Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica del Sistema Terra

Cloud Identification and Classification from Ground-Based and Satellite Sensors on the Antarctic Plateau

Relatore:
Prof. Tiziano Maestri

Presentata da:
Viviana Volonnino

Anno Accademico 2021/2022

Abstract

Cloud identification from satellites is considerably challenging in polar environments due to the similar radiative properties of surface and ice clouds.
CIC (Cloud Identification and Classification) is a machine learning algorithm adopted as the official software in the ESA Far-infrared Outgoing Radiation Understanding and Monitoring (FORUM) (*Palchetti et al. (2020)*) End2End simulator (FE2ES). CIC is based on Principal Component Analysis and performs cloud detection and multi-scene classification. It is adaptable to every type of sensor and is particularly suitable when a small number of elements are available for the Training Set. Assessment studies have already been conducted to evaluate the performances of the algorithm in multiple conditions. In *Maestri et al. (2019b)*, CIC was applied to simulated radiance all over the globe, while *Magurno et al. (2020)* used measured airborne interferometric spectra and in *Cossich et al. (2021)* the algorithm was tested on downwelling radiance collected at Dome-C in Antarctica.
CIC is applied to high spectrally resolved data taken from the ground and, for the first time, from satellites. Ground-based data are collected by the REFIR-PAD sensor (*Di Natale et al. (2020)*), covering the far and mid-infrared part of the spectrum. Collocated satellite data are measured by IASI (Infrared Atmospheric Sounding Interferometer) which collects upwelling radiance between 3.4 and 15.5 μm. The period under study spans from 2012 to 2020. CIC results applied to ground-measured spectra are compared to IASI and MODIS L2 cloud products.
Large discrepancies between the classifications are observed, indicating an overestimation of the cloud occurrence in the case of IASI and an opposite result in MODIS. A verification is obtained using collocated ground-based LIDAR measurements, which are available for subsets of the REFIR-PAD radiances. Finally, the CIC algorithm is trained with a subset of IASI data collocated with REFIR-PAD measurements. The training sets are defined also with the help of the Advanced Very High-Resolution Radiometer (AVHRR) on board of MetOp satellites. The AVHRR has 1 km resolution (at nadir) and its collocated measurements are used to evaluate the scene homogeneity in the satellite field of view. Statistical analyses are then performed on IASI spectra using the CIC classification. Results indicate a much better agreement with ground-based data, improving the cloud occurrence provided in IASI L2 products.

Sommario

Identificare le nubi da satellite è estremamente complicato in ambienti polari a causa delle simili proprietà radiative della superficie e delle nubi di ghiaccio. CIC (Cloud Identification and Classification) è un algoritmo machine learning adottato come software ufficiale del Simulatore End2End (FE2ES) della missione ESA FORUM (Far-infrared Outgoing Radiation Understanding and Monitoring). CIC si basa sull'Analisi delle Componenti Principali per rilevare le nubi e classificare le scene osservate. L'algoritmo è adattabile ad ogni tipo di sensore ed è particolarmente adeguato in situazioni in cui sono disponibili pochi elementi per costruire i Training Set. Diversi studi sono stati svolti per valutare le performance dell'algoritmo in diverse condizioni. In *Maestri et al. (2019a)*, il CIC è stato applicato a radianze simulate in tutto il globo, mentre in *Magurno et al. (2020)* sono stati utilizzati spettri raccolti da aerei e in *Cossich et al. (2021)* l'algoritmo è stato testato su radianze misurate dal basso a Dome-C in Antartide.
In questo lavoro, il CIC viene applicato a dati ad alta risoluzione spettrale misurati da terra e, per la prima volta, da satellite. I primi sono raccolti dal sensore REFIR-PAD (*Di Natale et al. (2020)*) e coprono il lontano e vicino infrarosso. I dati satellitari collocati sono invece misurati da IASI (Infrared Atmospheric Sounding Interferometer) e contengono la radianza in uscita tra 3.4 e 15.5 μm. Il periodo in esame va dal 2012 al 2020. I risultati del CIC sugli spettri misurati da terra sono comparati con i prodotti di nube L2 di IASI e MODIS.
Si osservano grandi discrepanze tra le classificazioni, che indicano una sovrastima di eventi nuvolosi nel caso di IASI e risultati opposti per MODIS. Una prima verifica è ottenuta utilizzando misure LIDAR collocate, disponibili per un sottoinsieme di radianze del REFIR-PAD. Infine, l'algoritmo CIC viene allenato con una parte di dati IASI collocati con misure REFIR-PAD. I Training Set sono definiti anche con l'aiuto del AVHRR (Advanced Very High-Resolution Radiometer) a bordo dei satelliti MetOp. L'AVHRR ha una risoluzione spaziale di 1 km (a nadir) e le misure collocate sono utilizzate per valutare l'omogeneità della scena osservata nel campo di vista del satellite. Varie analisi statistiche sono poi eseguite sugli spettri IASI, usando la classificazione del CIC. I risultati sono in accordo con i dati da terra e migliorano le percentuali di eventi nuvolosi indicate nei prodotti L2 di IASI.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clouds are a key component of the Earth System and are the most important regulator of the Earth's climate. It is estimated that on average clouds cover between 70% and 80% of the Earth's surface (*Whitburn et al. (2022)*). Their presence controls the weather, the water cycle and the Earth's radiation budget (ERB). In fact, clouds can cool the system, reflecting part of the solar radiation back to space and reducing the amount of energy available to the earth-atmosphere system. This mechanism is also called "cloud-albedo feedback". Clouds also absorb the infrared (IR) radiation emitted from the surface and the atmosphere below. The cloud top temperature is generally colder than the surface, thus the radiation emitted by clouds is lower that the amount absorbed. This "IR greenhouse effect" warms the atmosphere. The net result is extremely variable and depends on cloud micro-physical properties, the cloud top height and the cloud albedo. For instance, low clouds have temperatures very close to the surface so their contribution to the IR budget is negligible, while their high albedo results in a cooling effect. High clouds, such as thin cirrus clouds, have a larger impact on the IR radiation, causing a consequent warming of the surface and the underlying atmosphere. The great spatial and temporal variability makes cloud studies very challenging and, despite their high importance in the Earth system, they remain the biggest uncertainty in climate models nowadays.

As for polar regions, clouds influence on radiative balance becomes even more crucial since the energy balance changes in the region resonate all over the globe and influence global climate (*Adhikari et al. (2012)*). The two existing polar regions, the Arctic and Antarctica exhibit different behaviours due to their latitudes, topographic features and thus their role in the atmospheric and oceanic circulation (*Fyke et al. (2018)*). Antarctica is the coolest region on the planet and acts as an energy sink in the global balance. Modelling studies have proven how changes in cloud properties over Antarctica

impact regions all over the globe. However, the role of polar clouds in global atmospheric circulation is still not fully understood and well-modelled. For instance, clouds contribution to feedback mechanisms such as surface melting is unknown. During the summer the net positive surface energy balance causes surface melting which in turn changes the albedo properties of the surface, leading to a further increase in temperature and melting. Clouds can enhance this effect or reduce its consequences. Clouds affect also other components of the Earth system, such as the cryosphere. In fact, the two major regulators of cryosphere dynamics, atmospheric moisture content and temperature, respond to cloud changes. Moreover, clouds provide precipitation which is the major source of mass gain, together with blowing snow, in Antarctica (*Adhikari et al. (2012)*). Other studies have been conducted over the years, *Bromwich et al. (2012)* identified clouds' influence on the Southern Ocean's heat and freshwater budget and their impact on global ocean circulation and the global carbon cycle.
Cloud properties retrievals for climate studies are mainly constrained by the lack of observations in polar regions. This issue might be addressed with an increase in satellite observations, although frequent temperature inversions at the surface complicate remote sensing algorithms performance.

1.1 Antarctic Clouds

Studies in polar regions are very challenging due to extreme environmental conditions, in particular in Antarctica where temperatures reach also -60°C in the austral winter. As said before, any changes in the region's climate cause variations in the atmospheric and oceanic circulations, which act as compensating mechanisms to maintain the global energy balance. In turn, surface heat and radiative budget in Antarctica are highly modulated by the cloud cover. Another factor that influences the Antarctic weather, and thus cloud occurrence, is the topography and sea ice extent. Ice sheets elevation determines the surface temperature and thus the snow cover (*Fyke et al. (2018)*). A topographic map of the region is presented in *Fig.1.1.* The majority of land is covered by snow and ice. The eastern and western parts have different characteristics and the bulk of the Antarctic continent is located in the eastern part, which has also a higher surface elevation. The climate in East Antarctica is characterised by a perennial anticyclone, surrounded by a circumpolar trough of low pressure in the Southern Ocean. The sea-ice transition zone provides a suitable condition for cyclogenesis in the circumpolar trough, due to the baroclinic instability caused by the confluence of cold air from the

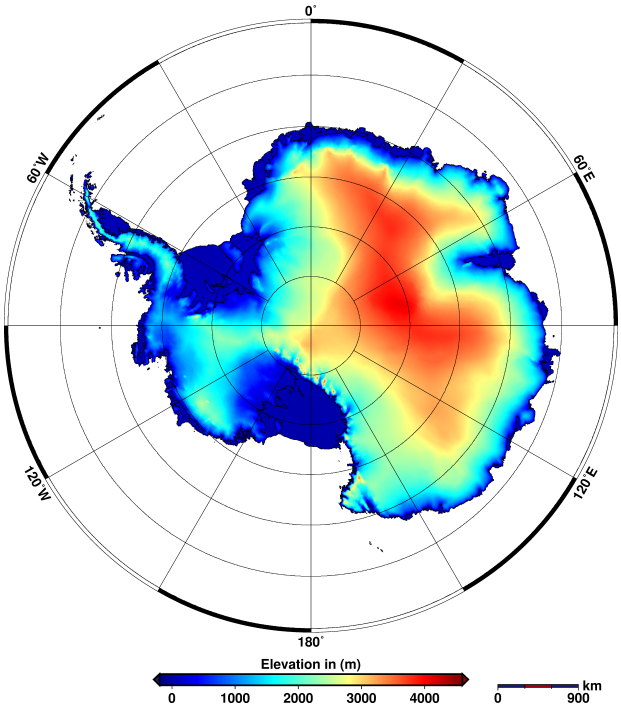


Figure 1.1: Antarctica's elevation map derived from CryoSat-2 measurements collected between 2011 and 2014 (*Helm et al. (2014)*).

continent and warmer maritime air. Also, cyclone activity is enhanced by midlatitude disturbances responsible for the latitudinal transport of heat and moisture toward the poles. This results in higher cyclonic activity in winter than in summer, because of the strongest temperature gradient between the midlatitudes and the poles. Numerous studies on Antarctic clouds have been performed over the years. *Adhikari et al. (2012)* analysed satellite data from 2006 to 2010 for the entire region, using CloudSat and CALIPSO measurements. They found a contrast between the cloud distribution in the Southern Ocean and the Antarctic continent. The first was characterized by persistent cloud cover, while the continent had a more moderate cloud occurrence. In particular, the eastern part of the continent had the lowest occurrence, between 20-30%. The vast Antarctic Plateau in East Antarctica experienced values of 25-30% between 1-3 km, and less than 10% above 5 km. Its high elevation and the low amount of water vapour are all factors contributing to the minor cloud occurrence. While the higher cloud presence in West Antarctica is mainly due to cyclone activity. Among the cloud types observed, low-level clouds ($H_{base} \leq 2km$) were the most common in all of Antarctica (especially around the circumpolar trough). Instead, deep and high-level clouds ($H_{base} \geq 6km$) were found to be associated with synoptical systems and have a higher frequency in the austral winter months in west Antarctica. In East Antarctica, midlevel clouds ($6 \leq H_{base} \geq 2km$) prevailed, especially in winter. Clouds observed in the East part of the continent were generally at lower heights with respect to those in the West. Moreover, thin clouds were the dominant type in the region, with a vertical extent of less than 1 km in 60% of the cases in West Antarctica and 45% in the Antarctic Plateau.
Another study was performed by *Cossich et al. (2021)*, using ground-based

interferometric spectra collected at Dome-C, on the Antarctic Plateau, during the years 2012-2015. The spectra were used to identify cloud signatures and classify the cloud thermodynamic phase. The authors found a cloud occurrence spanning from 23 and 31% over the four years, in accordance with results obtained in the study previously cited. A positive correlation was also found between the mean atmospheric temperature and the cloud presence, revealing a positive cloud forcing.

1.2 Cloud Observation and Detection Methodologies

The importance of clouds in weather and climate application implies the need for accurate and coherent time series of cloud properties (e.g. cloud type, cloud phase, cloud amount, cloud top height, optical thickness). Ground-based remote sensing measurements provide the most reliable way to observe clouds. Both passive and active sensors could be used to recognise the cloud signal. Lidar or radar systems are examples of active instruments capable of identifying clouds. They have different sensitivities to cloud droplet size and concentration, lidar signal is attenuated by thick clouds, while radars have less attenuation. However, ground measurements are difficult in remote places such as the poles and sometimes are not available throughout the entire year. Satellite measurements constitute a fundamental tool since they provide global coverage, daily cloud monitoring and rather high spatial resolution. Lidar and radar technology are currently carried on board of CALIPSO and CloudSat respectively. Although, the majority of weather satellite sensors are based on passive technology and measure the radiation emitted or reflected by the Earth. Clouds are often characterised as layers of higher reflectance and lower temperature with respect to the ground (*Ahmad, Quegan (2012)*). Different clouds imply different spectral signatures, i.e. thick clouds absorb almost all the radiation coming from the surface, while thin clouds might have some features similar to other atmospheric constituents. Remote sensing algorithms often use the visible part of the spectrum to exploit cloud reflectance properties. However, at the poles those techniques are unusable for half a year due to the absence of solar radiation. To overcome this problem, *Amato et al. (2014)* explains that meteorological satellites nowadays carry onboard infrared sensors that allow monitoring the Earth's emission at very high spectral resolution. Avoiding the misinterpretation of clouds signal is still a challenge and regional characteristics have to be taken into account for accurate cloud detection.

Cloud detection is also a preliminary step in many others remote sensing applications. Cloud-free scenes are often used as input for the retrieval of trace gas and surface properties. Even a small cloud amount in the satellite field of view alters the radiance at the top of the atmosphere measured by the sensor. Different techniques have been developed over the years to detect and isolate cloud spectral features. Classical algorithms are based on a series of tests involving few spectral channels and exploit reflectance or brightness temperature spectral variations (*Mahajan, Fataniya (2019)*). Simple comparisons in the infrared classify a scene as cloudy if the measured radiance differs from the clear-sky radiance by a certain value. For instance, the brightness temperature difference between bands at 11 and 12 μm can be used as an indicator for cirrus clouds (*Clarisse et al. (2013)*). These algorithms usually rely on predefined thresholds, which are dependent on the specific wavebands and the type of sensor used, the surface type, seasonal condition, latitude, sun elevation and atmospheric parameters (temperature, humidity, viewing angle). In short, they are time and space-dependent and it is impossible to find global thresholds.

Most visible and infrared sensors on board satellites have their own algorithm for detecting the presence of clouds in the field of view. The MODerate resolution Imaging Spectroradiometer (MODIS) is provided with the cloud product MOD35 (*Ackerman et al. (1998)*) and its algorithm relies on various statistics and tests based on different wavelengths. MODIS has 36 bands from 0.4 to 14.5 μm, covering the visible and middle infrared part of the spectrum. It uses 14 bands and follows several steps to consider all the possible variability for a globally efficient cloud mask. Thresholds applied depend on the surface type and the solar illumination and they are never global. In particular, five cloud-type groups exist: thick high clouds, thin clouds, low clouds, upper tropospheric thin clouds and cirrus. Since tests within a group may be used to detect also other cloud types, *Ackerman et al. (2010)* describes four groups of tests: IR threshold test, brightness temperature difference, solar reflectance threshold, NIR thin cirrus, IR thin cirrus. The final cloud mask is then determined from the results of each group. The product is also associated with a confidence value which falls in between one of these categories: confident clear, probably clear, probably cloud or confident cloud (*Ackerman et al. (1998)*). Among the other imagers which use threshold tests, SEVIRI (Spinning Enhanced Visible and Infrared Imager) cloud detection utilises ten tests based on the surface type.

The increasing number of channels available from the current instrumentation has led research towards new methodologies to face the massive amount of data and extract as much information hidden inside as possible. Among these there are innovative cloud detection methods based on statistical or

pattern recognition approaches. They learn the features of the cloudy and clear-sky conditions from elements whose sky conditions are known, and then sky conditions of other new data are inferred from these by relying on some of the learned properties. Statistical or pattern recognition approaches are suitable for high-dimensional data and naturally handle multispectral measurements (*Murino et al. (2014)*). Statistics-based algorithms can also answer the problem of the large variability of clouds and the underlying surface. *Amato et al. (2014)* proposed a classification method based on cumulative discriminant analysis. They retrieved the proper threshold based on training data. In this way, statistical conditions for clear and cloudy sky are estimated based on elements features.

These techniques played a key role in the development of machine learning algorithms. Machine learning is an area of artificial intelligence in which the system learns automatically based on given existing data. AI algorithms may be supervised or unsupervised, the firsts need labelled training sets as input to train the algorithm, while the seconds are able to find relationships between the elements by themselves. A list of the machine learning approaches for cloud detection is given in *Mahajan, Fataniya (2019)*. The most common AI algorithms in literature used for cloud detection are Neural Networks, Support Vector Machine, Deep Learning, decision tree, logistic regression (*Zhang et al. (2019)*) and more complex ones (*Kurihana et al. (2022)*). In *Whitburn et al. (2022)*, authors developed a cloud detection algorithm based on a supervised neural network (NN) which takes as input parameter IASI spectra. To build the training set they used the Level 2 IASI product as a reference and removed different channel corresponding to gas absorption from the data. Machine learning algorithms are more flexible than classical ones and easier to implement. However, their result depends on input data and it may be not consistent.

These techniques have been employed on satellite sensors, such as the IASI AVHRR (Advanced Very High-Resolution Radiometer), a built-in imager provided for IASI satellite to help the classification of clouds in the sounder's field of view. As explained in *EUMETSAT (2011)*, a clustering procedure divides each pixel observation into six classes (clusters). For each class, it is given the fraction of the IASI FOV covered. Together with this analysis, it is also provided the mean and standard deviation for each AVHRR channel. These quantities are intended to help in the assessment of the homogeneity in the IASI FOV. *Lavanant, Lee (2005)* used those data to retrieve a cloud mask, in particular, they applied different threshold values for clusters of channels 1 and 4.

IASI Level 2 product is provided with a cloud mask which is based on three different methods (*EUMETSAT (2017b)*). The first uses the AVHRR

collocated cloud mask and the cloud fraction embedded in IASI L1C data. When this fraction exceeds a certain threshold the scene is flagged cloudy. The percentage of AVHRR cloudy pixels in the IASI field of view is well reproduced in the tropical and mid-latitude regions. While its sensitivity decreases at high latitudes, especially during the winter period, likely due to the absence of visible light and the very cold surface temperatures, as well as in conditions of high albedo (e.g. snow and ice). The second test is based on a windows channels test, radiance simulated with NWP forecasts and RTTOV radiative transfer code is compared to IASI observations and differences are interpreted as cloud signals. The last one applies a neural network algorithm on IASI data and AVHRR clusters. IASI pixel is flagged as cloudy if at least one of the three tests detects a cloud. While a scene is clear if all the tests concluded the absence of clouds.

Other algorithms reduce the elevated amount of data by means of data mining techniques. Principal component analysis (PCA) is the one mostly used dimension reduction method. PCA transforms the original set of correlated variables into a set of uncorrelated variables called principal components (PCs). It reduces the dimension of the initial data, retaining only the components with higher variance and disregarding those that carry no information. In fact, PCs are ordered by the amount of variance explained by each associated eigenvalue, thus the first PC explains the highest variation in the data. However, it is difficult to relate PCs to physical features because each PC is a linear combination of the original variables. *Ahmad, Quegan (2012)* compared the MODIS cloud mask with a spectral analysis and a PC analysis. They found that PCA underestimates the number of cloudy pixels with respect to the spectral analysis, although it was in good agreement with the MODIS cloud mask and concluded that cloud and clear-sky pixels have different spectral signatures that could emerge through the statistical approach of PCA. Different authors applied PCA also to sounding measurements and further investigate the meaning of the first set of PCs. For instance, *Huang, Antonelli (2001)* found that the first eigenvector, associated with the largest variance, was correlated with the window spectral channels. In general, instruments measuring infrared spectra with a high-spectral resolution are great candidates for PCA, as IASI (*Clarisse et al. (2013)*), because many of the spectral channels are highly correlated and PCA can compress the data and remove the PCs associated with the uncorrelated random error.

1.3 FIR contribution

Satellites currently flying measure up to the middle infrared part of the spectrum. Although, the far-infrared (FIR) has a critical role in the Earth Radiation Budget. Between 40 and 65% of the total outgoing longwave radiation (OLR) is dominated by the FIR contribution (*ESA (2019)*). FIR is extremely useful in the detection of clouds, especially ice clouds. Scattering processes become the most important in this part of the spectrum and their relation with the crystal shape can be exploited to improve the detection of ice crystals. Also, at FIR wavenumbers, there is a higher BT sensitivity to the cloud particle phase. The FIR spectral band is particularly useful in those regions where the atmosphere is very dry, such as the poles. In fact, the rotational absorption band of the water vapour is the greatest contributor to the modulation of the outgoing radiation at these wavelengths. Up to now, studies on simulations have proven that cloud classification algorithms performances in polar regions are highly improved with the addition of FIR channels. The European Space Agency (ESA) promoted FORUM (Far-infrared Outgoing Radiation Understanding and Monitoring) as the 9th Earth Explorer mission, which will fill the observational gap in the far infrared and is expected to enhance cloud detection performances.

1.4 Thesis objectives and overview

Satellite measurements constitute a fundamental tool to collect continuous measurements in remote regions, such as the poles. They are capable of covering large areas while producing high spectral resolution observations. Numerous efforts have been made to generate reliable and flexible cloud detection algorithms. Although satellite sensors are not lacking in problems when used for cloud detection in polar regions and the assessment of cloud products can be particularly challenging, due to the scarcity of ground-based stations available for validation campaigns. For instance, active instruments, such as radar, often miss optical thin clouds, which are very common on the Antarctic Plateau (*Maestri et al. (2019a)*). Moreover, the coarse vertical resolution of the CPR onboard of CloudSat (500 m) and its limited sensitivity near the surface does not allow accurate detection of low clouds. MODIS tests and thresholds rely mainly on measurements at shortwaves to perform cloud detection. When solar radiation is not available, the cloud mask switches in the infrared channels, becoming less efficient and increasing the number of missed clouds. The similar radiative properties of the surface and ice clouds and the frequent temperature inversions in Antarctica, make IASI algorithms

based on brightness temperature differences unreliable, and many clear-sky scenes are erroneously flagged as cloudy.
In this thesis, an innovative machine learning algorithm CIC (Cloud Identification and Classification) is tested against downwelling spectra collected from the ground and IASI spectral radiances and results are compared to cloud detection products provided by IASI and MODIS satellites. The area of study is Dome-C, a base situated on the Antarctic Plateau, where is located the REFIR-PAD, a spectroradiometer that measures spectral radiance in the FIR and MIR channels.
The work is structured as follow:

- chapter 2 provides an overview of the Radiative Transfer theory, with a focus on the case of a cloudy atmosphere
- in chapter 3 the CIC algorithm setup and PCA theoretical framework are presented.
- chapter 4 illustrates the classification results obtained on downwelling spectra, collected by the REFIR-PAD instrument, over the years 2012-2020
- chapter 5 describes the collocation procedure of IASI data with ground-based measurements and the application of CIC to IASI L1 products. Statistics provided by MODIS and IASI L2 cloud products are also discussed.

Chapter 2

Radiative Transfer in Cloudy Atmosphere

The transmission of solar and Earth's radiation through the atmosphere follows the laws of radiative transfer, where the atmosphere acts as the transmitting medium. In this chapter, the basis of radiative transfer in a scattering and non-scattering atmosphere will be presented. Furthermore, it will be discussed the case of a cloudy atmosphere and the cloud influence on the IR Earth's radiation.

The radiative transfer is based on a very simple equation describing the interaction of radiation with matter. A beam of radiation I_λ passing a medium of thickness ds, will be reduced by a quantity dI_λ, given by:

$$dI_\lambda = -k_\lambda \rho I_\lambda ds \tag{2.1}$$

The loss of radiation depends on the initial intensity and the properties of the medium, characterised by its density ρ and the mass extinction cross-section k_λ, measured as $[m^2/kg]$. This quantity represents the area taking part in the extinction process at the specific wavelength λ, normalized over the mass or quantity of the absorbing/scattering material. Here extinction refers to the decrease in the incident radiation by both absorption and scattering processes.
On the other hand, the radiation intensity may also be increased by the emission of the medium itself and by scattering phenomena. A source function coefficient j_λ can be defined with the same physical meaning as the extinction cross-section, though in this case the initial radiation variation is:

$$dI_\lambda = j_\lambda \rho ds \tag{2.2}$$

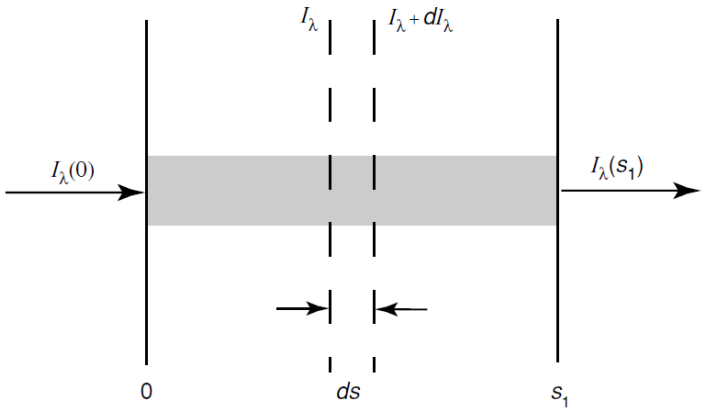


Figure 2.1: Scheme of radiation passing through a medium of thickness *ds*.

The source function can be written as the source function coefficient divided by the extinction cross-section

$$J_\lambda = j_\lambda / k_\lambda \tag{2.3}$$

and thus the full radiative transfer equation (RTE) obtained is given by:

$$\frac{dI_\lambda}{k_\lambda \rho ds} = -I_\lambda + J_\lambda \tag{2.4}$$

2.1 Schwarchild Equation

The solution of the RTE for a non-scattering medium, in local thermodynamic equilibrium, is described by Schwarzschild's equation. These conditions occur when only the infrared radiation emitted by the surface and the atmosphere is considered. The source function can be treated as a blackbody emission and thus be quantified by the Planck function, which depends only on temperature:

$$J_\lambda = B_\lambda(T) \tag{2.5}$$

The radiative transfer equation becomes:

$$\frac{dI_\lambda}{k_\lambda \rho ds} = -I_\lambda + B_\lambda(T) \tag{2.6}$$

The first and second terms describe respectively the absorption and emission process. Following that, k_λ accounts only for absorption processes and is called absorption cross-section, while $k_\lambda \rho$ is the absorption coefficient. To

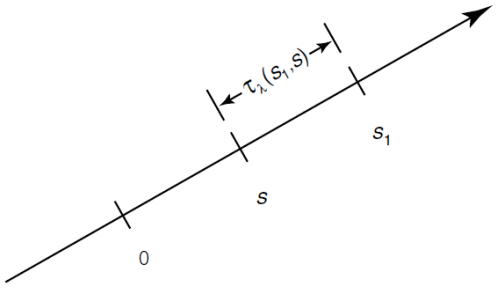


Figure 2.2: Representation of the optical thickness (*Liou (2002)*).

find a solution for this differential equation, it can be defined the monochromatic optical thickness of the medium between two points, namely s and s_1 illustrated in *Fig.2.2*

$$\tau_\lambda(s_1, s) = \int_s^{s_1} k_\lambda \rho ds' \tag{2.7}$$

or in differential form

$$d\tau_\lambda(s_1, s) = -k_\lambda \rho ds \tag{2.8}$$

In this way, the RTE can be rewritten as:

$$-\frac{dI_\lambda(s)}{d\tau_\lambda(s_1, s)} = -I_\lambda(s) + B_\lambda[T(s)] \tag{2.9}$$

Integrating both sides of the equation, the solution obtained is:

$$I_\lambda(s_1) = I_\lambda(0)e^{-\tau_\lambda(s_1,0)} + \int_0^{s_1} B_\lambda[T(s)]e^{-\tau_\lambda(s_1,s)} k_\lambda \rho ds \tag{2.10}$$

where $I_\lambda(s_1)$ represents the radiance reaching the observer at s_1. If the TOA is defined at s_1, the first term represents the emission of a black surface ($\epsilon_\lambda = 1$) at $s = 0$, attenuated by the atmosphere above with optical thickness $\tau_\lambda(s_1, 0)$. The second term is the sum of the grey-body emission from all the layers between 0 and s_1, attenuated by the layers above. This equation is also known as Schwarzschild's equation.

In many situations, the atmosphere can be approximated as plane-parallel in localised portions. This assumption implies that the variations of the atmospheric parameters and radiation intensity are allowed only in the vertical direction. The normal optical thickness, also called optical depth (OD), can be defined as:

$$\tau_\lambda = \int_z^{\infty} k_\lambda \rho dz' \tag{2.11}$$

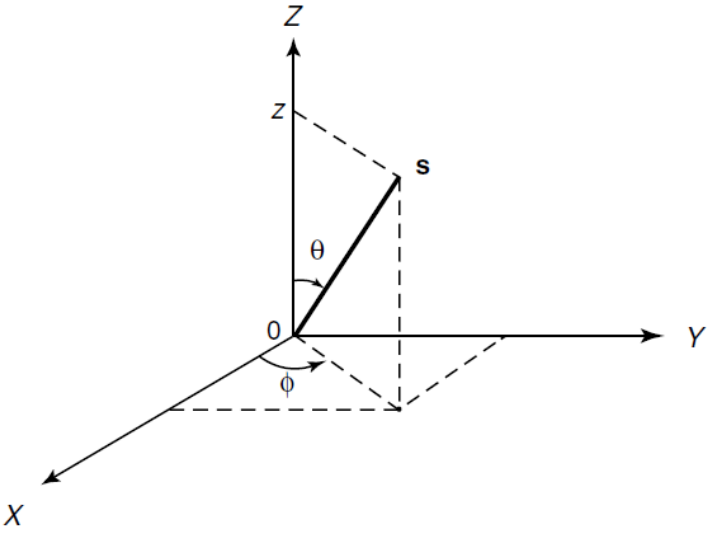


Figure 2.3: Geometry of a plane-parallel atmosphere, with the zenith and azimuthal angles represented by θ and ϕ respectively (*Liou (2002)*).

where $z = \infty$ is the top of the atmosphere and corresponds to $\tau_\lambda = 0$. So the variable z represents the vertical distance and two angular variables, θ and ϕ, are introduced to describe the zenith and azimuth angle respectively. Thus the general differential RTE equation is:

$$\mu \frac{dI(\tau;\mu,\phi)}{d\tau} = I(\tau;\mu,\phi) - J(\tau;\mu,\phi) \tag{2.12}$$

with $\mu = cos\theta$.

2.2 IR Radiative Transfer and Clouds contribution

From now on, the wavenumber domain will be used instead of wavelengths, since this part of the discussion will involve only the infrared part of the spectrum. In a plane-parallel atmosphere, absorption and emission processes in the IR are symmetrical with respect to the azimuthal angle and the radiance is a function of the vertical position and the zenith angle only. Schwarzschild's equation can be interpreted both as the radiance measured by a sensor at the TOA and as the downwelling radiance reaching the ground. In the first case, the zenith angle is $0 \leq \theta \leq \pi/2$ and $0 \leq \mu \leq 1$, while in the

second, $\pi/2 \leq \theta \leq \pi$ and $\mu = -\mu$. The total atmospheric optical depth, from the ground to the TOA, is τ_* and the solution of the RTE for upward and downward intensities are respectively:

$$I_\nu(\tau, \mu) = B_\nu(\tau_*)e^{-(\tau_* - \tau)/\mu} + \int_\tau^{\tau_*} B_\nu(\tau')e^{-(\tau' - \tau)/\mu}\frac{d\tau'}{\mu} \tag{2.13}$$

$$I_\nu(\tau, -\mu) = \int_0^{\tau} B_\nu(\tau')e^{-(\tau - \tau')/\mu}\frac{d\tau'}{\mu} \tag{2.14}$$

Two boundary conditions are here introduced. First, the surface has been treated as a black body in the infrared, with $\epsilon_\nu = 1$ (see *Eq.2.13*). Second, the downward emission at TOA is considered negligible ($B(\tau = 0) = 0$), so the first term in *Eq.2.10* vanishes and *Eq.2.14* depends only on the atmospheric contribution. Another way to express these solutions is by introducing the monochromatic transmittance or transmission function defined as the attenuation caused by the medium:

$$T_\nu(\tau/\mu) = e^{-\tau/\mu} \tag{2.15}$$

Its derivative acts as a weighting function in the RTE and can be written as:

$$W = \frac{dT_\nu(\tau/\mu)}{d\tau} = -\frac{1}{\mu}e^{-\tau/\mu} \tag{2.16}$$

The final solutions are

$$I_\nu(\tau, \mu) = B_\nu(\tau_*)T_\nu[(\tau_* - \tau)/\mu] + \int_\tau^{\tau_*} B_\nu(\tau')\frac{d}{d\tau'}T_\nu[(\tau' - \tau)/\mu)]d\tau' \tag{2.17}$$

$$I_\nu(\tau, -\mu) = \int_0^{\tau} B_\nu(\tau')\frac{d}{d\tau'}T_\nu[(\tau - \tau')/\mu)]d\tau' \tag{2.18}$$

The atmospheric contribution can be interpreted as the weighted sum of the Planck radiances from each layer. The weighting function peaks at the level from which the spectral signal is stronger, thus it is an indicator of which layer is responsible for the majority of the emission.

Up to now, the Schwarzschild equation has been considered for a clear atmosphere. If a cloud is present the atmospheric contribution should be split up into the radiation emitted by the layers below the cloud, the emission of the cloud itself and the contribution of the layers above the cloud. Considering τ_c the optical depth up to the cloud, the radiance measured at the TOA ($\tau = 0, \mu = 1$) is:

$$I_\nu(0, \mu = 1) = (1 - \epsilon_c)B_\nu(\tau_*)T_\nu(\tau_*) + (1 - \epsilon_c)T_\nu(\tau_c)\int_{\tau_c}^{\tau_*} B_\nu(\tau')dT_\nu + \epsilon_c B_\nu(\tau_c)T_\nu(\tau_c) + \int_0^{\tau_c} B_\nu(\tau')dT_\nu \tag{2.19}$$

where the cloud scattering has been neglected and $(1-\epsilon_c)$ is the cloud transmissivity. If the surface reflection is also accounted for, the first term of the equation should consider a surface emissivity different from 1 and an ulterior term should be accounted for, which is the radiance emitted after being reflected by the surface, which can be written as

$$(1-\epsilon_{\nu,surf})\int_{\tau_c}^{\tau_*} B_\nu(\tau')\left[\frac{T_\nu(\tau_*)}{T_\nu(\tau')}\right]^2 dT_\nu \tag{2.20}$$

On the contrary, the downwelling radiation ($\mu=-1$) in presence of a cloud becomes:

$$I_\nu(\tau_*,\mu=-1) = (1-\epsilon_c)\left[\frac{T_\nu(\tau_c)}{T_\nu(\tau_*)}\right]\int_0^{\tau_c} B_\nu(\tau')dT_\nu(\tau') + \epsilon_c B_\nu(\tau_c)\frac{T_\nu[(\tau_c)]}{T_\nu[(\tau_*)]} + \int_\tau^{\tau_*} B_\nu(\tau')dT_\nu \tag{2.21}$$

where the first and last terms describe the emission of the atmosphere, while the second term accounts for the cloud contribution. Different spectral channels can be used to observe different types of clouds. For instance, low-level clouds are visible only in transparent channels, in the so-called "atmospheric windows", since the weighting function peaks close to the surface and the cloud signature can still be recognised. High-level clouds, instead, can be observed also in absorbing channels because the cloud is generally above the maximum of the weighting function.

2.3 Multiple Scattering

In the general radiative transfer equation obtained in *Eq.2.12*, the change of intensity is the sum of four processes depicted in *Fig.2.4*:

- a reduction of the incident radiation due to extinction phenomena;
- an increase due to the single scattering of the direct solar flux from the direction $(-\mu_0,-\phi_0)$ to (μ,ϕ);
- an increase due to the multiple scattering of the already diffuse radiation, from a general direction $(-\mu',\phi')$ to the final direction (μ,ϕ);
- an increase due to the emission of the layer in the direction (μ,ϕ).

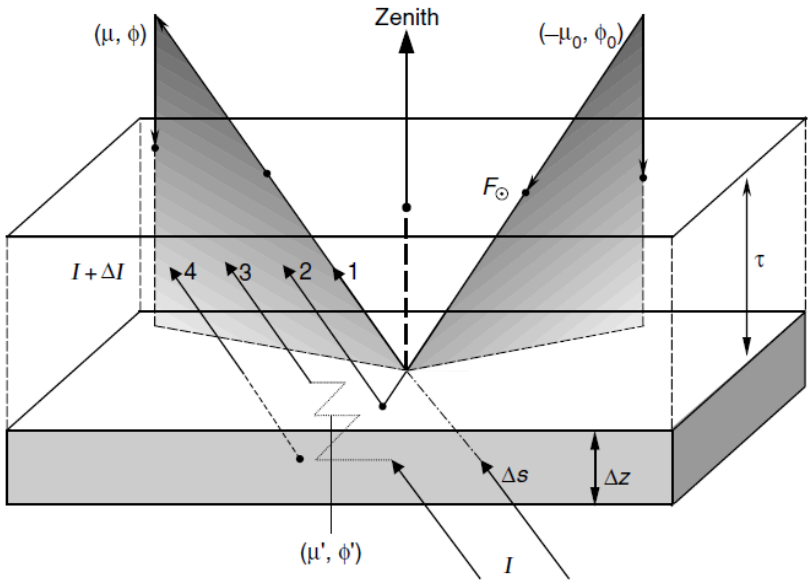


Figure 2.4: Scheme of upward radiation in a plane-parallel atmosphere. 1) attenuation by extinction; 2) single scattering of direct solar radiation; 3) multiple scattering; 4) emission from the layer (*Liou (2002)*).

The scattered intensity in the direction defined by the scattering angle Θ is given by:

$$I(\Theta) = I_0 \Omega_{eff} \frac{P(\Theta)}{4\pi} \tag{2.22}$$

where Ω_{eff} is the effective solid angle, computed as σ_{scatt}/r^2 and represents the fraction taking part in the scattering. While $P(\Theta)$ is the scattering phase function, which describes the probability of scattering in an angle Θ over the solid angle 4π. Now this expression can be included in the RTE to give a source function equal to:

$$J(\tau;\mu,\phi) = \frac{\tilde{\omega}}{4\pi}\int_0^{2\pi}\int_{-1}^{1} I(\tau;\mu',\phi')P(\mu,\phi;\mu',\phi')d\mu' d\phi' + \frac{\tilde{\omega}}{4\pi}F_0 P(\mu,\phi;-\mu_0,-\phi_0)e^{-\tau/\mu_0} + (1-\tilde{\omega})B[T(\tau)] \tag{2.23}$$

where $\tilde{\omega}$ is the single scattering albedo, defined as the ratio of the scattering coefficient to the extinction coefficient and quantifies the amount of energy diffused in all directions. The first term of the equation represents the multiple scattering, where the diffuse intensity is integrated over the 4π solid angle; the second accounts for the scattering of the solar irradiance F_0 coming from the angle $(-\mu_0, -\phi_0)$; while the last term is the emission of the

layer. For radiative transfer at IR wavelengths (100-2500 cm^{-1}), the solar flux F_0 is negligible and the multiple-scattering term becomes significant only in presence of clouds, otherwise the equation is reduced to the original Schwarzschild's equation.

Chapter 3

CIC: Cloud Identification and Classification algorithm

This chapter will provide an overview of the CIC (Cloud Identification and Classification) machine learning algorithm. Algorithms for cloud identification from hyperspectral infrared sounders measurements usually consist of brightness temperature thresholds or the evaluation of brightness temperature differences, as discussed in chapter 1. These techniques require many calibrations and are location and time-dependent. Machine learning algorithms, such as artificial neural networks, support-vector-machine and others, have been developed to overcome these limitations.
Principal component analysis (PCA) is an unsupervised statistical technique primarily used for dimensionality reduction in machine learning. Models become more efficient as the number of features is reduced. In theory, PCA produces the same number of PCs as there are features in the training set. However, not all PCs are equally important. The optimal number of PCs retained is dependent on the tradeoff between dimensionality reduction and information loss. The first principal component expresses the greatest amount of variance. Each additional component accounts for less variance and more noise and it is not correlated with the other ones, since PCs are orthogonal projections of data onto lower-dimensional space. Taking only a subset of PCs preserves the signal and discards the noise, reducing a large number of features to just a couple of principal components. A standard procedure to compute PCs is made of the following steps:

- feature standardization: each feature is set with a mean of 0 and a variance of 1;
- covariance matrix computation;

- eigendecomposition of the covariance matrix: computation of the eigenvectors and correspondent eigenvalues;
- sort the eigenvectors from the highest eigenvalue to the lowest;
- select the number of PCs.

There are also some disadvantages in PCA. The physical meaning of each feature is lost since each PC is a linear combination of the original features. In addition, PCA assumes a correlation between features and results are biased in datasets with strong outliers.

An innovative machine-learning algorithm CIC (Cloud Identification and Classification) was recently developed and described in *Maestri et al. (2019b)*. CIC is based on Principal Component Analysis and performs cloud detection and multi-scene classification. CIC allows the identification of the atmospheric scene observed (clear or cloudy) based only on the input spectra data, without the need for ancillary information or forecast model outputs. The algorithm uses a single threshold applied to a univariate distribution of the newly defined Similarity Index. This parameter defines the level of closeness or similarity of each spectrum analysed and a specific class. CIC is primarily used to distinguish cloudy scenes from clear-sky ones, but it is also able to characterise the cloud phase.
The first step is the definition of the training sets (TRs), consisting of a number T of spectra for each class. Then, the principal components (PCs) are computed for each TR and stored in a matrix. Each spectrum of the test set (TS) is then added to the different TRs, forming extended training sets (ETRs). Finally, the PCs of those ETRs are computed. The variations in the PCs of the extended training sets with respect to the original ones are evaluated by means of the Similarity Index. Small changes are interpreted as the spectrum belonging to that class, while large changes represent a spectrum containing different information and thus belonging to another class. CIC then associates the input TS spectrum with the class providing the smallest change.

3.1 Algorithm Description

CIC is based on the Principal Component Analysis (PCA) and is able to determine the atmospheric scene observed, separating the clear and cloudy sky conditions, and classify the type of cloud according to its thermodynamic phase. Three classes (clear sky, ice cloud and mixed-phase cloud) will be defined for the application of CIC to REFIR-PAD ground measurements. While only two classes (clear and cloudy) will be used for IASI spectra, due to the low number of observations available. In the following mathematical description of the algorithm, only two classes will be considered for clarity purposes. First, a training set is defined for each scene, using reference spectra of known class. A number T_{clear} and T_{cloud} of spectra are stored in columns in the training set matrices TR_{clear} and TR_{cloud}, respectively:

$$TR_i(\nu, t) \tag{3.1}$$

$$i \in [clear, cloud], \nu \in [1, .., \nu_{max}], t \in [1, .., T_i]$$

where each row corresponds to a specific wavenumber ν. After that, the principal components of each training set matrix are computed from the eigenvectors of the covariance matrix:

$$\epsilon TR_i(\nu, p) = eig(cov(TR_i(\nu, t))) \tag{3.2}$$

$$i \in [clear, cloud], \nu \in [1, .., \nu_{max}], t \in [1, .., T_i], p \in [1, .., P]$$

with $P = max(T_i, \nu_{max})$ defined as the total number of principal components, which is equal to the number of spectral channels. Each row of the square matrix in *Eq.3.2* contains a normalised eigenvector:

$$\sum_{\nu=1}^{\nu_{max}} \epsilon TR_i(\nu, p)^2 = 1 \tag{3.3}$$

In the same way, all the spectra of the test set (TS) are written in a matrix:

$$TS(\nu, j) \tag{3.4}$$

$$\nu \in [1, .., \nu_{max}], j \in [1, .., J]$$

with J equal to the number of TS spectra considered. Then, J Extended Training Set matrices are defined, containing the TR spectra concatenated with a single TS spectrum

$$ETR_{i,j}(\nu, t') = [TR_i(\nu, t)||row_j(TS(\nu, j))] \tag{3.5}$$
$$t' \in [1, .., T_i + 1]$$

In this expression, $row_j(TS(\nu, j))$ indicates the jth TS spectrum written as a one-dimensional array, and the matrix concatenation is expressed by the notation $||$. The PCs of the extended training sets are then computed as:

$$\epsilon ETR_i(\nu, p) = eig(cov(ETR_i(\nu, t))) \tag{3.6}$$

The eigenvectors represent the directions in the multidimensional space with a dimension equal to the number of spectra in the TR. When a new spectrum is included in the TR, the new directions will be rotated with respect to the initial ones. The variation depends on the amount of information added by the TS spectrum to the TR. Thus, a small change in the PCs after the addition of the test set spectrum denotes that the spectrum belongs to the class and it has the same features as the TR spectra. While a large difference indicates that the spectrum belongs to another class. The PCs variations are estimated through a newly defined parameter, called the Similarity Index (SI), computed for each class using the two eigenvectors matrices:

$$SI(i, j) = 1 - \frac{1}{2P_0} \sum_{p=1}^{P_0} \sum_{\nu=1}^{\nu_{max}} |\epsilon ETR_{i,j}(\nu, p)^2 - \epsilon TR_{i,j}(\nu, p)^2| \tag{3.7}$$

with P_0 equal to the number of PCs associated with the signal that constitute the information-bearing principal components (IBECs). This number of PCs is computed minimising the indicator function (IND), defined in the work of *Turner et al. (2006)* as:

$$IND(p) = \frac{RE(p)}{(P - p)^2} \tag{3.8}$$

where *RE(p)* is defined as the real error

$$RE(p) = \sqrt{\frac{\sum_{i=p+1}^{P} \lambda_i}{T_i(P - p)}} \tag{3.9}$$

λ_i is the *ith* eigenvalue of the covariance matrix and T_i is the number of spectra in the *ith* TR. In the work of *Turner et al. (2006)*, it is explained that the PCs retained following this procedure are the eigenvectors associated with the physical signal. CIC computes the number P_0 separately for the different TRs matrices, then a unique value P_0 is chosen to compute the SI in *Eq.3.7*.

Selecting the maximum value means taking all the information included in one TR and adding some noise in the other, as the opposite, choosing the minimum value leads to selecting all the information from the corresponding TR but excluding some in the other. The SI estimates of how much the PCs of a TR rotate after a new spectrum is added, thus it does not depend on eigenvalues but on eigenvectors only. All the principal components associated with the physical signal are accounted for with the same weight since all of them might carry crucial features for the classification. The Similarity Index defined in *Eq.3.7* is normalised. Its absolute value is at most equal to 2 and has to be summed over all the P_0 differences, reaching a maximum value of $2P_0$:

$$0 \leq \sum_{p=1}^{P_0} \sum_{\nu=1}^{\nu_{tot}} |\epsilon ETR_{i,j}(\nu,p)^2 - \epsilon TR_{i,j}(\nu,p)^2| \leq 2P_0 \tag{3.10}$$

therefore the SI value is

$$0 \leq SI(i,j) \leq 1 \tag{3.11}$$

where a $SI = 1$ means that the eigenvectors matrices of the TR and ETS are identical and the analysed spectrum adds no further information. While $SI = 0$ represents two very different sets of PCs, denoting that the test set spectrum describes new physical properties. The similarity index is the metric used by CIC to measure how much each spectrum of the test set resembles the characteristics of each TR.

3.2 Classification

3.2.1 Elementary Approach

The Similarity Index quantifies the change in the PCs every time a new spectrum of the TS is added to a TR. The classification result is obtained comparing the value of the SIs obtained for the two classes, associating the spectrum with the most similar class.
Considering the same two classes as before, clear and cloudy sky, if

$$SI(clear,j) > SI(cloud,j) \tag{3.12}$$

the *jth* spectrum is classified as clear, while if

$$SI(cloud,j) > SI(clear,j) \tag{3.13}$$

the spectrum is labelled as cloudy. These two conditions can be simplified introducing a new parameter called Similarity Index Difference (SID) which

acts as a binary classifier:

$$SID(j) = SI(cloud, j) - SI(clear, j) \tag{3.14}$$

where

$$\begin{cases} SID(j) > 0 & j \in \{cloudy\ spectra\} \\ SID(j) < 0 & j \in \{clear\ spectra\} \end{cases} \tag{3.15}$$

A third class can be introduced to decouple ice clouds from mixed-phase ones, because they have often very distinct spectral characteristics. Three TRs have to be defined, each representative of the variability within that class. For each TS spectrum three SIs are obtained (SI_{clear}, $SI_{ice\ cloud}$ and $SI_{mixed-phase\ cloud}$). From these, a vector of similarity index differences (SIDs) is defined from the mutual comparison of the three classes:

$$\begin{cases} SID(1) = SI_{clear} - SI_{ice\ cloud} \\ SID(2) = SI_{clear} - SI_{mixed-phase\ cloud} \\ SID(3) = SI_{ice\ cloud} - SI_{mixed-phase\ cloud} \end{cases} \tag{3.16}$$

The classification output is then obtained following the logical scheme in *Fig.3.1*. The white boxes represent the partial results and the green ones are the final classification outcome. A result is derived only if a class prevails over the other two, otherwise, the spectrum is considered unclassified.

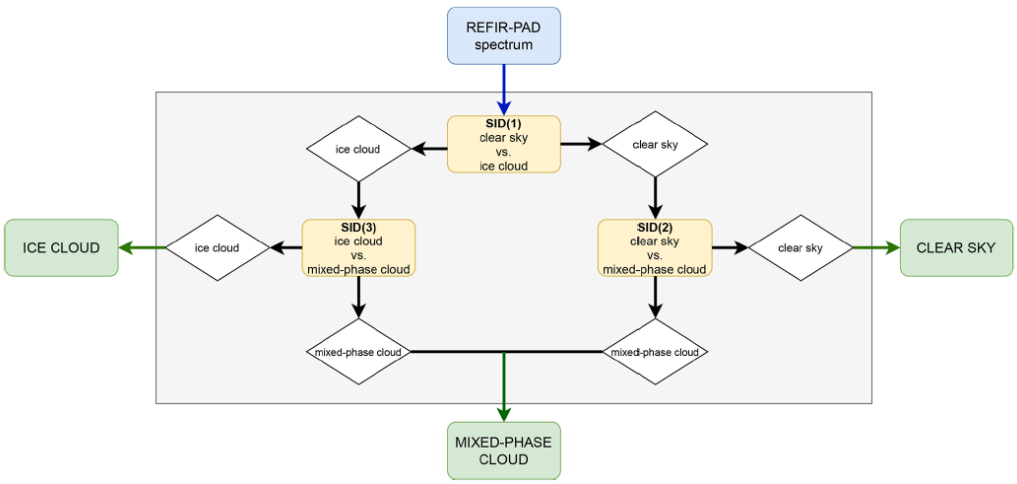


Figure 3.1: Logic diagram of the classification process performed by the CIC algorithm.

This classification approach is called elementary since it is based only on the SID sign. The process is repeated independently for each spectrum of the test set. Results are depended on the training sets composition, especially when the number of spectra composing the TRs is very small. If the TR does not represent the entire class variability, spectra might be misclassified, since not all the spectral features are well reproduced in the PCs. For instance, if the features are not clearly distinct, the SID distributions may overlap, resulting in clear spectra with positive SID or vice versa. Thus, the TR definition is a crucial step to improve the accuracy of the algorithm.

3.2.2 Distributional Approach

A perfect classifier would ideally generate a bimodal distribution of the SID parameter, denoting two homogeneous groups among the classified spectra. However, this class separation is difficult to obtain and the distributions exhibit an amount of overlap which depends also on the spectra that constitute the training sets. A distributional method has been developed to address this issue and optimise the algorithm performance. CIC is initially applied to the training set spectra which have known classes and their SIDs are calculated. The resulting distribution is used to define the most suitable delimiter between the classes, selecting a shift of the original zero that maximises the CIC performance on the training sets. A new parameter called Corrected Similarity Index Difference (CSID) is defined for each spectrum, which is computed applying the shift to the original SIDs:

$$CSID(j) = SID(j) - shift \tag{3.17}$$

The CSID becomes the new binary classifier and, as previously defined for SID parameter

$$\begin{cases} CSID(j) > 0 & j \in \{cloudy\ spectra\} \\ CSID(j) < 0 & j \in \{clear\ spectra\} \end{cases} \tag{3.18}$$

The shift optimal value can be obtained using different functions that potentially forecast the performance of the algorithm. In *Maestri et al. (2019b)* the Consistency Index (CoI) is suggested, which is defined as:

$$CoI(shift) = 1 - max\left(\frac{FP_{clear}}{T_{clear}}, \frac{FP_{cloud}}{T_{cloud}}\right) \tag{3.19}$$

where FP is the number of false positives for each class. FP_{clear} is equal to the number of cloudy spectra misclassified as clear, while FP_{cloud} indicate the

clear spectra misclassified as cloudy. The CoI is an indicator of the training set representativeness and shows how many TR elements would be classified correctly if they were part of the test set. The CoI value is close to 1 only if both clear and cloudy FP are rare and the TRs represent the full variability. Previous studies of *Maestri et al. (2019b)* and *Magurno et al. (2020)* have proven that the distributional approach increases the performance of the classification algorithm.

Chapter 4

REFIR-PAD ground-based measurements

In this chapter, the CIC is applied to high spectral resolution downwelling radiances at far infrared (FIR) and middle infrared (MIR) wavenumbers, collected at Dome-C on the Antarctic Plateau, between 2012 and 2020. 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 Multi-Process Atmospheric Studies), within the Italian National Program for Research in Antarctica (PNRA) (*Palchetti et al. (2020)*).

4.1 Instrumentation and Measurements

The REIFR-PAD instrument is installed in the Physical Shelter and provides spectral measurements in the zenith direction of downwelling radiance in the range 100-1500 cm^{-1}, with a spectral resolution of 0.4 cm^{-1}. It is able to detect atmospheric emission in the FIR and MIR regions of the spectrum. To obtain a complete spectrum, four calibration acquisitions and four sky observations are made. Each of them takes about 80 seconds, thereby the entire sequence lasts 14 minutes (5.5 minutes of calibrations, 5.5 minutes of sky observations and further delays due to the detector settling) (*Palchetti et al. (2020)*). The instrument operates full-time, alternating cycles of 5-6 hours of measurements, with 1-3 hours of analysis. In the same Physical Shelter, there is also a LiDAR (Light Detection and Ranging), an active remote sensing instrument that emits a beam of radiation in the visible band, at 532 nm. The LiDAR measures the backscattering and depolarization vertical profiles up to 7 km above the surface. The interpretation of its signals

provides information on the cloud layers. In clear sky conditions, the LiDAR backscatter signal decreases with altitude. When a cloud is present, the radiation detected by the instrument increases due to the backscatter of the layer. On the other hand, the light polarization indicates the cloud particle phase. Liquid water droplets retain the polarization state of the incident beam, whereas the light backscattered from non-spherical ice particles is partially depolarized as a result of internal reflections. Theoretical studies show that liquid water droplets change the polarization of 2-4%, while non-spherical ice particles can have a strong depolarization, between 30-40%. The threshold to determine the cloud's thermodynamic phase depends on the atmospheric state and cloud microphysical parameters. Moreover, clouds can be composed of different layers, each having a different depolarization value. In particular, mixed-phase clouds are generally observed to be composed of a layer of ice particles at the cloud top and a layer of water particles at the cloud bottom, where the temperature is higher. In this work, a depolarization of 15% is used as a threshold to discriminate ice clouds (signal > 0.15) from mixed-phase clouds (signal < 0.15).

A dataset of 233508 spectra collected by REFIR-PAD at Dome-C, during the years 2012-2020, is analysed using the CIC algorithm. *Tab.4.1* shows the number of elements available for each year. The algorithm setup follows the results obtained by *Cossich et al. (2021)*. The same Training Set and Test Tet elements, which were labelled with the help of LiDAR images, have been employed and the classification is performed on the spectral interval 380-1000 cm^{-1}, which maximises the classification scores.

Year	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Num. spectra	10280	11118	25288	24506	30904	29983	32715	34070	34003

Table 4.1: Number of DataSet elements collected over the years 2012-2020 by the REFIR-PAD instrument.

4.2 CIC Algorithm Set-Up

This section will describe the procedure used to select proper spectra to compose the Training Set and Test Set. This work aims to continue the classification performed by *Cossich et al. (2021)* for the period 2012-2015, with a larger dataset spanning from 2012 to 2022. The entire dataset has been recalibrated and new methods have been introduced to check the quality of the spectra. As a result, only some of the Training set and Test set elements are still available. However, to remain consistent with the previous results, no spectrum has been substituted to missing ones. The CIC classification considers three classes: clear-sky, ice cloud and mixed-phase cloud, and results are obtained following the scheme in *Fig.3.1.* The classification has been performed over 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. The channels selected depend on a variety of factors, such as the sensor characteristics, (noise and spectral resolution), the chosen training sets, the classes considered, the observation specifics and the atmospheric condition. In the case of Antarctica, the atmosphere is very dry over the entire year. As a consequence, FIR channels (up to 667 cm^{-1}) become less opaque and can be exploited to improve cloud detection.

4.2.1 Training Set

Spectra composing the Training Sets are selected from a subset of manually classified observations. This pre-classification is performed using the LiDAR instrument. REFIR-PAD spectra are co-located with LiDAR measurements, associating each REFIR-PAD observation with the closest LiDAR vertical profile. The time considered for the REFIR-PAD measurement is the beginning of the acquisition. Backscatter and depolarization LiDAR profiles were visually inspected to detect the presence of a cloud and determine its thermodynamic phase. In clear sky conditions the LiDAR signal decreases with height, while a sharp increase indicates the presence of a scattering layer such as a cloud. The depolarization profiles are then inspected to discriminate the cloud phase, values above 0.15 are representative of ice clouds, while mixed-phase clouds are identified in layers with depolarization below this threshold. Training Set spectra should describe all the observed variability in the area and, at the same time, be sensitive to the addition of a new Test Set spectrum. The optimum number of elements depends upon the trade-off between these two factors. Due to the intense change in environmental conditions, the Training Sets are defined independently for two macro seasons, a warm one from November to March and a cold one from April to October. Mixed-

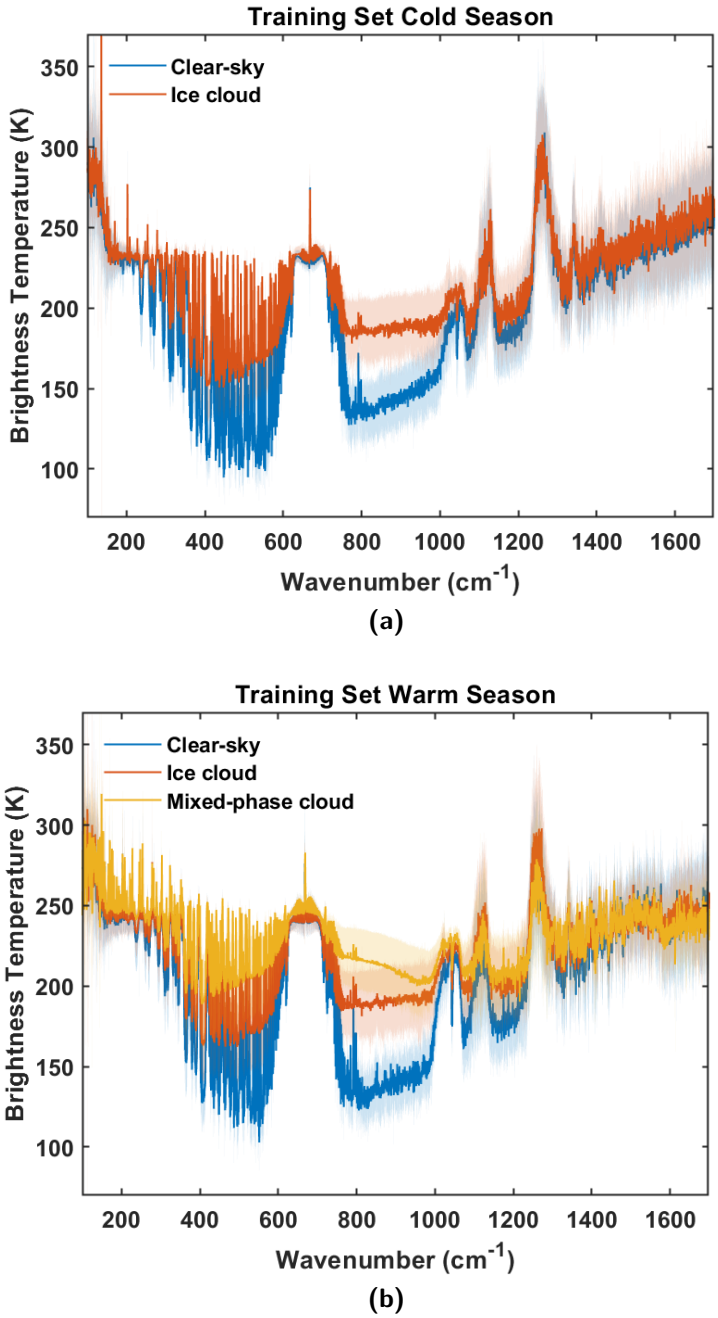


Figure 4.1: Mean Brightness Temperature spectra (solid lines) forming the Training Set, measured by the REFIR-PAD and respective standard deviations (shaded areas), grouped in accordance with the associated class during the cold (a), and the warm season (b).

phase clouds are observed only in the warm season, with a higher frequency in December and January. In cold months, extremely low temperatures prevent their formation. For this reason, three classes constitute the Training Set in the warm season (clear-sky, ice cloud, mixed-phase cloud), while only two classes (clear-sky and ice cloud) are used in the cold months. 119 spectra are selected to form the Training Set (59 in the warm season and 60 in the cold one). In the warm season, there are 23 clear-sky elements, 22 ice clouds and 14 mixed-phase clouds. While in the cold season, the training set is split into 40 clear-sky spectra and 20 ice clouds. *Fig.4.1* shows the mean Brightness Temperature (BT) spectra and their standard deviations for the classes used in each season. The two atmospheric windows, one in the far infrared, between 400 and 600 cm^{-1}, and the other in the middle infrared, between 800 and 1000 cm^{-1}, present the major discrepancies. Wavenumbers after the O_3 band are highly affected by instrumental noise. The standard deviation in the MIR window channels is larger for the cloudy spectra, in both seasons, which suggests a larger variability of the signal. Spectra are classified using the distributional approach, thus a first run of the algorithm on the Training Set elements has to be performed in order to find the best delimiter between each couple of classes. An example of results obtained on the clear-sky and ice cloud classes of the warm season is provided in *Fig.4.2a*. The classifica-

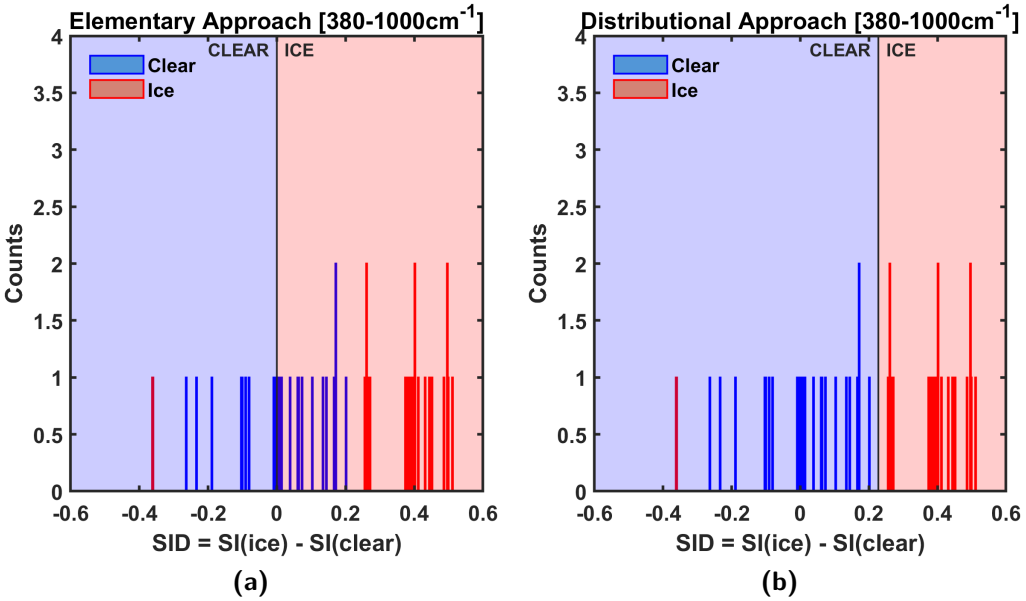


Figure 4.2: Example of SID distributions obtained for the Training Set elements of clear-sky and ice cloud classes in the warm season.

tion was performed using the spectral interval 380-1000 cm^{-1} (*Cossich et al. (2021)*). Positive SID values indicate spectra classified as ice clouds, while negative ones denote clear-sky elements. The colour of each bar represents the belonging class. There is a net separation between the two classes, which implies that the algorithm is able to correctly identify the spectral features separating the two scenes. The distributional approach allows to determine an ideal shift, other than zero, that maximises the classification results (in *Fig.4.2*).

4.2.2 Test Set

Once the Training Sets have been defined, CIC performances are optimised and assessed on an independent set of spectra composing the Verification and Test Set. The Verification Set was used by *Cossich et al. (2021)* to evaluate the best spectral interval, therefore, in this study, all the spectra have been merged to form the Test Set. Overall, 992 spectra were selected, collocated with the LiDAR measurements and visually classified using the backscatter and depolarization profiles. The Test Set for the warm season is composed of 52 clear-sky spectra, 125 ice clouds and 79 mixed-phase clouds. While the Test Set used for the cold season is made up of 271 clear-sky spectra and 465 ice clouds. The mean BT spectra are illustrated in *Fig.4.3*. Considering three general classes A, B and C, each spectrum belonging to class A can be classified correctly by the CIC as a member of class A, or incorrectly as a member of class B or C. Thus, the results can be interpreted as

- True Positive (TP): the spectrum belongs to class A and it is properly classified in class A.
- True negative (TN): the spectrum does not belong to class A and it is properly classified in its class of pertinence (B or C).
- False positive (FP): the spectrum belongs to class B or C but it is misclassified in class A.
- False negative (FN): the spectrum belongs to class A but it is misclassified in class B or C.

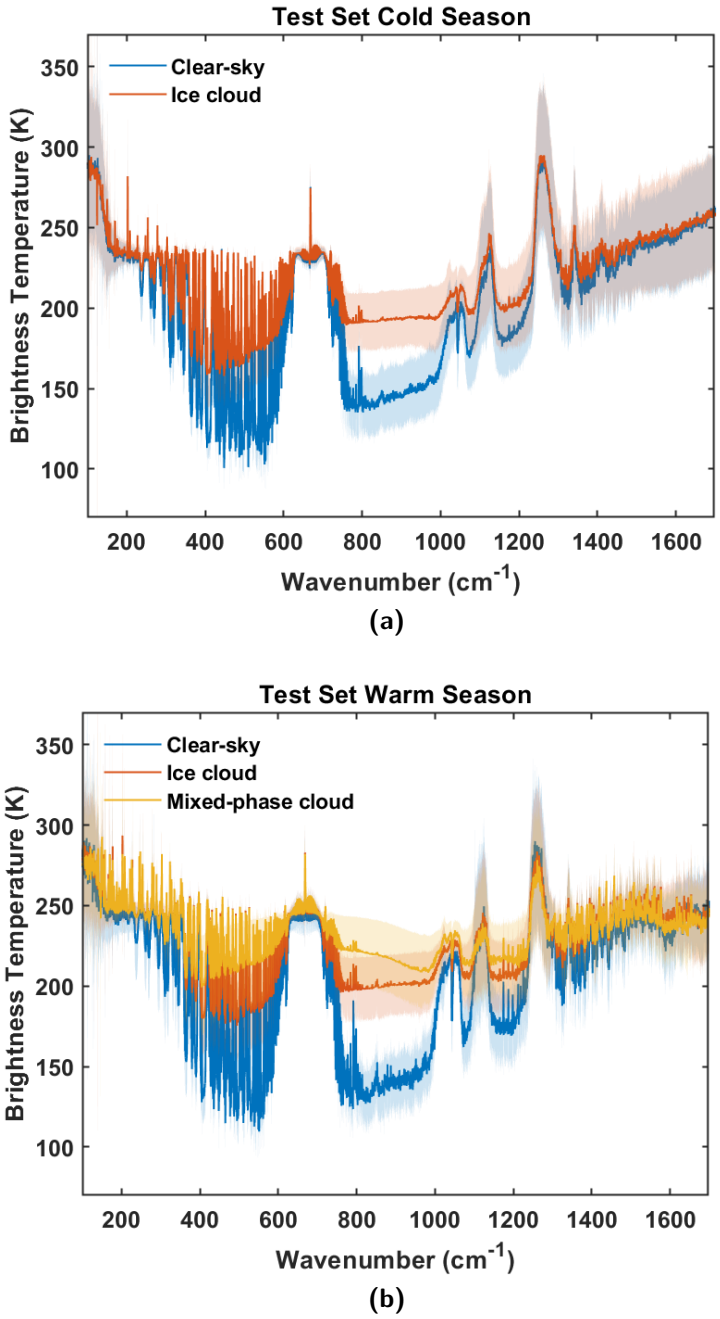


Figure 4.3: Mean Brightness Temperature spectra (solid lines) forming the Test Set, measured by the REFIR-PAD and respective standard deviation (shaded areas), grouped in accordance with the associated class during the cold (a), and the warm season (b).

Two main parameters are here defined to evaluate the algorithm performance: the Threat Score (ThS) and the Hit Rate (HR). For class A, the Threat Score is defined as:

$$ThS_A = \frac{TP}{TP + FN + FP} \tag{4.1}$$

where the true positive (TP) accounts for the total number of correctly classified spectra, while the false negative (FN) and false positive (FN) indicate the misclassified ones.
As for the Hit Rate, it is given by:

$$HR_A = \frac{N_A^{CIC}}{N_A^{true}} = \frac{TP}{TP + FN} \tag{4.2}$$

with N_A^{CIC} equal to the number of occurrences of class A that are correctly classified by the CIC, corresponding to the TP (true positive). While N_A^{true} is the total number of spectra belonging to class A, corresponding to TP + FN (false negative) of that class.
An additional indicator that can be used to evaluate the consistency of the spectra identified by the CIC as members of class A, is the Positive Predictive Value (PPV), also called precision, defined as:

$$PPV_A = \frac{TP}{TP + FP} \tag{4.3}$$

where the terms represent the True Positives (TP) and False Negative (FN). The PPV quantify the prevalence of correctly classified spectra within that class and represents the probability that an element classified as a member of class A actually belongs to class A.
Results obtained on the 992 Test Set spectra, using the spectral interval 380-1000 cm^{-1}, are presented in *Tab.4.2* for each class individually.

Field	**Num. spectra**	**ThS**	**HR**	**Num. misclass spectra**	**Misclassification**	**PPV**
Clear sky	323	0.91	0.92	25	7.8% - ice cloud 0% mixed-phase cloud	0.99
Ice cloud	590	0.93	0.98	10	0.68% clear sky 1.02% mixed-phase cloud	0.94
Mixed-phase cloud	79	0.80	0.86	11	0% clear-sky 13.92% ice cloud	0.92
Total	992	0.91	0.95	46	4.6%	0.95

Table 4.2: Results of the CIC classification obtained for the Test Set spectra.

Overall, 95% of spectra are correctly classified. Only a small percentage of cloudy spectra (ice clouds plus mixed-phase clouds) are misclassified as clear-sky (4 over 669 elements), and about 8% of the clear-sky spectra are erroneously identified as ice clouds. A very positive result is obtained in the case of mixed-phase clouds, where the CIC is able to identify the presence of the cloud in 100% of the cases, while 14% of the time the cloud phase is classified as ice instead of mixed-phase. Looking at the single classes, the clear-sky has a Hit Rate of 92%, 298 spectra over 323 are correctly classified. The 25 misclassified elements are labelled as ice clouds. The ice cloud class has the highest scores, with 98% (580) of spectra correctly classified. Of the remaining 10 spectra (2%), 4 are labelled as clear-sky and 6 as mixed-phase clouds. Mixed-phase clouds have a Hit Rate of 86%, which accounts for 68 spectra correctly classified and 11 elements misclassified as ice clouds. The Positive Predictive Value indicates that the clear class is composed of 99% clear-sky spectra, which is a great result considering that retrieval algorithms require a reliable flag of clear observations. The ice cloud and mixed-phase cloud are also very well defined, with 94% and 92% of correct elements respectively.

4.3 Results over the entire DataSet

The CIC is finally run over the full dataset, using the previously defined 380-1000 cm^{-1} spectral interval. Results are provided in terms of percentages, defining the occurrence of each class with respect to the total number of analysed spectra. An error can be associated with the percentage occurrence, exploiting the HRs derived in the Test Set analysis.
From *Eq.4.2*, the number of elements classified as members of class A as a function of the HR can be derived:

$$N_A^{CIC} = N_A^{true} \times HR_A \tag{4.4}$$

Hence, the number of misclassified spectra (N_A^{err}) can be defined as

$$N_A^{err} = N_A^{true} \times (1 - HR_A) \tag{4.5}$$

Now, since the number N_A^{true} is unknown for the dataset, *Eq.4.4* can be substituted in *Eq.4.5*, giving:

$$N_A^{err} = N_A^{CIC} \times \frac{(1 - HR_A)}{HR_A} = N_A^{CIC} \times \left(\frac{1}{HR_A} - 1\right) \tag{4.6}$$

The relative error (ϵ) for class A can be obtained by dividing the number of misclassified spectra of class A by the total number of spectra in the dataset (N_{tot}):

$$\epsilon_A = \frac{N_A^{err}}{N_{tot}} = \frac{N_A^{CIC}}{N_{tot}} \times \left(\frac{1}{HR_A} - 1\right) \tag{4.7}$$

The HR values associated with each class for the entire dataset are unknown. However, it is assumed that the CIC scores over the test set spectra represent the performances obtained over the full dataset. Thus, the values obtained in the Test Set for each class are used in the equation. The percentage classification error for class A is written as:

$$\epsilon_A\% = \epsilon_A \times 100 = \frac{N_A^{CIC}}{N_{tot}} \times \left(\frac{1}{HR_A} - 1\right) \times 100 \tag{4.8}$$

where N_A^{CIC} is the number of spectra identified by CIC as a member of class A, and N_{tot} is the total number of spectra in the entire dataset.

The percentages of the classification results over the entire dataset and the associated uncertainties are given in *Tab.4.3*.

	Clear-sky (%)	**Ice cloud (%)**	**Mixed-phase cloud (%)**	**Observation Time (%)**
Total	70.09 ±5.88	27.69 ±0.48	2.21 ±0.36	69.11
2012	56.07 ±4.70	38.25 ±0.66	5.68 ±0.92	27.38
2013	71.01 ±5.96	27.51 ±0.47	1.48 ±0.24	29.61
2014	69.97 ±5.87	27.97 ±0.48	2.06 ±0.33	67.36
2015	67.46 ±5.66	30.70 ±0.53	1.84 ±0.30	65.27
2016	67.50 ±5.66	30.55 ±0.53	1.96 ±0.32	82.32
2017	70.42 ±5.90	27.83 ±0.48	1.74 ±0.28	79.86
2018	72.48 ±6.08	24.35 ±0.42	3.17 ±0.51	87.14
2019	69.52 ±5.83	28.24 ±0.49	2.25 ±0.36	90.75
2020	76.39 ±6.41	22.13 ±0.38	1.48 ±0.24	90.57

Table 4.3: CIC classification results for the whole REFIR-PAD dataset (2012–2020) and for single years. Values and associated uncertainties are reported in percentages.

Almost 70% of the full dataset is composed of clear-sky elements, the rest 30% is divided between ice clouds (almost 28%) and mixed-phase clouds (2%). Uncertainty values associated with the clear-sky classification are higher due to the smaller Hit Rate of the class. The most cloudy year is 2012, with a cloud occurrence that exceeds 40%. On the opposite, 2020 has the lowest cloud percentage of less than 25%. The last column in the table indicates how long the REFIR-PAD was actively measuring each year. Values are calculated considering the instrument time resolution of 14 minutes. In 2012, the time in which the instrument was actually observing the sky is less than 30%, and the data is mostly distributed in the summer months. This result can explain the bias introduced in the percentages, which are significantly different from the rest of the years, and the larger number of mixed-phase clouds (more than 5%). *Fig.4.4* illustrates the distribution of data available over the entire time period 2012-2020. The number of observations is represented by the associated colour scale. There is an evident increase in measurements over the years, and the monthly distribution becomes more uniform. The first two years have less than 1800 measurements each month, which correspond to the overall 30% shown in the last column of *Tab.4.3*.

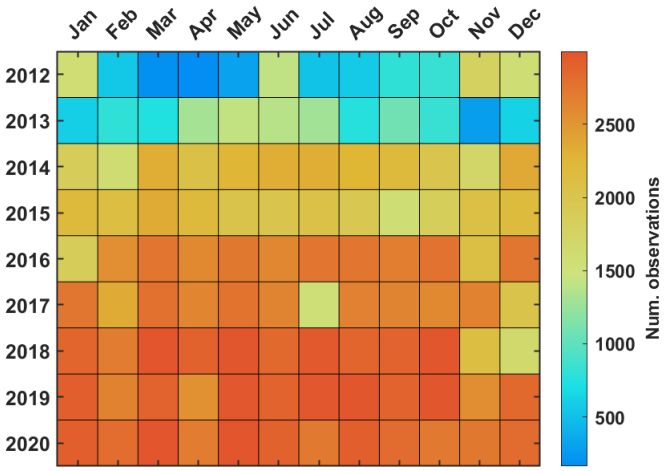


Figure 4.4: Representation of the dataset distribution over the years 2012-2020. Colours indicate the number of data available.

Some differences are found between 2012 and 2013. The first year has a larger number of observations, with respect to 2013, only in November, December and January, and these are also the only months for which data are more than 1000. A rapid indication of the uniformity of the data distribution is given from the difference between the minimum and maximum number of measurements in one year. This value is equal to 1622 in the first year, with a maximum of 1785 observations in November and a minimum of just 163 in April. In 2013, this discrepancy reduces to 1148 and data appear more well-distributed. The maximum is obtained in May, with 1431 REFIR-PAD measurements, and the minimum of 283 in November. Even though the time in which the in-

strument was actively measuring is almost the same in these two years and the number of observations available is equivalent (see *Tab.4.1*), only the cloud occurrence obtained in 2013 is in line with values of the other years. A possible explanation is given by the highest consistency of data counts in 2013 between the various months. As a matter of fact, 2012 is the most heterogeneous of all the years. After 2013, there are between 1500 and 3000 data per month. A t-Student test was then conducted to confirm that the largest cloud occurrence in 2012 was due to the different number of observations rather than a physical signal. The t value was calculated from the mean cloud occurrence in 2012 and the one observed from 2014 to 2020 (2013 was excluded to avoid any bias), together with the standard deviations of the two distributions. The test confirmed that the percentage obtained in 2012 is included in a confidence level between 95 and 95.5% and that the two distributions are statistically consistent. Hence, the higher mean cloud occurrence in 2012 is likely due to the low number of observations and their particular distribution over the months, which results in favour of cloud scenes.

4.3.1 Statistical Analysis

In this paragraph, results obtained for the entire dataset will be presented in terms of cloud occurrence and in correlation with the surface temperature. The cloud occurrence is calculated by summing the ice cloud and mixed-phase cloud values. *Fig.4.5a* shows the same annual percentages reported before. An indication of the annual variability is given by computing the minimum and maximum values observed within the year (shaded area). As previously said, 2012 and 2020 are the most and least cloudy years respectively. The number of observations increased after 2013 and the cloud variability remained quite stable, with a maximum around 60% in 2014 and a minimum just above 5% in 2013. Over the 9 years, the mean value is 31.02%. As for the monthly cloud occurrence, in *Fig.4.5b*, the largest mean values are in winter, with a peak in August at 38.78%. The lowest cloud occurrence is instead found in April at 20.74%. Both values account for ice clouds only, in fact, mixed-phase clouds are found only from November to March, with a maximum of almost 12% in December and January. The high values of maximum cloud occurrence in June and August are due to the results obtained in 2012. The month with the lowest variability observed in 9 years is April, followed by months between December and March. During the winter, the variability increases and this is probably due to the fact that very few observations in 2012 are performed within those months, increasing the number of clouds over the total measurements and thus the percentages reported in the graph.

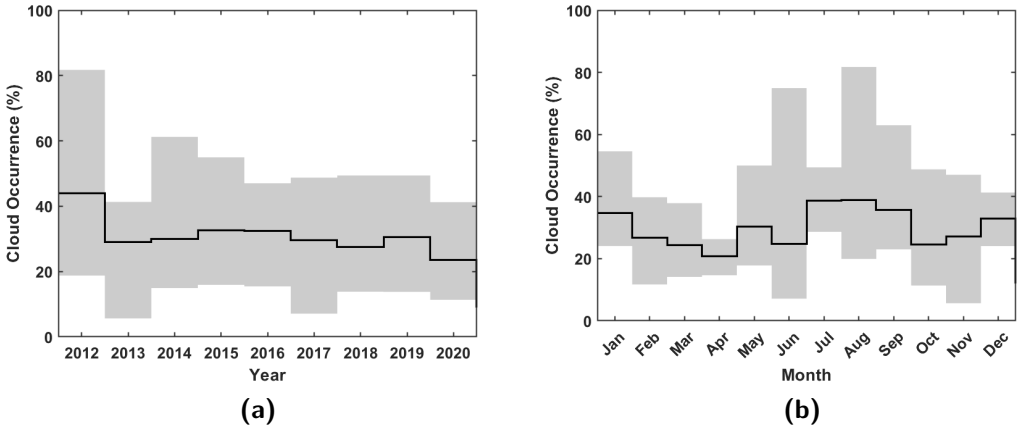


Figure 4.5: Mean annual (a) and monthly (b) cloud occurrence (solid lines) provided by the CIC over the entire dataset. Shaded areas represent the variability observed between the maximum and minimum values.

The seasonal cloud occurrence is presented in *Fig.4.6*. The highest value is observed during summer (December, January and February) around 33.80%, followed by winter (June, July and August) months. However, they remain quite stable between 20 and 40%. The seasonal variation over the years is illustrated in *Fig.4.7*. The minimum value in Spring (SON) is recorded in 2019 at 20.44%, while summer (DJF) and autumn (MAM) have a low in 2020 at 24.63% and 19.58% respectively. As for the winter season (JJA), the least cloudy year was 2017 with a cloud occurrence of 22.61%. The maximum cloud fraction in spring is registered in 2017 at 38.72%, while the other three seasons have a peak in 2012 at 40.47% in summer, 30.76% in autumn and 69.06% in winter, as a consequence of the issues addressed before.

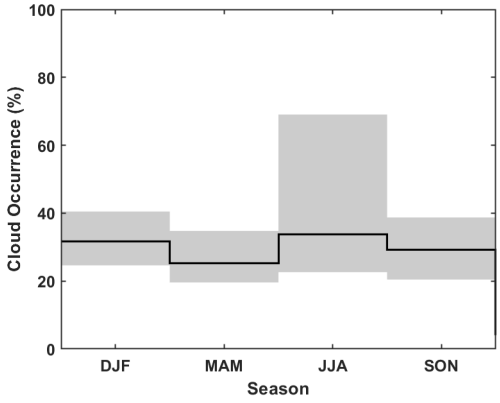


Figure 4.6: Mean seasonal cloud occurrence (solid lines) provided by the CIC over the entire dataset. Shaded areas represent the variability observed between the maximum and minimum values.

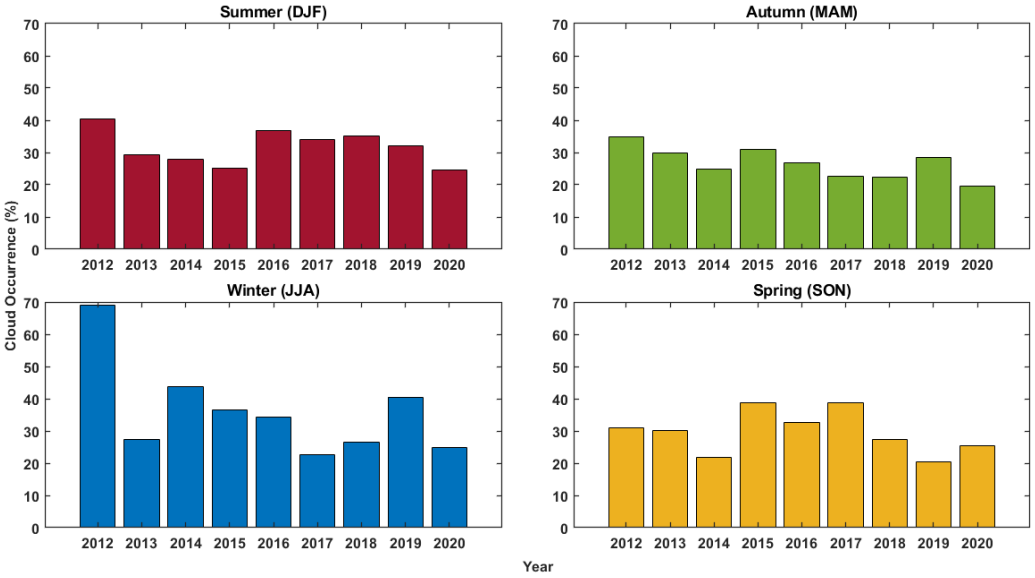


Figure 4.7: Mean seasonal cloud occurrence provided by the CIC from 2012 to 2020.

Finally, the relationship between the different classes occurrences and the surface temperature is investigated. Ground measurements are performed every hour at the Concordia Station, thus each REFIR-PAD measurement has been associated with the temperature measured in that hour. *Fig.4.8* shows results divided into the four seasons, Summer (DJF), Autumn (MAM), Winter (JJA) and Spring (SON). Over the nine years, the surface air temperature (corresponding to REFIR-PAD measurements) varies between a minimum of -81.2°C and a maximum of -14.3°C. The largest variation is found in spring. The maximum temperature reached in winter is -25°C while the minimum is around -80°C. During the summer, the maximum is around -11°C, and the minimum is just above -70°C and at temperatures below -60°C only clear-sky scenes were identified. A positive cloud forcing, due to an increase in the downwelling longwave radiation from the cloud layers, is highlighted by the increase in temperature observed at the surface in presence of clouds. This effect is a little blunt in autumn and spring in correspondence with temperature bins containing fewer observations. Mixed-phase clouds are found only at temperatures higher than -50°C. In addition, temperatures higher than around -30°C both in autumn and winter are associated only with clouds.

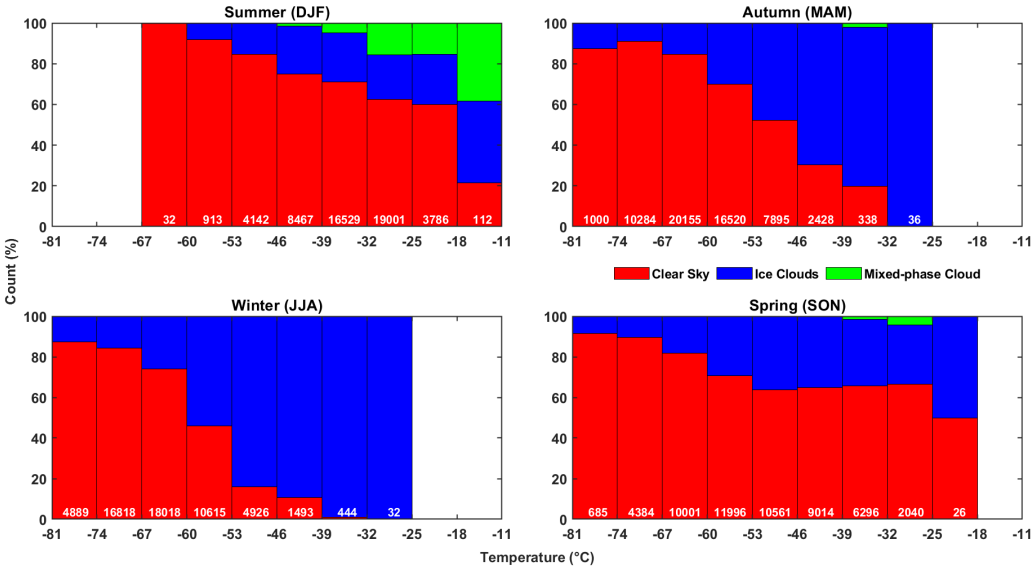


Figure 4.8: Occurrence of each sky condition as a function of the surface air temperature in the four seasons. The number of observations for each bin is reported at the base of each histogram.

The mean temperatures for the clear and cloudy scenes in each season are reported in *Tab.4.4*, together with the mean values in all-sky conditions. The largest difference between clear-sky and cloud temperatures is found in winter at 8.65°C, with a mean cloud forcing of almost 3°C. The clear-sky and cloudy temperatures in this season reach the lowest values (-66.62°C and -57.97°C respectively). The difference mitigates in summer to 2.85°C, due to a temperature rise for both scenes. In this season, the increase in the mean cloud temperature comes with a larger number of mixed-phase clouds observed, which leads to a larger effect on the surface temperature.

	$<T_{clear}>$	$<T_{cloud}>$	$<T_{all\text{-}sky}>$	**Difference Cloud - Clear**	**Difference All-sky - Clear**
DJF	-35.64°C	-32.79°C	-34.76°C	2.85°C	0.88°C
MAM	-62.04°C	-55.41°C	-60.36°C	6.63°C	1.68°C
JJA	-66.62°C	-57.97°C	-63.65°C	8.65°C	2.97°C
SON	-53.44°C	-49.41°C	-52.29°C	4.03°C	1.15°C

Table 4.4: Mean seasonal surface temperature measured at Concordia Station in correspondence of the sky scene identified by the CIC and their differences.

Chapter 5

Cloud Detection from Satellite

In the first part of the chapter, the CIC algorithm is applied to satellite data collected by IASI (Infrared Atmospheric Sounding Interferometer) over the years 2012-2015. First, the specifics of the IASI instrument and products are described. Then, the procedure followed to collocate IASI spectra with the ground observations is delineated and finally, results obtained from the CIC classification are compared to IASI L2 cloud products. The second part illustrates the cloud occurrence observed by MODIS compared with ground-based measurements, over the years 2012-2020.

5.1 MetOp satellites: the IASI instrument

IASI is an infrared sounder that measures the thermal radiation emitted by the Earth's surface and the atmosphere. It is composed of a Fourier transform spectrometer and an associated Integrated Imaging Subsystem (IIS). The Fourier transform spectrometer provides infrared spectra with high spectral resolution between 645 and 2760 cm^{-1} (3.6 μm to 15.5 μm). From those measurements, the atmospheric composition and temperature can be retrieved, forming the Level 2 (L2) product. The IIS consists of a broadband radiometer with a high spatial resolution, that measures between 833 cm^{-1} and 1000 cm^{-1} (12 μm and 10 μm) and its information is used for co-registration with the Advanced Very High-Resolution Radiometer (AVHRR), on board the same platform. The processing of IASI data is conducted by the EUropean organisation for the exploitation of METeorological SATellites (EUMETSAT). Infrared (IR) sounders such as IASI are very useful for studying cloud properties since they are not affected by day–night biases due to solar contamination. The high spectral resolution makes the instrument very reliable also for the determination of cirrus properties (*Hilton et al.*

(2012)). Moreover, the long time series of data available makes IASI an important contributor to climate studies. From the atmospheric emission spectra, IASI products provide temperature profiles in the troposphere and the lower stratosphere with a vertical resolution of 1 km, an accuracy of 1 K and a horizontal sampling of 25 km; water vapour profiles in the troposphere with the same vertical resolution of 1 km and horizontal sampling of 25 km and an accuracy of 10% on relative humidity; the total amount of ozone (O_3) with an accuracy of 5% and information about its vertical distribution with an accuracy of 10% (*EUMETSAT (2017b)*). Additionally, IASI is used for the determination of other trace gases such as nitrous oxide (N_2O), carbon dioxide (CO) and methane (CH_4), as well as land and sea surface temperature and emissivity, and cloud properties (fractional cloud cover and cloud top temperature/pressure).

5.1.1 IASI orbit and Field of View

Three IASI instruments have been operative over the years on board the MetOp satellites: IASI-A (IASI on board MetOp-A platform) launched in 2006 and switched off in 2021, IASI-B from 2012, and IASI-C launched in 2018. IASI is a polar-orbiting satellite, flying at an altitude of 817 km with an inclination of 98.7°. It is an across-track scanning system and each scan line has a swath width of 2200 km on the ground and contains 30 fields of view, 15 for each side of the nadir direction. The scan starts on the left side with respect to the flight direction of the spacecraft. Each scan is called

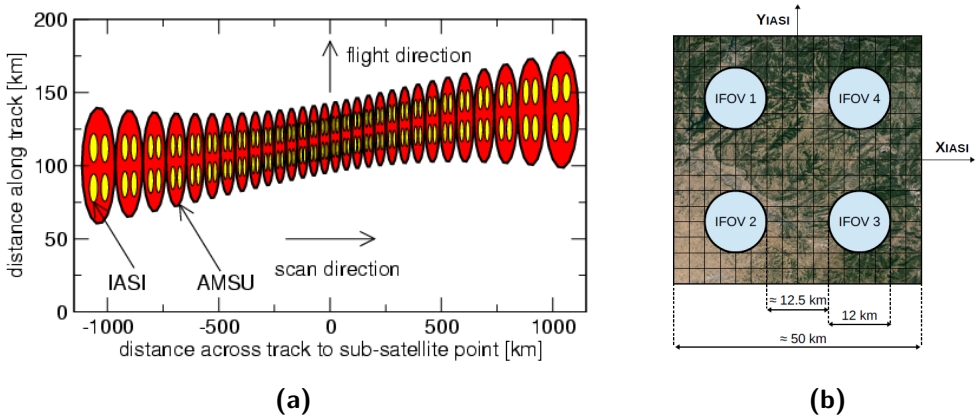


Figure 5.1: (a) IASI scan line geometry (*EUMETSAT (2019)*), (b) IASI EFOV (about 50x50 km), each IFOV spreads 12 km of the Earth's surface and is separated from its neighbouring IFOVs by 12.5 km (*García-Sobrino et al. (2017)*).

elementary of effective field of view (EFOV) and consists of a cell about 3.3 degrees x 3.3 degrees, or 50 km x 50 km at nadir, analysed simultaneously by a 2 x 2 array of detectors centred in the viewing direction and forming a matrix of four circular pixels called Instantaneous Field of View (IFOV). A representation of a full scan line and a single EFOV is given in *Fig.5.1.* Each IFOV has a diameter 14.65 mrad, corresponding to a ground footprint of 12 km at nadir, while the size at the edge of the scan line along the across-track direction is 39 km, as reported in *Tab.5.1.*

Characteristics	Value	Unit
Scan type	step and stare	–
Scan rate	8	second
Stare interval	151	ms
Step interval	8/37	second
Number of Earth scans / line - EFOV	30	-
Swath	± 48.333	degrees
Swath line	± 1100	km
IFOV - shape at nadir	circular	–
IFOV - size at nadir	12	km
IFOV - size at edge of scan line across track	39	km
IFOV - size at edge of scan line along track	20	km

Table 5.1: IASI scanning characteristics (*EUMETSAT (2019)*).

5.1.2 Products and Processing Levels

IASI measures radiance spectra composed of 8461 channels between 645 and 2760 cm^{-1} (15.5 μm and 3.63 μm), with a spectral resolution of 0.5 cm^{-1} after apodisation (contained in the L1c product). The spectral sampling interval is 0.25 cm^{-1}. Each spectrum is measured in three bands (summarised in *Tab.5.2*), each with a separate detector. The first band (from 645 to 1210 cm^{-1}) spans over long-wave channels, comprising the CO_2 and O_3 absorption bands and the LW window; from 1210 to 2000 cm^{-1} there are channels mainly sensitive to humidity; and finally the last band from 2000 to 2700

Band	Wavenumbers (cm^{-1})	Wavelengths (μm)
1	645 - 1210	8.26 - 15.50
2	1210 - 2000	5.00 - 8.26
3	2000 - 2760	3.62 - 5.00

Table 5.2: IASI's three spectral bands (*EUMETSAT (2019)*).

cm^{-1} contains the short-wave channels. IASI spectral radiances are included in the Level 1 product, whose processing chain comprises three sublevels. Level 1a contains spectra radiometrically calibrated from the raw Level 0 product using two calibration views (Black body and Cold space). The geolocation of IASI is estimated based on the results from the coregistration of AVHRR Level 1b data and the calibrated IIS image. Validation of the geometric calibration is carried out frequently, using scenes with high-contrast features, e.g. coastlines (*Hilton et al. (2012)*). IASI Level 1a spectra are then resampled, obtaining the Level 1b product. Finally, Level 1c accounts for the apodization and contains the AVHRR radiance analysis. The usability of an IASI spectrum is indicated by the Boolean flag *GQisFlagQual* for each band. The IASI Level 1c products are organised as successive scan lines, each forming a MDR (Measurement Data Record) structure. While the IFOVs within one scan line are referenced by the geolocation and the acquisition time.

The retrievals of geophysical parameters such as atmospheric vertical profiles, gases and cloud properties are included in the Level 2 product. In particular, cloud parameters derived from IASI include cloud fraction, cloud top temperature, cloud height and cloud phase (their detailed description is given in *EUMETSAT (2017a)*). The cloud phase is estimated for cloudy IASI IFOVs by evaluation of the infrared window regions between 8 μm to 9 μm and 11 μm to 12 μm. While cloud detection is performed using three distinct methods, to be able to detect clouds under all conditions. The first one is a NWP test, which uses simulated radiances in the window channels, computed using the radiative transfer code RTTOV, compared to the actual IASI observation. Large differences are interpreted as the presence of a cloud in the IASI field of view. The second uses the AVHRR collocated cloud mask within the IASI IFOV. IASI pixels with AVHRR cloud fractions (embedded in the L1c product) exceeding a configurable threshold are flagged cloudy. The third test applies artificial neural networks to IASI radiances and AVHRR cluster information (mean value and variance) and classifies the scenes into cloud-free, partly cloudy or fully cloudy. The three tests are summarized in *Tab.5.3*. Each IASI IFOV is declared cloud-free with high confidence only if all tests conclude the absence of clouds. If a cloud is detected by at least one of the tests, a cloud characterisation is attempted and where no clouds

Test Name	**Type of test**	**Measurements used**
NWP	Window channel test	IASI spectra; NWP forecast
AVHRR	Integrated fractional cloud coverage	AVHRR cloud-mask
ANN	non-linear classification	IASI and AVHRR measurements

Table 5.3: IASI Level 2 cloud detection tests (*EUMETSAT (2017b)*).

could be confirmed with confidence, the IASI IFOV is flagged as clear pixel with potential cloud contamination.

5.2 Collocation

The application of the CIC algorithm to IASI spectral radiances requires the availability of a set of labelled data, where the scene observed (clear-sky or cloudy) is known a priori. For this purpose, REFIR-PAD measurements have been used as reference data. The high scores obtained for the classification, described in the previous chapter, and the availability of LiDAR observations to confirm some of those results, make the instrument very reliable and suitable for the purpose. IASI data from 2012 to 2015 were downloaded and collocated with the ground-based measurements. The IASI cloud phase included in the L2 products was used as a first comparison with the REFIR-PAD observations, while the IASI spectral radiances, from the L1c products, were used as input to the CIC algorithm.

IASI orbits were downloaded programmatically from the EUMETSAT Data Store, through the EUMETSAT Data Access Client (EUMDAC) Python library. First, all the L2 orbits available for the four years under study were downloaded and collocated, and then L1c products were downloaded, using the same procedure, only for those orbits containing collocated pixels. The processing of IASI data, conducted by EUMETSAT, has changed over time due to instrumental and software modifications. In 2019, EUMETSAT reprocessed the 2007–2017 IASI Metop-A L1C data with the most recent version of the algorithm. There is now a homogeneous L1C dataset available and consistent with both the L1C product generated after 2017 and with IASI-B. Various changes have been assessed in the study performed by *Bouillon et al. (2020)*, thus in order to remove any bias, the new version of L1c products for MetOp-A (forming the IASI Level 1C Climate Data Record Release 1) was downloaded using an FTP access provided by EUMETSAT.

5.2.1 Criteria used for L2 and L1 products

A collocation of the ground-based measurements and the satellite FOV occurs if they observe the same place at approximately the same time. For each IASI orbit file, all MDRs (corresponding to the scan lines) have been analysed. A unique acquisition time is specified in the record header for each scan line or MDR struct, while the four pixels in the 30 IASI fields of view are distinguished by the correspondent latitude and longitude and the satellite zenith angle. The distance in km was calculated from each pixel

centre (correspondent to the given coordinates) to the REFIR-PAD location (located at -75.1° latitude and 123.33° longitude).
Knowing that the acquisition time for the REFIR-PAD instrument is about 15 minutes (see Chapter 4) and each IASI IFOV has a diameter of 12 km (see *Fig.5.1b*), a IASI pixel is flagged as collocated with a REFIR-PAD observation if:

- the IASI measurement is carried out maximum 15 minutes before the observation time of the REFIR-PAD;
- the maximum distance of the IASI pixel centre from the REFIR-PAD instrument is less than 6 km (the ground observation is inside the IASI IFOV).

Finally, a filter for satellite zenith angles below 6.7° has been introduced to avoid geometric distortions.

Year	Num. Observations
2012	21
2013	30
2014	52
2015	64

Table 5.4: Number of IASI granules collocated with REFIR-PAD observations over the four years under study.

In total, 167 IASI observations were identified as spatial and temporal collocated to the ground-based measurements, subdivided over the four years as described in the *Tab.5.4*. In 2012, the only operative satellite was MetOp-A. From 2013 also MetOp-B data has become available, increasing the number of collocated observations, which nearly tripled in 2015. Using the information contained in the L2 products, the cloud phase retrieved by IASI (corresponding to the field *CLOUD_PHASE*) and the one determined by the REFIR-PAD for each collocated measurement were saved into a matrix, together with the IASI filenames, the name of the MetOp platform, the acquisition time, the distance from Dome-C and the satellite zenith angle. The number of scan line, EFOV and collocated pixel were used to download the corresponding spectral radiances from the L1c products.

IASI L1C Radiances

The field containing the radiance spectrum in the IASI L1c product is the *GS1cSpect*, although those values have to be multiplied by scale factors. EUMETSAT has divided the IASI spectrum into five spectral regions and each region has a scale factor that is approximately proportional to the value of the

radiance in the region. They have to be read from *GIADR-SCALEFACTORS* records. The decoded spectra are obtained using the following expression:

$$SpectDecoded(w) = Spect(w) \times 10^{-SF} \tag{5.1}$$

where *Spect* is the original spectrum provided in *GS1cSpect*, *SF* is the specific scale factor (power of 10) to be applied within the corresponding band of the spectral sample number *w*. The computation of the wavenumber associated with the spectral sample number w for the IASI L1c spectra is given by the formula:

$$wavenumber(w) = IDefSpectDWn1b \times (IDefNsfirst1b + w - 2) \tag{5.2}$$

with *IDefNsfirst1b* equal to the number of the first sample of the IASI spectrum and *IDefSpectDWn1b* the sample width. More detailed information on the entire procedure can be found in *EUMETSAT (2019)*.
The spectra obtained were thus included in the file of collocated measurements.

5.2.2 IASI collocated dataset

At first, the collocated measurements were analysed in terms of a one-to-one comparison between the IASI cloud phase and the REFIR-PAD observed scene. Results are summarised in *Tab.5.5*.

		REFIR-PAD		
		Clear-sky	**Ice cloud**	**Mixed-phase cloud**
IASI	**Clear-sky**	33	6	4
	Ice cloud	83	37	1
	Mixed-phase cloud	1	1	1

Table 5.5: Matrix comparing the number of clouds detected by the REFIR-PAD and the IASI instruments for the entire collocated dataset of 167 observations.

The majority of scenes observed by the REFIR-PAD instrument are classified as clear-sky by the CIC and account for 70.06% of the total (117 measurements). On the other hand, IASI observed 43 clear-sky scenes, corresponding to 25.75%, while ice clouds are detected in 72.45% of the cases (121 observations). Almost 71.8% of clear-sky observations collected from the ground are classified as cloudy by IASI, 98.8 % of which as ice cloud. As for the ice clouds seen by the REFIR-PAD instrument, 84.09% are confirmed by IASI. While 66.67% of the mixed-phase clouds are classified as clear-sky scenes.

When a cloud is detected, the two instruments are in accordance with its thermodynamic phase 76% of the time. Considering the scene observed from the REFIR-PAD as the "truth", the IASI instrument can be considered quite reliable when no clouds are present, in fact, 76.74% of the clear-sky scenes observed by the satellite agree with the ground observations. Although, the ice clouds detected by the satellite are actually seen by the REFIR-PAD only 68.59% of the time. This problem can be due to the different field of view dimensions among the two instruments or the frequent temperature inversions recorded at Dome-C.

The two plots in *Fig.5.2* illustrate the mean spectra for each sky scene observed, labelled according to the CIC algorithm results on the REFIR-PAD collocated data (*Fig.5.2a*) and IASI cloud phase provided in the L2 products (*Fig.5.2b*). According to the IASI classification, the Brightness Temperatures (BTs) in clear-sky conditions are generally warmer than in presence of an ice cloud, especially in the atmospheric spectral window between 800-1200 cm^{-1}. The opposite is true according to the REFIR-PAD scene observed. The standard deviation reveals a significant variability, in both classifications. The values for the first two classes (clear and ice cloud) overlap and there is no net separation between the scenes. As for the mixed-phase clouds, they are identified as warmer both from satellite and from the ground. However, in the latter case, their standard deviation is slightly smaller and thus the mixed-phase cloud scenes observed are more homogeneous. In general, the ground measurements reveal a temperature at the surface colder in clear-sky conditions, which increases in presence of a cloud. The opposite is true from satellite measurements, except for mixed-phase clouds which form only in very warm conditions.
These results are obtained by averaging the BTs over the entire time period. Distinct features are exhibited when dividing the spectra into two macro seasons. A cold season can be defined from April to September, where clear-sky scenes are identified in conditions of lower surface temperatures with respect to the cloudy ones, from both ground and satellite measurements (*Fig.5.3*). Due to extremely low temperatures, only ice clouds are found in those months and their mean BTs are lower in the IASI classification and more similar to the clear-sky scenes.
The same behaviour observed before, averaging the BTs over the full time period, is found in the warm seasons, from October to March, shown in *Fig.5.4*. In particular, also the clear-sky and ice cloud temperatures are quite high in this season and their standard deviations reach values observed for mixed-phase clouds. The scenes identified as clear-sky by the IASI algorithm have surface temperatures just a few kelvins higher than the ice clouds ones.

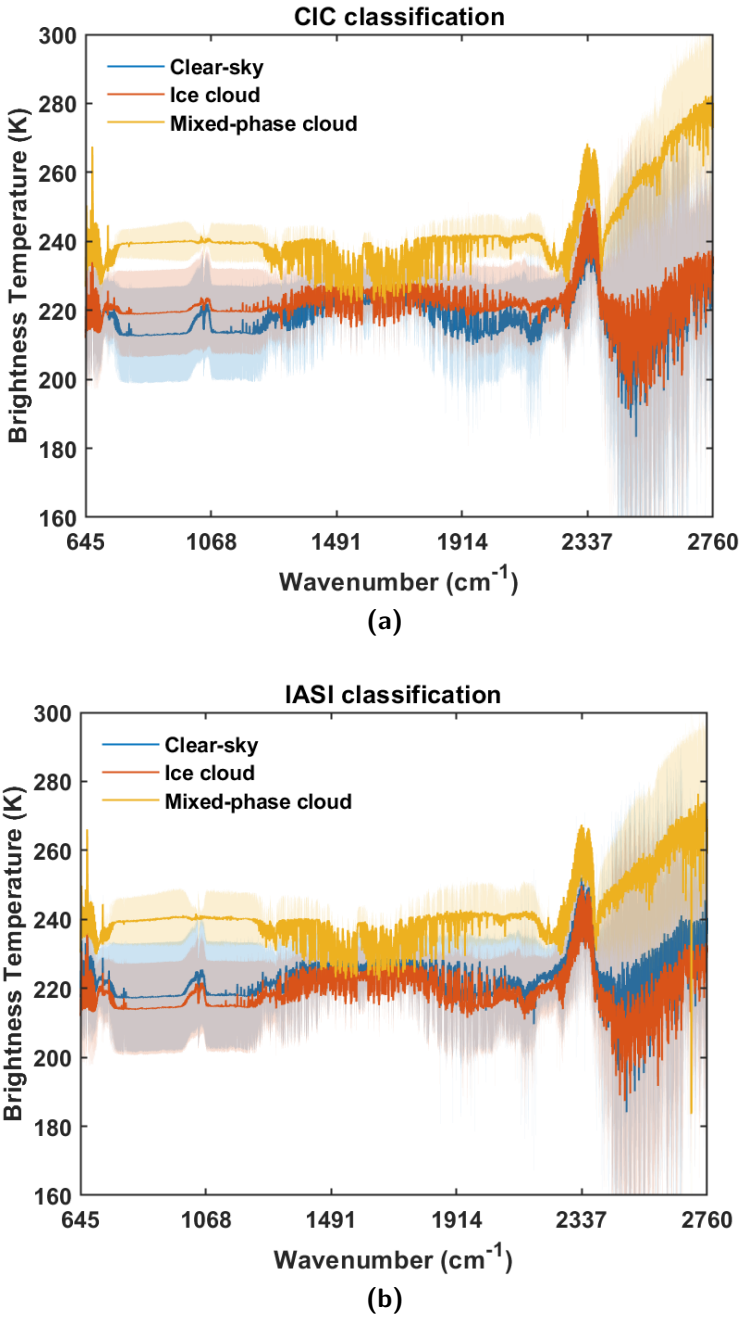


Figure 5.2: Mean Brightness Temperature spectra (solid lines) measured by IASI and respective standard deviation (shaded areas), grouped in accordance with the associated class identified by the CIC algorithm over the REFIR-PAD collocated measurements (a), and the IASI cloud detection algorithms (b).

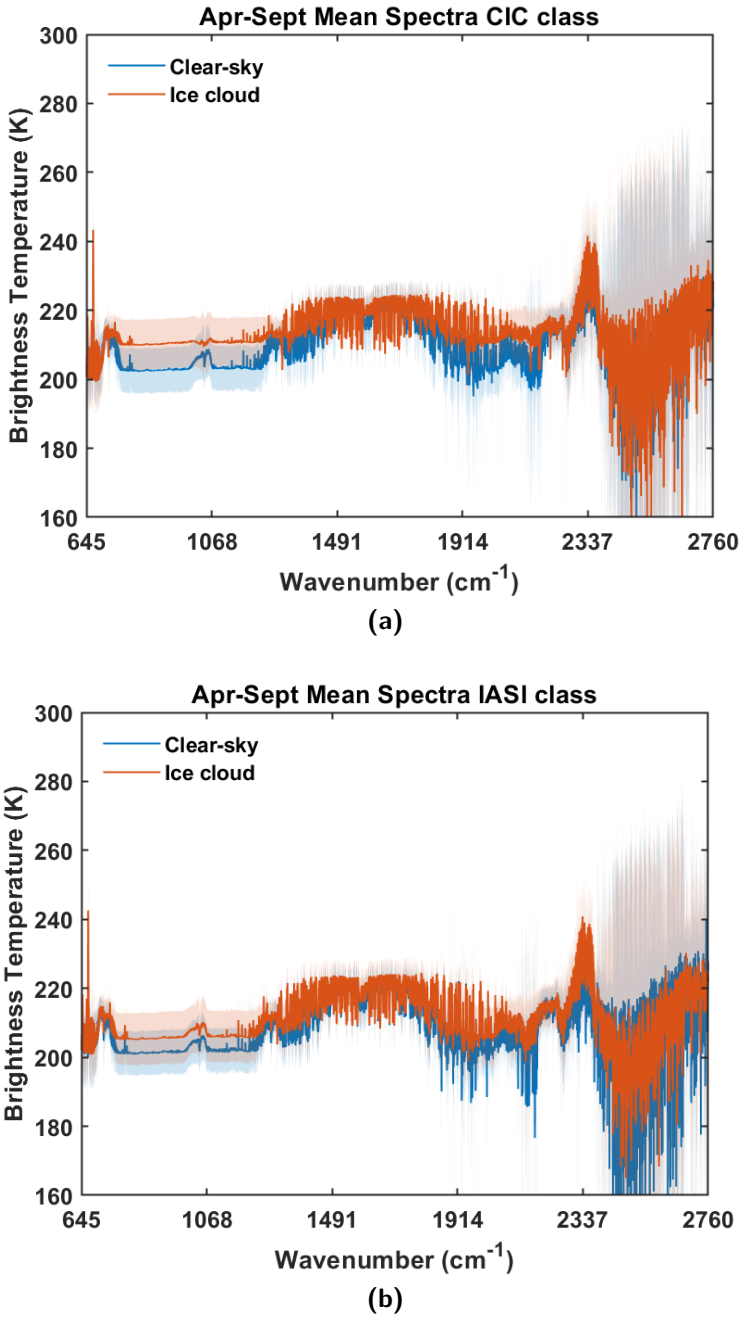


Figure 5.3: Mean Brightness Temperature spectra (solid lines) for the cold season, between April and September, measured by IASI and respective standard deviation (shaded areas), grouped in accordance with the associated class identified by the CIC algorithm over the REFIR-PAD collocated measurements (a), and the IASI cloud detection algorithms (b).

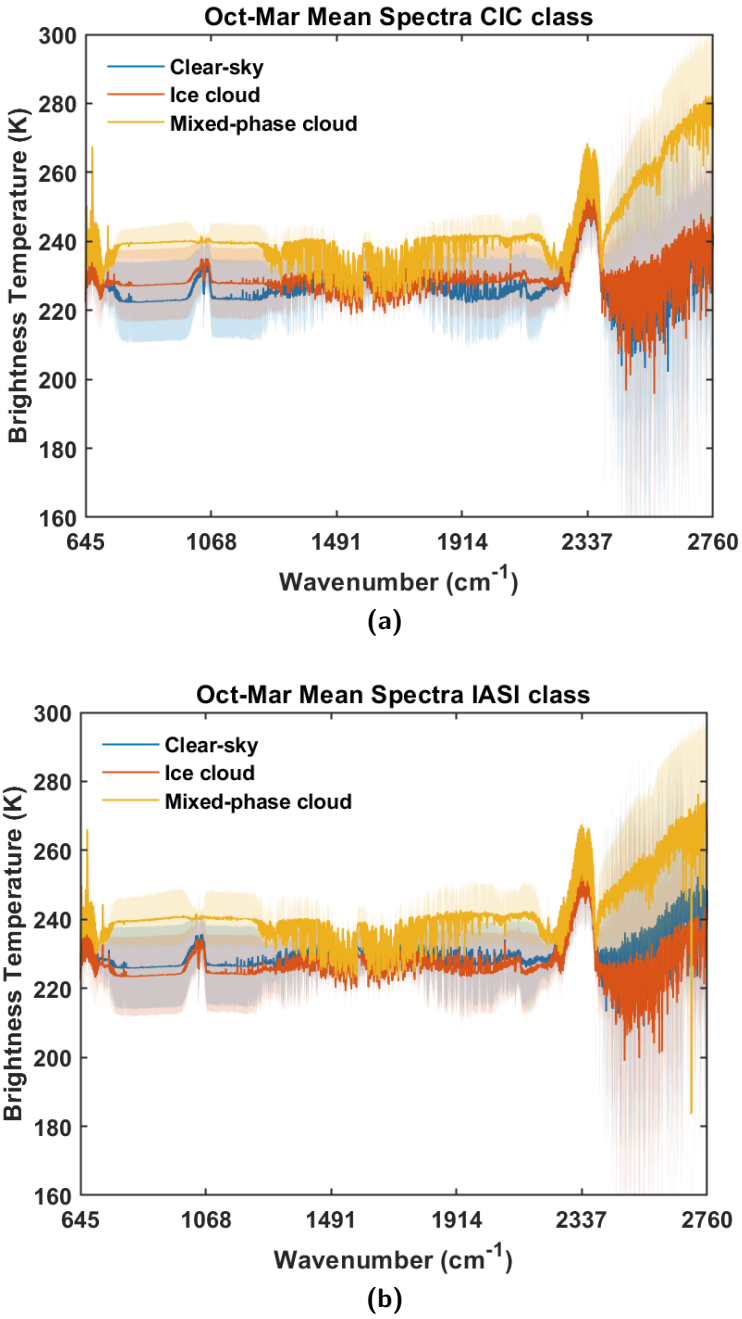


Figure 5.4: Mean Brightness Temperature spectra (solid lines) for the warm season, between October and March, measured by IASI and respective standard deviation (shaded areas), grouped in accordance with the associated class identified by the CIC algorithm over the REFIR-PAD collocated measurements (a), and the IASI cloud detection algorithms (b).

Spectral characteristics observed in *Fig.5.2a* for the clear-sky and ice cloud scenes emphasize the phenomenon of temperature inversion at the surface, on the Antarctic Plateau. The atmospheric vertical profiles of temperature, obtained from radiosondes launched every day at Concordia Station, are used to further investigate this point. Radiosounding measurements are available only at 12 UTC, thus only IASI measurements performed between 10 a.m. and 3 p.m. were used for this analysis and matched to the correspondent temperature profile, atmospheric conditions are considered to be quite stable in this time interval. The mean profiles obtained for the three different classes (clear-sky, ice cloud and mixed-phase cloud), classified by the CIC algorithm (applied to the ground measurements), were calculated for the two macro seasons and reported in *Fig.5.5.* In the cold season, spanning from

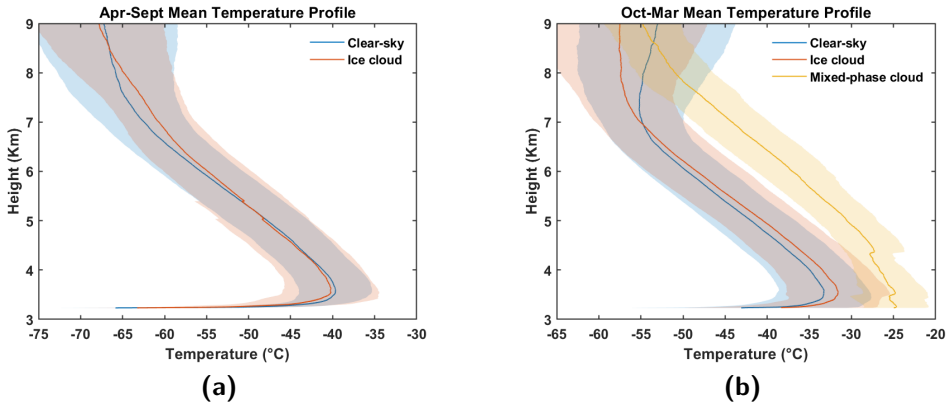


Figure 5.5: Mean Temperature profiles (solid lines) for the cold season, between April and September (a) and the warm season, between October and March (b) with the respective standard deviation (shaded areas), divided according to the sky scene classified by the CIC algorithm over the REFIR-PAD collocated measurements.

April to September, the temperature inversion is evident (*Fig.5.5a*). Considering a general cloud height around 4 km, as reported in Chapter 1, the mean cloud top temperatures are higher than those near the surface, even within the range of variability given by the standard deviation. For the same period, the mean surface temperature measured at the Concordia Station is -66.3°C for clear-sky scenes and -62.9°C in presence of an ice cloud. These results are consistent with the mean spectra obtained for the same season according to both the satellite and the ground classification (*Fig.5.3*). During the warm season, from October to March, the temperature inversion is less obvious and almost negligible in presence of mixed-phase clouds. The

mean temperatures measured at the surface are -44.7°C, -40.2°C and -28.4°C respectively in clear, ice cloud and mixed-phase cloud conditions. These values confirm the temperatures seen in the spectra grouped in accordance with the associated class of the REFIR-PAD observations.

5.3 AVHRR Scene Homogeneity

The differences in the scene observed by the satellite and the ground instrument may also depend on the sensor's field of view. If a cloud is detected from the ground, it has to be seen from the satellite too. However, the cloud phase detected might differ in case of mixed-phase clouds since liquid droplets are generally located in the lower layers. While, if the ground instrument sees clear sky, a cloud may still be present in the satellite FOV. The collocated Advanced Very High-Resolution Radiometer (AVHRR) has been used to determine the homogeneity of the scene within the IASI pixel. The AVHRR is a six-channel scanning radiometer with a spatial resolution of 1.1 km (at nadir) and provided of six spectral channels between 0.63 and 12.00 μm, three solar channels in the visible and near-infrared region and three thermal infrared channels. In this analysis only the infrared channels 4 and 5 will be employed, corresponding to 10.8 and 12 μm respectively. The IASI L1C products contain the results of the cluster analysis applied to the AVHRR collocated observations. The AVHRR pixels are clustered into homogeneous classes in the radiance space, using the K-mean classification algorithm. For each class and each AVHRR channel, the cluster product provides the coverage percentage of the class within the IASI pixel (*IDefCcsRadAnalWgt* product), the mean (*GCcsRadAnalMean*) and the standard deviation (*GCcsRadAnalStd*) of AVHRR brightness temperatures within the class. The latter two are in units of $W/m^2 \times sr \times m^{-1}$ for the infrared channels. Different methods have been suggested to perform a homogeneity analysis using this information. The following study is based on the work of *Farouk et al. (2019)*. An IASI FOV with several classes, each one having a small standard deviation and a mean radiance close to the one of the other classes, can be more homogeneous than a FOV with a single class but with a very large value of standard deviation. For this reason, the number of AVHRR clusters within each IASI pixel has not been used as a homogeneity criterion. The overall AVHRR cluster statistic is instead calculated, aggregating the information provided by all clusters in the IASI FOV.

Intercluster homogeneity

The intercluster homogeneity describes how much the single mean values of each class depart from the mean radiance in the pixel. It is defined independently for each channel as:

$$\sigma_{inter} = \sqrt{\frac{1}{\sum C_i} \sum_{i=1}^{N} C_i (L_i - L_{mean})^2} \tag{5.3}$$

where N is the number of clusters, L_i is the mean radiance for the cluster i in the channel under consideration and L_{mean} is the radiance weighted average. The weights are determined by C_i, which are the cluster fractions of each class i covering the IASI pixel. This result is interpreted in terms of homogeneity, i.e. a small σ_{inter} means that all the classes observe the same scene in that channel.

Intracluster homogeneity

This quantity is used to determine the homogeneity within each class. It is defined according to the following formula:

$$\sigma_{intra} = \sqrt{\frac{1}{\sum C_i} \sum_{i=1}^{N} C_i \sigma_i^2} \tag{5.4}$$

where σ_i is the standard deviation of the cluster i, provided in the IASI product.

Criteria

Two different thresholds have to be defined for the intercluster and intracluster deviations normalised over the mean radiance L_{mean}. Values suggested by *Farouk et al. (2019)* are too high for this study, probably because surface temperatures in presence of a cloud are not so distant from the clear-sky ones in Antarctica. An attempt to set calibrated thresholds is performed by assessing the normalised values of σ_{inter} and σ_{intra} against the cloud fraction provided by the AVHRR and included in the IASI L1c product. *Fig.5.6* shows results for channel 4 (10.8 μm). The deviation values are very small for both quantities, remaining below 0.04 for σ_{intra} and 0.06 for σ_{inter}. Moreover, there is no explicit relation between those quantities and the cloud fraction provided by the AVHRR.

Possible thresholds are set as:

- $\sigma_{intra}/L_{mean} = 0.010$
- $\sigma_{inter}/L_{mean} = 0.015$

where values smaller than those should represent homogeneous scenes. According to these parameters, the IASI collocated pixels were divided into homogeneous and non-homogeneous scenes. Results are shown in *Fig.5.7*. As expected, the pixels identified as non-homogeneous span within all cloud fractions and a significant part of them corresponds to cloud fraction equal to 100%. Thus, this analysis cannot be considered reliable in this case and those results are not utilised in the definition of scene observed by the IASI instrument.

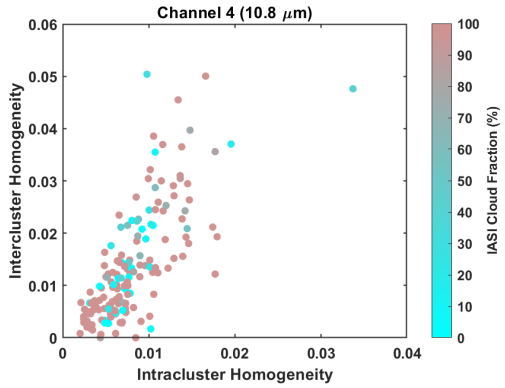


Figure 5.6: Scatter plot between σ_{intra} and σ_{inter} normalised over the mean radiance L_{mean}, calculated for the collocated IASI pixels. The colour scale represents the cloud fraction measured by the AVHRR.

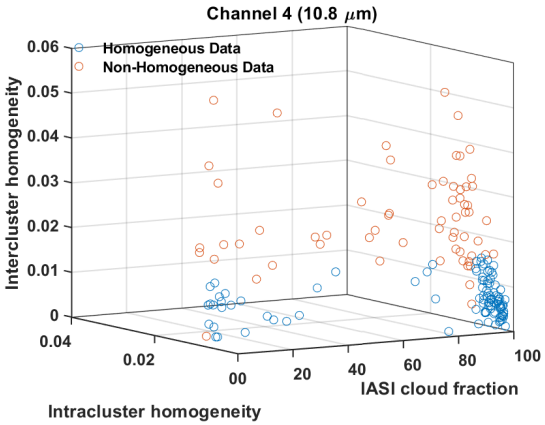


Figure 5.7: 3D scatter plot between the AVHRR cloud fraction, σ_{intra} and σ_{inter} normalised over the mean radiance L_{mean} for the collocated IASI pixels. Orange circles represent non-homogeneous data, while blue the homogeneous ones.

5.4 CIC applied to IASI Dataset

The IASI dataset is now used as input for the CIC algorithm. The 167 total collocated elements are divided into a Training set, a Validation set and a Test set and the scenes are labelled according to the REFIR-PAD observation. There are 114 clear-sky scenes and 50 cloudy ones, split into 44 ice clouds and 6 mixed-phase clouds.

5.4.1 Training Set

Due to the low number of collocated observations, only two classes are defined for the whole four years period, clear-sky and cloudy. The second includes all the cloudy scenes, without distinguishing the cloud phase. CIC results are sensitive to the composition of the Training Set, thus its elements should represent the entire variability within each class and characterise both the cold and the warm season. 50 spectra were manually chosen to populate the Training Sets, 25 clear and 25 cloud (22 ice clouds and 3 mixed-phase clouds), visually inspecting the variability within each month. Both thick and thin clouds were included in the cloud class. The monthly distribution of the selected observations follows the one of the entire dataset, as illustrated in *Fig.5.8*.

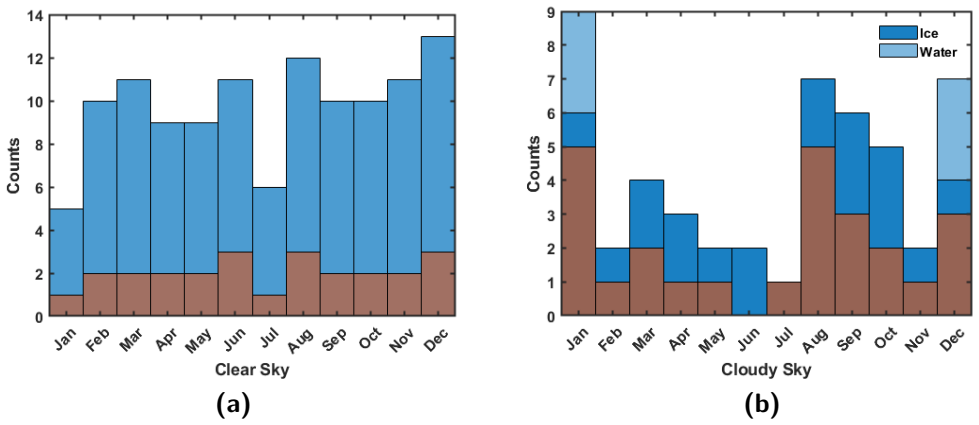


Figure 5.8: Monthly distribution of collocated IASI spectra in clear-sky (a) and cloudy (b) conditions. The total number is in blue, while the brown spectra are the ones selected for the Training Sets.

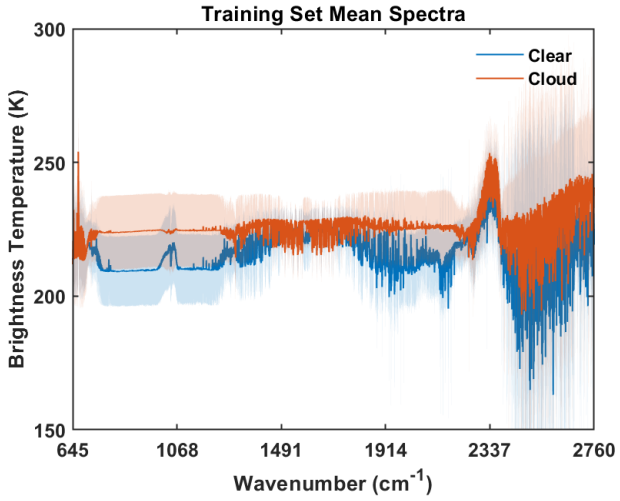


Figure 5.9: Mean clear and cloudy spectra forming the Training Set.

The mean spectra forming the two Training sets are depicted in *Fig.5.9*. Some different spectral features are evident in the figure and characterise the two classes, such as the O_3 band and wavenumbers around 2000-2200 cm^{-1}. While the standard deviations (shaded areas) indicate the variability within each class. Those spectra were then ingested by the CIC, following the distributional approach described in Chapter 3.2.2. For each spectrum, the Similarity index (SI) clear and cloud have been calculated, together with their difference

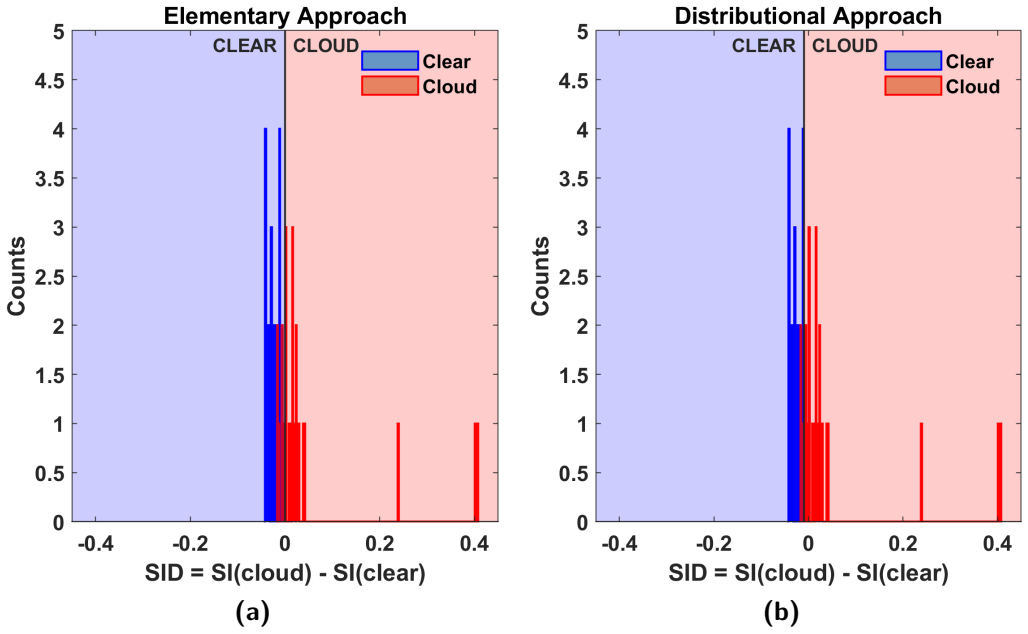


Figure 5.10: SID distributions for the Training Set elements, using the elementary (a) and the distributional approach (b), over the spectra interval 645-2250 cm^{-1} and using the minimum number of principal components.

(SID). Results obtained performing the classification over the spectral interval 645-2250 cm^{-1} (larger wavenumbers were excluded because are affected by the solar contribution) are shown in *Fig.5.10a*. A positive SID value indicates that the spectrum has been classified as cloud, while a negative one denotes a clear-sky spectrum. The colour of each bar represents the belonging class. This procedure is needed in order to find the best delimiter between the SID distributions. The shift that maximises the classification results is found at -0.0093 and is depicted in *Fig.5.10b*. Clear-sky elements (blue histogram) were classified with very similar SID values, generating a narrow distribution close to 0. On the other hand, the cloudy spectra (red histogram) span over a larger, even if limited, interval of SID values. Distinct groups suggest that different features emerged from the classification, producing slight variations in the SID values.

5.4.2 Validation Set

In machine learning, the validation set is used to adjust the parameters on a set of independent data, while the test set provides an unbiased evaluation of the final model. Once the Training sets have been analysed and the best delimiter has been calculated, the number of principal components, as well as the best spectral interval performing the classification, have to be defined. 36 spectra have been chosen to compose the Validation set, maintaining a sufficient representation for each month and year. 24 of which are clear-sky spectra and 12 cloudy (11 ice clouds and 1 mixed-phase cloud). Multiple runs of the CIC algorithm were performed on these elements by applying it to different spectral intervals. The end wavenumber was moved from 1400 to 2250, while the number of principal components was set equal to the minimum value between the two classes (which generally is around 2 or 3), the maximum (around 4-5) and fixed at 10 and 15. The maximum number of PCs allows to retain all the information in one class and add some noise to the other one. While a minimum number of PCs cuts part of the information in the class with the highest number of PCs. The algorithm performance is assessed by evaluating the Threat Score and the Hit rate, defined previously in *Eq.4.1* and *Eq.4.2*. Results are presented in *Fig.5.11*.
The Threat Score index shows that a number of principal components equal to 15 is stable for both classes, with a slight change in the cloud class over the spectral interval 645-2250 cm^{-1}. Looking at the clear-sky class, the number of PCs less stable is 10, while the minimum and maximum follow almost the same trend, with the minimum having a constant higher score, except in the spectral interval 645-1400 cm^{-1}. As for the cloud class, the scores are overall much lower, with a peak just above 0.4 when the end wavenumber exceeds

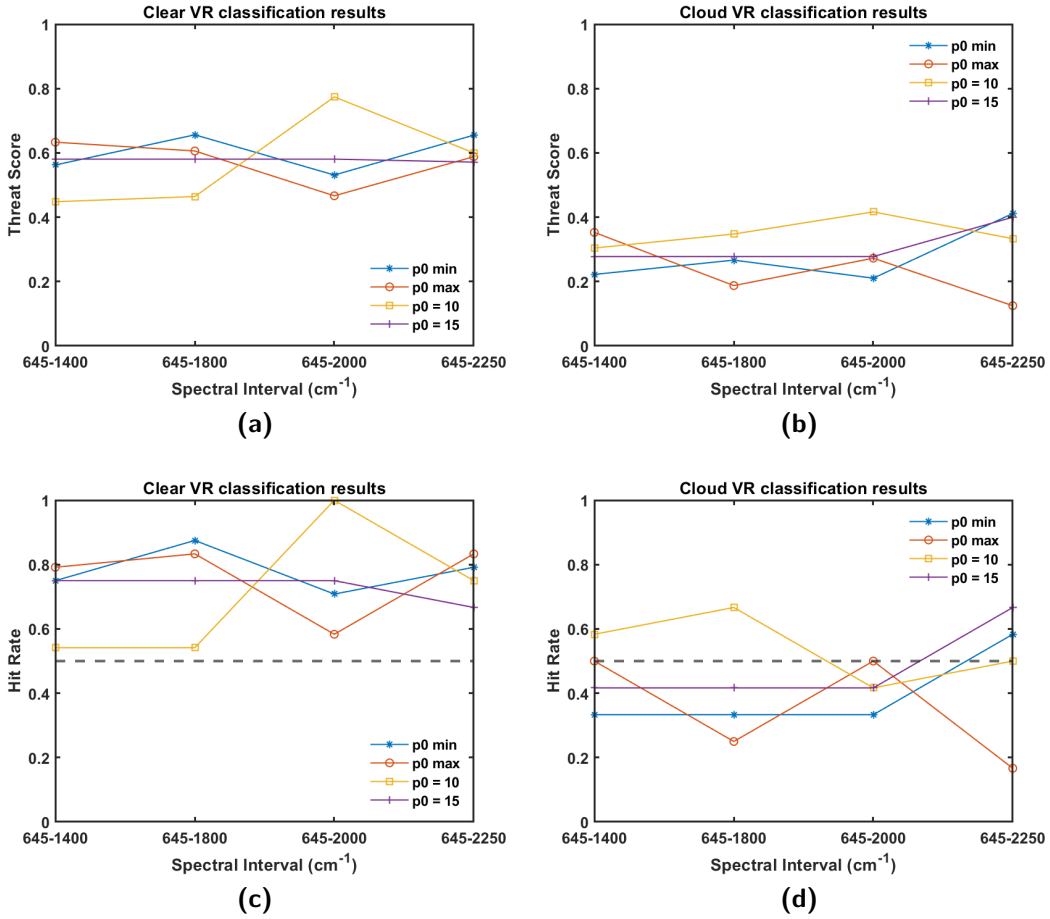


Figure 5.11: Classification results obtained on the Validation Set for different spectral intervals and number of principal components. Two indices are used: the Threat Scores clear (a) and cloud (b), and the Hit rate clear (c) and cloud (d).

2000 cm^{-1}. Moreover, there is no evident trend among the different PCs. Threat Scores are affected by the number of false positives and quantify the goodness of the classification within the class, while Hit Rate values (at the bottom of the figure) account only for the spectra correctly classified over the total number of spectra populating each class. Thus, HRs are directly linked to the percentages of the classification results. *Fig.5.11c* shows that using 10 PCs over the spectral interval 645-2000 cm^{-1} yields a perfect score in the clear-sky class, but drops to almost 40% in the cloud one, classifying correctly less than half cloud spectra. Overall, it appears that, when the number of PCs is fixed to 10 or 15, the score rises on the cloud class when they worsen on the clear spectra. and there is no spectral interval that gives a satisfactory

result in both classes for these configurations. A threshold of 0.5 on the Hit Rate has been defined in both classes in order to rule out configurations with a completely random classification or more chances to misclassify the elements.

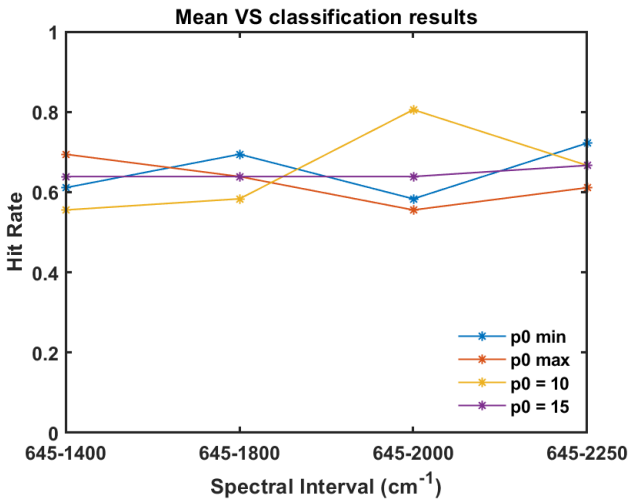
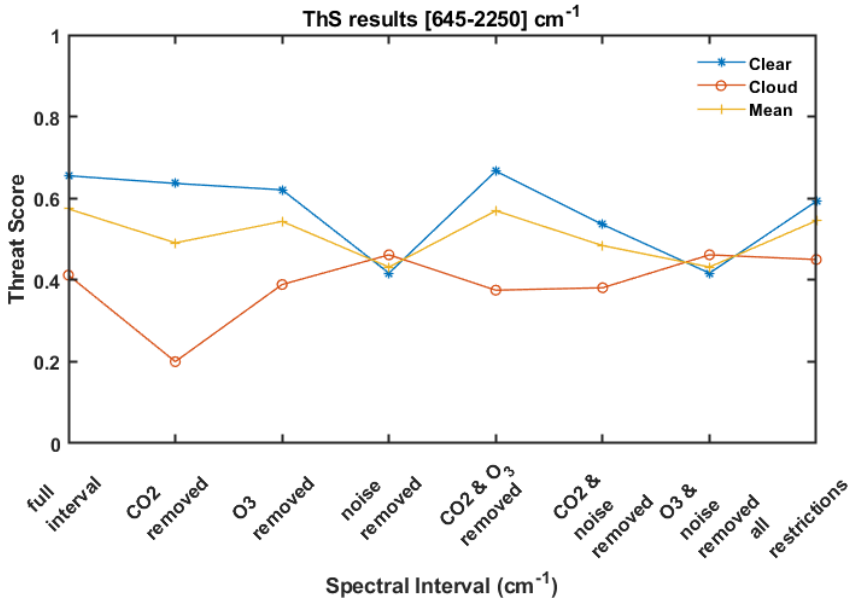
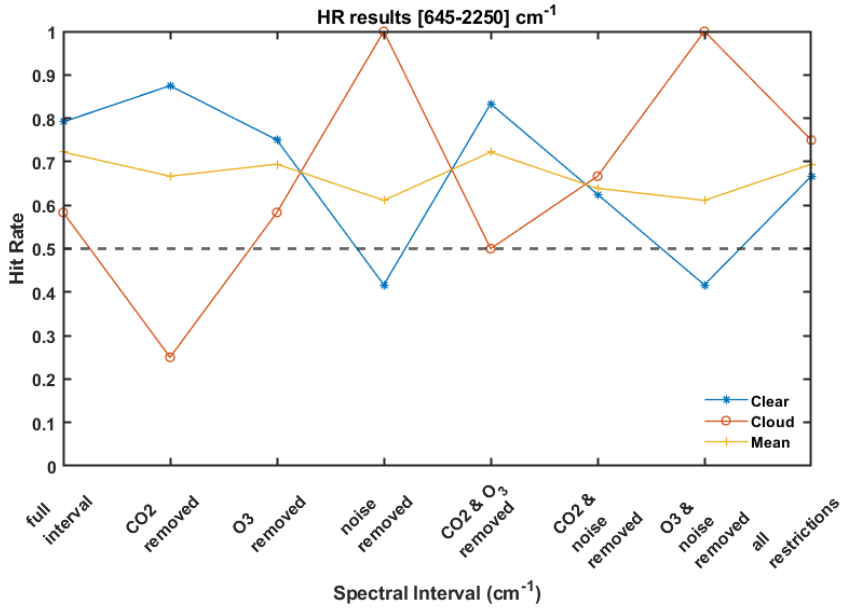


Figure 5.12: Classification results obtained on the Validation Set for different spectral intervals and number of principal components. The mean value of the Hit Rate is shown here.

The best spectral interval and number of PCs are chosen among the configurations passing the filter, looking at the weighted averages of the classes. The two indexes (ThS and HR) follow the same trend on the mean values, thus the Hit Rate (in *Fig.5.12*) has been used to assess the maximum classification results, to be consistent with the previous filter. The best performing configuration is given by the full spectral interval, 645-2250 cm^{-1}, and the minimum number of principal components. Further analyses have then been performed on the selected interval. In particular, the absorption bands of different gases and their correlation to the presence of clouds have been taken into account. The CIC was run on the Validation set spectra removing the wavenumbers correspondent to the CO_2 absorption (645-700 cm^{-1}), the O_3 absorption (1000-1100 cm^{-1}) and a combination of the above. Moreover, in the two spectral regions, 1145–1190 cm^{-1} and 1925–1980 cm^{-1}, the measurement quality decreases because the measurement noise increases at the edge of the two spectral bands (*EUMETSAT (2019)*). These channels are denoted as "noise" and removed from the full interval. *Fig.5.13* illustrates the results obtained for the two classes and their weighted mean. The usual threshold of HR $>$ 50% has been set in *Fig.5.13b*. As mentioned above, the mean trends of the two indices are consistent. The Threat Score presents higher values for the clear-sky class in every configuration, except when the noisy channels are removed from the full interval or from the full interval without the 1000-1100 cm^{-1} band. On the other side, the Hit Rate does not exhibit the same behaviour. This is probably due to the fact that the clear-sky class has twice the number of elements of the cloud class, which results in a small number of



(a)



(b)

Figure 5.13: Classification results obtained on the Validation Set for the full spectral interval between 645-2250 cm^{-1} and removing various spectral bands. The minimum number of principal components has been used. The Threat Score is depicted in (a) and the Hit rate in (b).

false positives at the denominator of the ThS_{clear}. The percentage of cloudy spectra correctly classified drops when the CO_2 band is removed, while it remains almost the same without the O_3 channels and rises to 100% eliminating the noisy channels. The opposite trend is observed for the clear-sky class. Keeping all the restrictions the mean score is close to the original one, however, the cloud class is better resolved and both classes have Hit Rates above 60%. Thus the selected spectra interval is [645-2250]cm^{-1}, excluding the intervals [645-700], [1000-1100], [1145-1190] and [1925-1980] cm^{-1}.

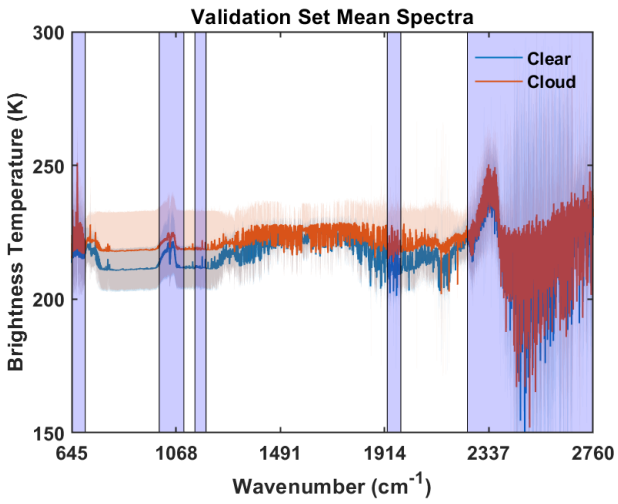


Figure 5.14: Mean clear and cloudy spectra forming the Verification Set. Shaded areas correspond to the excluded spectral intervals.

The mean spectra composing the Validation Sets are presented in *Fig.5.14*, divided in the two classes. The shaded areas in light blue represent the excluded wavenumbers. It can be noted that spectra resemble quite well the Training set in the selected channels (see *Fig.5.9*). The distinct spectral features appearing in the Training Sets around 2000 cm^{-1} are not found in the Validation spectra. This can be due to the particular elements chosen or to the fact that, using the minimum number of PCs, temperature becomes the main feature used by the CIC to separate the clear and cloudy spectra.

5.4.3 Test Set

The remaining 81 spectra compose the Test Set, divided into 68 clear-sky and 13 cloudy spectra. These data were ingested by the CIC algorithm, using the minimum number of PCs and the spectral interval 645-2250 cm^{-1}, without the absorption and noisy bands defined in the previous paragraph. The classification was performed by applying the shift previously found (-0.0093) through the SID distributions of the Training Set. The final scores are reported in *Tab.5.6* for the two classes (clear and cloud) and their weighted average. Although both classes have a Hit Rate higher than 50%, in the

	Threat Score	Hit Rate
Clear	0.53	0.56
Cloud	0.21	0.69
Mean	0.48	0.58

Table 5.6: Threat Scores and Hit Rate over the Test Set for the clear and cloud class and their weighted mean.

clear class it is unexpectedly lower that the value obtained for the Validation Set shown in *Fig.5.11c*. A possible explanation can be found in the different instruments' field of view and thus on the scene observed by the sensors. The homogeneity within the IASI field of view was not assessed due to the inconclusive results obtained from the AVHRR analysis. The satellite could have measured non-homogeneous scenes, coincidently with the clear-sky observations of the ground instrument. The CIC is sensitive even to small variations in the radiance due to faint cloud contamination. To address this issue, a simple analysis was performed using the wind data collected from the nearby Concordia station. A radiosonde is launched every day at 12 UTC, providing measurements of wind speed and direction along the vertical. For each clear-sky observation, the mean wind speed measured on the same day at about 4 km was derived. Considering the diameter of the IASI field of view equal to 12 km and the wind speed measured by the radiosonde, a maximum time interval was calculated and all the REFIR-PAD observations falling in that interval were analysed. If a cloud was detected in at least one observation, the IASI pixel was flagged as cloudy. Where velocities were too high, the time interval was lower than the REFIR-PAD time resolution (14 min) and there were no other observations to consider. Thus, a maximum wind speed of 20 km/h was imposed to include at least one observation before and after.

	Threat Score	Hit Rate
Clear	0.49	0.63
Cloud	0.42	0.63
Mean	0.46	0.63

Table 5.7: Threat Scores and Hit Rate over the Test Set for the clear and cloud class and their weighted mean. Scores refer to the REFIR-PAD observations over a longer time interval.

As a result, 22 clear-sky spectra were flagged as cloudy and all of them are part of the Test Set. The Training and Validation sets do not contain any "uncertain" observations. The Test Set is now composed of 46 clear-sky and 35 cloudy scenes, resulting in the scores defined in *Tab.5.7*. Results for the clear-sky class have improved from 56% to 63%. However, the performances on the cloud class have decreased from 69% to 63%. The opposite is true for the Threat Score because their values reflect also the number of misclassified spectra within the class.
A more thorough examination should consider the wind speed measured at the actual time of the IASI observation and at the proper cloud height. Probably there are cases in which the true time interval was shorter, therefore not all the REFIR-PAD observations considered fall within the IASI's field of view. The previous analysis might have included some fictitious cloudy scenes.
Removing those unknown cases, the results become (*Tab.5.8*) consistent with those obtained for the Validation Set. The total number of spectra considered now is 59, 46 of which are clear-sky and 13 cloudy spectra. The cloud class has a Threat Score of 0.3, due to the fact that 9 spectra are correctly classified by the algorithm (HR=69%). However, 37% of clear spectra are misclassified, corresponding to 17 elements counted as false positives, more than the total number of cloud spectra itself.

	Threat Score	**Hit Rate**
Clear	0.58	0.63
Cloud	0.30	0.69
Mean	0.52	0.64

Table 5.8: Threat Scores and Hit Rate over the Test Set for the clear and cloud class and their weighted mean. The unknown scenes were removed from the Test Set.

5.4.4 Results

Due to the low number of collocate measurements available, results presented in this paragraph refer to spectra belonging to both the Validation and the Test Set. The 22 "unknown" cases found before are analysed separately and are not included in the final statistic.

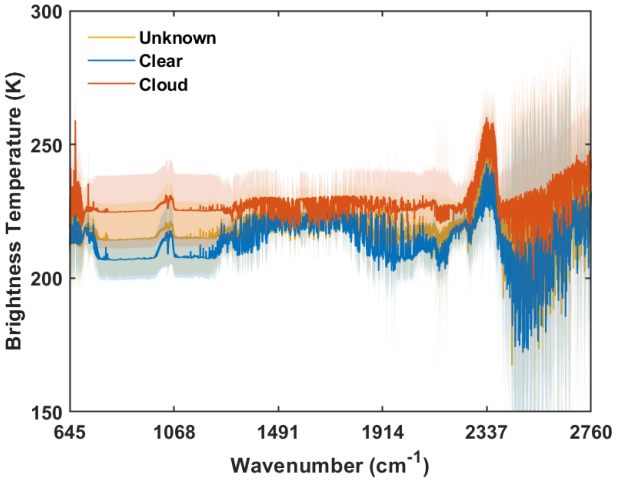


Figure 5.15: Mean clear and cloudy spectra forming the Verification and Test Set. Spectra whose scene observed is unknown are in yellow.

Fig.5.15 shows the mean spectra labelled according to the CIC classification and their standard deviation. The algorithm clearly separates the two classes based on the brightness temperature and it is also visible the spectral feature around 2000 cm^{-1}, recognised in the Training Sets. This result points out that using the minimum number of principal components allows to retain enough information. Cloudy spectra have a larger variability and there is only a small fraction of BTs overlapping the other class. The standard deviation of the clear-sky spectra is much smaller than the one obtained in the Training Set. Probably, it is because the CIC interprets the high BTs as cloud signals. *Fig.5.15* illustrates separately the "unknown" spectra. Their mean value is closer to the clear-sky brightness temperatures, though their large variability is an indicator of the non-homogeneity of the scenes observed, which makes their classification difficult. *Tab.5.9* provides a comparison one-to-one between the CIC classification and the REFIR-PAD observations. Matching scenes are found more than 65% of the time. 64% (45) of ground measurements of clear-sky and 72% (18) of cloudy are correctly identified by the CIC algorithm.

		REFIR-PAD	
		Clear	**Cloud**
IASI	**Clear**	45	7
	Cloud	25	18

Table 5.9: Matrix comparing the number of clouds detected by the REFIR-PAD and the IASI instruments using the CIC algorithm over the Verification and Test set. Unknown scenes are excluded.

Algorithm	Clear	Cloud	Unknown
REFIR-PAD	92 (78.63%)	25 (21.37%)	-
IASI	30 (25.64%)	87 (74.36%)	-
IASI-CIC	52 (44.44%)	43 (36.75%)	22 (18.80%)

Table 5.10: Comparison of the results obtained from the REFIR-PAD observations, IASI cloud detection and the CIC algorithm applied to IASI data (IASI-CIC) in percentages.

Statistical results according to the REFIR-PAD scene, the IASI cloud detection algorithm (provided in the L2 products) and the CIC applied to IASI spectra are presented in *Tab.5.10*. Overall, the CIC algorithm improves the cloud detection from satellite. The cloud occurrence decreases from 74.36%, as measured by IASI, to 36.75% with the CIC classification, which is closer to the 21.37% provided by the ground measurements. The percentage of clear-sky scenes increases by almost 20%, while the unknowns represent 18.8% of the total. A better assessment of the satellite field of view can help improve these results. Finally in *Fig.5.16* is illustrated the annual cloud occurrence obtained from the CIC algorithm, compared to the REFIR-PAD observations and IASI L2 products. IASI algorithms are not very reliable in polar regions, in fact, in the first two years all the scenes were flagged as cloudy. Results improve over 2014-2015, though they are still distant from the values provided by the ground measurements. On the other hand, all the percentages obtained with the CIC algorithm fall in the shaded area, which represents the annual variability, derived by calculating the maximum and minimum cloud occurrence measured by the REFIR-PAD between 2012

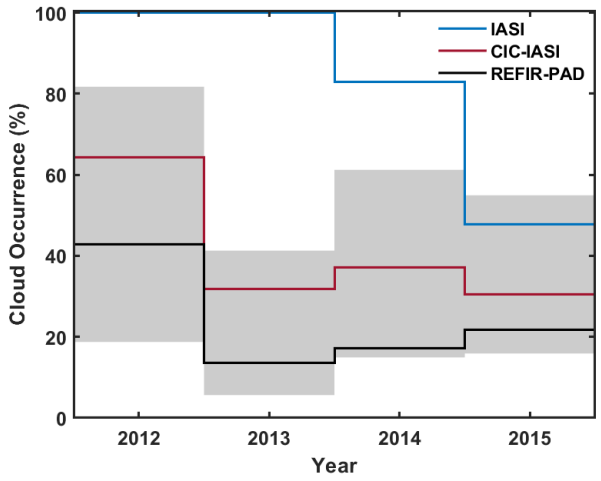


Figure 5.16: Annual cloud occurrence obtained applying the CIC algorithm to the Verification and Test Sets, compared to the REFIR-PAD and the IASI L2 products. Unknown cases are excluded. The shaded area represents the max and min variability over the entire REFIR-PAD dataset.

and 2015. Moreover, for the first three years, results follow the same trend provided by the ground instrument, while in 2015 the cloud occurrence drops to 30.43%, the lowest value recorded in the period.

5.5 MODerate-resolution Imaging Spectroradiometer (MODIS)

The MODerate resolution Imaging Spectroradiometer (MODIS) is a key instrument of the Earth Observing System (EOS) aboard the Terra (launched by NASA in 1999) and Aqua satellites (launched in 2002). MODIS provides global observations of Earth's land, oceans, and atmosphere every 1 to 2 days, acquiring data in 36 spectral bands in the visible and infrared regions (from 0.4 to 14.5 μm) (*MODIS Web*). In particular MODIS measures radiances in two visible bands at 250 m spatial resolution, five more visible bands at 500 m resolution, and the remaining 29 visible and infrared bands at 1000 m resolution (more specifics are provided in *Tab.5.11*).

Characteristics	Specifications
Orbit	705 km, 10:30 a.m. descending node (Terra) or 1:30 p.m. ascending node (Aqua), sun-synchronous, near-polar, circular
Scan rate	20.3 rpm, cross track
Swath	2330 km (cross track) by 10 km (along track at nadir)
Spatial resolution (nadir)	250 m (bands 1–2) 500 m (bands 3–7) 1000 m (bands 8–36)
Temporal resolution	1–2 days
Bands	36 spectral bands (490 detectors), cover wavelength range from 0.4 to 14.5 μm

Table 5.11: MODIS characteristics (*MODIS Web*).

5.5.1 Cloud Products

There are two MODIS Level 2 products that can be used for cloud detection: a Cloud Product in MOD06 and MYD06 (containing data collected from Terra and Aqua platforms respectively) and a Cloud Mask in MOD35 and MYD35.
The MODIS Cloud Product (M*D06) combines infrared and visible techniques to determine both physical and radiative cloud properties. The visible and near-infrared channels are used to derive cloud-particle phase, effective cloud-particle radius, cloud optical thickness and an indication of cloud shadows affecting the scene. Cloud-top temperature, height, effective emissivity, phase, and cloud fraction are produced by the infrared retrieval methods

both day and night at 1km resolution.
The MODIS cloud mask (M*D35) provides an estimate that a given MODIS field of view (FOV) is cloud-free at 1 km and 250 m spatial resolutions (at nadir). The 250-m cloud-mask flags are based on the visible channel data only. An indication of shadows affecting the scene is also provided. The algorithm employs a series of visible and infrared thresholds and consistency tests to specify the confidence level of cloud contamination inside the pixel. 14 spectral bands and 11 individual spectral tests are combined to give a cloud mask for each pixel and a level of confidence as: confident clear, probably clear, undecided, obstructed/cloudy. Each test returns a confidence ranging from 0 to 1. Similar tests are grouped together and the minimum confidence value is selected. More details about the type of tests and the procedure followed are described by *Ackerman et al. (2010)* and summarized in Chapter 1.

5.5.2 Statistical Analysis

Data containing the cloud mask and the cloud phase observed by MODIS from 2012 to 2020 were provided by the University of Wisconsin-Madison. All the MODIS fields of view, containing Dome-C, were used to produce the annual and monthly cloud occurrence. Data were filtered according to:

- a maximum distance of 1 km of MODIS pixel centre from Dome-C
- a satellite zenith angle below 8° (only nadir observations)

There are 2052 collocated observations, divided over the years as reported in *Tab.5.12*. Data are well distributed, with around 20 measurements every month of each year.

Year	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Num. spectra	194	228	232	233	231	235	229	237	233

Table 5.12: Number of MODIS elements collected over Dome-C in the years 2012-2020.

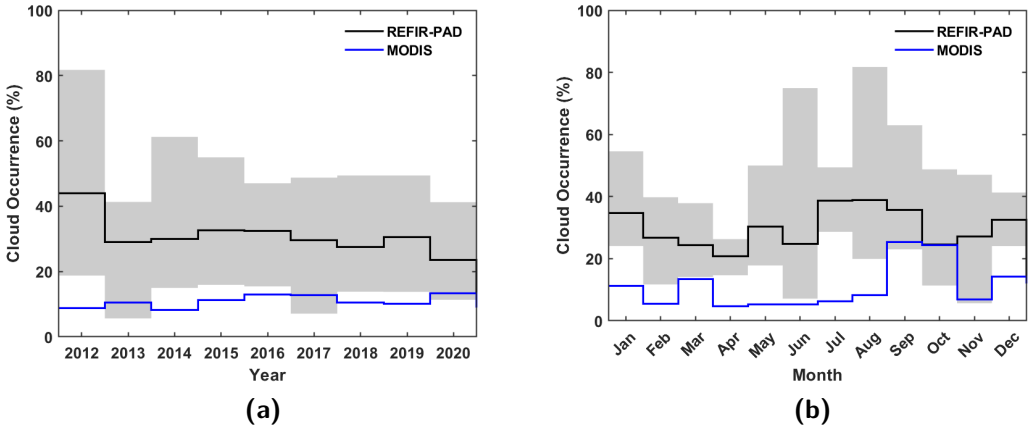


Figure 5.17: Comparison of MODIS mean annual (a) and monthly (b) cloud occurrence (blue lines) with the values provided by the CIC (black) over the entire dataset. Shaded areas represent the variability observed by the ground instrument between the maximum and minimum values.

Fig.5.17 shows the mean annual and monthly cloud occurrence detected by MODIS, compared to the ground-based observations obtained by the REFIR-PAD, described in Chapter 4. Values account for both ice clouds and mixed-phase cloud occurrences. Moreover, MODIS cloud phase retrieved using the visible and NIR channels was substituted with the infrared retrievals for the days with no solar radiation. The mean value for the entire dataset is 10.92%. The annual cloud occurrence (in *Fig.5.17a*) is quite stable throughout all nine years, however, does not follow the same trend as the ground instrument. The most cloudy year observed by MODIS is 2020, with a cloud percentage of 13.30%. The value falls in the interval of variability, though the same year is considered the least cloudy by the REFIR-PAD. The minimum is instead recorded in 2014 at 8.19%. Overall, the cloud occurrence remains below the values observed by the REFIR-PAD, with only 2013, 2017 and 2020 included in the shaded area. This is probably due to the low minimum value recorded in 2013 and 2017. 2017 and 2020 were also the least cloudy year in winter (JJA) and summer (DJF) respectively, according to the ground observations. *Fig.5.17b* illustrates the comparison of the monthly cloud occurrence detected by MODIS and the values obtained by the REFIR-PAD. In September, the difference between the two cloud occurrences is just 10%, although this month is the most cloudy according to MODIS with a value of 25.29%, in contrast with the result obtained from the REFIR-PAD, which placed the maximum in August. On the other hand, the minimum is recorded in April at 4.71% and is consistent with the results

shown by the REFIR-PAD. A stable low cloud occurrence is measured by MODIS in winter (JJA). This is due to the fact that the cloud phase is retrieved using the visible channels, however, when these are not available (no sunlight) the infrared channels are employed and the cloud mask becomes less efficient. As a consequence, the maximum observed from the ground in August is not detected.

The observations were then temporally collocated with REFIR-PAD measurements and analysed in terms of a one-to-one comparison. In particular, each satellite observation has to be performed within 15 minutes of a REFIR-PAD measurement to be considered collocated. *Tab.5.13* reports results in a confusion matrix. The total number of collocated measurements is 1118. The majority of observations are classified as clear-sky by both sensors and account for 68.52% of the total REFIR-PAD observations (766 measurements) and 89.62% of MODIS (1002 measurements). Moreover, almost 95% of clear-sky observations collected from the ground are correctly classified. The satellite detects a cloud in 9.92% of the cases (111 observations), much less than the 31.04% obtained from the ground. 79.54% of clouds seen from the ground are labelled as clear-sky scenes, 30.43% of which in winter (JJA). They both detect a cloud 20.46% of the cases, while the two instruments are in accordance with its thermodynamic phase 18.73% of the time. This result is reasonable considering that they are observing different layers of the cloud and the phase can change from the bottom to the top. The MODIS instrument can be considered quite reliable, in fact, there is a perfect match in almost 70.75% of the cases.

		REFIR-PAD		
		Clear-sky	**Ice cloud**	**Mixed-phase cloud**
MODIS	**Clear-sky**	726	273	3
	Ice cloud	40	63	5
	Mixed-phase cloud	0	1	2

Table 5.13: Matrix comparing the number of clouds detected by the REFIR-PAD and the MODIS instruments for the entire collocated dataset of 1118 observations (unclassified observations are not reported here).

MYD06

Finally, MYD06 products, collected from the Aqua satellite, have been analysed for the years 2012-2015. *Tab.5.14* shows the results obtained retaining only spatially collocated observations. A total of 40 satellite passes have been analysed and percentages have been calculated for each class. MODIS has a clear-sky occurrence 17.5% higher than the REFIR-PAD, while the number of clouds is less than half. This outcome is probably due to the problems encountered in winter. A simple comparison with IASI is obtained enlarging the area covered by the satellite field of view to 12 km. Results are even more in accordance with the ground statistic, which demonstrates the reliability of MODIS in the study of cloud variability and confirms that IASI misclassification is not related to its larger field of view.

Total		**Clear-sky**	**Ice Cloud**	**Mixed-phase**	**Unclass**
40	**REFIR-PAD (CIC)**	28 (70%)	9 (22.5%)	3 (7.5%)	-
	MODIS AQUA	35 (87.5%)	4 (10%)	0	1 (2.5%)
	MODIS AQUA 12km	31 (77.5%)	8 (20%)	0	1 (2.5%)

Table 5.14: Comparison of the results obtained from the CIC algorithm applied to REFIR-PAD measurements and MODIS AQUA cloud product (MYD06) at 1km and 12km spatial resolution.

Summary and Conclusions

Clouds play a critical role in the Earth System, especially in Antarctica where their effect on regional climate variability has consequences all over the world. The role of polar clouds in global climate is still not fully understood and well-modelled. Cloud properties retrievals are mainly constrained by the lack of observations in polar regions. Accurate measurements of cloud properties (e.g. cloud type, cloud phase, cloud amount, cloud top height, optical thickness) from ground-based sensors are difficult due to the extreme environmental conditions and the scarcity of observation sites. Satellite measurements constitute a fundamental tool since they provide global coverage and daily cloud monitoring at high spatial resolution. Various techniques have been developed to detect and isolate cloud spectral features. Classical algorithms are based on a series of tests that exploit reflectance or brightness temperature spectral variations. However, the high dimensionality of data and the large cloud variability have led research toward statistical approaches and machine learning methods. These algorithms typically learn the features of the cloudy and clear-sky conditions from elements whose sky conditions are known and use them to infer the type of scene observed in new data. Among the most common techniques, there are Neural Networks, Support Vector Machine, Deep Learning, decision tree and logistic regression. Other algorithms are used to reduce the elevated amount of data. Principal component analysis (PCA) is the most common method. It reduces the dimension of the initial data, retaining only the components with higher variance and disregarding those with no physical information.
Nevertheless, satellite cloud products have different issues when retrieved from data collected in polar regions. For instance, MODIS cloud detection algorithms are based on a series of tests and employ shortwaves and NIR bands. When solar radiation is not available, the cloud mask is not very efficient. On the other hand, the very similar radiative properties of the surface and clouds and the frequent temperature inversions in Antarctica, make algorithms based on brightness temperature differences unreliable, as in the case of IASI. Active sensors also present certain challenges when used for

cloud detection in polar regions. For instance, the coarse vertical resolution of the CPR onboard of CloudSat (500 m) and its limited sensitivity near the surface does not allow accurate detection of low clouds.

An innovative machine-learning algorithm CIC (Cloud Identification and Classification) was recently developed. CIC allows the identification of the atmospheric scene observed (clear or cloudy) based only on the input high spectral resolution radiances, without the need for ancillary information. The algorithm is based on PCA analysis and classifies each input spectrum based on the changes in the amount of information contained in each training set. The metric used is based on the Similarity Index, which defines the level of closeness of each analysed spectrum and a specific class. CIC is primarily used to distinguish cloudy scenes from clear-sky ones, but it is also able to characterise the cloud phase.

In this thesis, CIC is initially tested against high spectral resolution downwelling radiances at far infrared (FIR) and middle infrared (MIR) wavenumbers, collected at Dome-C on the Antarctic Plateau, between 2012 and 2020. Training sets are defined by the inspection of backscatter and depolarization LiDAR profiles. Results on the Test Set show that, overall, 95% of spectra are correctly classified. A very positive result is obtained in the case of mixed-phase clouds, where the CIC is able to identify the presence of the cloud in 100% of the cases, while 14% of the time the cloud phase is classified as ice instead of mixed-phase. Values obtained for the PPVs (Positive Predictive Value) indicate that the clear class is composed of 99% clear-sky spectra, while the ice cloud and mixed-phase cloud are defined with 94% and 92% of correct elements respectively. Statistical analysis over the entire dataset reveals a mean cloud occurrence of 31.02% and a monthly maximum value in August at 38.78%. A positive cloud forcing is found correlating each clear-sky and cloudy scene with the surface temperature measured from the ground station. For instance, temperatures higher than around -30°C, in autumn and winter, are associated only with clouds. The largest difference between the mean temperature values of clear and cloudy scenes is found in winter at 8.65°C.

CIC algorithm is then applied to satellite data collected by IASI (Infrared Atmospheric Sounding Interferometer) over the years 2012-2015. Data were collocated with REFIR-PAD measurements, which have been used as references. The homogeneity of the scene observed has been assessed both with the help of the cluster analysis performed over the AVHRR pixels and with data on wind speed collected at Dome-C. Multiple runs of the CIC algorithm were performed on a subset of elements to find the most performing configuration. The end wavenumber was moved from 1400 to 2250 cm^{-1}, while the number of principal components was set equal to the minimum and maximum

values between the two classes and fixed at 10 and 15. Further analyses have then been performed removing the absorption bands of different gases and IASI noisy channels. Finally, the Test Set was classified using the best configuration, corresponding to the minimum number of PCs and the spectral interval 645-2250 cm^{-1} with all the above restrictions. A punctual comparison between the results obtained and the REFIR-PAD scenes observed shows 65% of matching scenes. In particular, 64% of ground measurements of clear-sky and 72% of cloudy are correctly identified by the CIC algorithm. The IASI cloud phase included in the L2 products was also compared with these results. The IASI L2 cloud product can be considered quite reliable when no clouds are present, in fact, 76.74% of the clear-sky scenes observed by the satellite match with ground observations. Although, the ice clouds detected by the satellite are actually seen by the REFIR-PAD only 68.59% of the time. Overall, the CIC algorithm improves cloud detection from satellite. The cloud occurrence decreases from 74.36%, as measured by IASI, to 36.75% with the CIC classification, which is closer to the 21.37% provided by the ground measurements. The annual cloud occurrence confirms the reliability of the CIC algorithm, in fact, the values obtained fall within the annual variability measured by the REFIR-PAD.

Finally, MODIS cloud products from 2012 to 2020 are analysed and compared to the REFIR-PAD observations. The mean value for the entire dataset is 10.92%. The most cloudy year observed by MODIS is 2020, with a cloud percentage of 13.30%, while the monthly means show a maximum in September at 25.29%. A stable low cloud occurrence is measured by MODIS in winter (JJA). Observations were then temporally collocated with REFIR-PAD measurements and analysed in terms of a one-to-one comparison. Almost 94.78% of clear-sky observations collected from the ground are correctly classified. MODIS instrument can be considered quite reliable, in fact, there is a perfect match in almost 70.75% of the cases. The most evident issue of MODIS cloud detection is due to the fact that the cloud phase is retrieved using the visible and NIR channels, however, when these are not available (no sunlight) the infrared channels are employed and the cloud mask becomes less efficient. As a matter of fact, the analysis showed that 79.54% of clouds are missed by the satellite and labelled as clear-sky scenes, 30.43% of which are in winter (JJA). MYD06 products, collected from the Aqua satellite, have then been used to compare the occurrence of the different scenes observed at 1km and at 12km of spatial resolution. Both results are consistent with the ground observations, which confirm that IASI misclassification is not related to its larger field of view.

In conclusion, these studies demonstrate the potentiality of the CIC algorithm in polar regions, improving the satellite cloud detection provided by the current L2 products.

Bibliography

Ackerman S., Frey R., Strabala K., Yinghui L., Gumley L., Baum B., Menzel P. Discriminating clear-sky from cloud with MODIS. Algorithm theoretical basis document (MOD35). 2010.

Ackerman S. A., Strabala K. I., Menzel W. P., Frey R. A., Moeller C. C., Gumley L. E. Discriminating clear sky from clouds with MODIS // J. Geophys. Res. 1998.

Adhikari L., Wang Z., Deng M. Seasonal variations of Antarctic clouds observed by CloudSat and CALIPSO satellite // J. Geophys. Res. 2012.

Ahmad A., Quegan S. Cloud Masking for Remotely Sensed Data Using Spectral and Principal Components Analysis // Engineering, Technology Applied Science Research. 2012.

Amato U., Lavanant L., Liuzzi G., Masiello G., Serio C., Stuhlmann R., Tjemkes S. A. Cloud mask via cumulative discriminant analysis applied to satellite infrared observations: scientific basis and initial evaluation // Atmos. Meas. Tech. 2014.

Bouillon M., Safieddine S., Hadji-Lazaro J., Whitburn S., Clarisse L., Doutriaux-Boucher M., Coppens D., August T., Jacquette E., Clerbaux C. Ten-Year Assessment of IASI Radiance and Temperature // Remote Sensing. 2020.

Bromwich D. H., Nicolas J. P., Hines K. M., Kay J. E., Key E. L., Lazzara M. A., Lubin D., McFarquhar G. M., Gorodetskaya I. V., Grosvenor D. P., Lachlan-Cope T., van Lipzig N. P. M. Tropospheric clouds in Antarctica // Reviews of Geophysics. 2012.

Clarisse L., Coheur P.-F., Prata F., Hadji-Lazaro J., Hurtman D., Clerbaux C. A unified approach to infrared aerosol remote sensing and type specification // Atmos. Chem. Phys. 2013.

Bibliography

Cossich W., Maestri T., Magurno D., Martinazzo M., Di Natale G., Palchetti L., Bianchini G., Del Guasta M. Ice and mixed-phase cloud statistics on the Antarctic Plateau // Atmos. Chem. Phys. 2021.

Di Natale G., Bianchini G., Del Guasta M., Ridolfi M., Maestri T., Cossich W., Magurno D., Palchetti L. Characterization of the Far Infrared Properties and Radiative Forcing of Antarctic Ice and Water Clouds Exploiting the Spectrometer-LiDAR Synergy // Remote Sensing. 2020.

ESA . Earth Explorer 9 Candidate Mission FORUM — Report for Mission Selection. 2019.

EUMETSAT . IASI Level 1 Product Format Specification. 2011.

EUMETSAT . IASI Level 2 Product Format Specification. 2017a.

EUMETSAT . IASI Level 2: Product Guide. 2017b.

EUMETSAT . IASI Level 1: Product Guide. 2019.

Farouk I., Fourrié N., Guidard V. Homogeneity criteria from AVHRR information within IASI pixels in a numerical weather prediction context // Atm. Meas. Tech. 2019.

Fyke J., Sergienko O., Löfverström M., Price S., Lenaerts J. T. M. An overview of interactions and feedbacks between ice sheets and the Earth system // Reviews of Geophysics. 2018.

García-Sobrino J., Serra-Sagristà J., Bartrina-Rapesta J. Hyperspectral IASI L1C Data Compression // Sensors. 2017.

Helm V., Humbert A., Miller H. Elevation and elevation change of Greenland and Antarctica derived from CryoSat-2 // The Cryosphere. 2014.

Hilton F., Armante R., August T., Barnet C., Bouchard A., Camy-Peyret C., Capelle V., Clarisse L., Clerbaux C., Coheur P.-F., Collard A., Crevoisier C., Dufour G., Edwards D., Faijan F., Fourrié N., Gambacorta A., Goldberg M., Guidard V., Hurtmans D., Illingworth S., Jacquinet-Husson N., Kerzenmacher T., Klaes D., Lavanant L., Masiello G., Matricardi M., McNally A., Newman S., Pavelin E., Payan S., Péquignot E., Peyridieu S., Phulpin T., Remedios J., Schlüssel P., Serio C., Strow L., Stubenrauch C., Taylor J., Tobin D., Wolf W., Zhou D. Hyperspectral Earth Observation From IASI: Five Years of Accomplishments // American Meteorological Society. 2012.

Huang H.-L., Antonelli P. Application of Principal Component Analysis to High-Resolution Infrared Measurement Compression and Retrieval // J. Appl. Meteor. 2001.

Kurihana T., Moyer E. J., Foster I. T. AICCA: AI-Driven Cloud Classification Atlas // Remote Sensing. 2022.

Lavanant L., Lee A. C. L. A global cloud detection scheme for high spectral resolution instruments // Proceeding of the fourteenth International TOVS Study Conference, Beijing, China. 2005.

Liou K.N. An introduction to atmospheric radiation // Academic Press. 2002.

MODIS Web NASA. `https://modis.gsfc.nasa.gov/about/`. Latest access: 21-02-2023.

Maestri T., Arosio C., Rizzi R., Palchetti L., Bianchini G., D. Guasta M. Antarctic Ice Cloud Identification and Properties Using Downwelling Spectral Radiance From 100 to 1400 cm^{-1} // J. Geophys. Res.-Atmos. 2019a.

Maestri T., Cossich W., Sbrolli I. Cloud identification and classification from high spectral resolution data in the far infrared and mid-infrared // Atmos. Meas. Tech. 2019b.

Magurno D., Cossich W., Maestri T., Bantges R., Brindley H., Fox S., Harlow C., Murray J., Pickering J., Warwick L., Oetjen H. Cirrus Cloud Identification from Airborne Far-Infrared and Mid-Infrared Spectra // Remote Sensing. 2020.

Mahajan S., Fataniya B. Cloud detection methodologies: variants and development—a review // Complex & Intelligent Systems. 2019.

Murino L., Amato U., Carfora M. F., Antoniadis A., Huang B., Menzel W. P., Serio C. Cloud Detection of MODIS Multispectral Images // Journal of Atmospheric and Oceanic Technology. 2014.

Palchetti L., Brindley H., Bantges R., Buehler S. A., Camy-Peyret C., Carli B., Cortesi U., Del Bianco S., Di Natale G., Dinelli B. M., Feldman D., Huang X. L., C.Labonnote L., Libois Q., Maestri T., Mlynczak M. G., Murray J. E., Oetjen H., Ridolfi M., Riese M., Russell J., Saunders R., Serio C. FORUM: unique far-infrared satellite observations to better understand how Earth radiates energy to space // American Meteorological Society. 2020.

Bibliography

Turner D. D., Knuteson R. O., Revercomb H. E. Noise Reduction of Atmospheric Emitted Radiance Interferometer (AERI) Observations Using Principal Component Analysis // J. Atmos Ocean. Tech. 2006.

Whitburn S., Clarisse L., Crapeau M, August T., Hultberg T., Coheur P. F., Clerbaux C. A CO_2-free cloud mask from IASI radiances for climate applications // Atmos. Meas. Tech. 2022.

Zhang Q., Yu Y., Zhang W., Luo T., Wang X. Cloud Detection from FY-4A's Geostationary Interferometric Infrared Sounder Using Machine Learning Approaches // Remote Sensing. 2019.