N_{ice}HR_{ice} + N_{mixed}m_{mixed \to ice} \\ N^{CIC}_{mixed} = N_{clear}m_{clear \to mixed} + N_{ice}m_{ice \to mixed} + N_{mixed}HR_{mixed} \end{cases}, \tag{3.4}$$

where the quantities N^{CIC}_i represent the number of dataset spectra classified in the class i but not necessarily truely belonging to class i, and where the quantities N_i represent the number of dataset spectra truely belonging to class i. The quantities HR_i and $m_{j \to i}$ are learned on the Test Set and represent respectively the rate of correctly classified class i spectra (true positives and false negatives) and the rate of uncorrectly classified class j spectra (false positives). The true numbers of spectra N_{clear}, N_{ice} and N_{mixed} can be computed by solving the system. The final statistics for the years 2013-2020 follows:

	N_{clear}(%)	N_{ice}(%)	N_{mixed}(%)	Observation time (%)
Total	76.25	20.39	2.69	69.11
2013	75.79	22.43	1.78	29.61
2014	77.20	20.42	2.38	67.36
2015	76.39	21.09	2.52	65.27
2016	72.19	25.23	2.58	82.32
2017	74.94	22.22	2.84	79.86
2018	75.20	18.06	3.94	87.14
2019	76.29	20.01	3.18	90.75
2020	82.02	15.65	2.33	90.57

This statistics is compatible with what obtained in [13] using the classical CIC:

	N^{CIC}_{clear}(%)	N^{CIC}_{ice}(%)	N^{CIC}_{mixed}(%)
Total	70.09	27.69	2.21

The percentages reported by the authors of [13] represent the number of clear skies, ice clouds and mixed-phase clouds detected by the CIC. Thus, substituting these percentages to the variables N^{CIC}_{clear}, N^{CIC}_{ice} and N^{CIC}_{mixed} in the system 3.4, and accounting for the Hit Rates and misclassifications reported as always in [13], the resulting percentages are totally compatible with those obtained in the present work.

Chapter 4

Cloud Detection from Satellite Simulations

In this chapter the double-SI CIC is used to analyze the synthetic PREFIRE observations. PREFIRE (Polar Radiant Energy in the Far InfraRed Experiment) is a future mission selected under NASA's Earth Ventures Instrument (EVI) aimed at documenting, for the first time, the spectral, spatial, and temporal variations of polar far-infrared emission ([10]). The synthetic radiances are relative to the acquisition of one of the two PREFIRE satellites, for a total of 12 orbits, one for each month of the year 2021.

4.1 Dataset

The simulated radiances are produced with the Principal Component based Radiative Transfer Model (PCRTM; [9]), starting from ERA5 reanalysis. For simulating clear sky radiances, PCRTM requires the temperature, and gasses profiles in 101 levels from 0.005 hPa to the surface, along with the surface properties. The profiles used depend on latitude and season.

For the simulation of cloudy-sky radiances, PCRTM needs cloud phase, cloud optical depth, and cloud effective size for each level where the cloud is present. This data are provided by simulations performed by the Geophysical Fluid Dynamics Laboratory (GFDL) in the context of the first intercomparison project of global storm-resolving models, i.e., the Dynamics of the Atmospheric general circulation Modeled On Nonhydrostatic Domains (DYAMOND) ([16]). Typical surface emissivities are adapted from [8]. From the PCRTM radiances, 54 channels are derived using the appropriate spectral response functions of the PREFIRE Thermal InfraRed Spectrometer (TIRS). The wavelengths span from 0.84 μm to 53.16 μm.

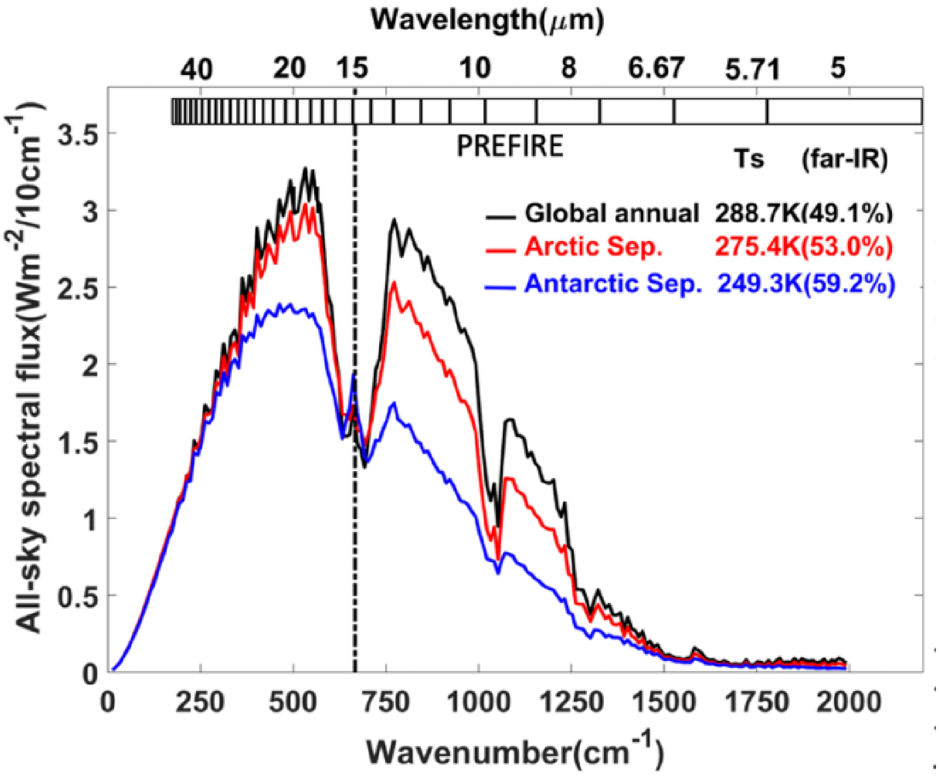


Figure 4.1: All-sky spectral fluxes for the entire globe and for the poles only. The striped bar on the top indicates the PREFIRE channels. ([10])

4.1.1 Orbits and field of view

The spectra are relative to 12 simulated orbits of the satellite, one for each month of the year 2021. The following table summarizes the geographical distribution of the data:

	Tropical Band	Polar Bands	Temperated Bands
No. of spectra	259893	253113	264582

An example of the field of view of the satellite is provided by figure 4.2, that is relative to the 1st of January and represents the Brightness Temperature in the atmospheric window. One can notice the high temperature of the Ocean ($> 290K$) and the low temperatures of high cloud layers over the continent.
Figure 4.3 compares the surface temperature used (among the other parameters) for simulating the radiances of the orbit of January with the atmospheric window Brightness Temperature. Only clear sky scenes have been used in order to highlight the similarity of the two quantities.

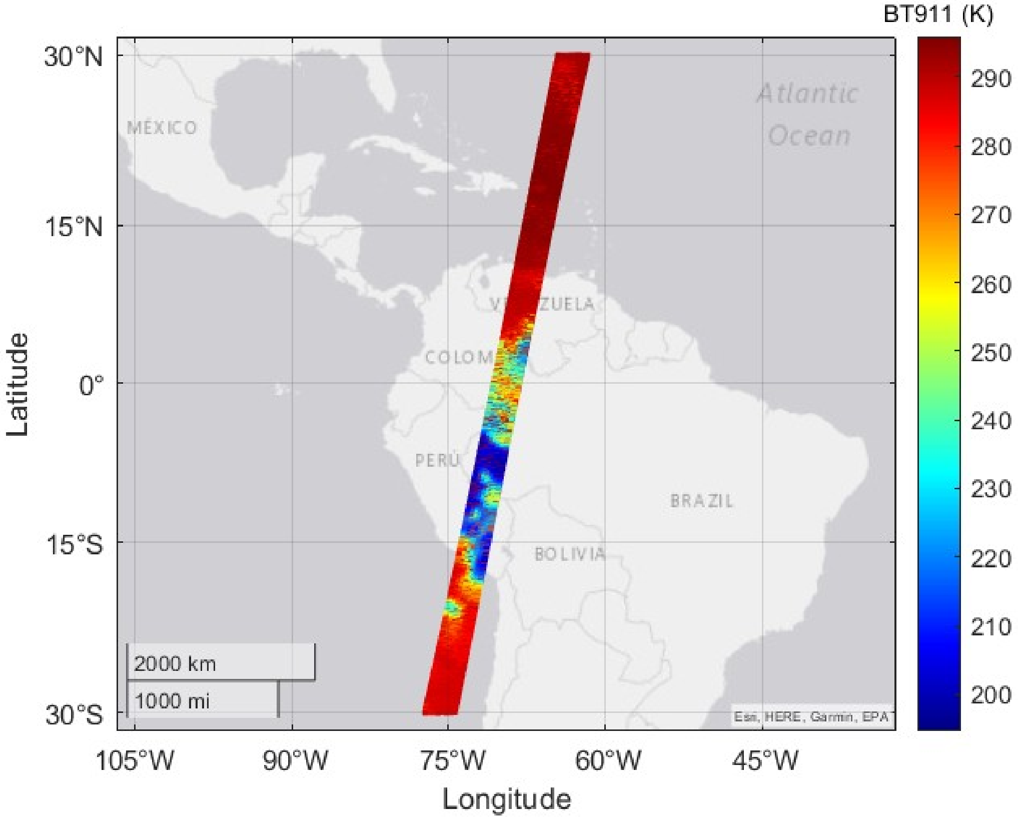


Figure 4.2: January's orbit: Brightness Temperature at 911 cm^{-1}

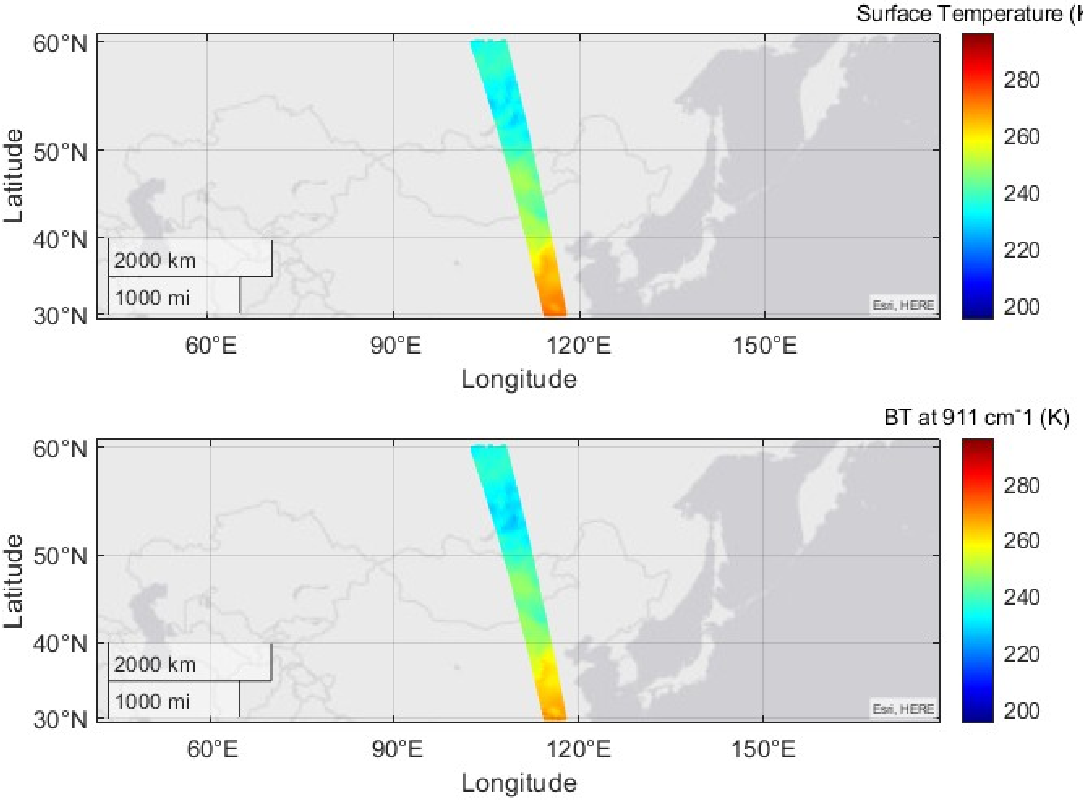


Figure 4.3: January's orbit: clear sky surface temperature (above) and Brightness Temperature at 911 cm^{-1} (below)

4.2 CIC Algorithm Set-Up

In order to train the CIC algorithm over the PREFIRE spectra, only channels from 188 to 1180 cm^{-1} are used. All the radiances are transformed into Brightness Temperature spectra.

4.2.1 Training Sets

To account for the variety of temperatures and climates that characterize the observed scene, the analysis is divided by climate band (Arctic, Northern Temperated, Tropical, Southern Temperated, Antarctic) and by period of the year (two six-month periods, one centered in mid July and one centered in mid January). It has been chosen to consider 4 classes, for a total of 4×5×2 = 40 Training Sets. During the Test Set analysis, the new observation will be classified using the parameters relative to the appropriate period and climate band.
Following the analysis of section 3.3, the Training Sets will be populated with 60 spectra chosen randomly from the appropriate period, true class and latitude. The following table represents the number of spectra relative to each period and each latitude band.

4.2.2 Class definition

The classes are defined based on some of the input physical quantities that PCRTM needs in order to simulate the final radiances. Each atmospheric layer $i \in \{1, ..., 101\}$ is characterized by an Optical Depth OD_i and by a cloud flag that can have the values 'clear', 'ice cloud' or 'liquid cloud'. If the OD_i of the i-th layer is zero the cloud flag signals the absence of cloud, while if OD_i is different from zero the cloud flag signals the presence of an ice cloud or of a liquid cloud. Hence for each scene a total ice Optical Depth can be defined:

$$OD_{ice} = \sum_{i=1}^{101} OD_{i,ice}. \tag{4.1}$$

Similarly, a liquid phase Optical Depth can be defined:

$$OD_{liq} = \sum_{i=1}^{101} OD_{i,liq}. \tag{4.2}$$

The total Optical Depth is then defined as the sum of the two quantities:

$$OD_{tot} = OD_{ice} + OD_{liq}. \tag{4.3}$$

For each latitude, band and period four classes are introduced:

- a class representing clear sky scenes;
- a class representing thick liquid (or mixed-phase) clouds;
- a class representing ice clouds;
- a class representing thin clouds.

The classes are defined based on the total Optical Depth characterizing the observed scene and based on the latitude band. For tropical and temperated bands the classes are defined as follows:

1. Clear sky: $OD_{tot} < 0.03$
2. Ice cloud: a number n of layers are present in the superior part of the atmosphere such that $\sum_{i=1}^{n} OD_{i,ice} > 3$ and $\sum_{i=1}^{n} OD_{i,liq} = 0$;
3. Thin cloud: $OD_{tot} \in [0.03, 3]$;
4. Thick liquid and mixed-phase cloud: otherwise.

The clear sky class is characterized by very low, but not necessarily zero, optical depth; it is expected a resulting radiance quite distinguishable from the cloudy scenes radiances, even if it could present spectral features similar to those of very thin clouds. Is is to be noted how a large variety of physical situations are classified as 'thick liquid and mixed-phase clouds'. Instead, the ice cloud class should be very well characterized since the physical situation it refers to is an optically thick ice cloud standing, eventually, above other clouds. Differently, for polar regions only thin ice clouds are accepted in the thin cloud class, and partially-liquid thin clouds are assigned to the class of mixed-phase clouds:

1. Clear sky: $OD_{tot} < 0.03$
2. Ice cloud: a number n of layers are present in the superior part of the atmosphere such that $\sum_{i=1}^{n} OD_{i,ice} > 3$ and $\sum_{i=1}^{n} OD_{i,liq} = 0$.
3. Thin ice clouds: $OD_{tot} \in [0.03, 3]$ and $OD_{liq} = 0$.
4. Mixed-phase cloud: otherwise.

Even in this case, the mixed-phase clouds class is associated to a variety of physical situations while the other classes are more precisely characterized.

It has been chosen not to differentiate clear sky scenes on the base of different land compositions nor sea or ice presence even though this information is input of PCRTM simulator. Such advancements will be object of further analyses if CIC algorithm won't be able to achieve good classifications using Training Sets built in this way.

Figures 4.4, 4.5 and 4.6 represent the total optical depth distribution for the Training Sets of the southern temperated band, the tropical region and the Antarctic region respectively.

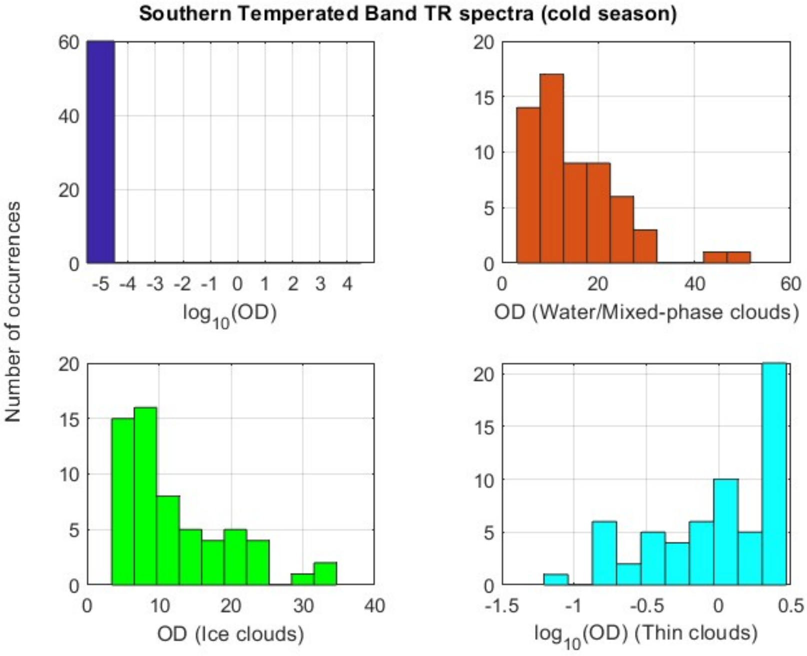


Figure 4.4: Southern Temperated Band TR total OD

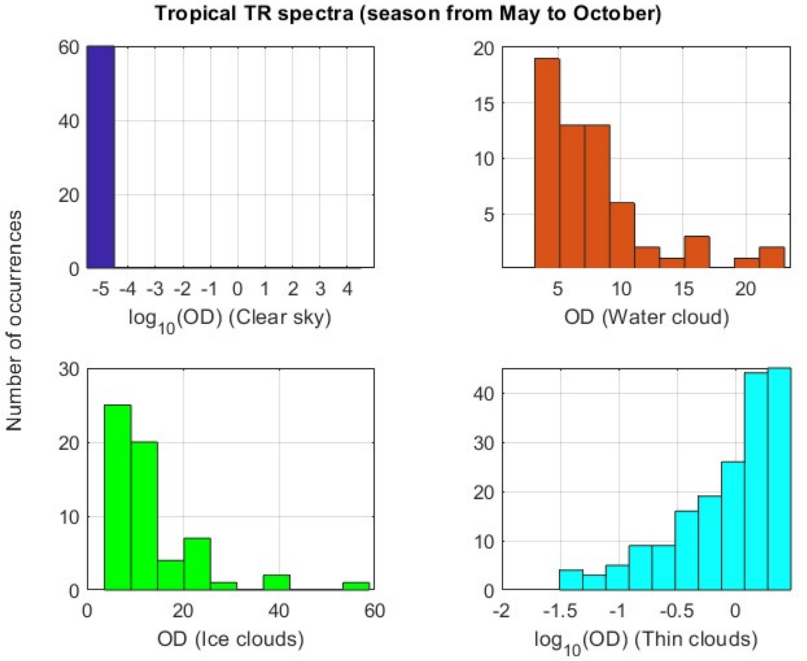


Figure 4.5: Tropical TR total OD

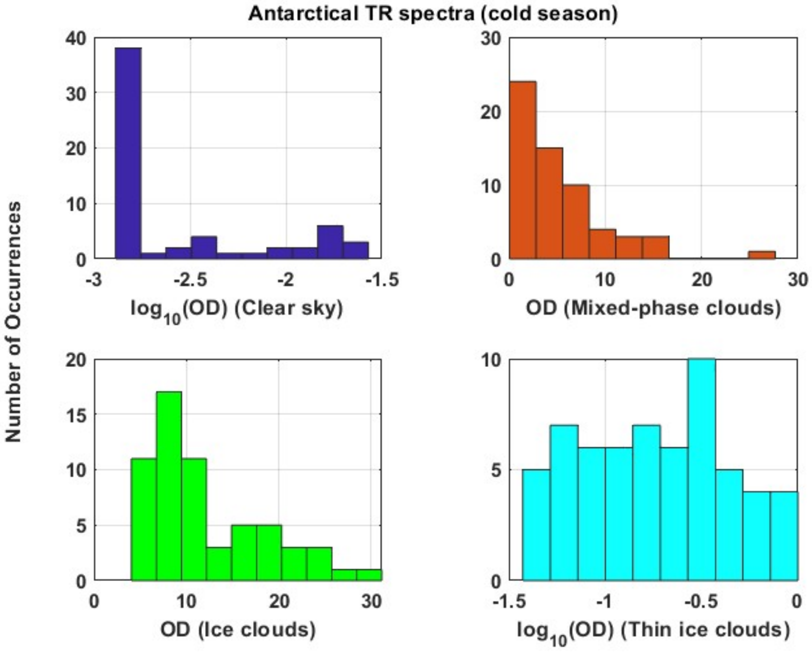


Figure 4.6: Antarctic TR total OD

Figures 4.9, 4.7 and 4.8 represent the average Training Sets spectra of Antarctic region, tropical band and southern temperated band for both the periods of the year. The Brightness Temperature (BT) at 11 μm is approximately equal to the temperature of the physical object emitting the observed radiance; it can be noted that the BT at 11 μm is always maximal for clear sky scenes except for Antarctica, where the highest values are reached by mixed-phase clouds. This fact is reasonable given the frequent thermic inversions that characterize Antarctica. The coldest scenes are always the ice cloud scenes as was expected. Another reasonable aspect of these average spectra is the fact that thin clouds usually have spectral features similar to clear sky scenes and to mixed-phase cloud scenes except for Antarctica, where the Thin Cloud class is populated only with optically thin ice clouds. Over Antarctica in the warm season, no thick ice clouds have been found; this should not surprise since the great number of spectra are relative to just twelve days of the year 2021.

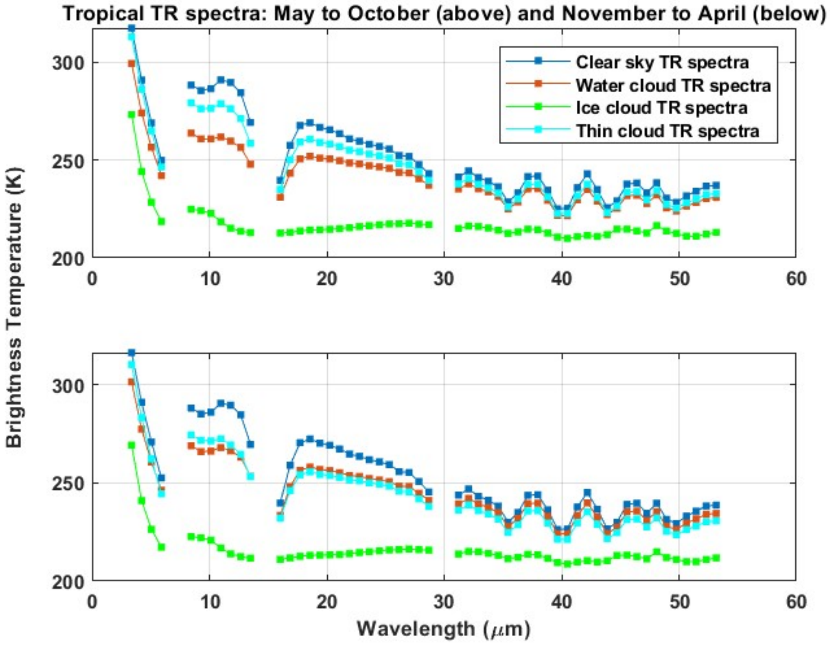


Figure 4.7: Equator TR

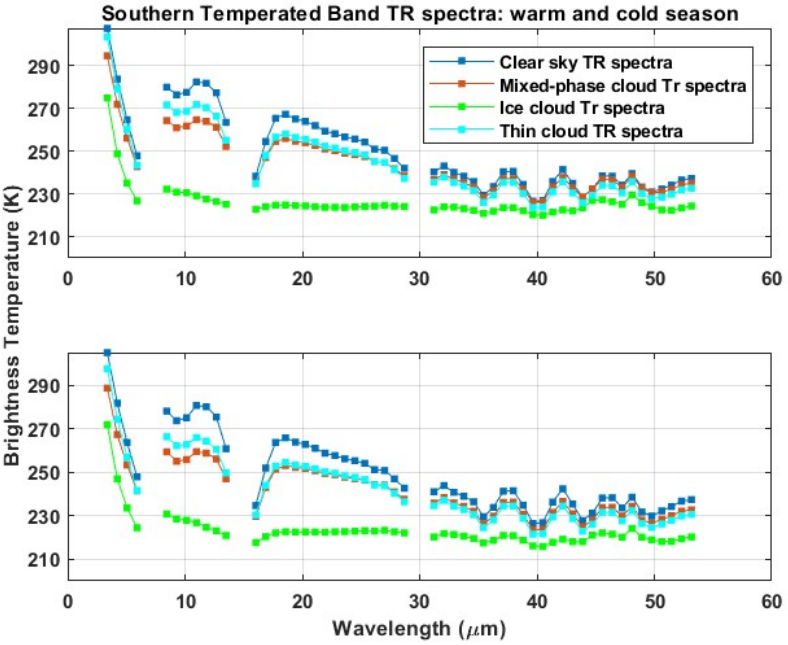


Figure 4.8: Southern Temperated Band TR

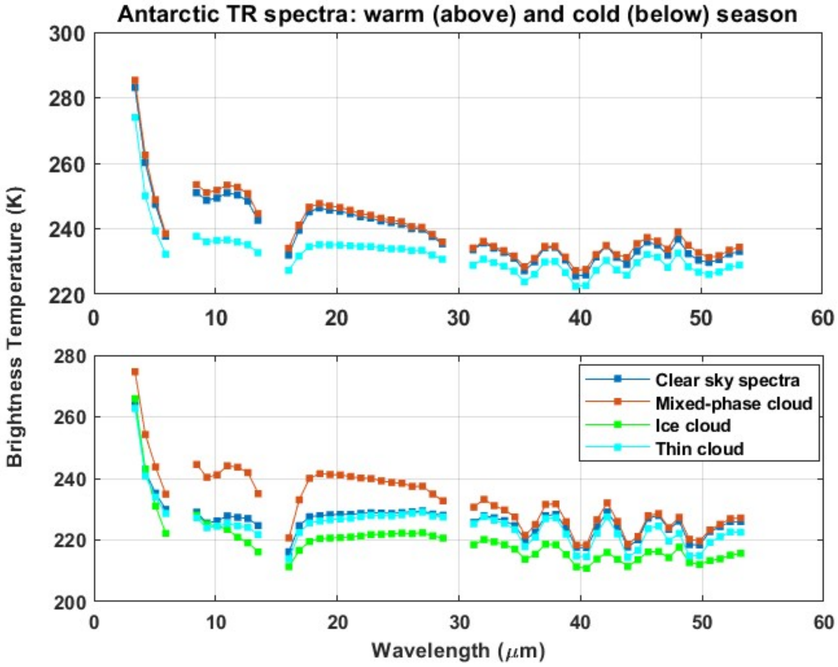
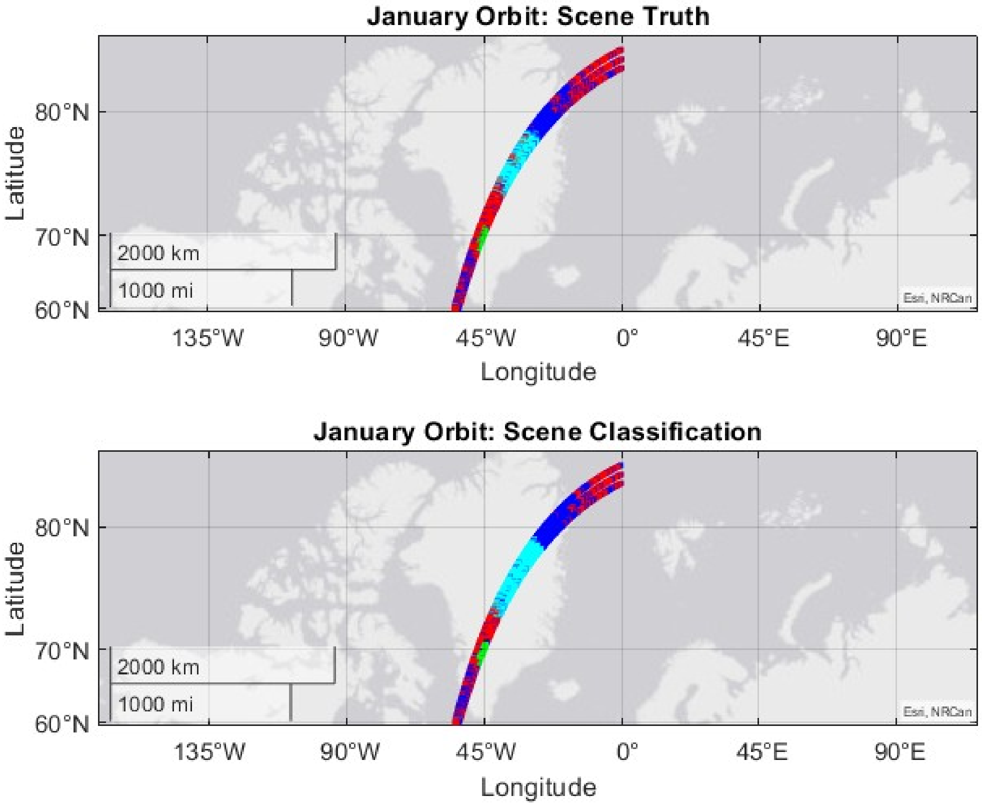


Figure 4.9: Antarctica TR

4.3 Results

The double-SI CIC is trained on the Training Sets defined in the previous section and used to analyze all the spectra in the dataset. The figures 4.10 and 4.11 represent subsets of PREFIRE orbits. The above images represent scene truth while the below images represent scene classification. It can be noticed how differences are present, especially among cloudy classes, and much less between the clear sky and some cloudy class.



January Orbit: Scene Truth

January Orbit: Scene Classification

Figure 4.10: Clear sky scenes (blue), mixed-phase cloud scenes (red), ice cloud scenes (green) and thin ice cloud scenes (cyan): truth (above) and classification (below).

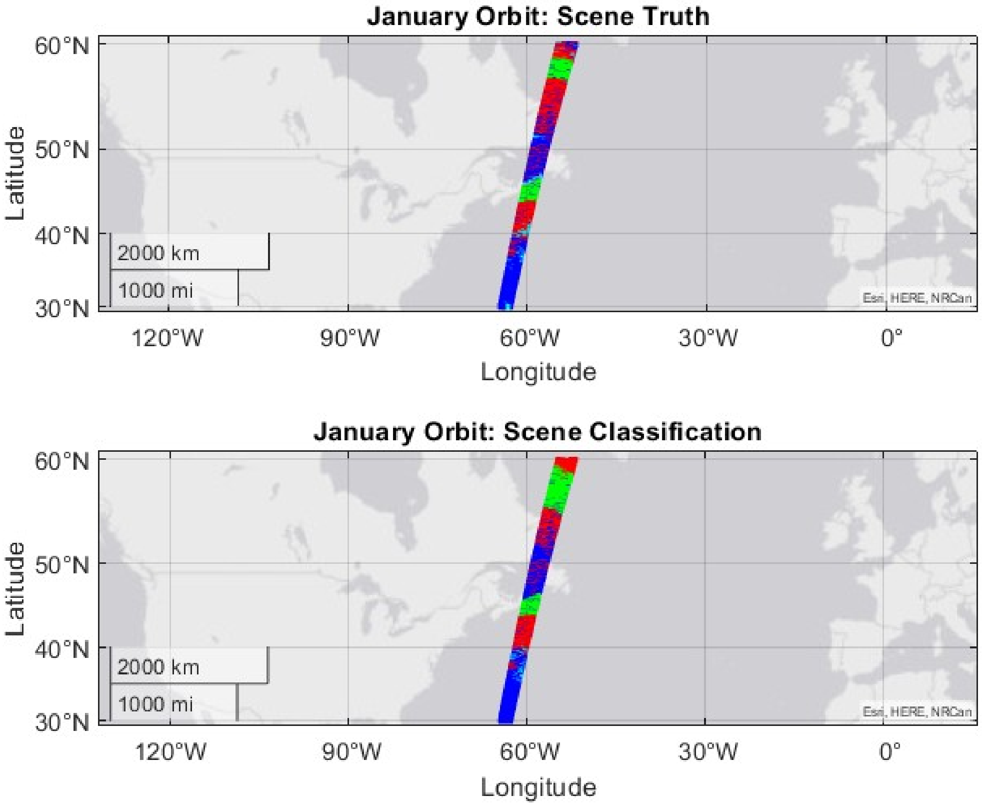


Figure 4.11: Clear sky scenes (blue), mixed-phase cloud scenes (red), ice cloud scenes (green) and thin ice cloud scenes (cyan): truth (above) and classification (below).

Global results

The following table reports the global results for clear sky and cloudy sky only.

Class	No. of spectra	Hit Rate	Misclassification
Clear sky	396081	94.2%	5.1% cloud 0.7% unclass
Cloud	381455	93.1%	5.4% clear sky 1.5% unclass

The results are very good for both cloud detection and clear sky identification. The result is even more notable if one notices that only 50 channels have been used and 60 training spectra.
The following table represents the global classification results with respect to single classes.

Class	No. of spectra	Hit Rate	Misclassification
Clear sky	396081	94.2%	1.4% liquid/mixed-phase 3.7% thin cloud 0.7% unclass
Liquid/mixed-phase	257973	80.5%	5.7% clear sky 1.7% ice cloud 10.6% thin cloud 1.5% unclass
Ice cloud	40085	92.8%	6.2% liquid/mixed-phase 0.8% thin cloud 0.2% unclass
Thin cloud	83397	77.1%	7.1% clear sky 13.2% liquid/mixed-phase 0.5% ice cloud 2.1% unclass

It can be noted how the class of ice clouds, the best characterized class, is the cloudy class with the highest hit rate (92.8%) while liquid and mixed-phase clouds and thin clouds have a relatively smaller hit rate (80.5% and 77.1%). This fact is not surprinsing since these two classes share different features with others. As an example 7% of thin clouds are misclassified as clear skies; it is plausible that this misclassified spectra are composed by the thinnest clouds only (ODs close to 0.03). On the other hand 13% of thin clouds are misclassified as liquid/mixed phase cloud and 10.6% of liquid/mixed phase clouds are misclassified as thin clouds. Even this fact is not surprinsing, since everywhere except in polar regions thin clouds contain liquid water whose information is brought by the measured radiances.
In fact, the separation between thin clouds and mixed-phase clouds gets better considering polar regions only:

Polar Bands results

Class	No. of spectra	Hit Rate	Misclassification
Clear sky	110735	93.7%	1.0% mixed-phase cloud 5.0% thin cloud 0.3% unclassified
Mixed-phase cloud	125731	86.5%	5.1% clear sky 0.8% ice cloud 7.0% thin cloud 0.6% unclassified
Ice cloud	3038	92.7%	6.3% mixed-phase cloud 0.2 % thin cloud 0.8% unclassified
Thin cloud	13609	85.6%	9.0% clear sky 4.4% mixed-phase cloud 0.2% ice cloud 0.8% unclassified

Although there are cases where the two classes are confused by the algorithm, mixed-phased clouds and thin clouds are better separated in polar regions. It is to be noted, in any case, that polar thin clouds are exclusively icy and mixed-phase clouds contain much more ice than thick clouds of other latitudes. A very notable result of the polar regions results is the high absolute thin cloud score, especially in its separation from clear sky scenes. In fact, 85% of thin ice clouds are correctly classified as thin ice clouds and only 9% of them is classified as a clear sky scene. This result is important since thin ice cloud detection from satellites is particularly challenging in polar regions due to the similar radiative properties of surface and cloud.

Chapter 5

Summary and Conclusions

Clouds have a significant impact on the planet's energy balance and the efficient detection and characterization of their presence and properties is necessary for understanding and modeling the climate system. The diffusion of passive instruments for radiance measurements led to the application and advancement of diverse machine learning and statistical techniques. Among the most commonly utilized methods are distance-based methods, Linear Discriminant Analysis algorithms and Neural Networks. CIC (Cloud Identification and Classification) is a recently proposed, innovative machine learning code adopted as the official classification algorithm in the ESA Far-infrared Outgoing Radiation Understanding and Monitoring ([15]) End2End simulator (FE2ES). CIC performs a classification by defining an eigenvectors-based Similarity Index that measures the information content brought into a Training Set when it is concatenated with a new observation.

In the present thesis work, a new metric for quantifying the information content change (eigenvalues-based Similarity Index) is proposed and studied. In addition, a new methodology is developed within the CIC algorithm framework, allowing for the simultaneous utilization of multiple Similarity Indices. Specifically, the Training Spectra are mapped into the 2-dimensional space generated by two different Similarity Indices; subsequently, an optimal separation line is defined and the new observations are then classified based on which half-plane they belong to.

Using a set of downwelling radiance spectra collected on the Antarctic Plateau (Dome-C region) in 2013, three versions of the CIC algorithms (eigenvectors-CIC, eigenvalues-CIC and double-CIC) are tested and compared with respect to their cloud detection power, cloud classification power and dependence on Training Set sizes. This investigation was carried out by collecting a number of spectra and assigning them a true class by visually inspecting the LIDAR images of the same scenario. This was done by exploiting the information on the backscattering signal and depolarization ratio provided by the LIDAR. A cross-validation procedure has been conducted by randomly creating Training Sets of different sizes, training the algorithms on them and, finally, by testing the algorithms on the remaining classified spectra. The procedure has been repeated ten times in order to obtain solid average values. For what concerns the cloud detection power, which is the most important score, all three methods showed very good scores (ranging between 94.0% and 97.5%) even for a small number of Training Sets spectra (10). The eigenvalues-CIC and the double-CIC showed better scores than the eigenvectors-CIC (+2% higher on average). On the other hand, for what concerns the cloud classification power (of secondary importance with respect to cloud detection), using 20 or more spectra for the Training Set, the classical eigenvectors-CIC showed slightly better scores (+1%) than the double-CIC. However, in the case with 10 training spectra, the classical CIC showed a 75% score while the double-CIC showed a 85%. The last result suggests that the double-CIC is even more suitable for situations where a small number of spectra for the Training Sets are available.

Given the overall better results of the double-Similarity-Index CIC, this version has been chosen for the classification of the entire dataset of spectra collected by the instrument REFIR-PAD on the Antarctic Plateau from 2013 to 2020 ([6]); the analysis resulted in statistics compatible with those of previous studies ([13]).

The double-Similarity-Index algorithm has been used also for the challenging analysis of the synthetic PREFIRE observations. A large collection of upwelling radiance (covering the entire Earth's thermal infrared), simulated using the PCRTM radiative transfer model, is used as the main dataset for this application. The very good cloud detection performances (more than 93% of clear and cloudy sky spectra correctly classified) obtained in this study constitutes an important result that witnesses the classification power and versatility of CIC algorithm. Notably, the classification results involving only the polar regions show that the 85% of thin ice clouds were correctly classified and that only the 9% of them were misclassified as clear sky spectra. This result is very encouraging since the detection of thin clouds in polar regions is very difficult due to the similar radiative properties of surface and cloud. Two are the overall important features of these results. The first one is the fact that only 60 training spectra were necessary; the second one is that such high scores were obtained when only 50 wavenumber channels were available.

These facts suggest that CIC algorithm is particularly suited to become an operative identification algorithm for satellite missions, where only a few initial measures can be labeled with the help of ground-based validation sites and then used as training spectra for subsequent measurements. Further investigations will be conducted in the near future to explore the performance of the CIC algorithm in classifying measured satellite radiances. Of particular interest is the application of this classification method to the data collected by the Infrared Atmospheric Sounding Interferometer (IASI) instrument. Additionally, feasibility studies will be undertaken to assess the availability of an adequate number of ground-based validation sites for the ongoing construction of Training Sets, with the ultimate goal of operationalizing the CIC algorithm. These studies will be conducted in the context of the FORUM (Far-infrared Outgoing Radiation Understanding and Monitoring) mission.

Bibliography

[1] U Amato, L Lavanant, G Liuzzi, G Masiello, C Serio, R Stuhlmann, and SA Tjemkes. Cloud mask via cumulative discriminant analysis applied to satellite infrared observations: scientific basis and initial evaluation. *Atmospheric Measurement Techniques*, 7(10):3355–3372, 2014.

[2] Cameron Bertossa, Tristan L'Ecuyer, Aronne Merrelli, Xianglei Huang, and Xiuhong Chen. A neural network-based cloud mask for prefire and evaluation with simulated observations. *Journal of Atmospheric and Oceanic Technology*, 2023.

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[4] Lieven Clarisse, P-F Coheur, Fred Prata, Juliette Hadji-Lazaro, Daniel Hurtmans, and Cathy Clerbaux. A unified approach to infrared aerosol remote sensing and type specification. *Atmospheric Chemistry and Physics*, 13(4):2195–2221, 2013.

[5] William Cossich, Tiziano Maestri, Davide Magurno, Michele Martinazzo, Gianluca Di Natale, Luca Palchetti, Giovanni Bianchini, and Massimo Del Guasta. Ice and mixed-phase cloud statistics on the antarctic plateau. *Atmospheric Chemistry and Physics*, 21(18):13811–13833, 2021.

[6] Gianluca Di Natale, Giovanni Bianchini, Massimo Del Guasta, Marco Ridolfi, Tiziano Maestri, William Cossich, Davide Magurno, and Luca Palchetti. Characterization of the far infrared properties and radiative

forcing of antarctic ice and water clouds exploiting the spectrometer-lidar synergy. *Remote Sensing*, 12(21):3574, 2020.

[7] Uğur Erkan. A precise and stable machine learning algorithm: eigenvalue classification (eigenclass). *Neural Computing and Applications*, 33(10):5381–5392, 2021.

[8] Xianglei Huang, Xiuhong Chen, Daniel K Zhou, and Xu Liu. An observationally based global band-by-band surface emissivity dataset for climate and weather simulations. *Journal of the Atmospheric Sciences*, 73(9):3541–3555, 2016.

[9] Xu Liu, William L Smith, Daniel K Zhou, and Allen Larar. Principal component-based radiative transfer model for hyperspectral sensors: Theoretical concept. *Applied Optics*, 45(1):201–209, 2006.

[10] Tristan S L'Ecuyer, Brian J Drouin, James Anheuser, Meredith Grames, David S Henderson, Xianglei Huang, Brian H Kahn, Jennifer E Kay, Boon H Lim, Marian Mateling, et al. The polar radiant energy in the far infrared experiment: A new perspective on polar longwave energy exchanges. *Bulletin of the American meteorological society*, 102(7):E1431–E1449, 2021.

[11] Tiziano Maestri, William Cossich, and Iacopo Sbrolli. Cloud identification and classification from high spectral resolution data in the far and mid infrared.

[12] Edmund R Malinowski and Darryl G Howery. *Factor analysis in chemistry*, volume 3. Wiley New York, 1980.

[13] Michele Martinazzo, Viviana Volonnino, Tiziano Maestri, Fabrizio Masin, Gianluca Di Natale, Giovanni Bianchini, Massimo Del Guasta, and Luca Palchetti. Cloud identification and classification from ground based and satellite sensors on the antarctic plateau. Technical report, Copernicus Meetings, 2023.

[14] Pietro Mastro, Pamela Pasquariello, Guido Masiello, and Carmine Serio. Cloud detection from iasi hyperspectral data: a statistical approach based on neural networks. In *Remote Sensing of Clouds and the Atmosphere XXV*, volume 11531, pages 40–49. SPIE, 2020.

[15] L Palchetti, H Brindley, R Bantges, SA Buehler, C Camy-Peyret, B Carli, U Cortesi, S Del Bianco, G Di Natale, BM Dinelli, et al. Unique far-infrared satellite observations to better understand how earth radiates energy to space. *Bulletin of the American meteorological society*, 101(12):E2030–E2046, 2020.

[16] Bjorn Stevens, Masaki Satoh, Ludovic Auger, Joachim Biercamp, Christopher S Bretherton, Xi Chen, Peter Düben, Falko Judt, Marat Khairoutdinov, Daniel Klocke, et al. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1):1–17, 2019.

[17] MODIS Cloud Mask Team, Steve Ackerman, Kathleen Strabala, Paul Menzel, Richard Frey, Chris Moeller, Liam Gumley, Bryan Baum, Crystal Schaaf, and George Riggs. Discriminating clear-sky from cloud with modis algorithm theoretical basis document (mod35), 2010.

[18] DD Turner, RO Knuteson, HE Revercomb, C Lo, and RG Dedecker. Noise reduction of atmospheric emitted radiance interferometer (aeri) observations using principal component analysis. *Journal of Atmospheric and Oceanic Technology*, 23(9):1223–1238, 2006.