DEVELOPMENT AND VALIDATION OF 3-D CLOUD FIELDS USING DATA FUSION AND MACHINE LEARNING TECHNIQUES

A Dissertation
Presented to
The Academic Faculty

By

Manon Huguenin

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
Guggenheim School of Aerospace Engineering

Georgia Institute of Technology

December 2018

DEVELOPMENT AND VALIDATION OF 3-D CLOUD FIELDS USING DATA FUSION AND MACHINE LEARNING TECHNIQUES

Approved by:

Dr. Dimitri Mavris, Advisor
Guggenheim School of Aerospace Engineering
Georgia Institute of Technology

Dr. Olivia Pinon Fischer
Guggenheim School of Aerospace Engineering
Georgia Institute of Technology

Dr. Patrick Taylor
Climate Science Branch
NASA Langley Research Center

Date Approved: November 30, 2018

A ma Bonne-Maman, qui attendait avec impatience ma première publication!

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Dimitri Mavris, for his guidance and support throughout the completion of my degree. Thank you for giving me the opportunity to join ASDL, and to be part of many challenging yet very exciting projects. Working here and exploring so many different research areas has been an invaluable experience that I could not have gained elsewhere, and I am very grateful for it.

My gratitude also goes to Dr. Olivia Pinon-Fischer for her constant help and commitment to this work, and more widely for her guidance and advice in this process as well as regarding my future endeavors. For all your time, all the enhancements and quality you brought to this thesis, and for your valuable advice, *merci de tout coeur!*

I would also like to thank Dr. Patrick Taylor from NASA Langley for his guidance throughout this process, and for the time he invested in following the work on this topic at ASDL since 2017. Thanks for your advice and insightful comments that helped me complete this thesis, and especially regarding climate science which I had little knowledge of.

I would like to recognize Gabriel Achour and Domitille Commun, my team members for the 2018 *Cloud Modeling by Data Fusion* Grand Challenge, for their work on this project which laid the foundations for this thesis. I would also like to thank Chelsea Johnson for her guidance in the early steps of this work, and her help in defining its scope.

Many thanks to my family, who have always been supportive of my academic endeavors, as far back as I can remember. Thank you all for taking the time to understand this thesis topic, and more generally showing so much interest in my studies and career.

Lastly, I would like to thank all my Atlantan friends for their never-ending support throughout this past year and a half. Living this far away from my homeland has not always been easy, but I've had the luck to meet heartwarming people that truly made me feel like home. Special thanks to my friend Yann who encouraged me to apply to GeorgiaTech and

join ASDL, and to my friend Florence with whom I shared the MSc thesis experience from the start. I could not have gone this far without you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The impact of climate change is projected to significantly increase over the next decades. Consequently, gaining a better understanding of climate change and being able to accurately predict its effects are of the upmost importance. Climate change predictions are currently achieved using Global Climate Models (GCMs), which are complex representations of the major climate components and their interactions. However, these predictions present high levels of uncertainty, as illustrated by the very disparate results GCMs generate. According to the International Panel on Climate Change (IPCC), there is high confidence that such high levels of uncertainty are due to the way clouds are represented in climate models.

Indeed, several cloud phenomena, such as the cloud-radiative forcing, are not well-modeled in GCMs because they rely on miscroscopic processes that, due to computational limitations, cannot be represented in GCMs. Such phenomena are instead represented through physically-motivated parameterizations, which lead to uncertainties in cloud representations. For these reasons, improving the parameterizations required for representing clouds in GCMs is a current focus of climate modeling research efforts.

Integrating cloud satellite data into GCMs has been proved to be essential to the development and assessment of cloud radiative transfer parameterizations. Cloud-related data is captured by a variety of satellites, such as satellites from NASA's afternoon constellation (also named the A-train), which collect vertical and horizontal data on the same orbital track. Data from the A-train has been useful to many studies on cloud prediction, but its coverage is limited. This is due to the fact that the sensors that collect vertical data have very narrow swaths, with a width as small as one kilometer. As a result, the area where vertical data exists is very limited, equivalent to a 1-kilometer-wide track.

Thus, in order for satellite cloud data to be compared to global representations of clouds in GCMs, additional vertical cloud data has to be generated to provide a more global coverage. Consequently, the overall objective of this thesis is to support the validation of GCMs

cloud representations through the generation of 3D cloud fields using cloud vertical data from space-borne sensors.

This has already been attempted by several studies through the implementation of physics-based and similarity-based approaches. However, such studies have a number of limitations, such as the inability to handle large amounts of data and high resolutions, or the inability to account for diverse vertical profiles. Such limitations motivate the need for novel approaches in the generation of 3D cloud fields. For this purpose, efforts have been initiated at ASDL to develop an approach that leverages data fusion and machine learning techniques to generate 3-D cloud field domains. Several successive ASDL-led efforts have helped shape this approach and overcome some of its challenges. In particular, these efforts have led to the development of a cloud predictive classification model that is based on decision trees and integrates atmospheric data to predict vertical cloud fraction. This model was evaluated against "on-track" cloud vertical data, and was found to have an acceptable performance. However, several limitations were identified in this model and the approach that led to it. First, its performance was lower when predicting lower-altitude clouds, and its overall performance could still be greatly improved. Second, the model had only been assessed at "on-track" locations, while the construction of data at "off-track" locations is necessary for generating 3D cloud fields. Last, the model had not been validated in the context of GCMs cloud representation, and no satisfactory level of model accuracy had been determined in this context.

This work aims at overcoming these limitations by taking the following approach. The model obtained from previous efforts is improved by integrating additional, higher-accuracy data, by investigating the correlation within atmospheric predictors, and by implementing additional classification machine learning techniques, such as Random Forests. Then, the predictive model is performed at "off-track" locations, using predictors from NASA's LAADS datasets. Horizontal validation of the computed profiles is performed against an existing dataset containing the Cloud Mask at the same locations. This leads

to the generation of a coherent global 3D cloud fields dataset. Last, a methodology for validating this computed dataset in the context of GCMs cloud-radiative forcing representation is developed. The Fu-Liou code is implemented on sample vertical profiles from the computed dataset, and the output radiative fluxes are analyzed.

This research significantly improves the model developed in previous efforts, as well validates the computed global dataset against existing data. Such validation demonstrates the potential of a machine learning-based approach to generate 3D cloud fields. Additionally, this research provides a benchmarked methodology to further validate this machine learning-based approach in the context of study. Altogether, this thesis contributes to NASA's ongoing efforts towards improving GCMs and climate change predictions as a whole.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Introduction and Motivation

Climate change is one of today's greatest global challenges, mostly because of the significant consequences it carries on global population and on the economy. According to the World Health Organization [1], climate change will cause approximately 250,000 deaths per year worldwide between 2030 and 2050. The direct damage costs to health are estimated to be between US$ 2-4 billion per year by 2030. Climate change also induces a global warming effect, which leads to environmental perturbations. As an example, in the last 130 years, global temperatures have arisen by around 1°C, which has an irreversible impact on the environment and biodiversity. Such examples are only a few of the heavy consequences that climate change will have on our lives and on the planet.

It is thus critical and urgent to have a better understanding of climate change in order to be able to accurately predict and model its effects. Climate change predictions are currently achieved using Global Climate Models (GCMs). A GCM is a complex mathematical representation of the major climate system components (atmosphere, land surface, ocean, and sea ice), and their interactions [2]. Such climate models divide the globe into grid cells, with various sizes depending on the model. The typical resolution of a grid cell is 1°x 1° in geographic coordinates [3], which is equivalent to about 100x100 km. GCMs grid cells thus have a typical size of 10,000 km^2.

GCMs build projections of the future states of global climate. One of the main measures for the state of climate is the *Equilibrium Climate Sensitivity* (ECS), which is the equilibrium annual global mean temperature response to a doubling of equivalent atmospheric CO_2 from pre-industrial levels. In other words, ECS is a measure of the strength of

the climate system's eventual response to greenhouse gas forcing [4].

The issue, however, as shown by *Schneider et al.*[5], is that ECSs, as computed by current climate models, are spread onto a wide range. According to the International Panel on Climate Change 2013 Report [3], there is high confidence that these high levels of uncertainty within the climate projections are due to the representation of clouds in climate models. Indeed, the direct climate forcing from clouds, *i.e.* the effect of clouds on the radiation balance of the Earth, also reffered to as cloud-radiative forcing [6], is not well modeled in GCMs, and very uncertain [7]. This uncertainty stems from the fact that the cloud processes, which are key to understanding the relationships of clouds with climate, are difficult to integrate into climate models. Indeed, clouds vary spatially and temporally, and involve processes on multiple scales, from microscopic to global. For example, the water vapor condensation process taking place within low-altitude clouds cannot be represented in the large grid cells of GCMs because the computational resources for doing so are not available to this day. The condensation processes are instead represented through physically-motivated parameterizations. This leads to difficulties in resolving low clouds, which are projected to be resolved in GCMs by 2060, when the required computational power is expected to be available [5]. Another cause of uncertainty is the representation of the warming effect of absorbing aerosols located within or above highly reflective clouds [7]. Such effects are generally omitted or underestimated in GCMs [3], which leads to additional errors and uncertainties in the estimation of cloud radiative forcing. An additional challenge for GCMs is the representation of the correct phase of the cloud condensate (the product of the water vapor condensation, *i.e.* water or ice, or fractions of both), particularly the representation of cloud ice. Indeed, clouds containing a non-negligible phase fraction of ice have been proven to have a non-negligible effect on cloud-radiative forcing [8]. Yet, few observations are available to evaluate models in terms of their representation of cloud phase [3]. As a result, additional relations and parameterizations have to be developed in order to accurately model the cloud phase composition.

For these reasons, improving the parameterizations required for representing clouds in GCMs is a current focus of climate modeling research efforts [3]. Such improvement can be achieved thanks to the integration of cloud satellite data into GCMs. Indeed, it has been recognized that such data, and in particular cloud data coming from active sensors onboard various satellites, is essential to the development and assessment of cloud radiative transfer parameterizations [9]. Several studies corroborate this statement, as they outline cases in which cloud satellite data have been useful for reducing the uncertainties of GCMs. In *Chand et al.* [7], satellite-based estimates of above-cloud aerosol information have been used as validation for the representation of cloud-radiative forcing. The data used came from active sensors, and proved more useful and accurate than the passive data previously available. A study by *Doutriaux-Boucher and Quaas* used satellite data to evaluate the LMDZ GCM (Laboratoire de Meteorologie Dynamique "Zoom" Global Climate Model) cloud phase parameterization. In particular, they improved its representation of the shortwave cloud radiative forcing by establishing statistical relationships between cloud top thermodynamical phase and cloud top temperature, using both satellite data and model results [10]. Another study by *Naud et al.* showed that satellite data is helpful for assessing the interactions between cyclone dynamics, atmospheric water vapor content, and frontal clouds, which are often neglected in GCMs [11], and stated that such interactions could be taken into account in GCMs by integrating the corresponding satellite datasets.

Cloud-related data is captured by a variety of satellites. Data from NASA's afternoon constellation (also named the A-train), in particular, is used in most of the aforementioned studies. As mentioned by *Naud et al.* [11], *"A-train-based analyses are useful tools for the evaluation of GCMs"*. The A-train is a constellation of Earth-observing satellites that closely follow one after another along the same orbital track and collect vertical and horizontal data along this track [12]. Figure 1.1 is a representation of the different satellites that compose the A-train, and of their associated ground tracks.

The three satellites that are associated with the collection of cloud-related data are

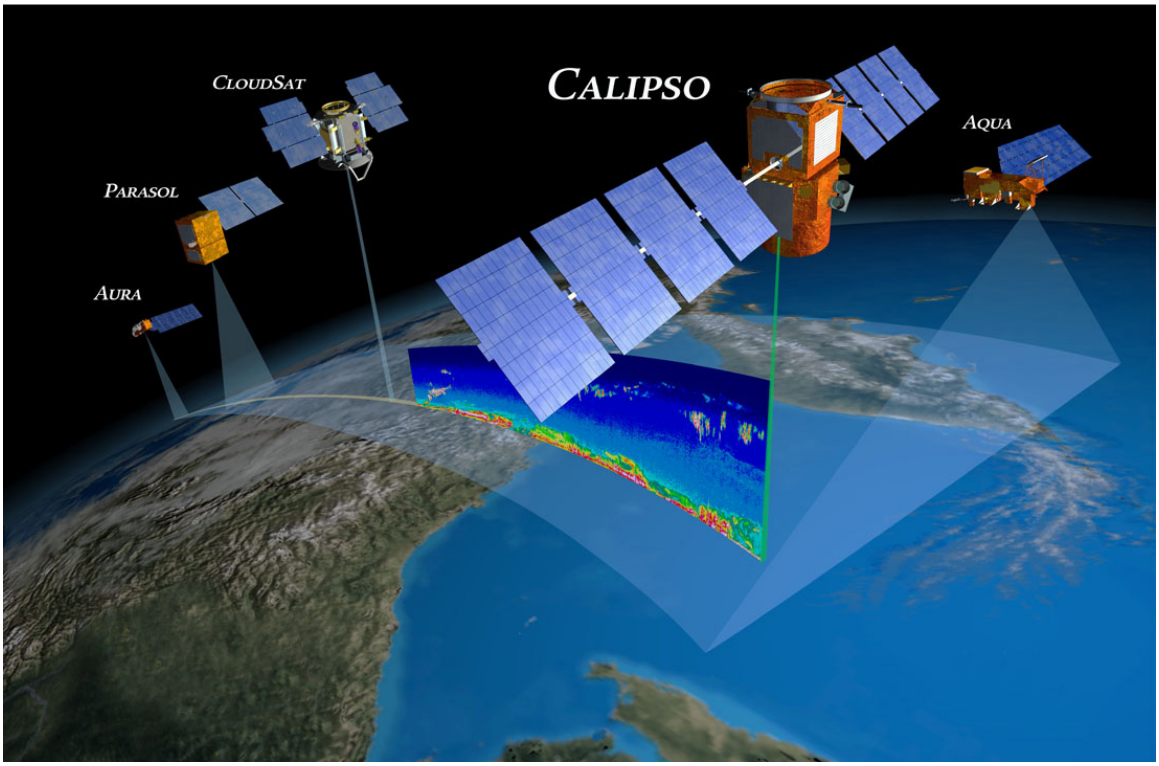

Figure 1.1: NASA's A-train [13]

Aqua, CloudSat and CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation). The MODIS (Moderate Resolution Imaging Spectroradiometer) sensor on Aqua collects radiance (reflected and radiant energy) data, which is a key property of cloud interaction. The Cloud Profiling RADAR (CPR) on CloudSat collects data relative to Climate Variability and Change and Weather, and so does the CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization) LIDAR on CALIPSO, which also collects information on Atmospheric Composition and Water and Energy Cycles. Altogether, these satellites provide useful data for cloud prediction. However, the availability of such data at various locations around the Earth is too scarce. This is due to the fact that the sensors that collect vertical data have very narrow swaths, with a width as small as one kilometer, due to their "pencil thin" RADAR or LIDAR beams. As a result, the area where vertical data exists is very limited, equivalent to a 1-kilometer-wide track.

In order for satellite cloud data to be usefully compared to representations of clouds in

GCMs, data have to be available at the same locations. As the GCMs grid cell's dimensions are typically 100 km x 100 km, the retrieved vertical cloud data from space-borne sensors would only cover 1/100th of the cell. Consequently, additional vertical cloud data has to be generated in order to fill the equivalent GCMs cells, so that GCMs cloud representations in each cell can be evaluated against such data, and improved based on this evaluation. This data generation process can leverage the A-train data to provide three-dimensional cloud field domains, spread on the typical area of GCMs grid cells.

The overall objective of this thesis can thus be formulated as follows:

Research objective: Support the validation of GCMs cloud representations through the generation of 3D cloud fields using cloud vertical data from space-borne sensors.

The generation of such 3D cloud fields has been attempted in several studies, as discussed in the following section.

1.2 Background

This section reviews existing approaches to the generation of 3D cloud fields.

1.2.1 Generation of 3D cloud fields using physics-based models

Retrieving three-dimensional cloud field domains is one of the goals of the Clouds and the Earth's Radiant Energy System (CERES) Project. CERES aims at creating physics-based cloud models by retrieving cloud properties and using them to derive cloud boundaries, phase, optical depth, and other key parameters [14]. CERES has created various cloud products, based on data from different sensors, not only on A-train satellites but on other additional satellites as well, such as the Tropical Rainfall Measuring Mission System (TRMM) sensor. The second edition of the CERES algorithms and products used data captured between 1998 and 2007, which represents very large amounts of data. CERES cloud products vary from 20-kilometer footprint data obtained directly from the Aqua and

Terra satellites, to globally-available cloud properties averaged over hours, days or months. These cloud products have proved useful in improving the understanding of the relationships between clouds and the radiation budget. Comparisons have been made with independent measurements from the space-borne sensor. While these comparisons have shown general consistency of the products, they have also outlined several discrepancies when compared to the independent data. The greatest discrepancies occur over ice-covered surfaces, with other non-negligible discrepancies occuring for several cloud properties [15]. Such discrepancies are thought to be caused by several known problems in the algorithms used to retrieve the cloud properties, such as the ozone absorption errors, the underestimation of thin cirrus optical-depth, and the representation of multi-layered clouds. Detailed comparisons with CALIPSO data have already been performed [16] and have led to enhancement of the CERES products. *Minnis et al.* stress that "more detailed comparisons with CloudSat and CALIPSO data will be particularly useful in future validation efforts". As such, CERES cloud products data, currently in their 4^{th} edition, are under continuous development and improvement.

However, the approach taken by the CERES project has several limitations. CERES products are obtained through various sources and types of data, but the algorithms developed to process such data are not easily adaptable to other data types and variables. Indeed, the CERES cloud processing algorithms are distributed into several components or boxes, as shown in Figure 1.2, and each one of this boxes features several sub-components and sub-algorithms, in order to process every variable linked to the retrieval of cloud properties. Consequently, if other A-train data had to be included and processed through this approach, the whole cloud processing scheme would have to be reorganized, as other variables would have to be integrated. New boxes and sub-boxes would have to be developed and added to the processing scheme, and existing boxes would have to be reorganized around them. The integration of new data would thus be quite tedious, and even more so if new data would have to be regularly added, as the process scheme would have to be constantly adapted.

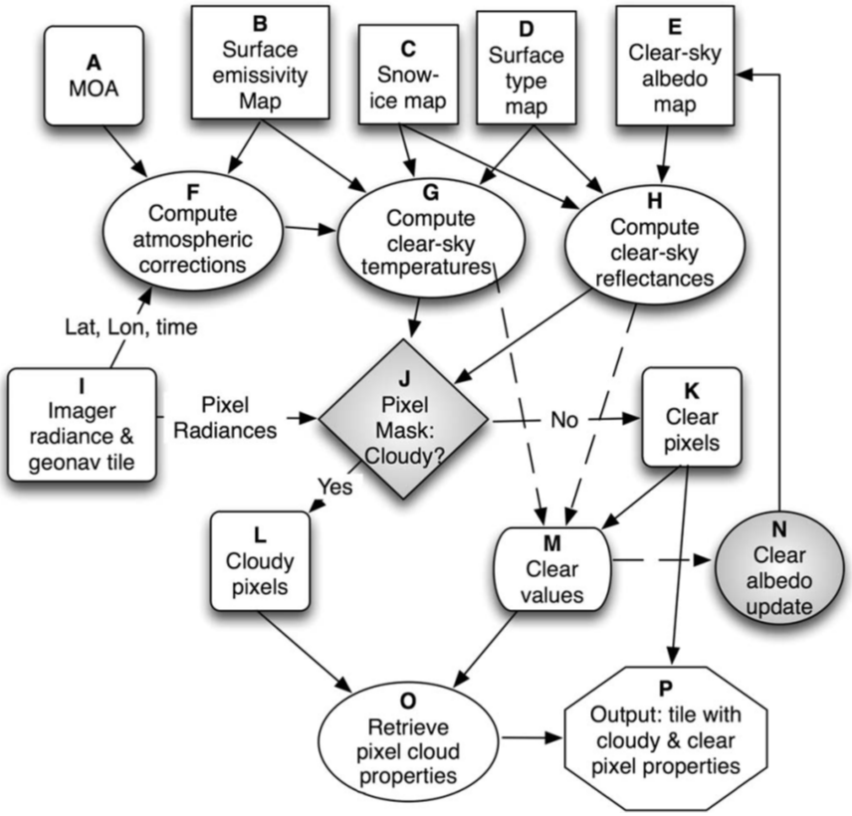

Figure 1.2: CERES cloud processing scheme [14]

Additionally, the CERES cloud property retrieval process has a high computational cost, which leads to a need to downscale the resolution of the input datasets. Indeed, the MODIS products have a 1-km resolution, but for computational purposes, only one reading out of four is considered by the CERES algorithms. Because this sample reduction is also performed on the other inputs datasets, the resolution of the inputs to the CERES algorithm ends up being down-scaled. According to *Minnis et al.* [14], this resolution downscaling induces a small error only.

1.2.2 Generation of 3D cloud fields using similarity to "off-track" data

The studies discussed below provide approaches based on similarities between specific cloud-related variables.

The generation of three-dimensional cloud fields has also been attempted by *Barker et al.* [9] using a matching algorithm. The matching algorithm is created in order to attribute vertical profiles to locations at which such profiles do not exist in the retrieved satellite data, *i.e.* surrounding "off-track" locations. The algorithm copies existing vertical profiles to "empty" locations by computing the squared differences between the observed spectral radiances at these locations, as presented in Figure 1.3.

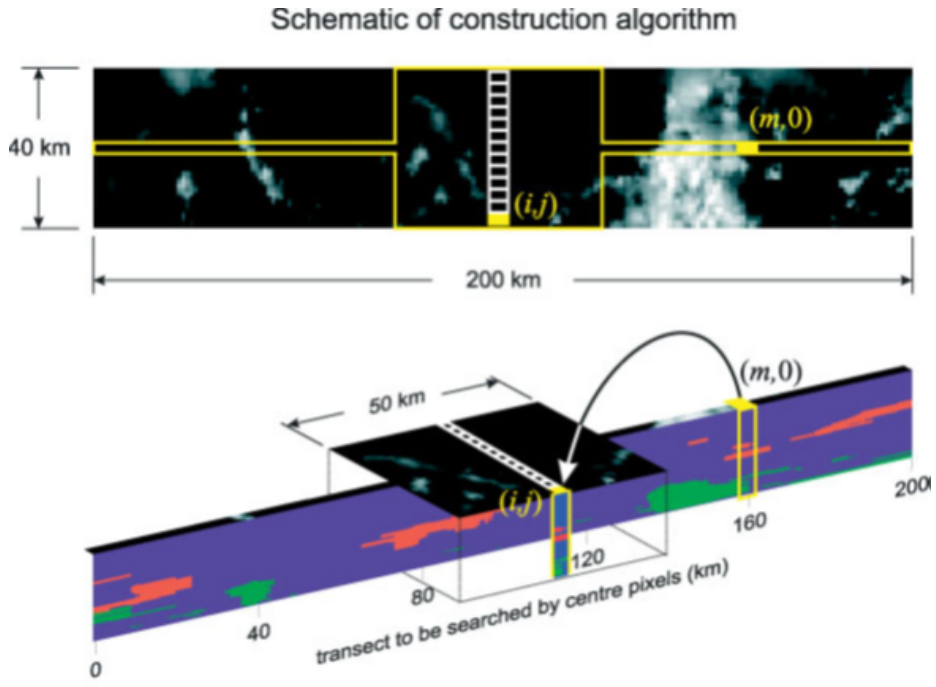

Figure 1.3: 3D cloud scene construction process [9]

Similar locations are thus matched by minimizing this error, as well as fulfilling certain criteria, such as having the same surface type, close solar zenith and solar azimuth angles values, and usually being within 10 kilometers apart. This algorithm is tested on compiled data from the A-train, and especially from the MODIS sensor, which provides the observed radiances at the target locations, with a 1-km resolution. The construction algorithm is applied to the Sun-up side of one orbit of the train, equivalent to locations between 60°S and 60°N, and captured on April 19^{th} 2007. The reconstruction of captured data was attempted in order to compare reconstructed to original data. Vertical profiles were retrieved from existing profiles no closer than 1, 5, 10 and 20 kilometers successively. When using existing

profiles close to the target location (about 1km), the results were satisfactory. However, using profiles located further away than 1 to 20 kilometers did not provide acceptable results. This shows that the approach is not adapted to retrieving profiles at mildly remote locations, *i.e.* at locations further away than 20 kilometers of the existing data. At "close" locations, the construction algorithm gives satisfactory results for high-altitude clouds (higher than 10 kilometers), but the performance for lower altitude clouds degrades very quickly as further locations are used.

One-dimensional radiative transfer calculations were then performed on cloud domains measuring 21 kilometers in the across-track direction, and 40 kilometers in the along-track direction, and obtained for eight partial orbits. The obtained broadband domain-averaged radiative fluxes are found to agree with the corresponding CERES fluxes.

Another approach for generating 3-D cloud fields to be used in large-scale models was also elaborated by *Barker et al.* [17]. In particular, they developed a stochastic algorithm to generate the vertical profiles, based on a parameterization of clouds different from the ones usually featured in GCMs. About 29,000 domains, each about 280-kilometers long, were created based on CloudSat and CALIPSO data, and used as inputs to various radiative transfer codes. The main purpose of the study was not to generate accurate vertical profiles for assessing cloud representations in GCMs, but rather to promote the use of 3D radiative transfer models in GCMs, as opposed to the current use of 1D models. The large domains that were computed for this study do not have a resolution that is adequate to computing cloud vertical properties.

The mentioned studies, both physics-based (CERES) and similarity-based (conducted by *Barker et al.*), have a number of limitations. First, both physics-based and similarity-based approaches to the generation of 3D cloud fields were implemented on a very limited number of cloud-related parameters. However, a considerably higher number of variables may have to be included in those models to increase their performance/validity. Yet, inte-

grating such quantity of data in the models and assessing its relations with cloud properties might prove computationally difficult using the existing physical-based and stochastic approaches. Furthermore, as highlighted earlier, the physics-based approach uses a complex cloud processing scheme, which would be costly to adapt to additional data and variables. Both approaches also have to make compromises in order to reduce the computational cost of retrieving the cloud profiles. As outlined earlier, the physics-based approach makes the choice of downscaling the resolution of the computed products in order to reduce the computational cost, even though this lower resolution leads to the averaging of cloud properties that were more detailed in the input datasets. The similarity-based approach does not downscale the resolution, but only uses small datasets for retrieving the profiles, corresponding to about an hour of consecutive samples. Because the datasets used in the approach represent a very small sample of the existing profiles, the predicted profiles are limited to the ones from the dataset. Indeed, this approach only uses the vertical configurations featured in the data sets it is based on, missing any of the other possible vertical configurations that exist but may not have been included in the data used. For example, the datasets may contain profiles with clouds at certain altitudes only, which means that this approach would be unable to generate profiles with clouds at other altitudes than the ones featured in the dataset. If the similarity-based approach were to be able to generate any possible vertical configuration, the datasets would have to feature every single one possible configuration, which would be computationally impossible as there is an infinity of such configurations. Consequently, the limited representation of the diversity of cloud profiles and properties present some serious limitations to the global generation of 3D cloud fields using this approach. Finally, the limitation highlighted when reconstructing from a simple sample would be expected to grow as this approach is used on a larger, more global, dataset.

The limitations and shortcomings identified from the literature highlight the need for additional approaches to the generation of 3D cloud fields. This leads to the following

overarching Research Question:

Overarching Research Question: What approach or combination of approaches would be best suited for the generation of 3D cloud fields?

From the review of the literature, one can identify the characteristics and capabilities such novel approach to the generation of 3D cloud fields would need to enable. Such approach needs to be:

- Able to predict "off-track" cloud profiles: such profiles should not be generated as copies of "on-track" ones, as it has been attempted by *Barker et al.*. Instead, "off-track" profiles should be generated whether or not similar profiles exist in the "on-track data", at close or remote locations
- Able to account for and handle **high amounts of data and predictive features**
- **Scalable**, i.e. support the integration and processing of multiple data sources
- **Flexible**, i.e. support the integration of data sources other than the one(s) originally considered, and once a preliminary approach has already been established
- Able to account for and handle **disparate sources and types of data**
- Able to handle sources of data that have **different resolutions or levels of granularity**
- Able to generate models in a reasonable amount of time and with a reasonable amount of computing resources

An approach that combines data fusion with machine learning techniques is anticipated to provide the aforementioned capabilities. Efforts conducted at the Aerospace Systems Design Laboratory at the Georgia Institute of Technology using such novel approach are discussed below.

1.2.3 Generation of 3D cloud fields using data fusion and machine learning techniques

In the past two years, a team of researchers at the Aerospace Systems Design Laboratory (ASDL) at the Georgia Institute of Technology has dedicated much effort in developing a novel approach that leverages data fusion and machine learning techniques to generate 3-D cloud field domains. Several successive graduate projects have helped shape this approach and overcome some of its challenges.

The first instantiation of a data fusion and machine learning-based approach was developed and implemented in the *Grand Challenges* led by *Johnson et al.* [18]. It is during this effort that the relationships between horizontal and vertical cloud data were first investigated. The main goal of this research was to develop prediction models for vertical cloud profiles based on various horizontal features, such as surface type, geolocation, or observed radiances, and to train these models using machine learning techniques. These features are contained within NASA's C3M dataset, which is a merged dataset containing information from CALIPSO, CloudSat, CERES and MODIS. This dataset can be obtained online, for dates ranging from July 2006 to April 2011 [19]. The resolution of the data was about 21 km^2 for each reading cell, which means that each reading corresponds to an average of the cloud properties over an area of 21 km^2. In order to first identify the relationships between horizontal and vertical cloud data, a visualization environment tailored to cloud data was developed. The identified relationships were then modeled using deterministic techniques. The target feature was the cloud fraction, *i. e.* the percentage of cloud present in each vertical measurement. The cloud fraction profile was modeled as a sum of Gaussians, and the model parameters were the characteristics of these Gaussians. The predictive models were trained using neural networks to predict vertical cloud profiles using the following predictive horizontal features: geolocation (latitude, longitude), surface type, solar zenith angle, and observed radiances (visible and infrared, 15 bands total). On-track vertical profiles were predicted and compared to the original ones. The comparison showed some agreement between the predicted and original profiles, but the model had some difficul-

ties in predicting certain configurations, especially for locations at which several layers of clouds were present. Additionally, the models were trained on data samples equivalent to a 200-kilometer spread of readings, as it was assumed that atmospheric conditions remained constant over 200 kilometers only, and the models had to be trained on similar conditions. This condition led to the use of very small samples for training and validating the models (about 10 points), due to the data resolution being 21 km^2.

Following this work, several efforts were dedicated to enhance the developed predictive models. In her Special Problem research [20], *C. Johnson* tested additional training methods, such as Decision Trees, Support Vector Machines, and Gaussian Processes, using the same datasets as the ones used in [18]. A comparison of the different models showed that decision trees were better suited for this specific problem. In order to be able to account for additional datasets and facilitate the development of predictive models, *V. Ngo* [21] developed a method for extracting the features of interest from the datasets used in [18] and [20]. Algorithms were developed in Matlab, and enabled to structure and convert the data to common Matlab structures for further analysis and predictive model development.

While these efforts represented a great first step towards the implementation of data fusion and machine learning techniques to develop a global cloud vertical dataset, the approaches proposed had a number of limitations. First, the predictive capability of the models was limited. This limitation was thought to have many origins:

- The resolution of the data used to build the predictive models was too low
- Some of the key variables necessary to build a highly accurate model were missing. Information about the atmospheric context, for example, was lacking
- The machine learning techniques implemented to build the past previous models needed to be improved or other techniques needed to be considered

Second, the extraction and structuring algorithm developed by Ngo was limited to the datasets investigated by the 2017 Grand Challenge. Consequently, it needed to be extended

to account for other relevant types of data or ones that have different resolutions.

Investigating such assumptions and limitations was the objective of the 2018 Grand Challenge project [22]. To this end, higher-resolution datasets were identified, processed, cleaned and merged. The GEOPROF-LIDAR, MODIS-AUX, PRECIP-COLUMN, and RAIN-PROFILE datasets were downloaded from the CloudSat Processing Center [23]. These datasets contain the features used as predictors in the aforementioned efforts, but with a higher resolution. Indeed, the readings have a cell area of 1 km^2, as opposed to 21 km^2 for the previous datasets.

The extraction and structuring algorithm first developed by Ngo was enhanced and adapted to such datasets. Atmospheric context was added to the model by introducing the MERRA-2 (second Modern-Era Retrospective analysis for Research and Applications) atmospheric dataset, which contains atmospheric parameters. Because this dataset was very different from the other ones (different source, time frame, dimensionality, etc.), a data fusion approach was implemented so as to obtain a single set of parameters to use as inputs to the predictive models. Next, predictive models of the vertical cloud fraction were generated, using horizontal parameters such as geolocation, elevation, radiance bands, liquid water percentage and surface type as predictors. Other models generated also included the atmospheric parameters as additional predictors. A 10-vertical-band classification model was used for parameterizing the cloud fraction, and models were trained using two different machine learning techniques: neural networks and decision trees. The training dataset corresponded to three consecutive days of satellite data collection, and the validation dataset was the following 4^{th} day. The models were thus trained on a much larger number of samples when compared to the previous efforts.

Model performance was assessed using Matthews Correlation Coefficient (MCC). MCC can be expressed with the following equation:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{1.1}$$

where TP, TN, FP and FN correspond to the following cases:

- True Positive (TP): an instance in the test set had a target feature value of 1 and the target feature value as predicted by the model is also 1
- True Negative (TN): an instance in the test set had a target feature value of 0 and the target feature value as predicted by the model is also 0
- False Positive (FP): an instance in the test set had a target feature value of 0 and the target feature value as predicted by the model is 1
- False Negative (FN): an instance in the test set had a target feature value of 1 and the target feature value as predicted by the model is 0

MCC is a balanced metric that can be used in classification problems where the numbers of True Positive and True Negative are heavily biased, as was the case with the targeted problem where readings showing no presence of clouds were much more frequent than readings showing cloud presence.

Using this measure, the models obtained using decision trees proved to perform better than models obtained with neural networks, thus verifying the statements made by *Johnson* [20]. The addition of climate context through atmospheric parameters also enhanced the performance of the models, for all machine learning techniques considered.

The product of this effort thus consisted in the development of a predictive model for vertical cloud fraction, based on decision trees and using MERRA-2 atmospheric data. This classification model predicts the presence of clouds in 10 superimposed altitude bands, the first band being the highest one. Figure 1.4 shows the schematized predictive performance obtained with the model on a sample of the vertical profiles used as validation data.

The MCCs associated with the prediction for each band are presented in Table 1.1. The closer the MCC is to 1, the better the prediction [24].

From the results of this last effort, several observations can be made. First, the model performance is lower for lower-altitude cloud bands than for higher-altitude ones, as demon-

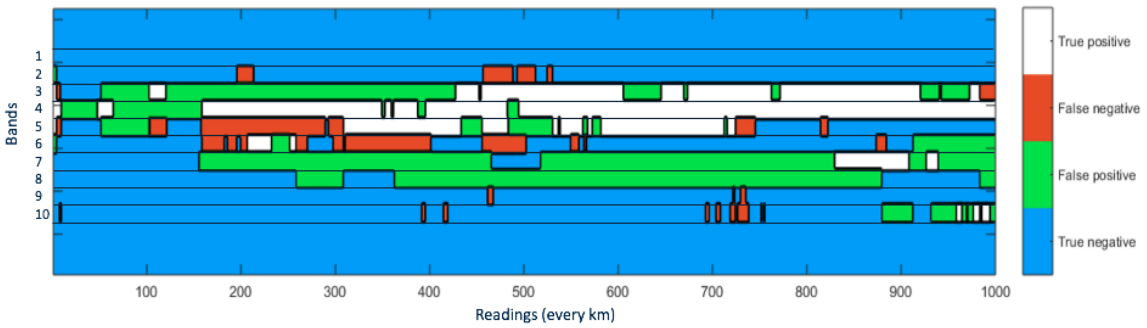

Figure 1.4: Visualization of the cloud predictive model performance on a parameterized profile

Table 1.1: MCC score obtained for each altitude band

Band	MCC
1	-0.0029
2	0.4944
3	0.4922
4	0.4333
5	0.4089
6	0.4210
7	0.3198
8	0.2329
9	0.2200
10	0.2821

strated by their lower MCC scores. This is due to the fact that, as the LIDAR waves cross more superimposed cloud layers to reach the bottom ones, absorption or scattering phenomena cause the measurements to become less and less accurate. Consequently, the predictive model is less accurate for those bands, because the data it is based on is less accurate for such bands. This observation concurs with the one stated by *Barker et. al* [9], who find that lower-altitude clouds are harder to retrieve than higher-altitude ones. This leads to the following Research Question:

Research Question #1: How can we better predict the presence of clouds in lower-altitude bands?

Additionally, the MCC score is very low for the top band. This is most likely due to the fact that cloud presence occurrences are very rare in this layer as compared to cloud absence, so it is hard for this performance measure to capture this very unbalanced scene. Globally, the performance of the model is acceptable, as it is for all layers well above 0 (except the top one), with "0" representing a random prediction. However, there is room for improvement. One of the most likely causes of this overall limited performance is the machine learning technique used, which might not be the most well-suited for the considered problem. This leads to the following Research Question:

Research Question #2: Which machine learning technique(s) would lead to an improved predictive capability?

The MCC scores displayed in Table 1.1 were obtained for specific locations, which correspond to "on-track" locations, as the predicted profiles have to be compared to existing ones in order to assess the performance of the models. Figure 1.5 shows the spatial distribution of "on-track" and "off-track" data.

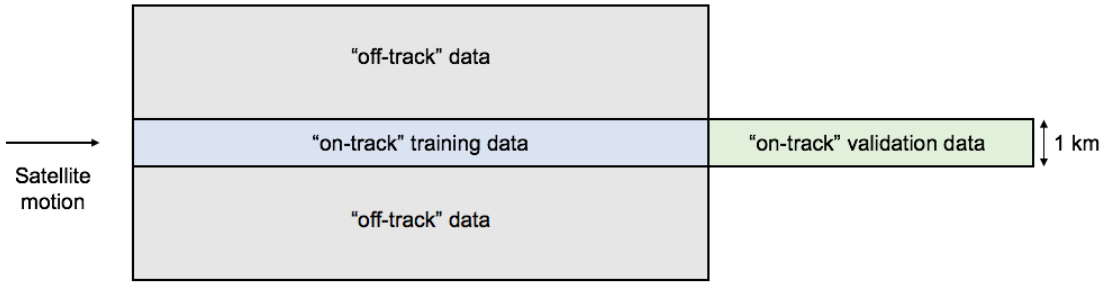

Figure 1.5: "On-track" and "off-track" data locations

The predictive models are limited to "on-track" locations only. While critical to the construction of 3D cloud fields, no "off-track" validation has been attempted up to this day. This leads to the following Research Question:

Research Question #3: What approach is best suited to validate the predicted models "off-track"?

Furthermore, while MCC represents a good objective measure of the performance of predictive models, no actual performance goal or threshold for the models has been set in this specific context of study. Hence, while previous efforts focused on developing models that were as accurate as possible, no requirement as to what level of model accuracy was necessary has ever been determined. This leads to the following Research Question:

Research Question #4: What level of accuracy is required from the predictive models to generate 3D cloud fields?

And the subsequent Research Question:

Research Question #4.1: What approach should be undertaken to determine a satisfactory level of model accuracy?

1.3 Thesis Structure

This present chapter motivated the need for this research, and defined and delimited its scope. The following chapter further formulates the problem this research aims to address in the context of the Research Questions formulated in Chapter 1. Hypotheses are then formulated for each research question based upon the challenges and shortcomings identified. Next, Chapter 3 briefly introduces the approach developed to address these Research Questions and validate each of their associated Hypotheses. Chapters 4, 5 and 6 detail the implementation of the approach, and discuss the results, in the context of the formulated Hypotheses. Finally, Chapter 7 provides a summary of the results and an overview of the benefits and contributions of this research. It further details additional steps to be considered for future work on this topic.

CHAPTER 2
PROBLEM FORMULATION

The previous chapter identified several gaps in the past works and studies aimed at creating a three-dimensional, global dataset containing horizontal and vertical cloud properties. Such gaps helped shape the following Research Questions, which provide specific context for this study:

- RQ #1: How can we better predict the presence of clouds in lower-altitude bands?
- RQ #2: Which machine learning technique(s) would lead to an improved predictive capability?
- RQ #3: What approach is best suited to validate the predicted models "off-track"?
- RQ #4: What level of accuracy is required from the predictive models to generate 3D cloud fields?
- RQ #4.1: What approach should be undertaken to determine a satisfactory level of model accuracy?

These research questions can be divided into two categories. The first two research questions relate to the **improvement of the predictive models** developed throughout the previous works on generating 3D cloud fields using data fusion and machine learning techniques. The last three research questions relate to the **validation and fidelity assessment** of the obtained predictive model in the specific context of this study: generating a global cloud vertical profile dataset to better assess and improve GCMs cloud representation.

The following sections detail the different approaches that could be implemented to answer these research questions, leading to the formulation of the corresponding hypotheses.

2.1 Model improvement

2.1.1 Baseline model overview

The baseline predictive model is the one developed by *Huguenin et al.* [22]. The model is trained using decision trees, with the horizontal predictors presented in Table 2.1. Such predictors come from the fusion of several datasets: GEOPROF-LIDAR [25], MODIS-AUX [26], PRECIP-COLUMN [27], and MERRA-2 (tavgl_2d_slv_Nx collection) [28].

Table 2.1: Predictors used in current predictive model

Predictor	Description
Geolocation	Latitude and Longitude of each reading ; featured in all datasets
Elevation	Ground elevation in meters, at reading location ; featured in GEOPROF-LIDAR
Surface Type	4 possible types: land, open ocean, inland water, sea ice ; featured in PRECIP-COLUMN
Radiances	Radiance bands 1 to 7, 17 to 19 and 26 ; featured in MODIS-AUX
Atmospheric variables	29 independent features from the MERRA-2 dataset (originally contains 47)

The predicted target is the cloud vertical profile. The profile is parameterized in 10 superimposed bands, as shown for a sample profile in Figure 2.1.

Each reading thus corresponds to 10 vertical values, each corresponding to one band. The values are binary, equal to either 0 (no cloud in the band at reading location) or 1 (cloud in the band at reading location).

This predictive model provides a basis for the improvements to be achieved in this work in order to address Research Questions #1 and #2, as discussed in the next sections.

2.1.2 Improvement of prediction performance for lower-altitude cloud bands

As highlighted in the first chapter, the performance of the predictive model is significantly lower for low-altitude bands than for high-altitude ones. This statement has led to the first

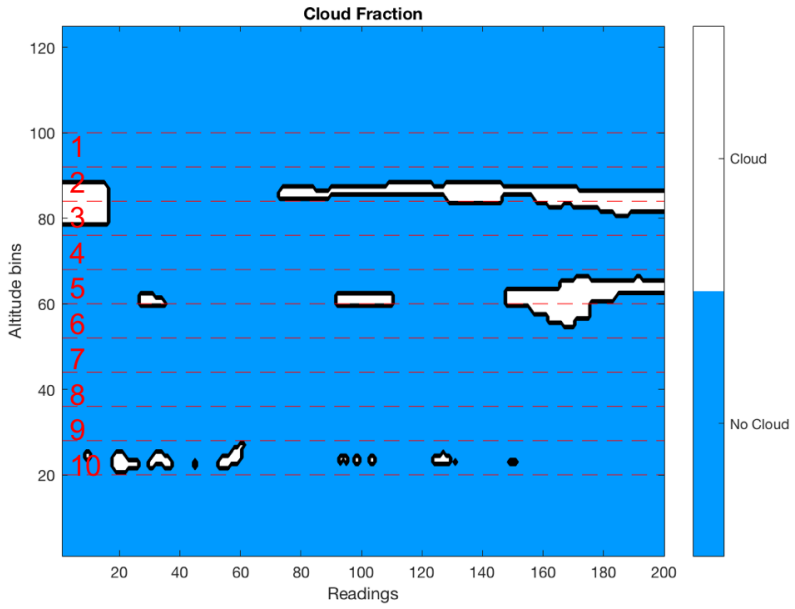

Figure 2.1: Cloud fraction parameterization

research question: ”How can we better predict the presence of clouds in lower-altitude bands?”

This problem has its roots in the accuracy of the data on which the model is based, as such data is less accurate for lower altitudes. This is due to the fact that LIDAR sensors, which capture the data that has been used for building the model, can be more sensitive to reflection and diffusion phenomena. As the data is captured from above, its accuracy diminishes as the altitude at which it is captured decreases.

Thus, in order to improve the prediction performance for lower-altitude bands, the data accuracy related to these bands has to be improved. The following solutions can be considered in order to reach this goal:

- Additional data sources have to be integrated, with a better accuracy for lower altitudes
- The model has to differentiate between high altitude and low altitude bands, so that it can account for the lack of accuracy of the lower-altitude bands

The following subsections detail the different options for implementing these solutions,

and evaluate whether or not they are well-suited for the problem of interest.

Additional data from ground-based stations

The LIDAR-captured data used so far was captured from space-borne sensors, which means that the first measurements taken by the sensors were the ones associated with the upper layers of the cloud profiles. This explains the lower accuracy for lower-altitude layers, as the sensor has to go through high layers to capture lower ones, which makes it more exposed to reflection and diffusion phenomena. This would be the opposite for data captured from below the clouds, as the sensors would capture the lower layers first. Thus, it is expected that integrating data captured from the Earth's surface (e.g. ground-based stations) would provide more accurate data for lower layers.

There is a number of ground-based stations collecting cloud and atmospheric data on a daily basis. Many are government or privately-owned and their collected datasets are not made available. However, some organizations provide access to results from a number of stations. For example, the National Center for Environmental Information (within the National Oceanic and Atmospheric Administration) provides cloud data (altitude of cloud layers, up to 6 layers) for land-based stations on a hourly basis [29]. However, the number of stations which took measurements at the same epoch as the A-train satellites is very limited: in 2011, which corresponds to the epoch of the datasets on which the model is based, only 39 stations, mostly European, have collected cloud-related data. Such data would do little to improve the model, as it is too scarce (too few locations) and not well distributed across the globe.

Considering these limitations, the integration of ground-collected cloud data will not be considered in this work.

Additional data from GeoProf RADAR

Another model improvement, which was already considered in previous studies, is the integration of additional data from the GEOPROF-RADAR dataset [30]. As mentioned previously, the baseline model developed by *Huguenin et al.* [22] only takes into account LIDAR and not RADAR data. Including such data in the model could lead to better model performance for the lower-altitude bands, as RADAR sensors are less sensitive to reflection and dispersion phenomena than LIDAR sensors.

The main challenge for integrating the RADAR data is that there is no *Cloud Fraction* variable in the dataset, while this the variable on which the target of the model is currently based. The closest variable featured in the GEOPROF-RADAR dataset is the *Cloud Mask*. While the Cloud Fraction is the percent of cloud in each pixel captured by the sensor (*i.e.* for each reading, in each altitude bin), the Cloud Mask is the probability that a cloud was accurately detected in the pixel location.

In order to include the RADAR data, these two variables (Cloud Fraction and Cloud Mask) need to be merged so as to generate a parameterized vertical profile similar to the one in *Huguenin et al.* [22]. Since the GEOPROF-RADAR dataset is available from the same source as the former datasets, ready to be integrated and provides data with a better accuracy for low-altitude clouds, this expected model improvement is considered in this research. The following hypothesis is thus formulated:

Hypothesis #1: If cloud vertical data from the GEOPROF-RADAR dataset is fused with the existing datasets, then the performance of the predictive model at lower-altitude bands will improve.

Taking into account upper bands prediction

Another way of improving the model's predictive capability for lower-altitude bands would be to account for the values associated with upper bands when predicting lower bands. For

example, the value in band 5 would be predicted using the aforementioned model predictors (Table 2.1), as well as the values obtained in bands 2, 3 and 4. Indeed, as the data is captured from above, representing vertically superimposed cloud layers using horizontal data/information only may not be the most suitable approach, especially for lower-altitude bands. Doing so assumes that each altitude band is independent, *i.e.* the predictive model for one band does not know if the other models have "already" predicted clouds. Hence, updating the models by including the value obtained for the upper bands to the predictors of the models for the lower bands could provide better results. However, the top layer would have to be excluded as it has proven very hard to predict.

This approach is expected to help the model differentiate between high altitude and low altitude bands, as the low-altitude would have more predictors, and could benefit from the better predictive performance of upper bands. This leads to the formulation of a second hypothesis for RQ#1:

Hypothesis #2: If the predicted values of higher-altitude bands are taken into account when predicting lower-altitude bands, then the predictive model(s) for these bands will be improved.

2.1.3 Determination of new ML techniques to be implemented

The statement that the overall performance of the predictive model can be improved for all bands has led to the formulation of the second Research Question: Which machine learning technique(s) would lead to an improved predictive capability?

Two types of techniques could be applied in order to enhance the model:

- training techniques: the current training technique is a decision tree, there may be other techniques better suited for this problem
- "preliminary" techniques: techniques to be applied before training the predictive model, in order to enhance the quality of the predictors and target variables

The following sections outline the different techniques that match these descriptions and are considered well-suited for this problem.

De-correlation of atmospheric variables

In [22], the correlation of atmospheric variables from the MERRA-2 dataset was studied in order to remove dependent features from the list of predictors. This was achieved by visually spotting similar patterns between different variables, and analyzing the physical relationships explaining such similarities. Such study enabled the number of atmospheric predictors to be reduced from 47 to 29. However, there was no further study of data correlation other than a visual one, although there might be other relationships between variables which cannot be visually identified. Consequently, a more thorough analysis of the different atmospheric variables should help identify additional relationships, and further reduce the number of predictors, which would in turn reduce the computational cost of training the models without penalizing its performance, and even possibly enhancing it. Such analysis could be performed using Principal Component Analysis (PCA) [31]. PCA extracts the important information from the inter-correlated variables, and represents it as a set of new orthogonal variables called principal components. Such components can be ordered by their variation, *i.e.* their ability to represent the inter-correlated variables. The most "influential" variables would thus be the ones used to represent the principal components with the greatest variations. These most influential variables would be the final atmospheric predictors to be included in the models.

This leads to the following hypothesis:

Hypothesis #3: If the number of predictors from the MERRA-2 dataset is reduced using Principal Component Analysis, then the model predictive performance will be improved.

Improvement of ML techniques

Another way of improving the predictive models would be to train them using other machine learning techniques better suited for this problem. So far, classification decisions trees have been used and proved efficient when compared to other methods, such as classification neural networks. Yet, other methods are specifically built for such classification problems that have the potential to improve the model predictive performance.

The following methods have been identified by *Sotiris* [32] as the best current methods to be applied to classification problems:

- Decision Trees: trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

- Rule learners: algorithm that aims at constructing the smallest set of rules that is consistent with the training data and accurately describes it. This technique is similar to decision trees, as decision trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree. However, rules can also be directly induced from training data using a variety of rule-based algorithms.

- Neural Networks: composed of large number of units (neurons) joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed; output units, where the results of the processing are found; and units in between known as hidden units. The network is trained on a set of paired data to determine input-output mapping. The weights of the connections between neurons are then fixed and the network is used to determine the classifications of a new set of data.

- Naive Bayesian Networks: very simple type of Bayesian networks, which are graph-

ical models for probability relationships among a set of features. Naive Bayesian networks are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent (which is not the case in other Bayesian networks).

- k-Nearest Neighbour algorithm (kNN): based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbours. The kNN locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label.

- Support Vector Machines (SVMs): SVMs revolve around the notion of a "margin", which is either side of a hyperplane, a "limit" that separates two data classes. SVMs maximize the margin and thus create the largest possible distance between the separating hyperplane and the instances on either side of it, in order to reduce the generalization error, i.e. the prediction error.

Table 2.2 assesses the performance of each of these techniques across several criteria.

As highlighted by *Sotiris*, SVMs and Neural Networks are well-suited for multidimensional problems and continuous features, as is the case for this problem. From Figure 2.2, SVMs have an equal to superior performance than NNs for most criteria. Neural networks have already been tested in previous efforts and did not provide acceptable results. SVMs appear as a good candidate approach to improve model performance.

Figure 2.2 also shows that Decision Trees have a good overall performance, comparable to that of SVMs for most criteria. Existing variations of the Decision Trees could thus be applied, in order to benefit from this good performance and compare to results obtained with SVMs. Random Forests are one of these variations, and easily implementable (see

Table 2.2: Comparing learning algorithms (**** stars represent the best and * star the worst performance) [32]

	Decision Trees	Neural Networks	Nave Bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete or binary or continuous attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)	*** (not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

following subsections).

The following subsections provide an overview of the selected techniques: Kernel methods (to which SVMs belong) and Random Forests.

Kernel methods

Kernel methods are a recent machine learning technique adapted to non-linear, 2-class

classification problems [33], which is specifically the type of problem studied here. They include several different methods and learning systems, of which the most well-known are the Support Vector Machines (SVMs). SVMs are supervised learning algorithms used for binary classification or regression [34]. Built-in tools are available in Matlab for implementing SVMs, with the same architecture as the Decision Trees and Neural Networks tools already used for training the predictive models in previous works. SVMs can thus be tested as another method for training the predictive models, and possibly improve their performance.

Random Forest

The other machine learning method to be considered and tested in this work is the Random Forest. Random Forests are a combination of decision trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [35]. It is thus based on the same technique currently used in the model, and extends it to multiple decision trees. Combining the results of multiple decision trees improves the model generalization and performance.

A built-in library for creating random forests is available in Matlab, and relies on the same methods that were used for creating the decision trees in the model developed in [22]. Random forests can thus be tested as another method for training the predictive models, and possibly improving their performance as well.

Support Vector Machines and Random Forests are thus the two selected training methods to be implemented in this study. This leads to the formulation of the following hypothesis:

Hypothesis #4: If Kernel methods or Random forests are used for training the model, the model performance will be improved.

2.2 Model validation

In [22], tools were developed for ”on-track” validation: the vertical cloud profile was predicted for locations which would be on the satellite orbital track for the following day. The predicted profiles were then compared to the real ones, which provided an “on-track” assessment of the model performance. However, as of now, no study has been achieved on the models performance at “off-track” locations. This statement has led to the following research question (Research Question #3): What is the best approach to validate the predicted models “off-track”?

Additionally, no in-context study has been achieved in terms of model accuracy. The model performance has been assessed using objective measures, such as the MCC, but not by taking into account the general objective to which this study pertains: generating 3-D cloud field domains in order to assess the representations of clouds in GCMs. This statement has led to the formulation of the following research questions:

- Research Question #4: What level of accuracy is required from the predictive models to generate 3D cloud fields?
- Research Question #4.1: What approach should be undertaken to determine a satisfactory level of model accuracy?

The following sections outline the methods and approaches that will be implemented in this study to address these research questions.

2.2.1 Prediction at off-track locations

In order to evaluate the model performance at off-track locations by direct comparison, the predicted vertical profiles have to be computed at these locations, and compared to existing vertical profiles at the same locations. Two types of dataset thus have to be identified:

- one providing the predictors in order to compute the profiles at off-track locations with the predictive model

- one containing original profiles at the same off-track locations to enable comparison

Such datasets also need to have the same resolution as the one of the datasets on which the model was trained, which is of about $1km^2$. Indeed, if the resolution is lower, the prediction quality will be impacted, as shown by *Huguenin et al.* in [22].

The first type of dataset can be provided by the *Level 1 and Atmospheres Archive and Distribution System* (LAADS) MODIS datasets [36]. Among the LAADS distribution, two specific datasets can be used to provide the predictors needed for performing the predictive model at off-track locations:

- the MOD02 dataset [37, 38], which contains the observed radiances
- the MOD03 dataset [39], which provides the Geolocation of the readings for the whole distribution

Both datasets have a 1 km^2 resolution, just as the training datasets. They feature readings at off-track locations, up to 500km on each side of the track. The different predictors can all be extracted from these datasets, and correspond to the ones used in the on-track, training part, as shown in Table 2.3. As the MERRA-2 dataset is available on a global scale, the atmospheric features taken from this dataset can be reused for computing the profiles at off-track locations.

Table 2.3: Corresponding predictors on and off-track

On-track predictors	Off-track predictors
Geolocation (from GEOPROF-LIDAR)	Featured in MOD02 and MOD03
Elevation (from GEOPROF-LIDAR)	Available in MOD03 (Height variable)
Surface Type (from PRECIP-COLUMN)	Available in the Land/SeaMask feature in MOD03
Radiances (from MODIS-AUX)	Available in MOD02, for the same bands as in the MODIS-AUX dataset
Atmospheric variables (from MERRA-2)	Atmospheric variables (from MERRA-2) available on a global scale

The second type of dataset has to feature cloud vertical profiles at the same locations as in the MOD02 and MOD03 datasets, so that the predicted profiles can be compared to existing ones. The LAADS does not provide such datasets, at least on a vertical scale. The MOD35 dataset [40] provides the Cloud Mask at the same locations, but this feature is horizontal, unlike the vertical Cloud Mask from the training dataset GEOPROF-RADAR [30]. This means that this feature only provides the cloud scene as "viewed from space", with no indication of the clouds altitude. The predicted vertical profiles thus cannot be strictly evaluated against data from the LAADS, in the sense that the LAADS distribution does not contain comparable **vertical** data.

Following the vertical profile and globality requirements, the Clouds and the Earth's Radiant Energy System (CERES) introduced in the first chapter appears as a potential dataset provider. CERES is based on physical relationships from which several cloud products can be obtained. The CERES products are the only such 3D global cloud datasets that can be freely accessed and obtained online. These cloud products have various spatial and temporal resolutions: they are available at the footprint level, which means at the same time and location the satellite sensor captures the data (*i.e.* at on-track locations); and available as well at a more global level, the data being averaged temporally over hours, days or months, and spatially, over 1°latitude x 1°longitude areas. The global dataset with the highest temporal and spatial resolution has been identified as the SYN1deg dataset [41]. Its readings are given hourly, for the whole globe [42], and with a spatial resolution of 1°latitude x 1°longitude, which corresponds to cells with a width of about 100 kilometers. This resolution is too low when compared to the one from the training and the LAADS datasets to enable a proper comparison between the computed profiles and the SYN1deg profiles. Indeed, the resolution of the computed profiles would have do be downscaled, which means that the profiles would have to be averaged over the greater cells of the lower resolution. As the profiles are discretized in binary values, "averaging" them would not make much sense.

Thus, there is no currently available dataset that provides vertical cloud profiles at off-track locations with the same resolution as the predicted profiles. Since a direct comparison of the predicted profiles to an existing dataset cannot be achieved, the model validation on a global scale will have to be achieved differently. Still, using the MOD02 and MOD03 datasets, the profiles can be computed at these off-track locations, and their coherence can be checked through the analysis of the predicted cloud amount as compared to on-track data: if the cloud percent predicted in each band is consistent with the cloud percent in actual samples, then the computed profiles can be considered as coherent. Additionally, as outlined above, the LAADS distribution contains the MOD35 dataset, which features horizontal profiles available at the same locations as the computed profiles. Thus, by summing up the computed profiles vertically, horizontal computed profiles can be obtained and compared to the MOD35 Cloud Mask. This would allow for the **horizontal** validation of the computed profiles. This leads to the formulation of the following hypothesis:

Hypothesis #5: If the improved predictive model is implemented at off-track locations with adequate predictors, then a coherent global 3D cloud field dataset can be generated.

2.2.2 Parameterization validation: performing radiative transfer code

The general objective to which this work pertains is to generate 3-D cloud field domains in order to assess the representations of clouds in GCMs. As stated earlier, cloud-radiative forcing is difficult to model in GCMs, so the representation of this phenomenon in GCMs should be evaluated against the cloud-radiative forcing derived from the generated 3D cloud field domains. The cloud-radiative forcing associated with the 3D domains should thus be computed, and validated against existing values. This would provide a more specific, in-context assessment of the required predictive models fidelity.

As discussed below, computing the cloud-radiative forcing characteristic values associated with the predicted profiles would provide unparameterized values to be compared to existing ones. While the predicted cloud vertical profiles are parameterized and discretized,

the computed total forcing would provide a single horizontal value at each location, regardless of the parameterization. Computing the cloud-radiative forcing associated with the profiles could thus help evaluate the vertical profile parameterization performed by the models.

Indeed, as developed earlier, the predictive model is based on a profile parameterization: the 10-band model. Through this parameterization, the cloud profiles are transformed from a continuous Cloud Fraction to discrete binary values. In previous studies, it has been shown that this parameterization does a good job at representing the majority of the various shapes of cloud over the globe and their vertical locations, but the precision loss it engenders has not been assessed. The predicted vertical profiles have been compared to the parameterized original ones, and not the continuous ones. The model performance has thus been evaluated without assessing the impact of this parameterization.

Because the predictive models have been so far evaluated without taking into account the context of GCMs cloud representation, the impact that such parameterization has on this representation is unknown. Analyzing the model performance in this context could thus help evaluate the model as a whole, taking into account the fact that the profiles have been parameterized, and compare the results to original, non-parameterized ones.

Such analysis can be performed using Radiative Transfer (RT) codes. RT codes compute the radiative fluxes associated with the cloud profiles. Specific characteristics of the vertical profiles, such as the number of layers and their altitude, are identified and extracted from the profiles and used as inputs to the RT code. Computing the total radiative fluxes associated with the profiles (parameterized or not) would provide single values that can be compared to real radiative fluxes, thus "relaxing" the parameterization. This comparison would help assess the fidelity of the model in the context of GCMs cloud representation, and evaluate the parameterization.

Several RT codes are available, among which are NASA's Fu-Liou code and the Intercomparison of 3D Radiation Codes (I3RC). Both codes are available online, and their

source code can be downloaded for free.

On the one hand, the Fu-Liou model [43, 44] is a RT code which has proven to be more accurate than several other RT models in terms of computation of RT fluxes [45]. The Fu-Liou is developed in Fortran 90, and is not supposed to require much computational resources to run. On the other hand, the aim of the I3RC is to bring "together the most advanced Radiative Transfer (RT) tools for cloudy atmospheres" [46]. The project is based on two main RT methods: the Spherical Harmonic Discrete Ordinate Method (SHDOM) of Evans [47] and the Monte Carlo (MC) method [48]. The I3RC is thought to be of higher fidelity than the Fu-Liou, but it requires computational resources that may be too difficult to obtain in the context of this thesis. Consequently, the Fu-Liou code will be used in this research, as it requires fewer computational resources, and still provides RT values.

Using the Fu-Liou code, the radiative fluxes associated with the discretized profiles can be computed and compared to the radiative fluxes at the same locations, as provided by the CERES products. Indeed, the SYN1deg dataset presented above contains such values, averaged over one hour, and with a 1°x 1° spatial resolution. The computed fluxes could thus be averaged over this area in order to match this lower resolution, and thus enable the comparison to existing values. This comparison should indicate whether or not the current parameterization is suited for creating a global cloud profile dataset to be used in the context of GCM cloud representation, or if more detailed, continuous profiles should be used. This leads to the formulation of the following hypothesis:

Hypothesis #6: If the radiative fluxes values are computed by running the Fu-Liou RT code on the global dataset obtained in this study, then the obtained radiative flux values will be similar to the CERES ones.

2.3 Chapter Summary

The problem and challenges that this research is attempting to address have been discussed through a thorough review of the relevant concepts, methods and past studies. This led to the formulation of several research questions and hypotheses. A synthesized view of the mapping between the Research Questions formulated in Chapter 1 and the Hypotheses formulated in this chapter are illustrated in Figure 2.2. The research questions to be answered and hypotheses to be tested shape the approach taken in this work. The following chapter details this approach, highlighting the relationship between its steps and the hypotheses formulated in this chapter.

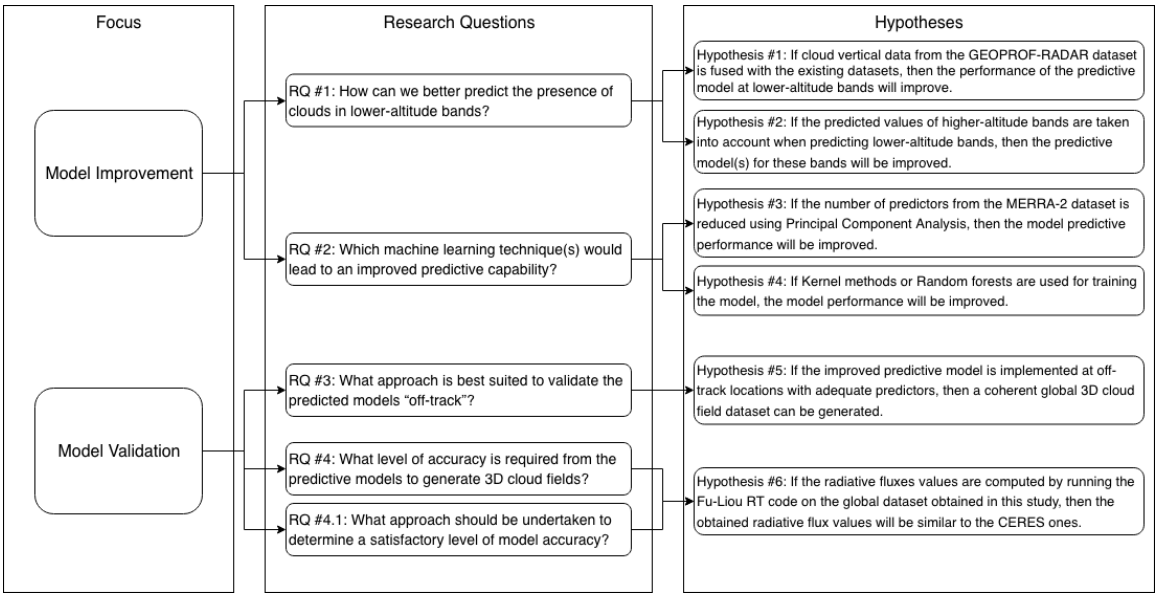

Figure 2.2: Mapping between Research Questions and Hypotheses

CHAPTER 3

PROPOSED APPROACH

This chapter details the approach proposed to test the hypotheses formulated in the previous chapter and address the research questions that motivate the present research.

3.1 General Approach

The proposed general approach is illustrated in Figure 3.1.

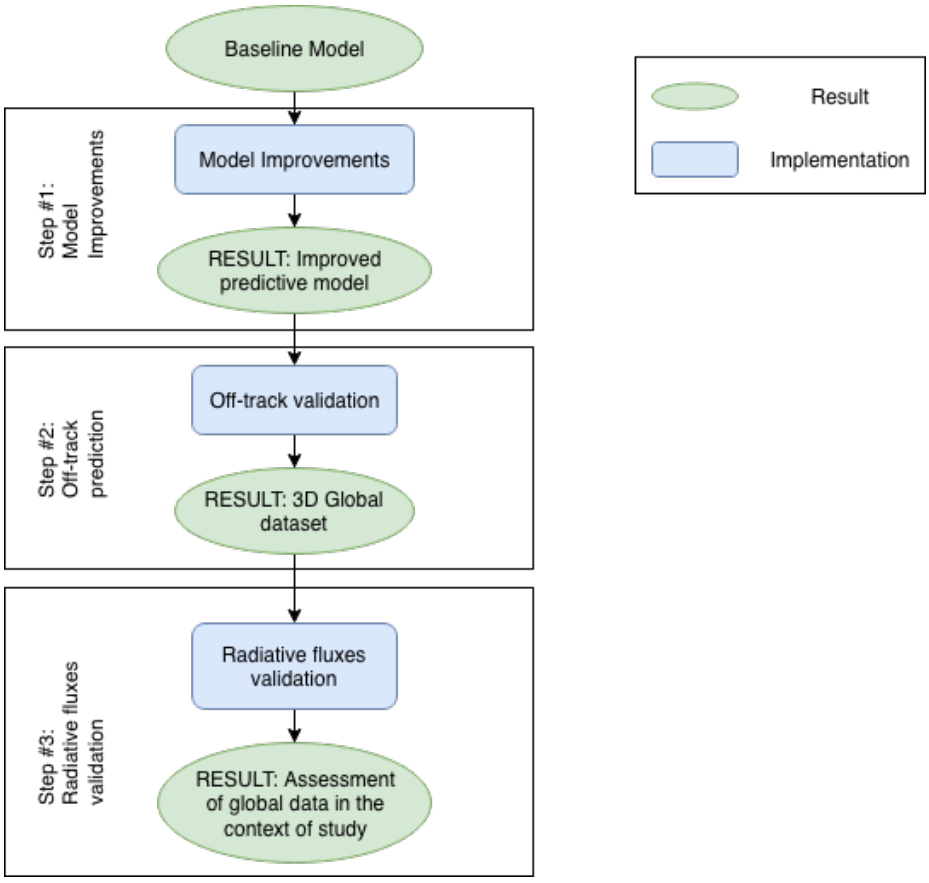

Figure 3.1: Proposed General Approach

This approach is based on what is called, for the purpose of this research, the "baseline" model, which is the model developed by *Huguenin et al.* in [22].

The first main step is to implement the improvements further detailed in Section 3.2 onto this baseline model. It is expected that this step will result in an improved predictive model, and one that will allow to test **Hypotheses #1, #2, #3** and **#4**, as defined in the previous chapter.

The second main step consists in validating the obtained improved model at "off-track" locations. This step will result in the construction of a 3D global dataset containing cloud profiles, and will allow one to test **Hypothesis #5**.

The third and last main step consists in validating the model in the context of study, by computing the radiative fluxes associated with the profiles and validating them against the corresponding fluxes from the CERES dataset. It is expected that this step will result in the assessment of the constructed 3D global dataset in the context of GCMs cloud representation, and will allow one to test **Hypothesis #6**.

The next sections discuss these three main steps in detail.

3.2 Step #1: Model Improvements

Figure 3.2 provides an overview of the approach to be undertaken to improve the baseline model. The steps of this approach can be performed in parallel, as they all correspond to a different type of improvement. For each step, the performance enhancement brought by the changes to the model has to be assessed so that each hypothesis corresponding to this step can be either validated or rejected.

3.2.1 Integration of the GEOPROF-RADAR dataset

The first step illustrated in Figure 3.2 corresponds to the integration of the GEOPROF-RADAR dataset into the predictive model. First, the structuring algorithm developed in previous efforts for processing and structuring the data has to be adapted to this dataset, so that the data can be exploitable. Next, the Cloud Fraction profiles from the GEOPROF-RADAR and LIDAR dataset have to be fused together in order to obtain one single vertical

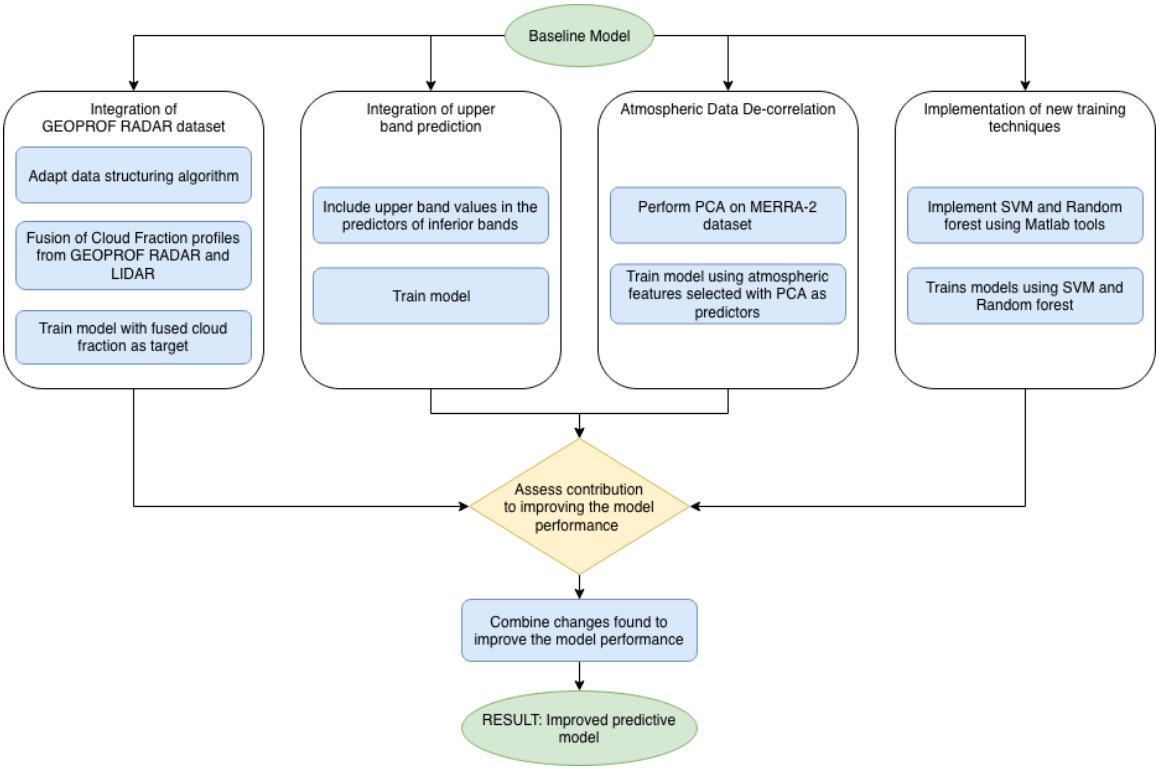

Figure 3.2: Approach for Model Improvement

Cloud Fraction feature taking into account information from both datasets. The resulting Cloud Fraction profile will then be used as a target when training the predictive model. The obtained predictive model will be assessed on "on-track" data, and such comparison will help validate or reject **Hypothesis #1**.

3.2.2 Integration of upper-bands prediction in the model

The second step corresponds to the integration of upper-bands prediction in the model. For this step, the model corresponding to each band has to be modified to account for the values of the upper bands in the predictors of the band. The resulting models will then be trained and validated against "on-track" data, and the obtained predictions will help evaluate **Hypothesis #2**.

3.2.3 Identification of independent atmospheric features

The third step consists in identifying the independent atmospheric features from the MERRA-2 dataset. Principal Component Analysis will first be performed onto these features to determine the principal components of these atmospheric variables, and the variables contributing the most to the components with the greatest variation. Such variables will then be used as the only atmospheric predictors (in addition to the non-atmospheric ones, such as elevation or surface type) to train the model. The computational time necessary to train the models will be assessed and the model validated, as a means to evaluate **Hypothesis #3**.

3.2.4 Implementation of additional machine learning techniques

The fourth step of this approach will consist in implementing additional machine learning techniques: Support Vector Machines and Random forests. Such methods will first be implemented in Matlab, using the available toolboxes featured in this environment. These methods will then be used to train and validate the predictive model. The results obtained with each method will be compared to those obtained with decision trees. This will allow for the validation or rejection of **Hypothesis #4**.

Once all steps have been performed, the changes which are found to effectively improve the model performance will be implemented for good, resulting in an improved predictive model.

3.3 Step #2: Off-track prediction

Figure 3.3 illustrates the steps to be taken in order to run the improved model obtained as a result of Step #1 at "off-track" locations and analyze the resulting dataset.

This approach uses data from the MOD02 and MOD03 datasets, as detailed in Chapter 2. First, those datasets have to be processed and fused in order to extract the predictors

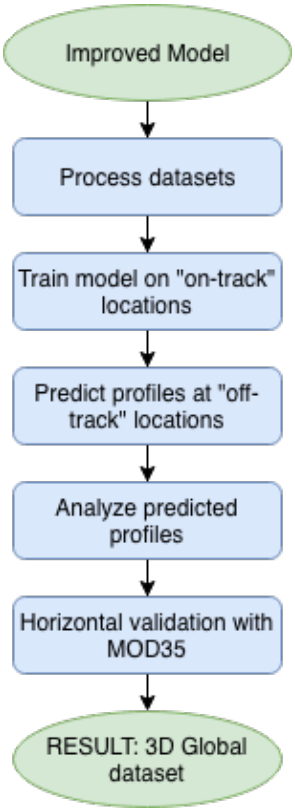

Figure 3.3: Approach for off-track prediction

needed to run the model, and combine them in a common structure. The predictive model is then trained on “on-track” locations and used to predict vertical profiles at “off-track” locations. The obtained profiles are then analyzed to ensure that they are coherent when compared to the training datasets. Some sample profiles are visualized, and the predicted amount of cloud is computed. The profiles are then summed up in order to get one single horizontal profile, and this profile is then compared to the MOD35 Cloud Mask. These steps enable the validation or rejection of **Hypothesis #5**. If **Hypothesis #5** is validated, this main step will result in a 3D global dataset containing cloud information.

3.4 Step #3: Radiative fluxes validation

Figure 3.4 illustrates the approach taken to validate the model using radiative fluxes as a mean to assess the resulting 3D global dataset developed in Step #2.

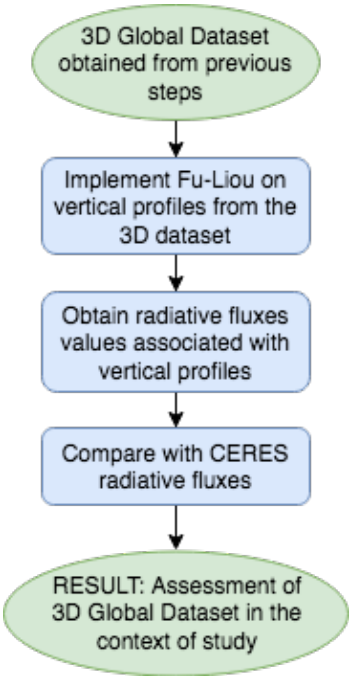

Figure 3.4: Approach for Radiative fluxes validation

Hence, the steps of this approach are based on the 3D global dataset constructed as a result of Step #2. First, the Fu-Liou radiative transfer code is implemented on the vertical profiles obtained in the 3D global dataset. The output of the Fu-Liou represents the radiative fluxes associated with these vertical profiles. These radiative flux values are then compared to the corresponding values in the CERES dataset. Doing so allows one to assess the performance of the models in terms of how accurately they represent radiative phenomena. This eventually helps validate or reject **Hypothesis #6**.

Combined altogether, these steps form the general approach to be undertaken in this research. Their implementation will allow one to validate or reject the hypotheses formulated in Chapter 2 and will eventually contribute to the general objective of this research.

The following chapters describe the implementation of each of these steps in more detail and further discuss the results they generate. A discussion regarding the validation or rejection of the different hypotheses formulated as part of this research is also provided.

CHAPTER 4
IMPROVING CLOUD VERTICAL PROFILE PREDICTIVE MODEL

This chapter details the steps taken for improving the baseline predictive model, following the approach outlined in Chapter 3. The results generated will help test **Hypotheses #1, #2, #3** and **#4**.

4.1 Integration of GEOPROF-RADAR dataset

This step is meant to test **Hypothesis #1** by developing a predictive model integrating data from the GEOPROF-RADAR dataset.

4.1.1 Vertical profiles fusion

First, the structuring algorithm used for extracting the features of interest from the different datasets must be adapted to handle the GEOPROF-RADAR dataset. Because this dataset comes from the same source as the ones already used in the baseline model, and has the same format (HDF files), size and dimensionality (14 files per day, same number of readings per file) [30] as the LIDAR dataset [25], adapting the algorithm is straightforward.

The information extracted from the RADAR dataset for the purpose of this thesis is the feature representing the cloud vertical profile *i.e.* the Cloud Mask. All other useful information, such as geolocation, is identical to the information brought by datasets already used in the baseline model. The *Cloud Mask* indicates whether or not a cloud was accurately detected by the satellite sensor. Its values range from -9 (detection error) to 40 (strong probability of detection). The documentation [30] provides Table 4.1 for interpreting the Cloud Mask values.

The documentation also indicates that for most applications using this dataset, Cloud Mask values equal to 30 and 40 should be considered as indicative of the presence of cloud,

Table 4.1: Cloud Mask values and corresponding interpretation [30]

Mask value	Interpretation	Estimated % of false detection
-9	Bad or missing radar data	
5	Significant return power but likely surface clutter	
6-10	Very weak echo	44%
20	Weak echo	5%
30	Good echo	4.3%
40	Strong echo	0.6%

while other values should be indicative of the absence of cloud at the reading location. The Cloud Mask feature can thus be considered as binary one, with 0 being assigned to values below 30 and 1 to values equal to 30 or above.

Such discrete values have to be merged with the continuous Cloud Fraction values from the LIDAR dataset. Indeed, the Cloud Fraction indicates the percent of cloud detected at the reading location, thus ranging continuously from 0 to 100. In [22], the Cloud Fraction vertical profile was discretized and split into 10 bands. Vertical readings were averaged over each band, and the average was compared to a certain condition (equal to 20) in order to determine whether or not the amount of cloud in the band was significant enough to be represented in the model. Here, the same condition (adapted to the scale) is applied to the binary Cloud Mask data. The Cloud Mask and Cloud Fraction are thus transformed into two 10-band binary vertical profiles. These two profiles are then superimposed in order to get a single vertical profile, containing information from both the LIDAR and RADAR datasets. The superimposition is achieved by assigning 0 to locations at which the binary Cloud Mask and Cloud Fraction are both equal to 0. Thus, if at one location the Cloud Mask is indicative of a cloud, but not the Cloud Fraction, then the location will be considered as cloudy in the profile generated by the superimposition. Table 4.2 shows the percent of cloud present in the vertical profiles of February 25, 2011, which is the day that has provided validation data for the models. Adding the RADAR data thus significantly increases the amount of cloud

detected: for the example in Table 4.2, about 5% of locations were wrongly indicative of cloud absence when only considering the LIDAR information. Also, if only the RADAR data would be considered, then about 5% of locations would also be wrongly indicative of cloud absence. Combining the LIDAR and RADAR data thus increases the amount of cloud in the profiles when compared to each dataset without combination.

Table 4.2: Cloud Percent in vertical profile over one day (Feb 25, 2011)

Profile	Cloud percentage
Profile with LIDAR data	13.12 %
Profile with RADAR data	13.24 %
Profile with LIDAR and RADAR data	18.13 %

Figures 4.1, 4.2 and 4.3 show the vertical profiles on the same sample set of locations, obtained respectively with the LIDAR dataset, the RADAR dataset and the combination of LIDAR and RADAR by superimposition. These sample profiles show that the additional cloud information brought by the RADAR dataset is mostly located in the lower part of the profile.

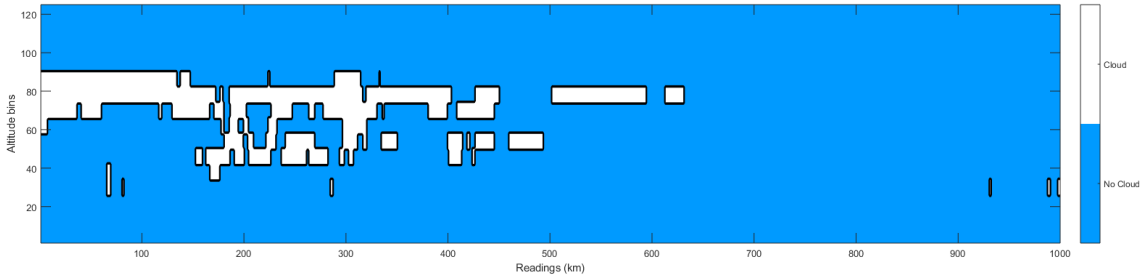

Figure 4.1: Vertical profile sample with LIDAR data

Indeed, as shown in Table 4.3, adding cloud data from the GEOPROF-RADAR dataset brings significant change to the profile starting around bands 5-6. As the RADAR sensor is more accurate for lower-altitude clouds than the LIDAR sensor, it detects more clouds in the lower bands, and thus produces a more accurate profile.

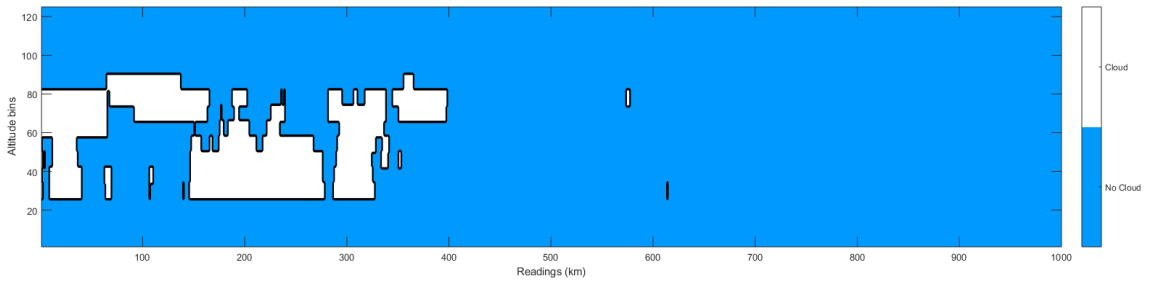

Figure 4.2: Vertical profile sample with RADAR data

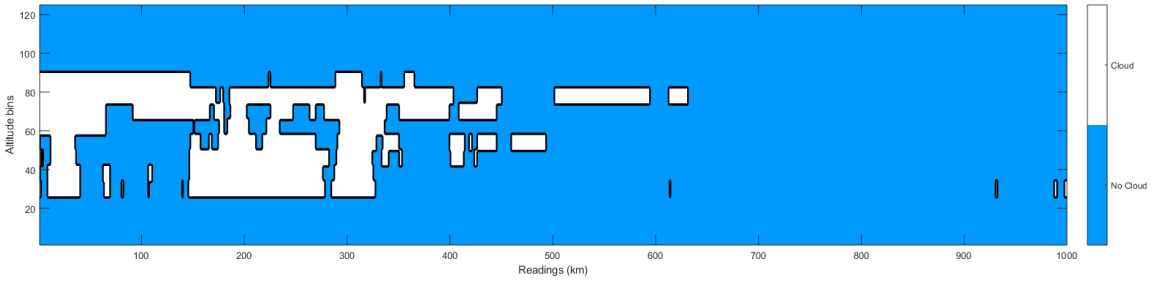

Figure 4.3: Vertical profile sample with LIDAR and RADAR data

Table 4.3: Cloud Percent in each band of vertical profile over one day (Feb 25, 2011)

Band	Cloud percentage for profile with LIDAR data	Cloud percentage for profile with RADAR data	Cloud percentage for profile with LIDAR and RADAR data
1	0.26%	0 %	0.26 %
2	6.24%	0.19%	6.25 %
3	9.29%	1.85%	9.32 %
4	10.69%	5.11%	11.05 %
5	14.34%	9.56%	15.83 %
6	16.61%	14.21%	19.51 %
7	18.68%	20.83%	24.46 %
8	17.33%	23.57%	26.50 %
9	16.96%	28.32%	32.04 %
10	20.76%	28.79%	37.93 %

4.1.2 Model training and validation

As in the previous works by *Huguenin et al.* [22], the predictive models associated with each band are trained using decision trees. The predictors are the same as the ones used in

[22], and the training and validation features are the combined binary profiles, computed over 3 days (February 22 to 24, 2011) for the training set and the 4^{th} day (February 25, 2011) for the validation set.

Matthew's Correlation Coefficient (MCC) is then computed for each band model on the validation set, and compared to the MCCs obtained with the baseline model. These scores are presented in Table 4.4.

Table 4.4: MCC scores obtained for each altitude band with LIDAR profile and LIDAR and RADAR combined profile

	Band	MCC for LIDAR profile	MCC for combined profile	Percentage of change
High altitude	1	-0.0029	-0.0027	-6.9%
	2	0.4944	0.5056	+2.27%
	3	0.4922	0.4826	- 1.95%
	4	0.4333	0.4731	+9.19%
	5	0.4089	0.4183	+2.3%
	6	0.4210	0.4506	+7.03%
	7	0.3198	0.4566	+42.78%
	8	0.2329	0.4158	+78.53%
	9	0.2200	0.3719	+69.05%
Low altitude	10	0.2821	0.4437	+57.28%

These results show that the model performance is enhanced for all bands (except the third band). Nevertheless, the performance degradation for band 3 is less than 2%, while for most other bands the performance is significantly improved, up to 57% for the last band. This performance improvement is increasingly significant from band 6, going downwards. Thus, using data with a higher accuracy for the lower bands leads to a significantly better prediction performance for these bands.

Another observation on the results is that the MCC for the first band is still very low, close to 0. This is due to the fact that there rarely are any clouds in the first, top band. As shown in Table 4.3, the cloud percentage is about 0.26% in the band. The cloud profile for this band thus has a very low number of positive values. This has an impact on the MCC

for this band, due to the fact that its computation (Equation 4.1) involves the number of true positives and false negatives as factors:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{4.1}$$

From this expression, if the total number of positives in the validation set is very low, then the number of true positives and false negatives will be very low as well, and the MCC will consequently be close to zero. The MCC is thus not a good performance measure for the top band of the model. Instead, the accuracy can be used for assessing the model performance for this altitude band (Equation 4.2):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

As shown in Table 4.5, the accuracy for the top band is already almost equal to 1 with the baseline model, using only the LIDAR profile. As there are rarely any clouds in the profile (see Table 4.3), it is not difficult for the model to predict the absence of cloud in the band. As the RADAR data does not bring any significant change to the vertical profile for the top band (Table 4.3), the predictive model produces results with a similar accuracy.

Table 4.5: Accuracy obtained for the top altitude band with LIDAR profile and LIDAR and RADAR combined profile

Accuracy for LIDAR profile	0.9941
Accuracy for combined profile	0.9946

The addition of RADAR data to the model thus does not impact the model performance for the top band. In the next sections, the model performance for the top band will be measured using accuracy rather than MCC.

The results from the model training and validation, as discussed in this section, thus validate **Hypothesis #1**: the predictive model is improved by integrating cloud vertical

data from the GEO-PROF RADAR dataset.

4.2 Integration of upper-band prediction

This step is meant to test **Hypothesis #2** by taking into account information about the presence of clouds within upper bands when predicting lower bands values.

As discussed, the baseline model developed in [22] is composed of 10 independent sub-models, each associated with one band, in order to model several superimposed layers of clouds, as pictured in 4.4. However, in reality, cloud layers are not independent from each other, as superimposed layers may have an effect on several features, such as radiances or atmospheric features, which are used as predictors in the built predictive model.

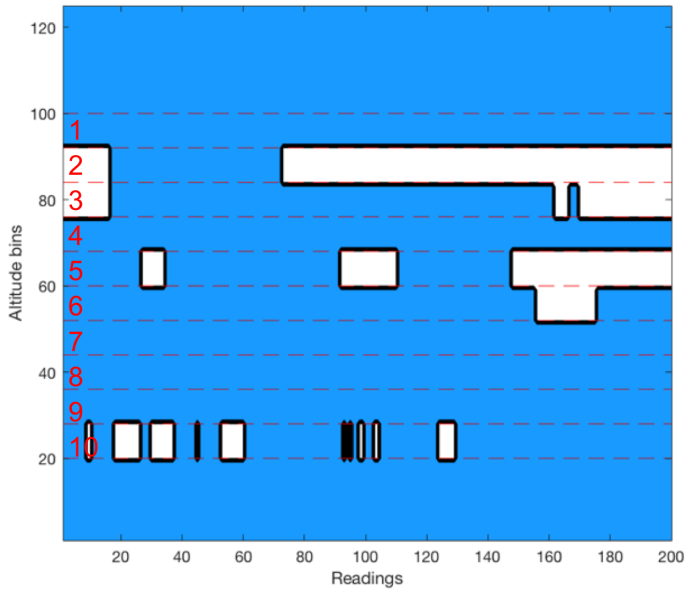

Figure 4.4: Example of vertical profile discretized in 10 bands

In order to alleviate the independence of the band models, the values associated with upper bands can be taken into account when predicting cloud profiles in lower-altitude bands. For example, the value in band 5 is predicted with the currently-used predictors, as well as the values obtained in bands 2, 3 and 4. Band 1 is voluntarily omitted, as the very low quantity of cloud in the band would not bring much information to the models associated with lower bands.

The following solution is implemented onto the baseline model: the predictors matrix is modified for each band, so that it takes into account the values of the bands above. Then, the resulting model is trained, similar to the baseline model, using decision trees. The validation is performed over the same profile as before (February 25, 2011). The MCC is computed for each band, and compared to the ones obtained with the baseline model, as presented in Table 4.6.

Table 4.6: MCC scores obtained for each altitude band with baseline model and with model including upper-band prediction

Band	MCC for basis model	MCC with upper band prediction	Percentage of change
2	0.4944	0.4944	0%
3	0.4922	0.4696	-4.59%
4	0.4333	0.4110	-5.15%
5	0.4089	0.3812	-6.77%
6	0.4210	0.4018	-4.56%
7	0.3198	0.3045	-4.78%
8	0.2329	0.2434	+4.51%
9	0.2200	0.1971	-10.41%
10	0.2821	0.2570	-8.90%

For most bands, the performance is slightly lower than the one obtained with the baseline model. There is only improvement for the 8^{th} band, and this improvement is not significant. This solution is thus inefficient at improving the predictive capability of the model. The same data was also used with Random Forests in lieu of Decision Trees, as discussed later in this chapter. The use of Random Forests did not lead to any performance improvement either. For these two techniques, although the performance for the top bands is good, it is not high enough for the predictions to be good predictors for the lower bands.

Therefore, the proposed solution is not suitable for this specific application, and will consequently not be implemented in the improved model. These results provide ground for rejecting **Hypothesis #2**.

4.3 Identification of influential atmospheric features

This step is meant to test **Hypothesis #3** and consists in performing Principal Component Analysis (PCA) on the MERRA-2 dataset in order to identify independent atmospheric features.

4.3.1 PCA implementation on MERRA-2 dataset

When developing the baseline model [22], a correlation study of the atmospheric variables contained in the MERRA-2 dataset was performed in order to remove dependent features from the list of predictors, as unnecessary predictors impede on the model's performance and computational time to train. Through this study [22], relationships between the atmospheric features were identified visually. This allowed for the identification of 18 dependent features, which consequently led to a reduction in the number of atmospheric predictors from 47 to 29.

However, this study did not allow for the identification of non-visual relationships. A more thorough analysis, one that leverages Principal Component Analysis (PCA), can be performed in order to identify such relationships. PCA extracts the important information from the inter-correlated variables, and represents it as a set of new orthogonal variables called principal components, and which are ordered so that the first few retain most of the variation present in all of the original variables [49]. This method is performed here on the MERRA-2 dataset [28] used in [22], which brings the atmospheric features to the predictive model.

PCA is implemented using the *pca* function available in Matlab 2017 [50]. This function returns the principal components (PCs) corresponding to the input dataset, as well as the associated variation of each PC. The PCs are expressed in terms of the atmospheric features of the dataset, so the contribution of each feature to each of the 47 principal components is known.

Because the intent is to identify the "most independent" atmospheric features in the MERRA-2 dataset, the total contribution of each feature to the variables of each PC needs to be computed. The total contribution of a feature is defined as follows:

$$TC = \sum_{i} |variance\ of\ PC\ i| * (contribution\ of\ feature\ to\ component\ i) \quad (4.3)$$

Once the total contributions are computed for all features, they can be ordered from the most to the least "influential" on the dataset. The ranking of the first 30 features is presented in Table 4.7.

From this ranking, the most influential features are the ones associated with pressure data, followed by those associated with temperature and then wind. The features can then be used as predictors in the cloud profile predictive model, according to their importance.

4.3.2 Model training and validation

Following the ranking established in the previous section, the model is trained using different numbers of predictors: 8, 12, 16, 20, 24 and 30 most influential features. The different models obtained are then validated against the same dataset as before. The resulting MCC scores for each band of the model are presented in Tables 4.8 and 4.9, and plotted in Figure 4.5.

From Figure 4.5, it appears that no ideal number of predictors stands out for all bands: the number of predictors maximizing the MCC for one band may minimize the MCC of another band, or at least not bring the best score. The number of atmospheric predictors thus cannot be set to the same value for all bands.

However, several groups of bands can be identified:

- The top bands (1 to 3) show a better performance for about 24 predictors
- The middle bands (4 to 7) perform better for a low number of predictors
- The bottom bands (8 to 10) perform better with a greater number of predictors

Table 4.7: Ranking of the first 30 atmospheric features by contribution to the PCs

Rank	Atmospheric Feature
1	Cloud Top Pressure (CLDPRS)
2	Tropopause Pressure based on blended estimate (TROPPB)
3	Tropopause Pressure based on thermal estimate (TROPPT)
4	Planetary Boundary Layer top pressure (PBLTOP)
5	Surface pressure (PS)
6	Sea level pressure (SLP)
7	Height at 250 hPa (H250)
8	Height at 500 hPa (H500)
9	Lifting condensation level (ZLCL)
10	Height at 850 hPa (H850)
11	Height at 1000 hPa (H1000)
12	Total column ozone (TO3)
13	Cloud Top Temperature (CLDTMP)
14	Surface skin temperature (TS)
15	2-meter air temperature (T2M)
16	10-meter air temperature (T10M)
17	Total precipitable water vapor (TQV)
18	Dew point temperature at 2 m (T2MDEW)
19	Wet bulb temperature at 2 m (T2MWET)
20	Air temperature at 850 hPa (T850)
21	Air temperature at 500 hPa (T500)
22	Tropopause temperature using blended TROPP estimate (TROPT)
23	Air temperature at 250 hPa (T250)
24	Eastward wind at 250 hPa (U250)
25	Eastward wind at 500 hPa (U500)
26	Eastward wind at 850 hPa (U850)
27	Northward wind at 250 hPa (V250)
28	Eastward wind at 50 meters (U50M)
29	10-meter Eastward wind (U10M)
30	2-meter Eastward wind(U2M)

Table 4.8: MCC scores obtained with different numbers of atmospheric predictors

Band \ Number of atmospheric features	Baseline	8 features	12 features	16 features	20 features	24 features	30 features
2	0.4944	0.4608	0.4168	0.4117	0.3601	0.4935	0.4843
3	0.4922	0.4912	0.4706	0.4703	0.5069	0.4976	0.4829
4	0.4333	0.4295	0.4094	0.3894	0.3807	0.4100	0.3980
5	0.4089	0.3786	0.3750	0.3908	0.3789	0.3865	0.3703
6	0.4210	0.4029	0.3452	0.3772	0.3859	0.4294	0.3865
7	0.3198	0.3364	0.3175	0.3114	0.2990	0.3143	0.2859
8	0.2329	0.2193	0.2593	0.2384	0.2214	0.2274	0.2274
9	0.2200	0.2051	0.2246	0.2312	0.2443	0.2319	0.2403
10	0.2821	0.2724	0.2798	0.2945	0.2837	0.2918	0.3028

Table 4.9: Percentages of change in MCCs obtained with different numbers of atmospheric predictors when compared to baseline

Band \ Number of atmospheric features	8 features	12 features	16 features	20 features	24 features	30 features
2	-6.79%	-15.70%	-16.72%	-27.16%	-0.18%	-2.04%
3	-0.20%	-4.39%	-4.45%	+2.97%	+1.10%	-1.89%
4	-0.88%	-5.52%	-10.13%	-12.14%	-5.38%	-8.15%
5	-7.41%	-8.29%	-4.43%	-7.34%	-5.48%	-9.44%
6	-4.30%	-18.00%	-10.40%	-8.38%	+1.99%	-8.19%
7	+5.19%	-0.72%	-2.63%	-6.50%	-1.72%	-10.6%
8	-5.84%	+11.34%	+2.36%	-4.94%	-2.36%	-2.36%
9	-6.77%	+2.09%	+5.09%	+11.05%	+5.41%	+9.23%
10	-3.44%	-0.81%	+4.40%	+0.57 %	+3.44%	7.34%

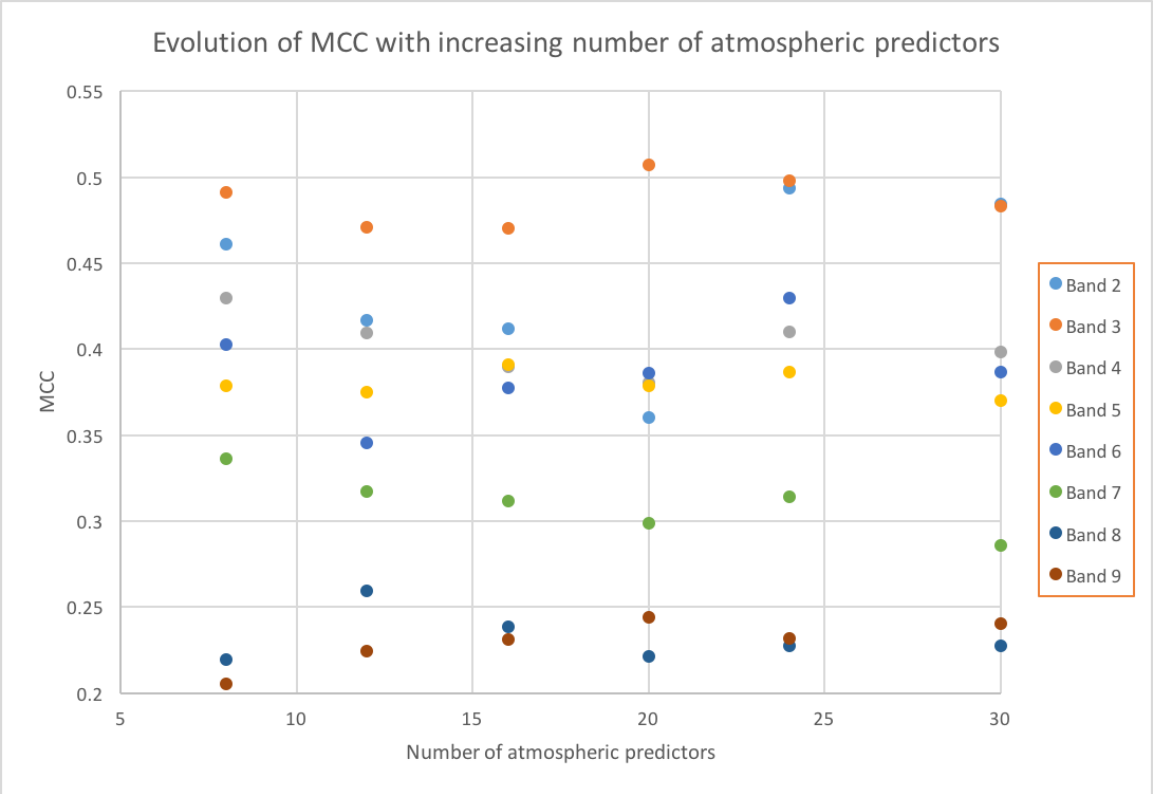

Figure 4.5: Evolution of MCC with the number of atmospheric predictors for each band of the model

Following this observation, the number of atmospheric predictors for each band can be chosen accordingly, as presented in Table 4.10.

Thus, the number of atmospheric features used as predictors can be reduced for the top and middle bands, as compared to previous works. These features are selected thanks to PCA, which validates **Hypothesis #3**: if the number of predictors from the MERRA-2 dataset is reduced using Principal Component Analysis, the model predictive performance will be improved.

4.4 Implementation of additional training techniques

This step is meant to test **Hypothesis #4** and consists in implementing additional techniques to train the cloud profile predictive model.

In [22], classification decisions trees were used and proved efficient when compared

Table 4.10: Chosen number of atmospheric predictors for each band of the model

Band	Chosen number of atmospheric predictors
1	24
2	24
3	24
4	8
5	8
6	8
7	8
8	24
9	24
10	24

to classification neural networks. However, as outlined in Chapter 2, other methods are specifically built for such classification problems. Among these methods, Kernel methods (more specifically Support Vector Machines) and Random forests have been identified as promising training techniques.

In this section, both methods are implemented to train the model and evaluated.

4.4.1 Support Vector Machines

The Matlab pre-implemented functions *fitcsvm* and *predict* are used for the implementation of SVMs.

For this application, SVMs proved very slow to train: it took several hours for training only one machine, *i.e.* obtaining a model for one single band. Additionally, the performance obtained was very low, as the MCCs were very close to 0 for all bands. This result is quite surprising, as SVMs were thought to be a well-suited training techniques for this application. Their very low performance may be due to the computing environment in which the predictive models are run. Indeed, the models are implemented and trained using Matlab, which was chosen for its user-friendliness and is used as a standard in the research community and in particular by the NASA Langley Research Center, who introduced the research objective for this thesis. However, as can be hinted by the high time of

training, it appears that Matlab is not performing well at training the models using SVMs. This could come from the fact that SVMs have a very low speed of learning with respect to the number of predictors and samples [32], and they are trained here on a high number of predictors and samples. Using an environment with a better computational speed (e.g. Python) in future efforts could thus lead to different results, most likely better than the ones obtained with Matlab.

For the scope of this research, SVMs are not considered as a suitable training technique for this specific problem with the environment used, and will not be implemented in the final improved model.

4.4.2 Random Forests

A Random Forest is a set of multiple decision trees trained on random subsets of the model predictors. They tend to correct the decision trees tendency to overfit to the training set, and are thus expected to provide better results than the ones obtained with one decision tree, *i.e.* with the baseline model.

Matlab toolboxes and pre-implemented functions are also used for implementing them: the *TreeBagger* tool in particular, is used for building the forests. The model is trained for different forest sizes, *i.e.* different numbers of trees in the forest, in order to determine an optimal size for the forest in terms of predictive performance and computational time. The results obtained are presented in Table 4.11, compared to the baseline results in Table 4.12, and plotted in Figure 4.6.

From Figure 4.6 and Table 4.12, the use of Random Forests for training the model leads to a significant improvement for all bands of the model. It can be noted that there is a high evolution of the performance between 1 and 20 trees in the random forest, while the performance is almost stagnating between 20 and 50 trees. As the computational time of training the model is directly linked to the size of the forest, and the performance improvement is not significant beyond 20 trees in the forest, the forest size should be set to 20 in order to

Table 4.11: MCC scores obtained for each band with different forest sizes

Band \ Forest size	1 tree (Baseline model)	5 trees	10 trees	20 trees	50 trees
2	0.4944	0.5530	0.5254	0.5483	0.5700
3	0.4922	0.5667	0.5823	0.5955	0.6175
4	0.4333	0.5059	0.5202	0.5492	0.5557
5	0.4089	0.4543	0.4612	0.4729	0.4642
6	0.4210	0.5041	0.5269	0.5537	0.5636
7	0.3198	0.4360	0.4205	0.4257	0.4580
8	0.2329	0.2974	0.3239	0.3528	0.3585
9	0.2200	0.2563	0.2962	0.3176	0.3036
10	0.2821	0.3466	0.3729	0.3833	0.3968

maximize performance while limiting computing cost.

From these observations, Random Forests have proved to be very efficient at enhancing the performance of the predictive model, when compared to the baseline model. This provides ground for validating **Hypothesis #4**: if Random Forests are used for training the model, the model performance will be improved. The Kernel methods are no longer considered in this hypothesis, when compared to its initial formulation provided in Chapter 2.

4.5 Improved model

The previous sections have supported the validation or rejection the first four hypotheses defined in Chapter 2:

- **Hypothesis #1** is validated: If cloud vertical data from the GEOPROF-RADAR dataset is fused with the existing datasets, then the performance of the predictive model at lower-altitude bands is improved.
- **Hypothesis #2** is rejected: If the predicted values of higher-altitude bands are taken into account when predicting lower-altitude bands, then the predictive model(s) for

Table 4.12: Percentages of change in MCCs obtained for each band with different forest sizes, as compared to baseline model

Band \ Forest size	5 trees	10 trees	20 trees	50 trees
2	11.87	6.29	10.92	15.31
3	15.15	18.32	20.99	25.45
4	16.75	20.04	26.75	28.24
5	11.11	12.81	15.66	13.54
6	19.74	25.18	31.54	33.88
7	36.36	31.49	33.14	43.24
8	27.68	39.08	51.47	53.90
9	16.52	34.62	44.38	37.99
10	22.89	32.22	35.88	40.68

these bands are not improved.

- **Hypothesis #3** is validated: If the number of predictors from the MERRA-2 dataset is reduced using Principal Component Analysis, then the model predictive performance is improved.

- **Hypothesis #4** is validated: If Random forests are used for training the model, the model performance is improved.

Following these results, an improved predictive model can be built, synthesizing the methods that contributed to enhancing the model's performance. The final predictive model is thus composed of the baseline model, to which the following improvements are added:

- Integrating GEO-PROF RADAR data to the lower bands of the model

- Using 3 different groups of atmospheric predictors:

  - 24 atmospheric predictors for bands 1 to 3

  - 8 atmospheric predictors for bands 4 to 7

  - 24 atmospheric predictors for bands 8 to 10

- Training the model using a Random Forest of 20 trees

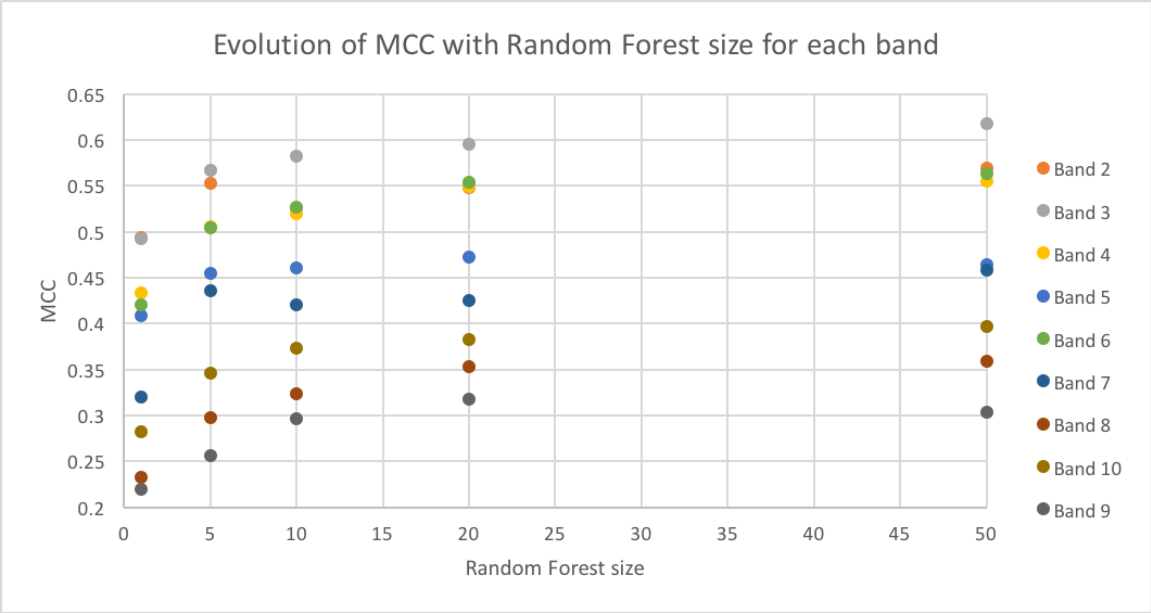

Figure 4.6: Evolution of MCC with forest size for each band of the model

This improved model is validated against the same dataset previously used for validation, and the resulting scores are compared to the ones obtained with the baseline model, as shown in Table 4.13.

As stated in section 5.1, the MCC is not a good performance measure for the top band, because of cloud scarcity in the band, and thus scarcity of positive points. The accuracy is used for measuring the performance in the top band, and is presented in Table 4.14.

From the results featured in Tables 4.13 and 4.14, the performance is improved for all bands when compared to the baseline model. Additionally, the performance obtained by combining the different changes made to the model is superior to the performance obtained with any single change made to the model taken individually, as presented in the previous sections. Figures 4.7 and 4.8 contrast a sample real profile and the corresponding one predicted by the improved model. Figures 4.9 and 4.10 illustrate the prediction performance improvement between the two models.

The work detailed in this chapter has thus led to the development of an improved predictive model, and provided answers to **Research Questions #1** and **#2**. The improved predictive model will be used in the next steps, in order to investigate **Research Questions**

Table 4.13: MCC scores obtained for each altitude band with basis model and with improved model

Band	MCC for baseline model	MCC for improved model	Percentage of change between baseline and improved models
2	0.4944	0.5857	18.47%
3	0.4922	0.5674	15.28%
4	0.4333	0.5594	29.10%
5	0.4089	0.5203	27.24%
6	0.4210	0.5750	36.58%
7	0.3198	0.5657	76.89%
8	0.2329	0.5372	130.7%
9	0.2200	0.5278	139.9%
10	0.2821	0.5570	97.44%

Table 4.14: Accuracy obtained for the top altitude band with the baseline model and with the improved model

Accuracy with baseline model	0.9941
Accuracy with improved model	0.9971

#3 and **#4** and the corresponding **Hypotheses #5** and **#6**.

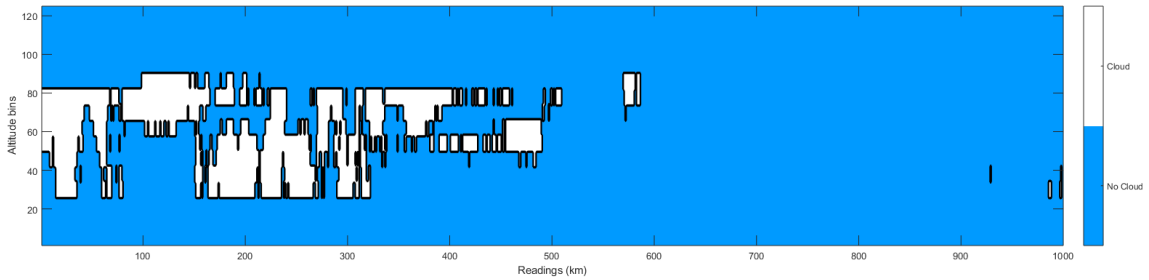

Figure 4.7: Real profile

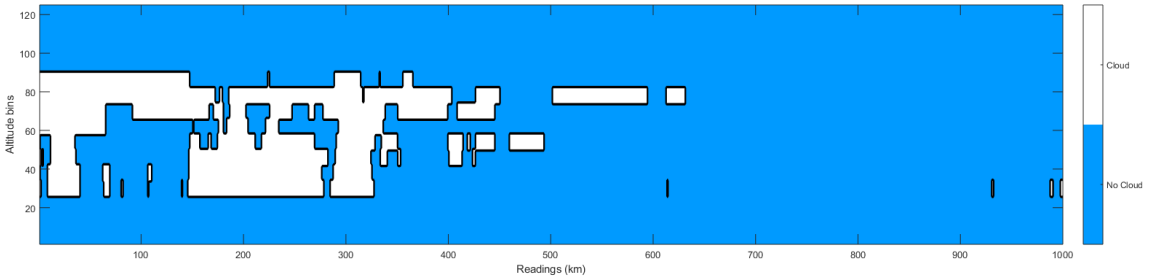

Figure 4.8: Predicted profile with improved model

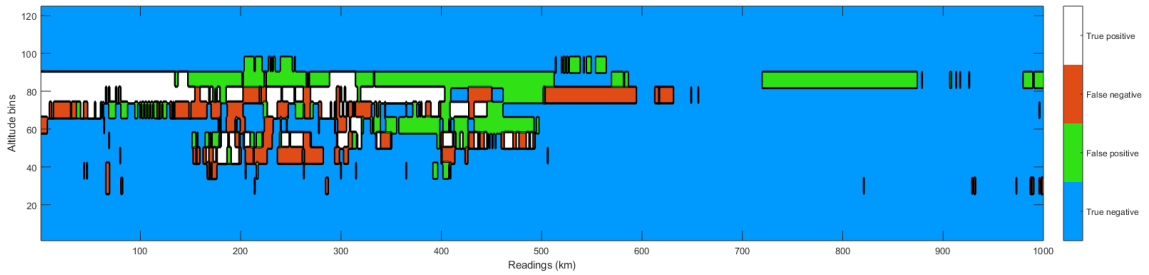

Figure 4.9: Performance visualization of baseline model

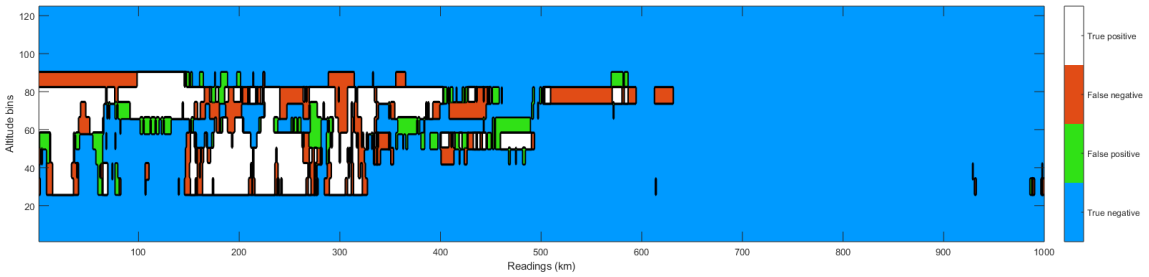

Figure 4.10: Performance visualization of improved model

CHAPTER 5
OFF-TRACK PREDICTION

This chapter details the steps taken for implementing the model developed in Chapter 4 at off-track locations, following the approach outlined in Chapter 3. The results obtained will help validate (or reject) **Hypothesis #5**.

The main goal of this step is to construct a 3D Global dataset containing cloud vertical profiles at off-track locations, using the improved predictive model developed in Step #1 of the proposed approach. For this purpose, predictors have to be extracted from the new datasets (MOD02 and MOD03), so that the model can then be performed on these predictors. The resulting profiles will then be analyzed, in order to check the consistency of the 3D global dataset constructed with these profiles when compared to existing cloud profile datasets. Last, horizontal validation of the computed profiles against the MOD35 Cloud Mask feature will be performed, in order to further assess the coherence of the computed dataset when compared to another existing one.

5.1 Datasets

The datasets identified in Chapter 2 as potential providers for the needed predictors at off-track locations are the MOD02 [37, 38] and the MOD03 [39] datasets, as well as the MERRA-2 [28] dataset.

Indeed, the LAADS datasets (MOD02, MOD03) and the MERRA-2 dataset contain the predictors needed for running the predictive model, as stated in Chapter 2 (see Table 2.2).

The LAADS datasets contain data with a 1 km^2 resolution, similar to the previous datasets that were used for training the model. However, it contains observations spread out on more than 2,000 kilometers cross-track, while the previous datasets only contained one observation per reading. As shown in Figures 5.1 and 5.2, there is much more data

available in the LAADS datasets, due to the fact that the observations are also available for off-track locations.

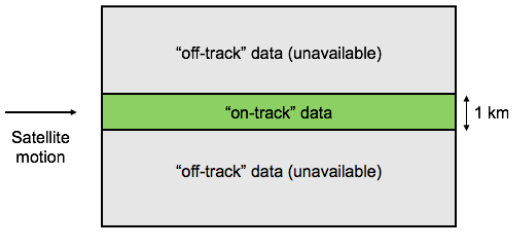

Figure 5.1: Training data available

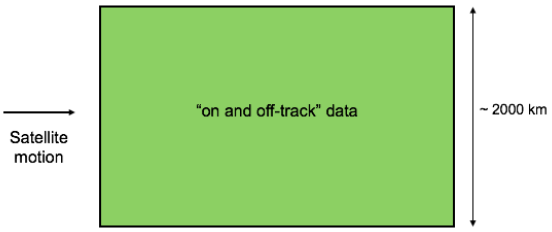

Figure 5.2: MOD02 and MOD03 data available

As shown in Table 5.1, the datasets required to construct the 3D dataset have very different formats and resolutions when compared to the datasets that were used to train and validate the predictive model in Step #1.

Table 5.1: Datasets specifications

	Datasets used in Step 1	MOD02 and MOD03	MERRA-2
File format	HDF	HDF	NC4
Spatial resolution	1 km^2	1 km^2	1/2° lat x 2/3 ° long (about 4,000 km^2)
Number of observations per reading	1	1355	360x180
Temporal resolution	14 files/day	288 files/day (roughly one file per 5 minutes of data acquisition)	24 files/day
Data size	about 40 MB/day	about 34 GB/day	about 500MB/-day

A solution for adapting the MERRA-2 atmospheric features to the higher resolution of the other training datasets has been developed by *Huguenin et al.* [22], so that the atmospheric features could be merged with other predictors in a common matrix of predictors. The main challenge in terms of data processing is to adapt the MOD02 and MOD03

datasets to match the format of the datasets used for training. Consequently, the data fusion algorithm previously developed in [22] has to be adapted.

5.2 Data Fusion

The aim of the Data Fusion (DF) algorithm is to create a single set of predictors to be used as an input to the predictive model, by extracting the features of interest from the different datasets, and adapting them to the highest resolution. It has been first developed by *Ngo* [21] and enhanced by *Huguenin et al.* in [22].

Here, the first change that has to be made to the DF algorithm is to adapt it to handle 3-dimensional datasets, as each feature has several observations for each reading (see Table 5.1). In previous works, the datasets were 2-dimensional has they contained only one observation per reading. Switching from a 2-dimensional structure to a 3-dimensional structure does not cause much difficulty in terms of coding, but does increase the computational time and memory use.

Similar to what was done in [22], features from the MERRA-2 dataset are fused with the new datasets by taking the "closest atmospheric condition" to each observation.

As the radiances in the MOD02 dataset are captured from different angles over the cross-track width, the values have to be corrected according to the capturing angle. The documentation on the MOD02 [38] implies that this correction has already been performed, and that the only correction that remains to be performed is the one using radiance scales and offsets, which are given in the MOD02 dataset. The access to the code for performing this correction is given in Appendix A.

Finally, the input features are resampled on the along-track and on the cross-track axes, in order to limit computational time and memory use. Indeed, computing the full-resolution profiles led to a memory error in Matlab. In order to obtain a dataset with locations equally spread out, and still with a high resolution, the locations are resampled and one out of five is kept. The resulting resolution is about 5 km on the along-track axis, and about 5 km on

the cross-track axis.

The output of the DF algorithm applied to the MOD02, MOD03 and MERRA-2 datasets is thus a single set of predictors at off-track locations, equivalent to one day of data acquisition (February 25, 2011).

5.3 Training and Prediction

Once the global set of predictors is built, the predictive models are trained on the on-track data, similar to what was done in Step #1. The predictive models are thus not trained on the off-track data, but only used to predict the cloud profiles at these locations for the purpose of validating the profiles at off-track locations.

The predictive models are then run using the set of predictors at off-track locations as input to the models. The output of the models is thus the cloud presence in each band of the vertical profile at off-track locations. Combining these outputs leads to a 3-dimensional dataset representing the predicted vertical cloud scene at both on- and off-track locations, for one day of data acquisition.

5.4 Analysis of dataset coherence

Once the 3D global dataset containing the vertical profiles is built, those profiles have to be analyzed in order to assess their coherence, *i.e.* their consistency with existing profiles. As stated in Chapter 2, the profiles cannot be validated against an existing dataset of vertical profiles, as the only available datasets do not have the required resolution for direct comparison with the dataset built in this work. Consequently, the computed dataset cannot be validated in the strict sense, but the consistency of the profiles can still be checked in several ways, in order to ensure that the computed profiles "look like" clouds.

First, the amount of cloud in each band predicted by the model can be compared to the amount of cloud in existing profiles, with the same conditions, *i.e.* for the same day of acquisition. Table 5.2 shows the percent of cloud present in each band, for the on-track,

available profiles taken from the GEOPROF-LIDAR and RADAR datasets as detailed in Chapter 4, and for the off-track computed profiles.

Table 5.2: Cloud Percent in each band of vertical profile over one day (Feb 25 2011)

Band	Cloud percentage for on-track profiles (from GEOPROF-LIDAR and RADAR)	Cloud percentage for on and off-track computed profiles
1	0.26 %	0 %
2	6.25 %	4.02 %
3	9.32 %	6.56 %
4	11.05 %	8.2 %
5	15.83 %	13.46 %
6	19.51 %	15.52 %
7	24.46 %	17.61 %
8	26.50 %	13.94 %
9	32.04 %	13.84 %
10	37.93 %	35.27%

On the whole, the cloud presence percentages in each band for the predicted profiles are of the same magnitude order as the percentages for the actual on-track profiles. It can be noted that the predicted profiles percentage is lower than the on-track percentage for all bands, and especially for bands 8 and 9. However, these figures do not allow for direct comparison, as the amount of data is not the same: there are a lot more target locations for predicted profiles than there are for actual on-track ones. Additionally, the conditions (atmospheric, surface type) may vary slightly on the cross-track width, leading to changes in the predicted cloud profiles. The cloud percentages thus only indicate that the models predict a coherent amount of cloud presence.

Another way to check the coherence of the computed dataset is to visualize the computed profiles. Figure 5.3 illustrates a sample of the computed 3-dimensional profiles.

From this sample, one can notice that the cloud presence indicators are aggregated, like the 2D cloud scenes shown in Chapter 4, and not randomly spaced out. The visual shape of the computed 3D cloud scene is thus coherent.

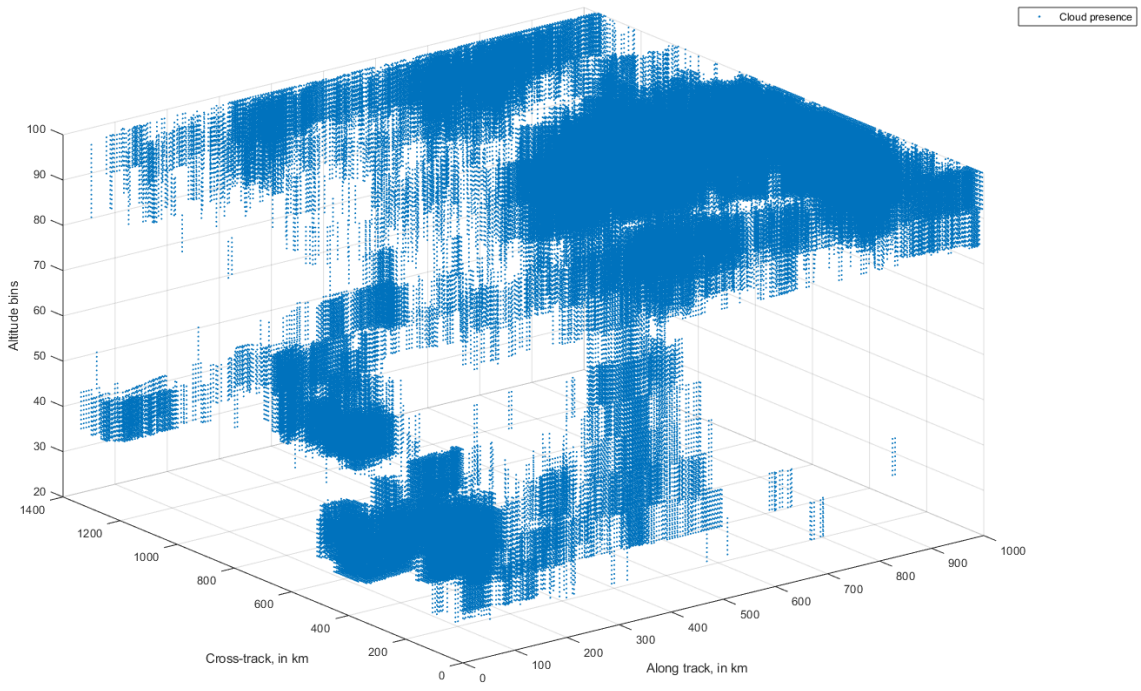

Figure 5.3: Example of computed 3D cloud scene

5.5 Horizontal validation with MOD35 Cloud Mask

As stated earlier, the strict validation of the computed 3D profiles is not possible, as there is no available dataset containing vertical profiles at the same locations. Hence, for validation purposes, the profiles have to be transformed to a quantity that can be compared to existing datasets.

5.5.1 Data processing and fusion

One solution is horizontal validation: unlike vertical information, horizontal cloud information is available on a global scale. Indeed, the MOD35 dataset [40] contains the horizontal "Cloud Mask", which describes the horizontal cloud cover. For example, at a given point, *i.e.* at given latitude and longitude, the horizontal Cloud Mask indicates whether a satellite detects cloud at the given location. The altitude is thus not taken into account in the horizontal Cloud Mask.

In order to get the horizontal Cloud Mask from the computed 3D dataset, the ten profiles

associated with each of the 10 cloud bands have to be summed up, in order to get one single profile. The profiles are superimposed in the following way: at each location, if there is cloud present in at least one of the ten bands, then cloud is considered present in the horizontal Cloud Mask at this location. The computed horizontal Cloud Mask can then be compared to the Cloud Mask feature from the MOD35 dataset.

The MOD35 dataset is, just as as MOD02 and MOD03, a dataset from the LAADS distribution. It thus contains information available at the same locations as MOD02 and MOD03, with the same format and temporal and spatial resolutions. The horizontal Cloud Mask is retrieved by an algorithm using a series of visible and infrared threshold on radiances, and can only be computed using radiometrically accurate radiances [40]. Consequently, there may be holes in the MOD35 Cloud Mask.

However, for the validation day used in this work (February 25^{th}, 2011), only a number of radiances could be exploited at the locations, leading to fewer Cloud Mask values being available. Indeed, only 161 files were available as opposed to the 288 if all radiances were exploitable. The MOD35 Cloud Mask is thus available at fewer locations than the computed horizontal Cloud Mask. For the purpose of comparing the two Cloud Masks, only locations at which the two masks are available will be considered.

The MOD35 Cloud Mask can take four different binary values, which are presented in Table 5.3. These binary values have to be retrieved from the raw dataset using an additional algorithm, given in the annex of [51].

Table 5.3: MOD35 Cloud Mask values

Binary value	Corresponding aspect	Corresponding value in transformed horizontal profile
00	Cloudy	1
01	Probably cloudy	1
10	Probably clear	0
11	Confidently clear	0

This binary values are then transformed into 0s and 1s as indicated in Table 5.3, in

order to obtain a transformed profile with values matching those of the summed predicted profiles. This data processing step leads to two horizontal profiles, containing only 0s and 1s, one obtained with the summed computed profiles, and the other obtained with the MOD35 Cloud Mask.

5.5.2 Cloud Mask comparison

The two horizontal profiles can then be compared using the same metrics as in the previous part of this work. Matthews Correlation Coefficient (MCC) obtained for this location is equivalent to 0.3557. This score is lower than the one obtained by comparing the computed profiles to real, vertical on-track data, as the profiles had an MCC of around 0.5 (Chapter 4). This might be due to the fact that the Cloud Mask from MOD35 is not real retrieved data, but the output of a radiance-based model.

Samples of the predicted horizontal Cloud Mask and the MOD35 one can be visualized at the same locations, and the prediction performance can also be visualized by superimposing these profiles. Figures 5.4 and 5.5 show the predicted cloud mask and the MOD35 cloud mask, respectively, on the same sample horizontal profile. Figure 5.6 shows the prediction performance of the model when compared to the MOD35 Cloud Mask.

As can be seen in Figure 5.6, the general shape of the horizontal Cloud Mask is retrieved: the variations between the two profiles mostly take place at the limits of cloud presence. The model performance is also quite even across the track, as seen in Figure 5.7. The performance does not decrease as the locations get further from the track, as was the case for other models (see description of the work from *Barker et al.* in Chapter 1).

The greatest part of the wrong predictions (more than 90%) is made up of False Negatives, *i.e.* the locations at which the predicted horizontal Cloud Mask is not cloudy while the MOD35 Cloud Mask is. The predicted profiles thus tend to feature fewer clouds than the MOD35 ones. However, since the computed profiles are compared to profiles generated using another model, the model performance studied here is relative, and not absolute, as

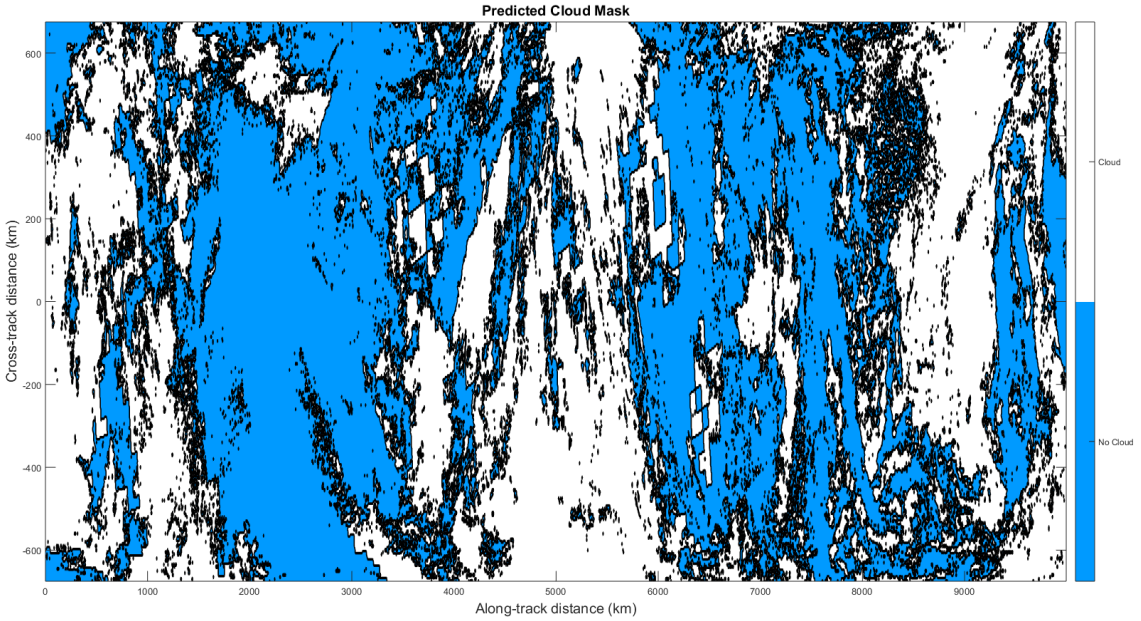

Figure 5.4: Sample predicted Cloud Mask

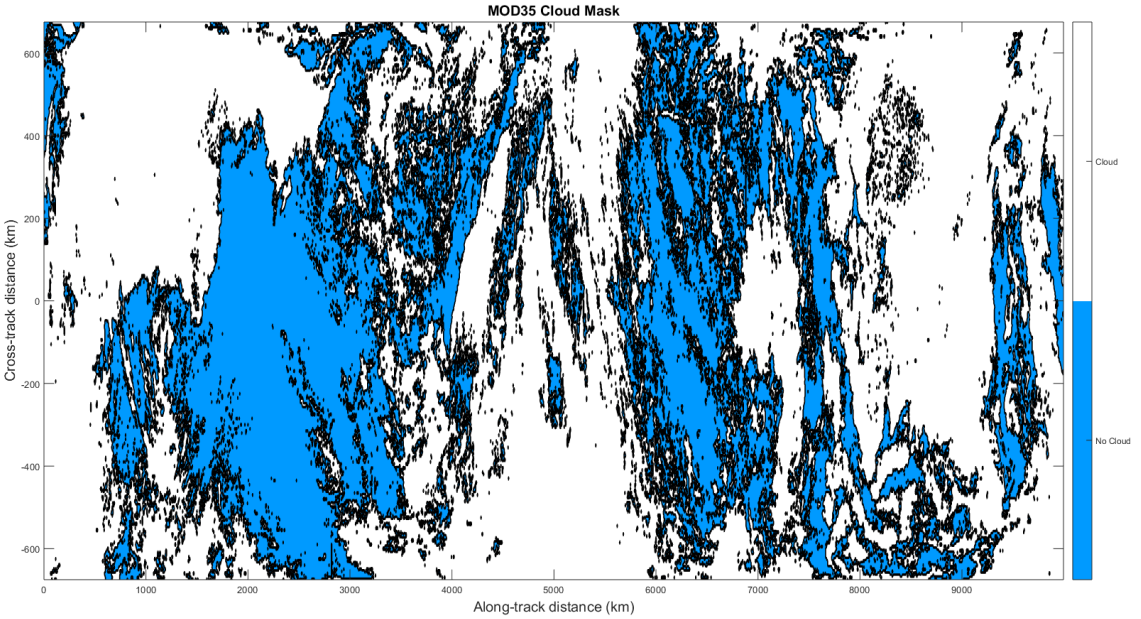

Figure 5.5: Sample Cloud Mask from MOD35

was the case in Step 1. The validity of this performance assessment thus depends on the quality of the MOD35 horizontal profiles.

This can be illustrated by the influence of the "probably cloudy" and "probably clear" zones in the MOD35 Cloud Mask. Indeed, as shown in Table 5.3, the MOD35 Cloud

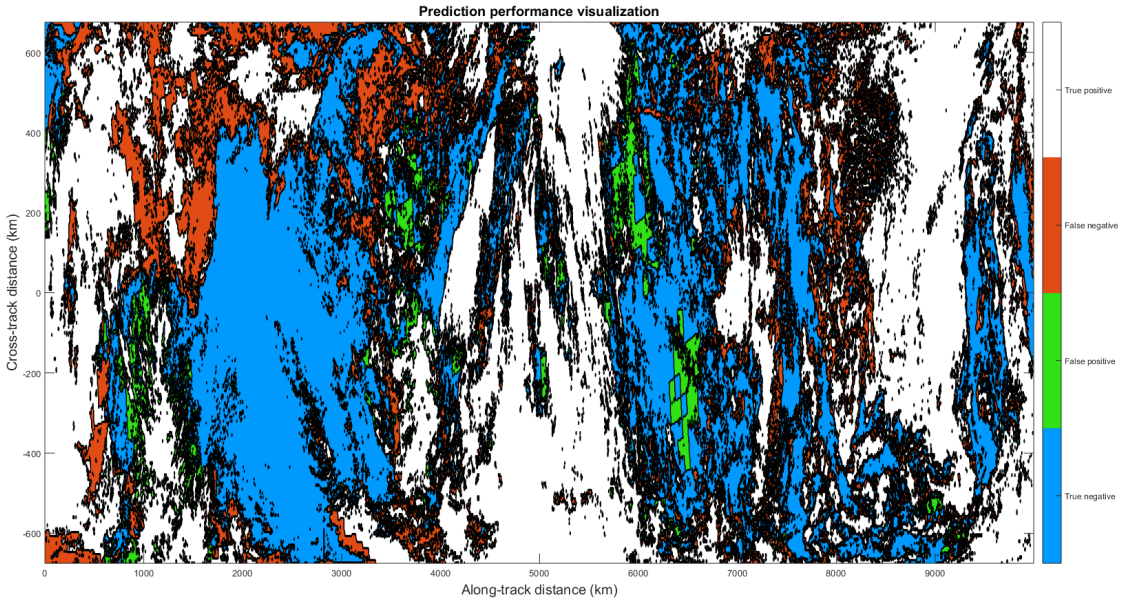

Figure 5.6: Prediction performance visualization

Mask contains not only strictly cloudy and clear locations, but also "probably cloudy" or "probably clear" locations. In our analysis, we have considered that "probably cloudy" corresponded to cloudy, and "probably clear" to clear. However, depending on the quality of the MOD35 profiles, it is possible that "probably cloudy" locations are actually clear, or the other way round. Since the predicted profiles tend to feature fewer cloudy locations, performing the same comparison but this time considering that "probably cloudy" locations are actually clear should provide better results. Only the strictly cloudy locations will be considered as cloud for comparison with the computed profiles, which means that the MOD35 profiles will have fewer clouds, just as the computed ones.

With this method, the obtained MCC is 0.4397, which corresponds to a rise in about 23% when compared to the MCC from the previous comparison. The number of False Negatives is significantly reduced (by more than 20%), and the performance visualization on a similar cloud scene shows noticeable improvement, as seen in Figure 5.8.

This example thus further reinforces the fact that the method used for assessing the performance of the predictive model developed in this work is relative to the quality of the MOD35 horizontal profiles.

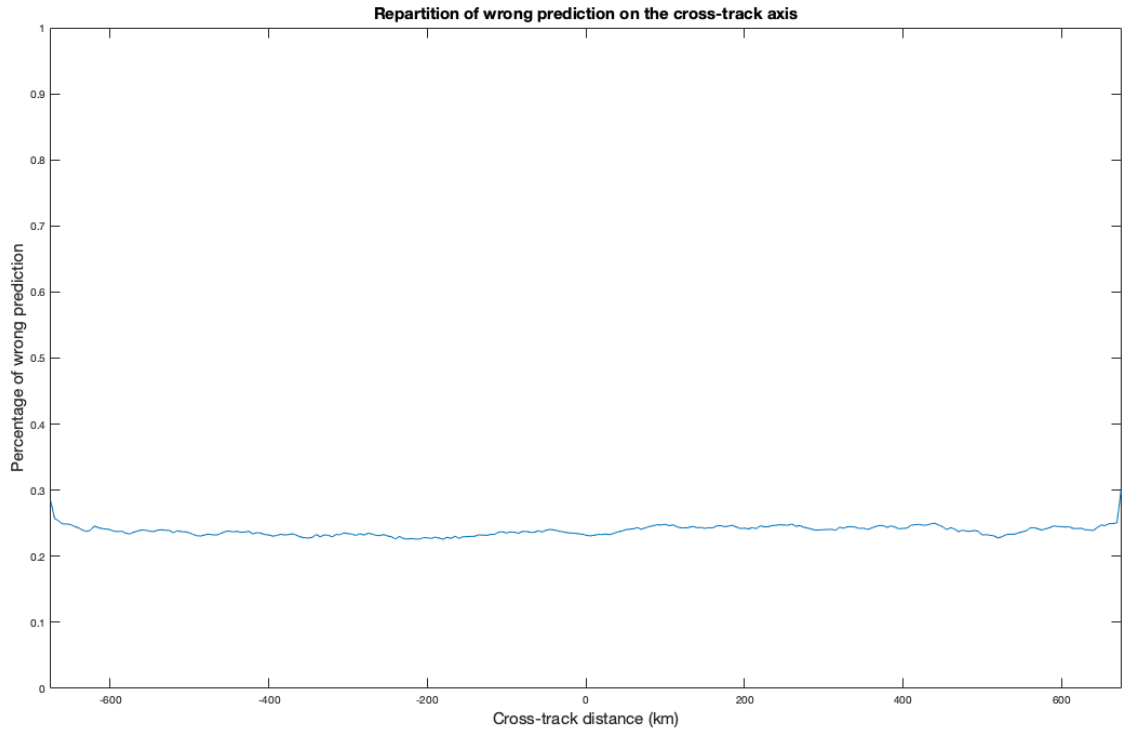

Figure 5.7: Error repartition across the track

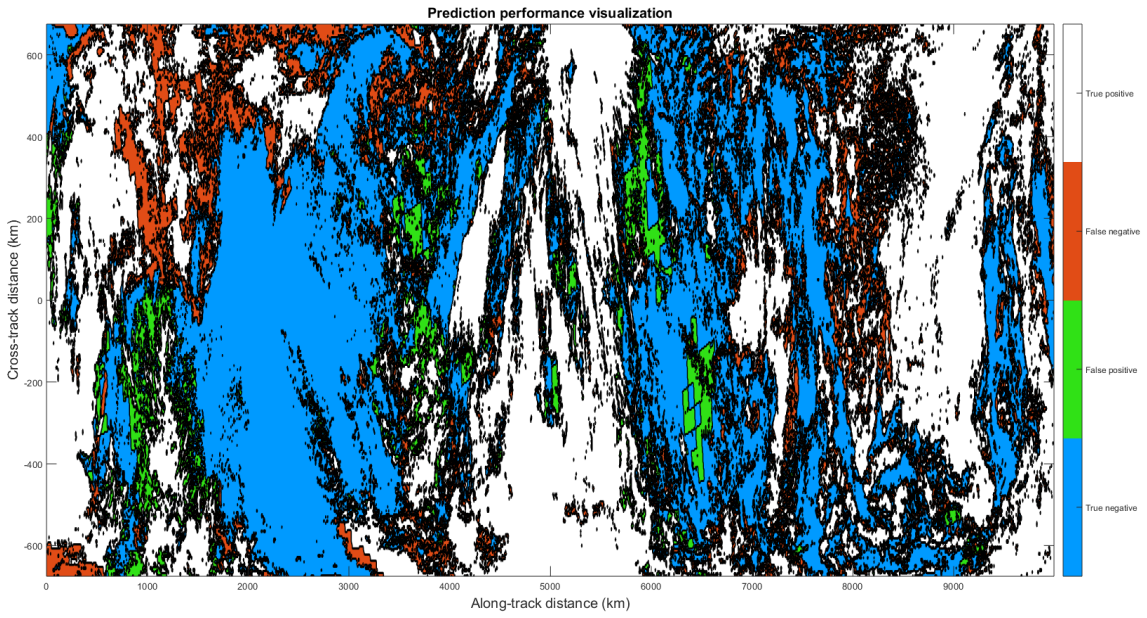

Figure 5.8: Prediction performance visualization with strictly cloudy MOD35 Cloud Mask

Altogether with the consistency check of the profiles presented earlier, the horizontal validation of the computed profiles helps validate **Hypothesis #5** formulated in Chapter

2: if the improved predictive model is implemented at off-track locations with adequate predictors, then a coherent global 3D cloud field dataset is generated.

5.6 Chapter conclusion

The cloud presence percentage and the computed cloud scene visualization thus help check the coherence of the computed 3-dimensional dataset. The horizontal validation against the MOD35 Cloud Mask provides a relative assessment of the computed datasets as compared to another existing model. The computed profiles are thus generally consistent with the MOD35 Cloud Mask. The work detailed in this chapter has thus led to the construction of a 3D global dataset, containing data for both on and off-track locations, and provided answers to **Research Questions #3** through the validation of **Hypothesis #5**.

Since the horizontal validation of the predictive model performed in this section is relative, and thus does not provide an independent, absolute assessment of the model, additional validation and comparison to other values is required in order to ensure the validity of the model. For this purpose, the 3D cloud profile dataset constructed in this chapter will be used in the next step of this work, in order to investigate **Research Question #4** and the corresponding **Hypothesis #6**, and provide an adequate, in-context validation of the computed profiles.

CHAPTER 6
RADIATIVE FLUXES VALIDATION

This chapter details the steps taken to perform the in-context validation of the computed 3D profiles, following the approach outlined in Chapter 3. The results obtained will help validate (or reject) **Hypothesis #6**.

As mentioned in Chapter 2, the chosen Radiative Transfer code for validating the 3D profiles is the Fu-Liou model [43, 44],due to its accuracy and the fact that it does not require much computational resources to run. In order to run the Fu-Liou on the computed profiles, the environment for running Fu-Liou must first be set, and the different inputs needed to run the code must be identified and obtained from the computed profiles. Once this is done, the code can be run on the computed profiles, and the associated radiative fluxes obtained.

6.1 Fu-Liou Environment setting

The Fu-Liou code can be downloaded from the CERES/ARM Validation Experiment (CAVE) website [52]. It contains all the needed libraries, an example input file as well as a User's Manual [53].

The Fu-Liou code is written in Fortran, thus a Fortran compiler is needed to run it. The compiler used in the User's Manual provided with the code mentions *gfortran* as the compiler used by the Fu-Liou developers for running the provided examples. *gfortran* is the name of the GNU Fortran project [54], which aims at developing a free Fortran 95/2003/2008 compiler for the GNU Compiler Collection, available on Linux distributions. For this work, the Fu-Liou code is thus run on a Ubuntu distribution.

To date, the latest release of *gfortran* is *gfortran* 9, which is still in development. However, as confirmed by Fred Rose, the author of the User's Manual, contacted for the purpose of this work, the Fu-Liou code can only be run with versions of *gfortran* anterior to 6.3.

Consequently *gfortran-6* is installed on the machine.

Once this environment is set up, the example provided is tested to ensure that the results obtained match the ones provided in the User's Manual. For this, the local environment first has to be set up to compile F90 and F77 source codes. This can be done by executing the following commands in the directory where the code is located:

```
$ export F90COMP=''--c''
$ export FCOMP=''--c''
$ export F90=gfortran
$ export F77=gfortran
```

*Note: these commands are the *bash* commands for setting up the environment. The README.txt file from the Fu-Liou distribution only provides the *csh* commands.

Once the local environment is set up, the libraries can be compiled and the example run using the single command:

```
$ make
```

Following these directions, the example is successfully run. The results obtained are found to be identical to the ones given in the User's Manual. The environment for running the Fu-liou is thus correctly set up.

6.2 Inputs preparation

Once the Fu-Liou environment is set up, the necessary inputs for running it must be identified and retrieved from the computed cloud profiles. The different inputs are grouped into 4 sets: Atmosphere Structure Inputs, Cloud Inputs, Surface Inputs and Aerosol Inputs. In the context of this research, only the Cloud Inputs are considered. The other inputs sets will thus be set to values that do not interfere with the results.

From the User's Manual [53], the Fu-Liou inputs related to clouds are the following:

- Cloud Fraction (ranging from 0 to 1)
- Maximum number of overlapped layers
- Top and base pressure of each layer (hPa)
- Cloud Phase (Water or Ice)
- Particle size (μm)
- Visible Optical Depth (in logarithmic mean)

The Cloud Fraction can directly be derived from the computed profiles, as the Cloud Fraction values are either 0 or 1. As the Fu-Liou only considers the cloudy zones (1s), and not the clear ones (0s), the Cloud fraction in the Fu-Liou will thus always be set to 1.

Next, as the predictive model developed in the previous chapters contains 10 bands, the maximum number of overlapped cloud layers is set to 10. However, after a number of trials and errors, and discussions with developers and users of the Fu-Liou code, it appears that the Fu-Liou can only handle no more than 4 overlapped layers. The cases at which more than for 4 layers of clouds are overlapped will thus have to be adapted to this limitation. This can be done by "fusing" adjacent bands when cloud is present in all of them. For example, if cloud is present in bands 2, 3, 6,7 and 8, then this condition can be represented by two bands only: one containing bands 2 and 3, and the other containing bands 6,7 and 8. The maximum number of overlapped layers can thus be set to 4 without much issue.

The four remaining inputs cannot be directly extracted from the computed profiles, as they are not taken into account in the predictive model. However, some can be derived from the characteristics of the cloud profiles, while others can be taken from standardized average values. The following subsections detail the chosen methods for retrieving such parameters.

6.2.1 Top and Base Pressures

The Cloud Top and Base pressures are the pressure values corresponding to the altitudes of the top and the bottom, respectively, of the cloud layer. Such pressures are not featured in the profiles, but can be retrieved from the altitudes of the model bands. Indeed, the 10 bands of the model have specific altitudes: they are originally formed of 8 altitude bins each, and the height in meters of these bins are featured in the GEOPROF-LIDAR dataset. The altitudes of the Top and Base of the model bands can thus be obtained, and these altitudes can then be converted to pressure values using an atmospheric model, such as the 1976 US Standard Atmospheric Model [55]. Table 6.1 shows the Cloud Base and Top pressure values obtained for each band, using interpolation of the atmospheric values featured in [56].

Table 6.1: Retrieved Cloud Top and Base pressure for each model band

Band	Cloud Base (hPa)	Cloud Top (hPa)
1	71	54
2	96	73
3	130	99
4	175	134
5	237	182
6	317	245
7	417	327
8	542	431
9	695	559
10	882	717

The Top and Base pressures can thus be retrieved for each band, and used as inputs to the Fu-Liou code.

6.2.2 Cloud Phase

The next input to be determined is the Cloud Phase, which can be either water or ice in the Fu-Liou code. The predictive model did not take into account the phase of the clouds,

so this input has to be determined arbitrarily in order to run the Fu-Liou on the computed profiles.

One solution consists in assuming that all bands contain either only water clouds or only ice clouds. However, this would most likely bring large errors to the computation of radiative fluxes as ice and water have very different interactions with radiation. Another solution consists in considering that one band has only the most common phase found at the band altitude. Indeed, a study by *Hu et al.* [57] provides statistics of water and ice presence in cloud depending on temperature. The Top and Base pressure corresponding to such temperatures can be retrieved with the 1976 US Standard Atmospheric Model as well, so the bands can be associated with temperature. The study finds out that water and ice are equally frequent at about -30°C (243.15 K), which corresponds to Band 7 of the model using the 1976 US Standard Atmospheric Model.

The Cloud Phase can thus be approximately determined using the following method: if the band is higher than 7 (*i.e* between 1 and 6), then the Cloud Phase for this band is set to Ice only, otherwise the Cloud Phase is set to Water only.

6.2.3 Particle size

Particle size is another input to the Fu-Liou code that cannot be directly derived from the computed profiles. The chosen solution for getting this input is to rely on average values from studies on cloud particle size. Existing studies find particle sizes of about 10 μm [58], or ranging between 2-10 μm to 50 μm [59]. The particle size for ice clouds is usually larger than the one associated with water clouds. In order to find the most appropriate particle size for the simulations, multiple values of this input can be tested, within the ranges provided by existing studies, and with a higher particle size for ice clouds. The particle size giving the best results can then be selected as the most appropriate input value.

6.2.4 Visible Optical Depth

The last input to be determined is the optical depth, which corresponds to the "vertical optical thickness between the top and bottom of a cloud" [60]. Optical depth can be retrieved from satellite data by different models and algorithms, but is not featured in the predictive model developed here. A study by *Marchand et al.* [61] provides the distributions of retrieved optical depth by different models, for different locations around the globe. The majority of the retrieved values for all models approximately range between 1 and 15. This range of values, as for the particle size, can be tested with the Fu-Liou code in order to determine the most appropriate one.

6.2.5 Inputs summary

The previous subsections have thus discussed how the cloud-related inputs required to run the Fu-Liou code were determined. Table 6.2 summarizes the needed inputs and their corresponding values.

Table 6.2: Fu-Liou cloud inputs values

Input	Value
Cloud Fraction	only 1
Max number of overlapped layers	4
Top and Base pressure	Pressure associated with cloud band
Cloud Phase	Ice for bands 1 to 6, Water for bands 7 to 10
Particle size	Value between 2 and 50 μm
Visible Optical Depth	Value between 1 and 15

6.3 Running the Fu-Liou code

Once the Fu-Liou environment has been set and the necessary inputs have been determined, the Fu-Liou code can be run on the computed profiles. This section will first demonstrate

that the Fu-Liou code runs on sample profiles from the computed dataset, and then provide a method for running it on the whole computed dataset.

6.3.1 Demonstration on a sample of computed profiles

In order to check the coherence of the outputs provided by the Fu-Liou when using the inputs defined in the previous section, chosen sample profiles from the computed dataset are tested, and the associated outputs are compared. Tables 6.3, 6.4, 6.5 give the set of inputs for each tested sample profile. Note that the third sample corresponds to the superimposition of the first two sample profiles. Doing so allows one to analyze the results obtained for two superimposed cloud layers, and compare them to the ones obtained for each layer when run separately.

Table 6.3: Sample 1: Fu-Liou cloud inputs values

Input	Value
Cloud Fraction	1
Number of overlapped layers	1
Top and Base pressure: Ice band (hPa)	134 ; 237
Particle size: Ice band	50 μm
Visible Optical Depth	10

Table 6.4: Sample 2: Fu-Liou cloud inputs values

Input	Value
Cloud Fraction	1
Number of overlapped layers	1
Top and Base pressure: Water band (hPa)	327 ; 542
Particle size: Water band	10 μm
Visible Optical Depth	10

For each profile, the Fu-Liou computes different types of radiations: Shortwave, Longwave and Window, in upward and downward directions. The flux values are given at different locations of the atmosphere: Top of Atmosphere (TOA), Surface, as well as at different

Table 6.5: Sample 3: Fu-Liou cloud inputs values

Input	Value
Cloud Fraction	1
Number of overlapped layers	2
Top and Base pressure: Ice band (hPa)	134 ; 237
Top and Base pressure: Water band (hPa)	327 ; 542
Particle size: Ice band	50 μm
Particle size: Water band	10 μm
Visible Optical Depth	10

pressure levels between TOA and Surface [53]. There are four simultaneous computation modes available: Clear Sky (No Clouds), Total Sky, Pristine (No Aerosol or Clouds), and Total No Aerosol (Clouds, No Aerosol). The considered modes must thus be either Total Sky or Total No Aerosol, as they are the only ones taking into account the cloud profile.

The outputs associated with each sample profiles are shown in Figures 6.1, 6.2 and 6.3. The fluxes values are here given at different pressure levels, from Top (first level) to Surface (6^{th} level), for all computation modes and all types of radiations.

Several observations can be made on these results. First, the Clear and Pristine results are virtually identical for all three samples. Indeed, only the cloud inputs were modified, so the Clear and Pristine modes are not affected as they do not take into account the cloud scene. This verifies that the samples are well-implemented and the cloud inputs are independent from other types of inputs.

Next, the variations of downward longwave radiation values are coherent across the samples. Indeed, the longwave downward flux at the surface level (thus below all the clouds) is higher for Sample 2, which contains a low-altitude cloud, than for Sample 1, which contains a higher altitude one. Indeed, clouds always have a positive radiative effect on longwave downward radiations, and this effect is higher for low-altitude clouds [62], as is the case here. The combined profile from Sample 3 generates slightly more radiation as compared to Sample 1, as there is more cloud in the former than in the latter.

```
================================================================================
                      SHORTWAVE Down--------------    Shortwave Up-----------------
 #  Presure  Height  Clear   Prist   Total  TotNOA    Clear   Prist   Total  TotNOA
Lev  [hPa] [meters]  Down    Down    Down    Down     Up      Up      Up      UP
  1    0.10  66296. 1365.03 1365.03 1365.03 1365.03  139.24   51.92  617.97  997.46
  2   70.00  18905. 1330.89 1330.14 1331.58 1333.32  140.04   50.43  625.65 1012.37
  3  200.00  12242. 1316.24 1315.08 1013.12 1311.70  136.55   45.05  425.59 1116.75
  4  500.00   5780. 1226.40 1237.80  600.16 1236.88  119.38   29.94   86.54 1127.18
  5  850.00   1502. 1037.03 1117.88  487.85 1191.30   62.57   10.01   37.26 1141.99
  6 1012.76      2.  909.85 1063.24  417.76 1163.43   -0.00    0.00    0.00 1163.43
                     LONGWAVE Down---------------    Longwave  Up-----------------
 #  Presure  Height  Clear   Prist   Total  TotNOA    Clear   Prist   Total  TotNOA
Lev  [hPa] [meters]  Down    Down    Down    Down     Up      Up      Up      UP
  1    0.10  66296.   0.00    0.00    0.00    0.00  275.73  279.78  126.36  126.37
  2   70.00  18905.  13.33   13.31   13.33   13.31  274.39  278.53  122.09  122.10
  3  200.00  12242.  27.94   27.69  127.56  127.56  282.50  286.50  141.25  141.34
  4  500.00   5780. 141.39  139.21  201.10  200.33  332.24  334.67  332.48  334.74
  5  850.00   1502. 289.56  283.91  313.65  309.94  400.24  400.64  400.36  400.73
  6 1012.76      2. 356.46  350.81  370.06  365.96  422.72  422.66  422.85  422.81
                     WINDOW Down-----------------    WINDOW  Up-------------------
 #  Presure  Height  Clear   Prist   Total  TotNOA    Clear   Prist   Total  TotNOA
Lev  [hPa] [meters]  Down    Down    Down    Down     Up      Up      Up      UP
  1    0.10  66296.   0.00    0.00    0.00    0.00  101.88  104.73   22.93   22.93
  2   70.00  18905.   1.65    1.65    1.65    1.65  103.46  106.40   21.98   21.98
  3  200.00  12242.   2.03    1.97   22.84   22.84  107.07  110.00   27.96   28.02
  4  500.00   5780.   5.08    4.05   31.32   30.88  111.64  113.57  111.83  113.65
  5  850.00   1502.  30.79   26.15   50.03   46.91  118.92  119.25  119.03  119.34
  6 1012.76      2.  62.49   57.32   74.81   71.03  121.54  121.48  121.66  121.62
```

Figure 6.1: Outputs obtained for Sample 1

Finally, when considering both Total modes for shortwave radiations, differences of trends across the pressure levels can be noticed, while the values are the same for longwave radiation. This is due to the fact that while clouds do interact with both shortwave and longwave radiations [63], aerosols interact with shortwave radiation only. As the predictive model developed in this work does not take into account aerosols, additional aerosol information should be retrieved and taken into account accordingly when running the Fu-Liou. Such information could be retrieved from either MODIS or MERRA-2 products. Once this information is retrieved, the computed shortwave and longwave radiation values should be used for comparison with existing ones.

Altogether, the results associated with these three samples show that the Fu-Liou code runs correctly on the inputs defined in the previous sections, and that the outputs are coherent. They also show that the shortwave and longwave radiations, in both upward and downward directions, should be the ones primarily considered for the validation of the

			SHORTWAVE Down				Shortwave Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66295.	1365.03	1365.03	1365.03	1365.03	139.24	51.95	541.58	1013.73
2	70.00	18904.	1330.89	1330.14	1331.23	1333.35	140.04	50.45	547.18	1028.40
3	200.00	12241.	1316.24	1315.08	1319.50	1327.17	136.55	45.08	548.83	1036.93
4	500.00	5780.	1226.06	1237.45	825.54	1309.28	119.38	29.94	202.23	1167.61
5	850.00	1502.	1036.95	1117.80	582.19	1236.37	62.56	10.01	44.32	1179.07
6	1012.76	2.	909.80	1063.18	497.79	1203.51	0.00	0.00	0.00	1203.51
			LONGWAVE Down				Longwave Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66295.	0.00	0.00	0.00	0.00	274.73	278.77	182.15	182.40
2	70.00	18904.	13.33	13.31	13.33	13.31	273.38	277.51	178.87	179.12
3	200.00	12241.	27.94	27.68	27.92	27.68	281.13	285.13	183.39	183.53
4	500.00	5780.	141.42	139.24	260.72	260.73	332.34	334.76	290.57	291.14
5	850.00	1502.	289.56	283.91	347.18	345.91	400.24	400.64	400.55	400.87
6	1012.76	2.	356.46	350.81	389.89	387.97	422.72	422.66	423.05	423.03
			WINDOW Down				WINDOW Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66295.	0.00	0.00	0.00	0.00	101.88	104.73	44.06	44.15
2	70.00	18904.	1.65	1.65	1.65	1.65	103.46	106.40	43.83	43.92
3	200.00	12241.	2.03	1.97	2.01	1.97	107.06	109.99	44.98	45.04
4	500.00	5780.	5.08	4.05	63.92	63.93	111.64	113.57	83.53	84.00
5	850.00	1502.	30.79	26.15	78.34	77.30	118.92	119.25	119.21	119.47
6	1012.76	2.	62.49	57.32	93.12	91.37	121.54	121.48	121.84	121.82

Figure 6.2: Outputs obtained for Sample 2

			SHORTWAVE Down				Shortwave Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66296.	1365.03	1365.03	1365.03	1365.03	139.24	51.93	740.53	1007.72
2	70.00	18905.	1330.89	1330.14	1332.01	1333.32	140.04	50.43	750.19	1022.66
3	200.00	12242.	1316.24	1315.08	1108.11	1319.45	136.55	45.05	648.80	1134.36
4	500.00	5780.	1226.41	1237.80	486.42	1195.56	119.38	29.94	128.97	1115.52
5	850.00	1502.	1037.03	1117.88	343.73	1155.50	62.57	10.01	28.71	1119.30
6	1012.76	2.	909.85	1063.23	292.72	1134.97	0.00	0.00	0.00	1134.97
			LONGWAVE Down				Longwave Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66296.	0.00	0.00	0.00	0.00	275.75	279.80	126.27	126.27
2	70.00	18905.	13.33	13.31	13.33	13.31	274.42	278.55	122.00	122.00
3	200.00	12242.	27.94	27.69	127.55	127.54	282.52	286.52	138.59	138.59
4	500.00	5780.	141.39	139.21	260.83	260.84	332.34	334.76	290.58	291.14
5	850.00	1502.	289.56	283.92	347.20	345.94	400.24	400.64	400.55	400.87
6	1012.76	2.	356.46	350.81	389.91	387.99	422.72	422.66	423.05	423.03
			WINDOW Down				WINDOW Up			
# Lev	Presure [hPa]	Height [meters]	Clear Down	Prist Down	Total Down	TotNOA Down	Clear Up	Prist Up	Total Up	TotNOA UP
1	0.10	66296.	0.00	0.00	0.00	0.00	101.88	104.72	22.86	22.86
2	70.00	18905.	1.65	1.65	1.65	1.65	103.45	106.40	21.91	21.91
3	200.00	12242.	2.03	1.97	22.83	22.83	107.07	110.00	26.14	26.15
4	500.00	5780.	5.08	4.05	64.02	64.03	111.64	113.57	83.53	84.00
5	850.00	1502.	30.80	26.16	78.36	77.33	118.92	119.25	119.21	119.47
6	1012.76	2.	62.49	57.32	93.14	91.39	121.54	121.48	121.84	121.82

Figure 6.3: Outputs obtained for Sample 3

computed dataset, and additional aerosol information should be retrieved in order to enable validation with shortwave radiations.

6.3.2 Constructed 3D dataset

The previous section shows that the Fu-Liou code can be run on profiles from the computed dataset. Consequently, the radiative flux values at each location featured in the 3D computed dataset can be retrieved.

However, the Fu-Liou code cannot take several locations as an input: only one vertical profile can be run at a time. The computed dataset containing more than 17 million entries, it is virtually impossible to run them time after time by hand. The process of running the Fu-Liou code on all the entries of the dataset must then be automated. This will not be implemented for the purpose of this thesis due to time constraints, but a general approach for implementing this is provided below.

First, an algorithm should be implemented in order to associate each vertical profile with the Fu-Liou inputs, as described in Section 6.2. More specifically, this algorithm should take as inputs the numbers of the cloud-filled bands at the given location, and its outputs should be the features given in Table 6.2, associated with each band. If there are adjacent cloud-filled bands in the profile (e.g.: bands 6 and 7 are filled), then the algorithm should be able to fuse these bands and return the features associated with the fused band. The returned Fu-Liou inputs at each location should then be saved in a common structure. This algorithm will most likely be developed in Matlab, as the current structures containing the computed dataset are Matlab structures.

Next, an algorithm for automatically filling out the F90 input file should be implemented. The simple.f90 example file provided with the Fu-Liou distribution should be used as a template input file, and the algorithm should change the cloud-related section of this file. The algorithm should be able to add or remove cloud overlapping layers from the input file when needed, depending on the inputs previously retrieved from the computed

dataset.

Finally, an algorithm to automatically run the Fu-Liou code on all the profiles featured in the computed dataset will have to be implemented. The result of each run of the Fu-Liou should be saved in a common structure. The algorithm should integrate the one previously described, so that the input file is re-created at each new run with the inputs from the considered profile, rather than creating more than 17 million input files and then performing the computation. Since the Fu-Liou code is run using the bash command *make*, this algorithm should most likely be written using bash commands.

Such approach should enable the computation of the radiative fluxes associated with the 3D computed dataset, as it captures steps taken to achieve the Fu-Liou demonstration on sample profiles discussed in section 6.3.1.

6.4 Outputs comparison

Once the Fu-Liou outputs are generated for all the profiles of the 3D computed dataset, they have to be compared to existing values in order to validate the predicted profiles.

As mentioned in Chapter 2, the CERES SYN1deg dataset contains radiative flux values at global locations, updated every hour [42]. The radiative flux values are given for the Clear Sky and the Total Sky mode. Consequently, the Total Sky values of the CERES SYN1deg dataset can be compared to the ones from the Fu-Liou outputs computed using the Total Sky mode. Additionally, the Clear Sky mode can be used for computing the difference between the Total Sky and Clear Sky flux, which corresponds to the cloud radiative effect, and comparing it to the cloud radiative effect from the computed profiles. The CERES SYN1deg dataset also contains the different types of fluxes (Shortwave, Longwave and Window), both in upward and downward directions, and for different pressure levels as well, so each type of radiation can be compared to the ones obtained with the computed dataset.

However, the CERES SYN1deg dataset does not have the same resolution as the com-

puted profiles: its spatial and temporal resolutions are lower, with cells of 1°x 1°, which corresponds to about 100 km x 100 km, available for every hour. Thus, in order to compare the Fu-Liou outputs to the CERES values, the Fu-Liou radiative fluxes will have to be averaged over the area of the CERES SYN1deg cells, and the profile times will have to be retrieved. Doing so shall enable the validation of the outputs against existing values.

6.5 Chapter conclusion

The Fu-Liou environment has been set up, and the necessary inputs associated with the profiles from the 3D computed dataset have been identified. Methods for retrieving these inputs have been provided, and the Fu-Liou code was run on a sample set of inputs, corresponding to sample profiles from the computed dataset, hence allowing to check the coherence of the corresponding outputs. A method for performing the Fu-Liou calculations on the whole computed dataset was provided, as well as one for comparing the results to an existing dataset.

The work detailed in this chapter has thus provided a methodology for performing in-context validation of the computed dataset, and has benchmarked this method against sample profiles from the dataset. Doing so provides answers to **Research Question #4.1**, by detailing an approach that can be used in order to determine a satisfactory level of accuracy for the predictive model.

Research Question #4 and **Hypothesis #6** have not been fully investigated, but it is expected that the approach detailed above should, once implemented, provide sufficient results to fully validate (or reject) **Hypothesis #6**, and eventually address **Research Question #4**.

CHAPTER 7
CONCLUSIONS AND FUTURE WORK

Climate change predictions are currently achieved using Global Climate Models (GCMs), which are mathematical representations of the major climate system components. Several GCMs are used in current climate studies, but their results and predictions are quite different from one to the other, leading to high levels of uncertainty in climate projections. There is high confidence that these high levels of uncertainty are due to the representation of clouds in climate models. Indeed, cloud representations in GCMs rely on parameterizations and theoretical relations which are often unable to capture the complex physical phenomena leading to cloud formation. Such parameterizations can be improved thanks to the integration of cloud satellite data into GCMs. However, while certain types of cloud-related data are widely available, vertical cloud data is scarce, and thus needs to be generated globally so that it can be integrated into GCMs. The research objective of this thesis is thus to support the validation of GCMs cloud representations through the generation of 3D cloud fields using cloud vertical data from space-borne sensors.

For this purpose, a new approach for generating cloud vertical profiles, leveraging data fusion and machine learning techniques was developed, which fulfilled the following criteria:

- Able to predict off-track cloud profiles
- Able to account for and handle high amounts of data and predictive features
- Scalable, i.e. supporting the integration and processing of multiple data sources
- Flexible, i.e. supporting the integration of data sources other than the one(s) originally considered

- Able to account for and handle disparate sources and types of data
- Able to handle sources of data that have different resolutions or levels of granularity
- Able to generate models in a reasonable amount of time and with a reasonable amount of computing resources

Such characteristics are not all featured in the existing approaches reviewed in this thesis, but have been proven to enhance the predictive models resulting from implementing this approach.

The developed approach consists in first improving an existing model developed in previous successive works [18, 20, 21, 22]. The resulting predictive model has been successfully improved, and validated against on-track, available data. Next, the predictive model has been used to construct a 3D cloud profile dataset at off-track locations, which has been validated against available horizontal data. Finally, the basis for in-context validation of the computed dataset, and more globally of the approach itself, has been laid out.

7.1 Research Questions and Hypotheses review

The review of existing approaches for generating cloud vertical data helped identify several research gaps to be investigated in order to develop a new, more efficient approach. The following Research Questions have been formulated from such gaps:

- RQ #1: How can we better predict the presence of clouds in lower-altitude bands?
- RQ #2: Which machine learning technique(s) would lead to an improved predictive capability?
- RQ #3: What approach is best suited to validate the predicted models “off-track”?
- RQ #4: What level of accuracy is required from the predictive models to generate 3D cloud fields?

- RQ #4.1: What approach should be undertaken to determine a satisfactory level of model accuracy?

Six Hypotheses have been associated with these Research Questions, and have been investigated in this thesis in order to check their validity.

First, in order to enhance the performance of the model in predicting cloud presence in lower-altitude bands, vertical data from the GEOPROF-RADAR dataset has been integrated to the model. This has resulted in an improved predictive performance, which was especially significant for the lower-altitude bands. Consequently, this has allowed to validate **Hypothesis #1**, as defined in Chapter 2.

Second, the model has been modified so that it would take into account the values of higher-altitude bands when predicting the values associated with inferior bands. This has not resulted in a significant change in the global prediction performance of the model, nor in the performance specific to the lower-altitude bands: **Hypothesis #2** has thus been rejected. Altogether, the steps taken to check these two hypotheses have provided answers to the first research question.

Next, the number of atmospheric features used as predictors by the model has been reduced using Principal Component Analysis (PCA), in order to limit the total number of predictors to the most significant ones. The predictors have been classified according to their influence on other atmospheric features, and the predictive performance has been tested using different numbers of predictors. The performance has been significantly improved using a reduced number of predictors, as compared to previous works on the model. This has allowed to validate **Hypothesis #3**. Additionally, both Kernel methods, in the form of SVMs, and Random Forests, have been implemented in order to train the predictive model. While Kernel methods have not proven to be a good fit for this specific application, Random Forests have brought a significant improvement in the predictive performance of the models. This has allowed to validate **Hypothesis #4**. These steps have thus provided two answers to the second Research Question, as both PCA and Random Forests have led to an

improvement in the predictive capability of the model.

Then, using the improved model from the previous steps, cloud profiles at off-track locations have been generated and analyzed. Horizontal validation against an existing dataset has been performed, and has led to the conclusion that the generated global 3D cloud field dataset was coherent. **Hypothesis #5** has thus been validated, which provided an answer to the third research question.

Finally, a methodology for performing in-context validation of the computed dataset has been developed, and tested on samples from the dataset. Radiative fluxes values associated with these samples were computed by running the Fu-Liou code, and the results analyzed. Guidelines for comparing the computed radiative fluxes values with existing ones have also been provided. Such work has thus provided answers to **Research Question #4.1**, by detailing an approach to be used in order to determine a satisfactory level of accuracy for the predictive model. **Research Question #4** and **Hypothesis #6** have not been fully investigated in this work, but the given approach, if later implemented, should be able to provide meaningful insights.

7.2 Future work

As mentioned above, in-context validation of the 3D cloud profile dataset constructed in this thesis has not been fully performed. In Chapter 6, a method for validation using radiative flux values has been detailed and tested on sample profiles, but not implemented. This method should thus be implemented as described, and performed on the computed dataset. Then, information on aerosols at the same locations as the computed profiles should be retrieved and used as inputs to the Fu-Liou code. Next, the shortwave and longwave radiation values associated with the computed profiles should be matched and compared to the ones featured in the CERES SYN1deg dataset, both for the upward and the downward fluxes. Doing so would enable in-context validation of the computed profiles, and, more generally, of the predictive model. This in-context validation should provide a basis for assessing the

representations of cloud-radiative phenomena into GCMs, as cloud vertical distribution and associated radiative fluxes have been identified as the main sources of misrepresentation of clouds in GCMs.

The investigation of the cloud-related Fu-Liou inputs has also brought to light some possible limitations of the predictive model. Indeed, the model developed in this thesis only predicts the vertical distribution of clouds at all locations, and does not take into account physical parameters such as Optical Depth or Particle size. However, these parameters are key inputs for retrieving the radiations associated with the cloud profiles, and thus crucial parameters for assessing the representations of cloud-radiative phenomena into GCMs. Thus, the comparison of the Fu-Liou outputs to existing radiative fluxes data should provide an insight of the model limitations regarding these parameters. The methodology presented in Chapter 6 of this thesis suggests using standard values for these parameters, within ranges given by the available literature. Another way of retrieving some of these parameters would be to use information from the MERRA-2 dataset presented in this thesis. This dataset contains information on atmospheric parameters such as cloud temperature and pressure, which could be used instead of standard values as is the case now to determine top and base pressure or cloud phase. Hence, using this additional information, other values for the different input parameters to the Fu-Liou code could be determined. Considering these multiple values, sensitivity studies on the influence of each input should be conducted, in order to assess the physics-related limitations of the model.

If these physics-related limitations are too high, the model should be adapted to give better values of these physical parameters and relations between the profiles and these parameters should be further investigated. This would lead to an improved in-context predictive capability.

Additionally, the in-context validation of the computed vertical profiles would allow to assess the impact of the parameterization of the vertical profiles. In this research, the profiles have been discretized in ten superimposed bands, as in the works by *Huguenin et al.*

[22]. The choice of ten bands arises from the need to have a limited number of independent models (as one band corresponds to one model) to be trained, while describing the original cloud profile efficiently. There is thus a trade-off between the number of models, which we want to limit, and the precision of the parameterization. As shown by *Huguenin et al.*, using ten bands for parameterizing the profile provided a good visual description of multiple types of profiles, and gives a reasonable number of models to be trained. However, no study has been achieved that this parameterization is the most adequate one for the context of study. A sensitivity study on the number of bands used to parameterize the profiles should be conducted, in order to determine the configuration providing the best results for in-context validation, *i.e.* leading to the generation of radiative flux values closest to the existing ones.

Finally, another sensitivity study should be conducted on the computing environment and programming language used for implementing and validating the models developed in this research. Matlab has been used to obtain the major part of the results presented in this dissertation, fuse the data and train the predictive models. Matlab was chosen based on its user-friendliness, its data and variable visualization advantages, and the fact that it is widely used in the research community, as for example by the NASA Langley Climate Science Branch who first introduced this topic at ASDL. However, some other environments and languages may prove better suited to processing the data and training the model. For example, the poor performance obtained with SVMs in this thesis, and with classification neural networks in previous works, could be directly related to the environment in which the model was trained. Conducting a sensitivity study on different environments could thus also lead to an improved predictive capability as compared to the one achieved in this thesis.

Appendices

APPENDIX A

CODE

The code developed through this work and used to obtain the results from Step #1 and Step #2 of the approach presented in this dissertation is available through GitHub, at the following address: `https://github.com/manonhuguenin/3D-Cloud-Fields`

This code is not exhaustive, but sufficient to obtain the most important results presented in this dissertation.

The numerous datasets used in this research are not featured in the GitHub, in order to limit storage space use. All of these sets can be directly downloaded by any potential user, following the references given in this dissertation [30, 26, 27, 25, 28, 42, 37, 39, 38, 40]. Once these datasets have been dowloaded, they have to be stored in a folder named "Data" in the same directory as the code. For Step #1, "Data" must contain the satellite data arranged by day and then by type, as shown in an example in Figure A.1.

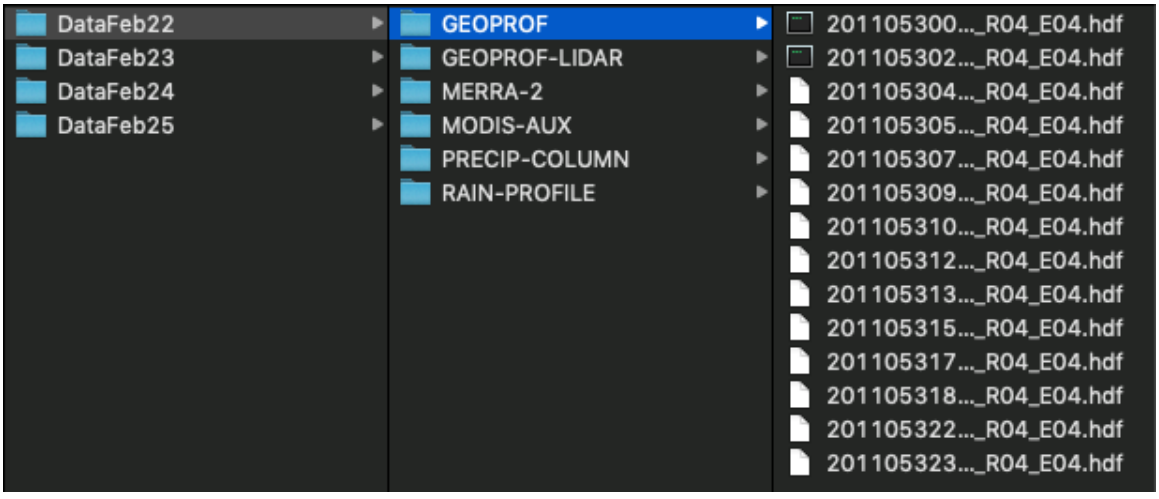

Figure A.1: Data folder organization for Step #1

For Step #2, "Data" must contain the satellite data arranged by type (MOD02, MOD03 and MOD35), as shown in an example in Figure A.2.

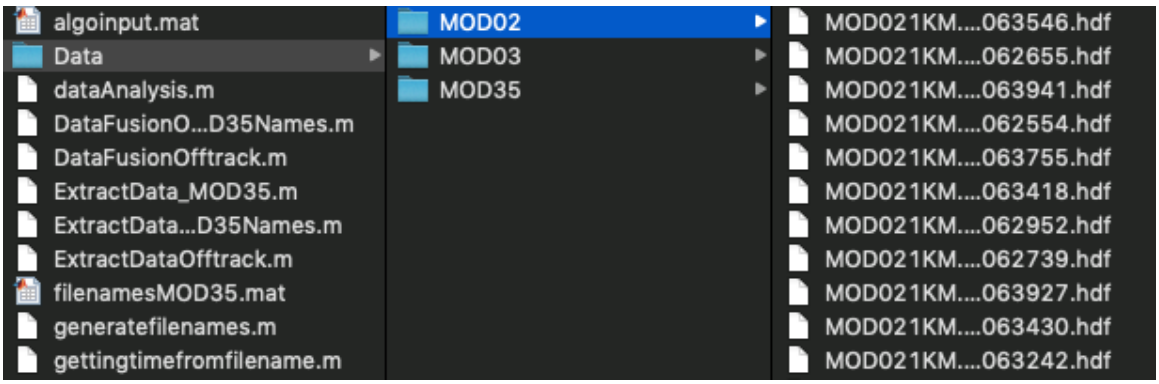

Figure A.2: Data folder organization for Step #2

REFERENCES

[1] World Health Organization. (2017). Climate change and health. Available at http://www.who.int/mediacentre/factsheets/fs266/en/.

[2] J. Alvich, *Climate Modeling*, [Online], Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Available at http://www.gfdl.noaa.gov/climate-modeling{%}5Cnpapers2://publication/uuid/66FC00BF-F75E-42ED-8833-6DEE1013FB96, 2009.

[3] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen, "Evaluation of Climate Models," in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013, pp. 741–866, ISBN: 9781107415324.

[4] G. Hegerl, F. W. Zwiers, P. Braconnot, N. Gillett, Y. Luo, J. M. Orsini, N. Nicholls, J. Penner, and P. Stott, "Understanding and Attributing Climate Change," in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, 3-4, vol. 80, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007, pp. 213–238, ISBN: 978 0 521 70596 7.

[5] T. Schneider, J. Teixeira, C. S. Bretherton, F. Brient, K. G. Pressel, C. Schär, and A. P. Siebesma, "Climate goals and computing the future of clouds," *Nature Climate Change*, vol. 7, no. 1, pp. 3–5, 2017.

[6] V. Ramanathan, R. Cess, E. Harrison, P Minnis, B. Barkstrom, E Ahmad, and D Hartmann, "Cloud-radiative forcing and climate: Results from the earth radiation budget experiment," *Science*, vol. 243, no. 4887, pp. 57–63, 1989.

[7] D. Chand, T. L. Anderson, R. Wood, R. J. Charlson, Y. Hu, Z. Liu, and M. A. Vaughan, "Quantifying above-cloud aerosol using spaceborne lidar for improved understanding of cloudy-sky direct climate forcing," *Journal of Geophysical Research Atmospheres*, vol. 113, no. 13, pp. 1–12, 2008.

[8] C.-H. Ho, M.-D. Chou, M. Suarez, and K.-M. Lau, "Effect of ice cloud on GCM climate simulations," *Geophysical Research Letters*, vol. 25, no. 1, pp. 71–74, 1998.

[9] H. W. Barker, M. P. Jerg, T. Wehr, S. Kato, D. P. Donovan, and R. J. Hogan, "A 3D cloud-construction algorithm for the EarthCARE satellite mission," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 657, pp. 1042–1058, 2011.

[10] M. Doutriaux-Boucher and J. Quaas, "Evaluation of cloud thermodynamic phase parametrizations in the LMDZ GCM by using POLDER satellite data," *Geophysical Research Letters*, vol. 31, no. 6, n/a–n/a, 2004.

[11] C. M. Naud, D. J. Posselt, and S. C. Van Den Heever, "Observational analysis of cloud and precipitation in midlatitude cyclones: Northern versus Southern hemisphere warm fronts," *Journal of Climate*, vol. 25, no. 14, pp. 5135–5151, 2012.

[12] NASA. (2015). The afternoon constellation. Available at `atrain.nasa.gov/`.

[13] *NASA A-Train*, [Online], Available at `https://www.nasa.gov/images/content/329173main_CALIPSO-a-train-full.jpg`, 2018.

[14] P. Minnis, S.-M. Szedung, D. F. Young, P. W. Heck, D. P. Garber, C. Yan, D. a. Spangenberg, R. F. Arduini, Q. Z. Trepte, W. L. Smith, J. K. Ayers, S. C. Gibson, W. F. Miller, G Hong, V Chakrapani, Y Takano, L. Kuo-Nan, X. Yu, and Y. Ping, "CERES Edition-2 Cloud Property Retrievals Using TRMM VIRS and Terra and Aqua MODIS Data - Part I: Algorithms," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 11, pp. 4374–4400, 2011.

[15] ——, "CERES Edition-2 Cloud Property Retrievals Using TRMM VIRS and Terra and Aqua MODIS Data - Part II: Examples of Average Results and Comparisons With Other Data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 11, pp. 4374–4400, 2011.

[16] Q. Z. Trepte, P. Minnis, C. Trepte, and S. Sun-Mack, "Improved cloud detection in CERES edition 3 algorithm and comparison with the CALIPSO vertical feature mask," in *13th Conference on Atmospheric Radiation and Cloud Physics*, 2010, pp. 1–7.

[17] H. W. Barker, J. N. Cole, J. Li, and K. von Salzen, "A parametrization of 3-D subgrid-scale clouds for conventional GCMs: Assessment using A-Train satellite data and solar radiative transfer characteristics," *Journal of Advances in Modeling Earth Systems*, vol. 8, no. 2, pp. 566–597, 2016.

[18] C. Johnson, V. Ngo, and M. Rines, "Characterizing the Interaction of Radiance and Retrieved Vertical cloud Structure," 2017.

[19] E. Kize, *CERES CCCM (C3M) Product Information*, [Online], Available at `https://ceres.larc.nasa.gov/products.php?product=CCCM`, 2018.

[20] C. Johnson, "Generating 3D Cloud Fields Through Prediction Using A-Train Data and Machine Learning Techniques," 2017.

[21] V. Ngo, "Improving Data Quality and Data Mining for Vertical Cloud Prediction Models with Higher Resolution Data," 2017.

[22] M. Huguenin, G. Achour, and D. Commun, "Cloud Modeling by Data Fusion," 2018.

[23] NASA, *Cloudsat data processing center*, Online, available at `http://www.cloudsat.cira.colostate.edu`, 2018.

[24] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS ONE*, vol. 12, no. 6, pp. 1–17, 2017.

[25] CloudSat Project, A NASA Earth System Science Pathfinder Mission, "Level 2 Radar - Lidar GEOPROF Product VERSION 1.0 Process Description and Interface Control Document," Tech. Rep., 2007.

[26] ——, "CloudSat MODIS-AUX Auxiliary Data Process Description and Interface Control Document," Tech. Rep., 2017.

[27] ——, "CloudSat 2C-PRECIP-COLUMN Data Product Process Description and Interface Control Document," Tech. Rep., 2017.

[28] Global Modeling and Assimilation Office, Earth Sciences Division, NASA Goddard Space Flight Center, "MERRA-2: File Specification," Tech. Rep., 2016.

[29] N. C. for Environmental Information (NCEI), *Data access*, [Online], Available at `https://www.ncei.noaa.gov/access-ui/data-search;sdate=2011-02-01;edate=2012-08-01?datasetId=global-hourly`.

[30] CloudSat Project, A NASA Earth System Science Pathfinder Mission, "Level 2 GEOPROF Product Process Description and Interface Control Document Algorithm version 5.3," Tech. Rep., 2007.

[31] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[32] B. K. Sotiris, I Zaharakis, and P Pintelas, "Supervised machine learning: A review of classification techniques.(2007)," *Informatica (03505596)*, vol. 31, no. 3, p. 24, 2007.

[33] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008. eprint: `0701907v3`.

[34] C Campbell, "An Introduction to Kernel Methods," in *Radial basis function networks*, Physica Verlag Rudolf Liebing KG Vienna, Austria, Austria ©2001, 2001, pp. 155–192, ISBN: 3-7908-1367-2.

[35] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. eprint: /dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 (http:).

[36] *Atmospheres archive modaps services*, [Online], Accessible at:https://modaps.modaps.eosdis.nasa.gov/services/distribution/archive_atmos.html, 2006.

[37] *L1b ev 1km file specification – terra*, [Online], Accessible at: https://ladsweb.modaps.eosdis.nasa.gov/filespec/MODIS/6/MOD021KM, 2014.

[38] MODIS Characterization Support Team, "MODIS Level 1B Product User's Guide," NASA/Goddard Space Flight Center, Tech. Rep., 2009.

[39] *Mod03 - geolocation - 1km*, [Online], Accessible at:https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/modis-L0L1/MOD03/, 2018.

[40] W.P. Menzel and S. Ackerman, "MODIS Cloud Mask (MOD 35)," *Data Products Handbook*, vol. 2, no. Aqua/MODIS, pp. 118–119,

[41] B. A. Wielicki, B. R. Barkstrom, E. F. Harrison, R. B. I. Lee, G. L. Smith, and J. E. Cooper, "Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing Sytem Experiment," *Bulletin of the American Meteorological Society*, no. 77, pp. 853–868, 1996.

[42] N. A.S. D. Center, "CERES_SYN1deg_Ed4A: Data Quality Summary," Tech. Rep., 2017.

[43] Q. Fu and K. N. Liou, "On the Correlated k -Distribution Method for Radiative Transfer in Nonhomogeneous Atmospheres," *Journal of the Atmospheric Sciences*, vol. 49, no. 22, pp. 2139–2156, 1992.

[44] ——, "Parameterization of the Radiative Properties of Cirrus Clouds," *Journal of the Atmospheric Sciences*, vol. 50, no. 13, pp. 2008–2025, 1993.

[45] M. Liu, Y. J. Kim, and Q. Zhao, "Numerical experiments of an advanced radiative transfer model in the U.S. Navy operational global atmospheric prediction system," *Journal of Applied Meteorology and Climatology*, vol. 51, no. 3, pp. 554–570, 2012.

[46] R. F. Cahalan, L. Oreopoulos, A. Marshak, K. Franklin Evans, A. B. Davis, R. Pincus, K. H. Yetzer, B. Mayer, R. Davies, T. P. Ackerman, H. W. Barker, E. E. Cloth-

iaux, R. G. Ellingson, M. J. Garay, E. Kassianov, S. Kinne, A. Macke, W. O'Hirok, E. E. Takara, T. Varnai, W. Guoyong, and T. B. Zhuravleva, "The i3rc - Bringing Together the Most Advanced Radiative Transfer Tools for Cloudy Atmospheres," *AMERICAN METEOROLOGICAL SOCIETY*, no. September, pp. 1275–1294, 2005.

[47] K. F. Evans, "The spherical harmonics discrete ordinate method for three-dimensional atmospheric radiative transfer," *Journal of the Atmospheric Sciences*, vol. 55, no. 3, pp. 429–446, 1998.

[48] G. I. Marchuk, G. A. Mikhailov, M. Nazareliev, R. A. Darbinjan, B. A. Kargin, and B. S. Elepov, *The Monte Carlo methods in atmospheric optics*. Springer-Verlag, 1980.

[49] I. T. Jolliffe, "Principal Component Analysis, Second Edition," *Encyclopedia of Statistics in Behavioral Science*, vol. 30, no. 3, p. 487, 2002.

[50] *Pca*, [Online], Accessible at: `https://www.mathworks.com/help/stats/pca.html`, 2018.

[51] S. Ackerman, R. Frey, K. Strabala, Y. Liu, L. Gumley, and B. Baum, "MODIS clear sky - cloud algorithm," Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin - Madison, Tech. Rep. October, 2010.

[52] *Langley Fu & Liou Access*, [Online], Accessible at: `https://www-cave.larc.nasa.gov/cgi-bin/lflcode/accesslfl.cgi`, 2016.

[53] F. G. Rose, "Users Guide Ed4 LaRC FuLiou," NASA_LaRC, Hampton, VA, Tech. Rep., 2015.

[54] *gfortran - the GNU Fortran compiler, part of GCC*, [Online], Accessible at: `https://gcc.gnu.org/wiki/GFortran`, 2018.

[55] "U.S. Standard Atmosphere, 1976," National Oceanic, Atmospheric Administration, National Aeronautics, and Space Administration, Washington D.C., Tech. Rep., 1976.

[56] R. Carmichael, *A sample atmosphere table (si units)*, [Online], Accessible at: `http://www.pdas.com/atmosTable1SI.html`.

[57] Y. Hu, D. Winker, M. Vaughan, B. Lin, A. Omar, C. Trepte, D. Flittner, P. Yang, S. L. Nasiri, B. Baum, R. Holz, W. Sun, Z. Liu, Z. Wang, S. Young, K. Stamnes, J. Huang, and R. Kuehn, "CALIPSO/CALIOP cloud phase discrimination algorithm," *Journal of Atmospheric and Oceanic Technology*, vol. 26, no. 11, pp. 2293–2309, 2009.

[58] S. Iwasaki, K. Maruyama, M. Hayashi, S. Y. Ogino, H. Ishimoto, Y. Tachibana, A. Shimizu, I. Matsui, N. Sugimoto, K. Yamashita, K. Saga, K. Iwamoto, Y. Kamiakito, A. Chabangborn, B. Thana, M. Hashizume, T. Koike, and T. Oki, "Characteristics of aerosol and cloud particle size distributions in the tropical tropopause layer measured with optical particle counter and lidar," *Atmospheric Chemistry and Physics*, vol. 7, no. 13, pp. 3507–3518, 2007.

[59] A. J. Heymsfield, C. Schmitt, and A. Bansemer, "Ice Cloud Particle Size Distributions and Pressure-Dependent Terminal Velocities from In Situ Observations at Temperatures from 0 to 86C," *Journal of the Atmospheric Sciences*, vol. 70, no. 12, pp. 4123–4154, 2013.

[60] *Cloud optical depth*, [Online], Accessible at: `http://glossary.ametsoc.org/wiki/Cloud_optical_depth`, 2012.

[61] R. Marchand, T. Ackerman, M. Smyth, and W. B. Rossow, "A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS," *Journal of Geophysical Research Atmospheres*, vol. 115, no. 16, pp. 1–25, 2010.

[62] A. Viúdez-Mora, M. Costa-Surós, J. Calbó, and J. A. González, "Modeling atmospheric longwave radiation at the surface during overcast skies: The role of cloud base height," *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 1, pp. 199–214, 2015.

[63] P. F. Coley and P. R. Jonas, "Back to basics: Clouds and the earth's radiation budget," *Weather*, vol. 54, no. 3, pp. 66–70, 1999.