Geostationary passive retrieval of ice water path with quantile regression neural networks

Master's thesis in Engineering Mathematics and Computational Science

ADRIÀ AMELL TOSAS

DEPARTMENT OF SPACE, EARTH AND ENVIRONMENT

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

Geostationary passive retrieval of ice water path with quantile regression neural networks

Adrià Amell Tosas

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Space, Earth and Environment
Chalmers University of Technology
Gothenburg, Sweden 2021

Geostationary passive retrieval of ice water path with quantile regression neural networks
Adrià Amell Tosas

Supervisors: Simon Pfreundschuh, Department of Space, Earth and Environment
Patrick Eriksson, Department of Space, Earth and Environment
Examiner: Patrick Eriksson, Department of Space, Earth and Environment

Master's thesis 2021
Department of Space, Earth and Environment
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Gothenburg, Sweden 2021

Geostationary passive retrieval of ice water path with quantile regression neural networks
Adrià Amell Tosas
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

Accurate characterisation of clouds can help determine their influence on weather hazards, climate effects, and the hydrological cycle. The retrieval of atmospheric ice mass can contribute to this characterisation and one way to quantify it is with the ice water path (IWP). The polar orbiting CloudSat satellite provides profiles of the ice in clouds, but it suffers from a long revisit period and its cross-track footprint is less than 2 km. This poses a challenge in the study of atmospheric ice variability, both in time and space. Imagers aboard geostationary satellites can provide pictures of one side of the Earth in short time intervals, which can help overcome this limitation.

Current algorithms to retrieve IWP from geostationary passive remote sensors are based on physical approaches or target a specific type of cloud. Quantile regression neural networks (QRNNs) can capture complex relationships for different conditional quantiles of a dependent variable. This can be used to predict a retrieval value, as well as the case-specific uncertainty. This work employs QRNNs to retrieve the IWP distribution using data from the SEVIRI instrument aboard the geostationary Meteosat-9 satellite calibrated against DARDAR, a product derived from combining CloudSat and CALIPSO measurements. The QRNNs are trained and evaluated on a large African region, covering both land and ocean areas.

A multilayer perceptron and a convolutional neural network are compared, and it is seen that the use of spatial information improves both the retrieval and the associated uncertainty. Models trained using all SEVIRI channels show better overall performance in several metrics, although models that use only infrared channels show a relatively similar performance. Moreover, the retrieval with infrared channels shows to satisfactorily retrieve the IWP throughout the diurnal cycle. Models that use visible and infrared channels likely suffer an artefact in the diurnal cycle, but it cannot be completely assessed due to the coverage of the reference data. Monthly means and diurnal variations are compared with a physical-based IWP retrieval for two tropical areas, and it is found correlations that favour the QRNN models. Finally, it is suggested that the models trained on Meteosat-9 observations may also be used on observations from other Meteosat satellites, expanding the usage of the models beyond the lifetime of Meteosat-9.

Keywords: quantile regression, neural networks, ice water path, SEVIRI.

Acknowledgements

First of all, I would like to express my appreciation to Patrick Eriksson for his patient guidance in this work. I would also like to extend my gratitude to Simon Pfreundschuh for his enthusiastic encouragement and worthwhile technical meetings. I also wish to thank Ingrid Ingemarsson who made the experience of working from home easier through virtual fika breaks.

Adrià Amell Tosas, Gothenburg, June 2021

Contents

Abbreviations

API Application Programming Interface
CiPS Cirrus Properties from SEVIRI
CLAAS CLoud property dAtAset using SEVIRI
CNN Convolutional Neural Network
CPP Cloud Physical Properties algorithm
CRPS Continuous Ranked Probability Score
HRV High-Resolution Visible channel
IQR Interquartile Range
IR InfraRed: SEVIRI channels without a solar contribution
IWC Ice Water Content
IWP Ice Water Path
LST Local Solar Time
MAE Mean Absolute Error
ME Mean Error
MLP Multilayer Perceptron
MSG Meteosat Second Generation
NaN Not a Number
PDF Probability Density Function
QRNN Quantile Regression Neural Network
RMSE Root Mean Square Error
RSS Rapid Scanning Service
SatCORPS Satellite ClOud and Radiation Property retrieval System
SEVIRI Spinning Enhanced Visible and InfraRed Imager
UTC Coordinated Universal Time
VISIR VISible and InfraRed: all SEVIRI channels

List of Figures

List of Tables

1

Introduction

Clouds are an aggregate of minute liquid droplets or ice crystals suspended in the atmosphere which are key to the planet. In the global average, roughly 70 % of the Earth's surface is covered with clouds, with oceans covered by approximately 10–15 % more cloudiness than land (Stubenrauch *et al.*, 2013). They are a crucial part of the energy balance, climate, and weather on Earth. Some contribute to cooling by reflecting the energy of the Sun, while others keep the planet warm by retaining some of the heat emitted by the Earth, its thermal infrared radiation. Cloud systems help spread the energy received from the Sun evenly. They are also part of the hydrological cycle and thus influence the water resources.

Small changes in the distribution of clouds can reflect changes in the climate and vice versa. Generally speaking, the effect of clouds depends on their coverage on Earth, thickness, particles, and water and ice contents. Realistic computer simulations of current and future climate require all these factors of clouds to be accurately represented, but important uncertainties remain (Boucher *et al.*, 2013). Ground-based measurements can make significant contributions, but are limited mostly to land areas. Satellite remote sensing instruments do not suffer from this limitation, and they can provide observations on a global scale, including remote ocean and land regions.

The ice phase is estimated to be involved in 30 % of the precipitation events in the tropics, with this percentage increasing with latitude (Field & Heymsfield, 2015). Therefore, the atmospheric ice mass is a considerable factor for the modelling community, but it has been recognised that it is a significant challenge in remote sensing (Waliser *et al.*, 2009). Optical and infrared sensors can detect ice clouds with great sensitivity, but the attenuated signal measured from the clouds means that mainly cloud-top information is obtained with these passive sensors. To quantify the ice water content (IWC), the amount of ice per cubic meter of air, lower frequencies are required to sense through the clouds, either with passive microwave radiometers or radars. The former sensors can only provide implicit vertical information, and vertical profiling of the clouds requires active sensors.

The launch of CloudSat (Stephens *et al.*, 2002) in 2006 was a landmark for atmospheric ice quantification with remote sensing thanks to its Cloud Profiling Radar instrument. Before CloudSat, the ice water path (IWP), which is the vertically integrated IWC, could be retrieved from geostationary and polar orbiting satellites, though with large errors. Nevertheless, CloudSat suffers from a long revisit period, as it is a polar orbiting satellite, together with a limited spatial cross-track sampling. This poses a challenge in the study of the variability of atmospheric ice, both in time and space. The retrieval of

IWP using geostationary satellites can facilitate this study owing to their large spatial and temporal coverage.

The development of methods to retrieve IWP using the passive sensors aboard geostationary satellites can be non-trivial, as these sensors are intended to measure other properties, and their development can be limited by human knowledge. Another limitation is the sensitivity of the observation systems together with measurement errors. This last limitation makes the assignment of a single value in an IWP retrieval impossible. Thus, even in the absence of errors, a retrieval should have an uncertainty estimate.

The posterior probability distribution from the Bayesian framework can be used for this type of problems. This distribution is obtained by combining prior knowledge, described by the prior probability distribution, with the observations. Monte Carlo methods can be used to obtain the posterior distribution from which a retrieval value and a case-specific uncertainty can be computed. However, each retrieval with Monte Carlo methods can require a prohibitive computational cost or the traversal of a large retrieval database.

Machine learning methods are not limited by human knowledge because they identify patterns from data and make decisions with minimal human intervention. As opposed to Monte Carlo methods, they can make efficient estimations at the expense of a training step, but often no uncertainty or a general error for the model is given. Quantile regression neural networks (QRNNs) can capture nonlinear patterns from the data and estimate a retrieval value as well as an associated uncertainty consistent with uncertainties from Monte Carlo methods (Pfreundschuh *et al.*, 2018).

1.1 Aims and scope

This work aims to answer two questions. Firstly, whether quantile regression neural networks can provide a satisfactory IWP retrieval from passive observations with high spatial and temporal coverage, taking into account both the estimated value as well as the uncertainty. The second objective is to find whether taking into account spatial information improves the IWP retrieval. The latter is investigated by comparing neural networks that take into account spatial information with networks that do not.

For these purposes, observations from the Spinning Enhanced Visible and InfraRed Imager (SEVIRI) instrument (Schmid, 2000) aboard Meteosat-9, the second satellite in the Meteosat Second Generation geostationary satellites, are calibrated against DARDAR (Delanoë & Hogan, 2010), a data product that combines CloudSat observations with lidar measurements. Data from these two sources is publicly available and is ready for further processing. In April 2011 CloudSat suffered a spacecraft battery anomaly which forced the satellite to operate only in daylight, that is obtaining only daytime observations. Consequently, the period of time considered is bounded by this event. For reasons explained in Section 2.2, only data from 6 May 2008 onwards is considered. Thus, data from almost three years is used. The region of study is presented in Figure 1.1. To minimise the human intervention in the decision of which data should be used in the models, a naive approach is taken: the channels used are either all SEVIRI channels, except its high-resolution visible channel, or only channels without a solar reflectance contribution, that is, the infrared channels.

These two sets of channels considered for the models are motivated by nighttime IWP retrievals. An ultimate goal is to provide diurnal variations of IWP. Retrievals that strongly rely on visible channels, that is, channels that require solar reflectance, are not applicable at nighttime. Therefore, this leads to compare the performance of retrievals that only use infrared channels.

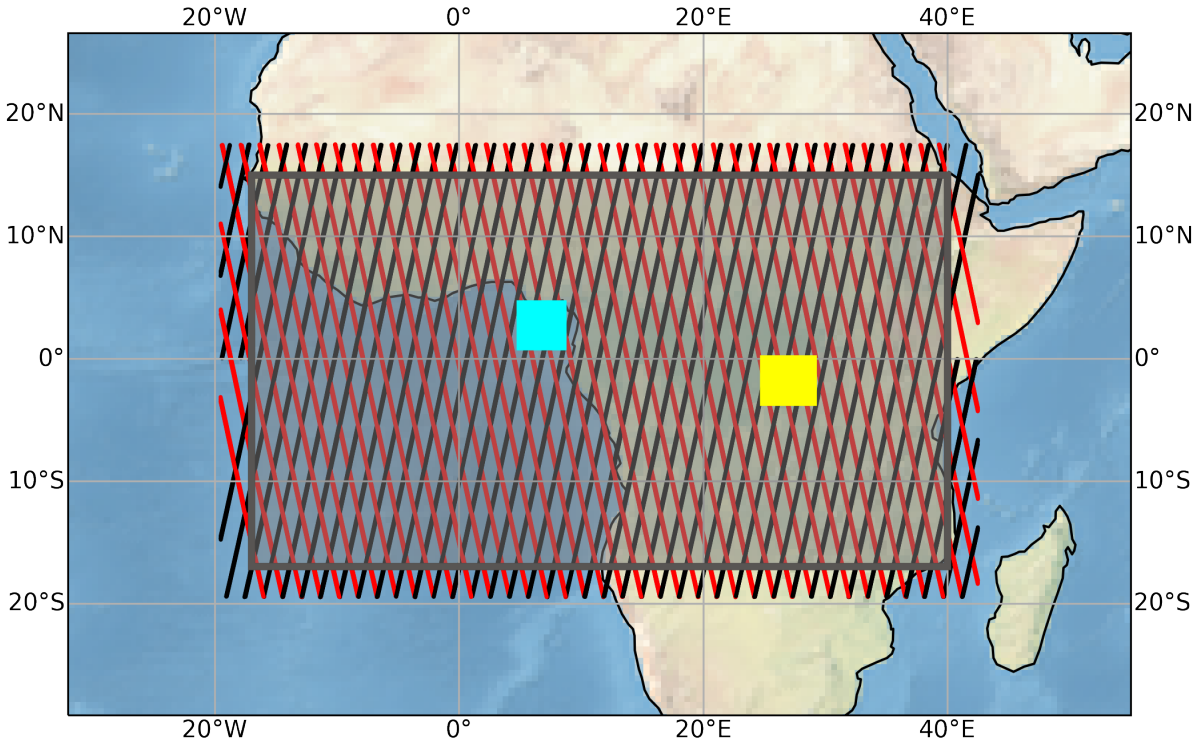


Figure 1.1: The region considered in this work delimited by the grey rectangle, ranging between −17° and +15° in latitude and between −17° and +40° in longitude. The blue and yellow areas correspond to the ocean and land areas, respectively, of Table 1.1. The red and black lines indicate the DARDAR swath for daytime and nighttime measurements, respectively. The DARDAR swath is not to scale: it is more narrow in reality.

Additionally, monthly IWP means and monthly IWP diurnal variations are compared with CLAAS-2.1 (Finkensieper *et al.*, 2020a), a publicly available dataset that provides IWP retrievals and is presented in Section 2.4. The comparison is performed in two tropical areas. One area expands over ocean on the Gulf of Guinea (hereinafter ocean area), and the other over the rainforests on the western Democratic Republic of Congo, covering part of Lomami National Park and Maiko National Park (hereinafter land area), both presented in Table 1.1 and illustrated in Figure 1.1. The size of both areas were determined to obtain data that matched the spatial size of the training data.

Table 1.1: The areas in the comparison with CLAAS.

Area	**Latitude**		**Longitude**	
	min.	max.	min.	max.
Ocean	1	4.5	5.0001	8.5
Land	−3.55	0	24.92	29

1.2 Related works

Previous studies on the retrieval of IWP from geostationary satellites can be divided into two groups: physical-based approaches or machine learning approaches, where this work falls in the latter. Both the Algorithm for the Physical Investigation of Clouds with SEVIRI (Bugliaro *et al.*, 2011) and the Cloud Physical Properties (CPP) algorithm (Roebeling, Feijt & Stammes, 2006) use the solar reflectance from the 0.6 µm and 1.6 µm SEVIRI channels to estimate the cloud optical thickness τ and particle effective radius r_{eff}, based on the method from Nakajima and King (1990). Assuming no vertical variations and an ice density ρ_{ice}, they compute the IWP as

$$\text{IWP} = \frac{2}{3}\rho_{\text{ice}}\tau r_{\text{eff}} \tag{1.1}$$

therefore they result a physical-based approach to the IWP retrieval, and only applicable to daytime observations.

The Satellite ClOud and Radiation Property retrieval System (SatCORPS) (Minnis *et al.*, 2008; Trepte *et al.*, 2019) comprises a set of algorithms to retrieve cloud properties for both daytime and nighttime adapted for polar-orbiting and geostationary imagers, including the Meteosat Second Generation series. The retrievals estimate τ and r_{eff} with physical approaches, and the IWP is obtained similar to Equation (1.1). Yost *et al.* (2021) remark that the Ice Cloud Optical Depth from Infrared using a Neural network method (Minnis *et al.*, 2016), which retrieves τ, may alleviate shortcomings for thick clouds at nighttime.

Finally, the Cirrus Properties from SEVIRI (CiPS) (Strandgren *et al.*, 2017) uses the thermal channels from SEVIRI to train neural networks against ice cloud properties retrieved with the Cloud Aerosol Lidar with Orthogonal Polarization (CALIOP) instrument (Winker *et al.*, 2009) and provides IWP estimates. This machine learning method applicable to daytime and nighttime retrievals targets only thin ice clouds.

2

Data sources

The construction of an IWP retrieval model requires data to train and evaluate it. In particular, reference data that is used as the ground truth, and input data to the model from which the target values are retrieved. In this work, the DARDAR-cloud product is used as ground truth and the Meteosat-9 SEVIRI data as input data. All data available between 6 May 2008 and 31 March 2011 was used to train and evaluate the models. This chapter further details how the data was acquired, processed and collocated, and also introduces the CLAAS dataset.

2.1 DARDAR

CloudSat (Stephens *et al.*, 2002) is a NASA Earth observation satellite launched 28 April 2006. It is a polar orbiting sun-synchronous satellite, designed to have a period of 99 minutes and revisit period of 16 days, crossing the equator at approximately 13:45 local solar time, and consequently 01:45 in the night. It carries the Cloud Profiling Radar (CPR), which has a cross-track resolution of approximately 1.4 km and operates at 94 GHz. One of its data products is IWC measurements.

The CALIPSO satellite (Winker, Pelon & McCormick, 2003) is another polar orbiting sun-synchronous Earth observation satellite from a joint mission from NASA and CNES, launched in the same vehicle as CloudSat. It was designed to fly in tandem with CloudSat only a few seconds apart. The CALIPSO satellite carries a lidar instrument which provides vertical profiles of clouds at 30 to 60 m resolution. Clouds can rapidly change and transform; the proximity between CloudSat and CALIPSO allows observations of the same clouds with a radar and a lidar within a small time difference.

The DARDAR-cloud product (Delanoë & Hogan, 2010) synergistically combines the CloudSat and CALIPSO measurements to provide at the CloudSat horizontal resolution and CALIPSO vertical resolution different cloud properties. When the CPR signal is unavailable, for example in optically thin ice clouds, the lidar facilitates an accurate retrieval. One of the properties provided by this product is IWC.

In this work, the DARDAR-cloud version 2.1.1 was used, which provides IWC for heights between −1020 m and 25 080 m above sea level at a 60 m resolution. An initial analysis on the IWC values from DARDAR-cloud revealed that some measurements presented IWC for negative heights, and in particular very negative heights, which is likely an artefact. For the region considered, it was found that the relative contribution of the IWC to the total IWP for altitudes below 2 km was less than 10^{-6} for any measurement. It is also reasonable to assume that there is no frozen precipitation below this altitude

for this region. Consequently, and to avoid any other possible artefact, the IWP was obtained by integrating the DARDAR-cloud IWC from 2 km. No further preprocessing on the data values was performed. As from April 2011 the DARDAR product only exists for daylight operations, all DARDAR data available before this month was used in order to not introduce any bias that had to be taken into account in the models. The data was downloaded from AERIS/ICARE Data and Services Center (2021).

2.2 Meteosat-9 SEVIRI

The Meteosat Second Generation (MSG) series from EUMETSAT consists of four geostationary satellites numbered 1 to 4 which provide images of the full Earth disc every 15 minutes. The Spinning Enhanced Visible and InfraRed Imager (SEVIRI) (Schmid, 2000) is the primary instrument of the MSG series and provides observations of the Earth in 12 spectral channels. The SEVIRI instrument has four visible and near infrared channels, including a high-resolution visible (HRV) channel which is a broadband channel between 0.4 µm and 1.1 µm, and eight infrared channels. Excluding the HRV channel, which has a resolution of 1 km, the SEVIRI channels have a resolution of 3 km at the nadir point and are presented in Table 2.1. The signal for the visible and near infrared channels comes from solar reflectances and from thermal emissions for the infrared channels. Despite the 3.9 µm channel is categorised as infrared, its signal has a solar contribution during daytime (Kerkmann, 2004). This work excluded the HRV channel.

Table 2.1: The SEVIRI channels, excluding the HRV channel.

Spectrum	**Visible**		**Near infrared**	**Infrared**							
Channel (µm)	0.6	0.8	1.6	3.9	6.2	7.3	8.7	9.7	10.8	12.0	13.4

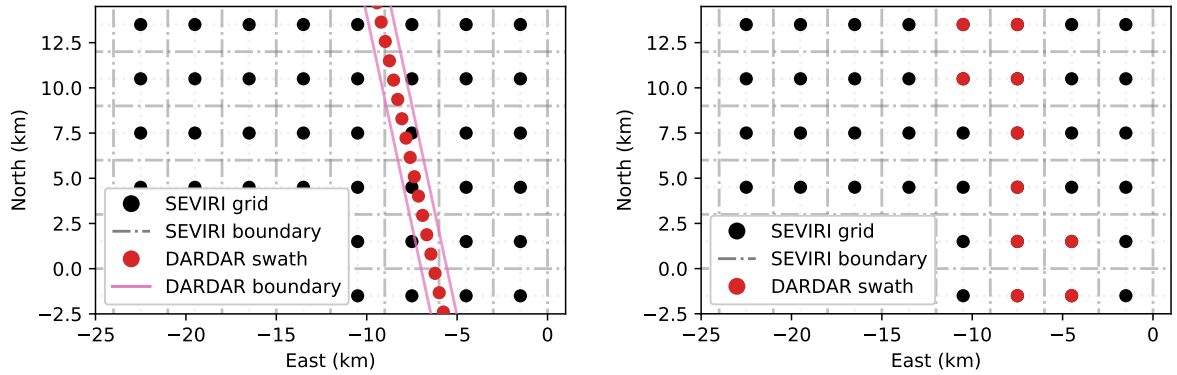
For the period of time considered for DARDAR, that is before April 2011, Meteosat-8 and Meteosat-9 (which are also named MSG1 and MSG2, respectively) were the only MSG satellites that were operating. Currently, for the operational time of CloudSat before the battery malfunction, there is much more Meteosat-9 data available than for Meteosat-8 in the EUMETSAT Data Store (EUMETSAT, 2021a). Both satellites can provide observations at the same time or cover missing observations from the other. To reduce the complexity and any potential difference between satellites, only Meteosat-9 observations were chosen as the data source.

Meteosat-9 was located at 0° longitude before April 2011. The High Rate SEVIRI Level 1.5 Image Data 0 – MSG – degree product provides geolocated and radiometrically pre-processed image data every 15 minutes (EUMETSAT, 2017, 2019). All dates available for this product were downloaded from the EUMETSAT Data Store using its API. There is data processed with two different algorithms for dates before 6 May 2008. Consequently, only data from this date onwards was used to not have to take into account possible differences between the algorithms.

2.3 Collocations

To use the DARDAR data as the reference data for retrieving IWP from SEVIRI observations the two datasets have to be both spatially and temporally collocated. SEVIRI has a regular sampling pattern, which consists of a grid with a spatial resolution of $0.05° \times 0.05°$ latitude by longitude. On the other hand, DARDAR measurements are along a track with a cross-track resolution of approximately 1.4 km.

The Meteosat-9 SEVIRI data is provided in a binary format, the MSG native archive format. The Satpy Python library (Raspaud *et al.*, 2021) incorporates a reader for this file format, and it reprojects the observations in a regular 3 km × 3 km grid when the data files are read. The DARDAR files are provided in HDF format, which can be opened with different libraries, and no reprojection is performed when reading them. Given that both datasets have different spatial resolutions, the initial situation is illustrated in Figure 2.1a. To spatially collocate the datasets, the DARDAR swath was resampled to the location of the SEVIRI regular grid. The resampling consisted of a nearest neighbour algorithm with a constant radius of influence, which is the maximum distance at which a pixel in the SEVIRI grid could be assigned a DARDAR data point. Given that the SEVIRI pixels are squares of side 3 km, the maximum distance from their centres to a vertex is approximately 2.13 km. This was the value chosen for the radius of influence, and the result is illustrated in Figure 2.1b.



(a) Before resampling.

(b) After resampling.

Figure 2.1: The SEVIRI grid is reprojected by the Satpy library upon reading the files. The DARDAR swath in (a) indicates the position of each measurement in the source data files. The boundaries indicate the spatial resolution of each dataset. In (b), the DARDAR swath was resampled to the SEVIRI grid assigning the nearest neighbour to each pixel with a radius of influence of approximately 2.13 km.

The SEVIRI instrument takes almost 15 minutes to scan the full Earth disc. This produces a dataset every 15 minutes. Therefore, each SEVIRI observation corresponds to a range of timestamps. On the other hand, each DARDAR measurement has a single timestamp. All DARDAR measurements roughly taken during a SEVIRI scan are assigned to that scan. An analysis of the SEVIRI metadata revealed that there is a gap of approximately 2 minutes and 30 seconds between the end of one scan and the start of the next one. This implies that DARDAR measurements taken within these time windows are not assigned a SEVIRI observation following this method. Within this time, CloudSat travels about one third of the region considered, and therefore this

is non-negligible. Each DARDAR measurement taken within these time windows was assigned the SEVIRI observation with the closest timestamp. For simplicity and after the temporal collocation, the SEVIRI observations were considered to have a single timestamp, corresponding to approximately the scan start time.

The collocated data was divided into non-overlapping square images[1] of side 128×128 pixels, with the DARDAR swath centred in the image. The latitude at which the image starts was randomly selected from a uniform distribution in order to reduce any potential bias; it was ensured that the region of study was always covered, so the data covered slightly more of it. Data files that contained at least one channel full of NaNs were discarded. A data split which respected the IWP distribution was performed. The split randomly assigned 60%, 20% and 20% of the images to a training, validation and test set, respectively (Table 2.2). Each image contained the information from the 11 SEVIRI channels considered, the IWP from the resampled DARDAR swath, the coordinates of each pixel as well as the time of the Meteosat-9 SEVIRI observation.

Table 2.2: Number of files and size for each split.

Data group	Train	Validation	Test
Number of files	20 812	6937	6937
Disk space (GB)	15.3	5.1	5.1

Both data sources provide their timestamps in UTC, so no time conversion was required for the temporal collocation. Considering a time variation from $-12\,\mathrm{h}$ to $12\,\mathrm{h}$ and a corresponding longitude from $-180°$ to $180°$, the local solar time (LST) for a given longitude can be approximated from UTC time as

$$\mathrm{LST} = \mathrm{UTC} + \frac{12\,\mathrm{h}}{180°} \times \mathrm{longitude} \tag{2.1}$$

with LST and UTC expressed in hours. With this approximation, the collocated data covers two time periods: a daytime period between roughly 13:15 and 14:15 LST and a nighttime period between roughly 01:15 and 02:15 LST. Note that these times were computed from the whole images produced, which include pixels without ground truth, as illustrated in Figure 2.1. The resampled DARDAR swath covers pixels roughly between 13:20 and 14:00 LST and between 01:20 and 02:00 LST.

2.4 CLAAS

The cloud property dataset using SEVIRI edition 2.1 (CLAAS-2.1, and CLAAS hereinafter) (Finkensieper *et al.*, 2020a) is a publicly available cloud properties data record from SEVIRI measurements between 2004 and 2015. One of the properties in this dataset is the IWP provided at SEVIRI native temporal and spatial resolution. The IWP retrieval for CLAAS is performed using the CPP algorithm. CLAAS further provides precomputed IWP monthly mean diurnal cycles and monthly means for the full Earth

[1] Here image is a 3D array of values with spatial dimensions that correspond to geographical coordinates and a channels dimension that corresponds to the SEVIRI channels and resampled DARDAR IWP.

disc. These two products were the datasets used in the comparison with the QRNN models.

For the year 2012, all CLAAS metadata indicates that Meteosat-9 was the data source, making this year convenient for the comparison. According to Figure 10.e from Benas *et al.* (2017), CLAAS indicates that the two tropical areas considered for the comparison are of high IWP on average.

3

Modelling

The retrieval of IWP from passive remote sensing data is performed with mathematical models. An IWP retrieval model is a function f that when data x is inserted in the function the IWP is estimated in some form from the output of $f(x)$. As reasoned in Chapter 1, uncertainties are required. Therefore, f should also estimate these values.

Models can be more transparent, such as a nonlinear equation following a physical approach, or a rather black-box machine learning model, as in artificial neural networks. The former is adequate if interpretability of the model is key, since it can reveal the physical causes of the retrieved IWP. The latter can capture patterns that are hard to identify, but artificial neural networks are often difficult to characterise and are mainly restricted to empirical evaluation.

3.1 Artificial neural networks

3.1.1 A mathematically idealised animal brain

Artificial neural networks are models inspired by the biological networks constituting an animal brain. Neurons are the main component of these biological networks. Mathematically speaking, they are connected to process data and therefore exhibit some dynamics. The input to a neuron can activate it, in which case an electrical impulse is sent to the nervous system. As an animal grows, it learns to process patterns and respond to different stimuli: the biological networks change over time establishing new or different connections between neurons.

Artificial neural networks use idealised neuron models. Artificial neurons are the mathematical functions used as the elementary units of an artificial neural network. An artificial neuron processes incoming signals and transfers an output signal to neurons connected to it. The output signal is the result of computing the input signal with a function, the activation function, which is often nonlinear. The input signal is composed of all the signals sent by the neurons connected to the neuron in question, usually computed as a weighted sum with a bias. That is, for a neuron i with activation function g_i its output signal x_i has the following form

$$x_i = g_i\left(\sum_{j \in \mathcal{N}_i^{\text{in}}} w_{ij} x_j + \theta_i\right) \tag{3.1}$$

where $\mathcal{N}_i^{\text{in}}$ is the set of neurons with incoming connections for neuron i, w_{ij} are the weights, and θ_i the bias. There are several concepts of artificial neural networks, but

one of the most common consists of neurons arranged in layers connected only in one direction. They are called feedforward networks, as the input signal to the network is forward propagated through the layers. Neural networks can also have connections that shortcut certain layers, known as residual connections. The layers between the input and output layers are called hidden layers and the neurons therein hidden neurons. Networks with many hidden layers are called deep neural networks, and those with many hidden neurons per layer wide.

The design of a neural network is a complicated problem due to the endless possibilities to choose from. It can even be claimed to be more an art than a science (Erenshteyn, Foulds & Galuska, 1994), as it is often a compromise between the knowledge, expertise and intuition of the investigator, the available software libraries and the computational power. The design starts with a choice of paradigms to include in the network architecture. The choices are then tuned by mainly determining the activation functions, the number of layers, the number of neurons in each layer, and the values of the learnable parameters, which include weights and biases. Tuning, in this context, stands for maximising a satisfactory answer to the question posed, which often includes a satisfactory generalisation to data not used in training.

3.1.2 Mathematical principles

There are a few techniques to maximise a satisfactory answer. One technique is supervised learning, in which the training data contains a list of input patterns together with a list of reference values. Training is performed by minimising the value of a function, the loss function, which represents the cost of estimating values different than the reference values. Minimisation of the loss function can be achieved with optimisation algorithms, in particular iterative algorithms.

Given a loss function $\mathcal{L}$ with a set of learnable parameters $\mathcal{P}$, the gradient descent method takes steps in the opposite direction of the gradient of the loss function at the point at which it is evaluated, as it is the steepest descent direction. In mathematical terms and for a given iteration, the learnable parameters are updated with the following rule at iteration t

$$u^{(t+1)} = u^{(t)} + \Delta u^{(t)}, \quad \Delta u^{(t)} = -\eta \frac{\partial \mathcal{L}}{\partial u^{(t)}}, \quad \text{where } u \in \mathcal{P} \tag{3.2}$$

where η is called the learning rate. It requires that $\mathcal{L}$ is differentiable, where the derivative with respect to $u^{(t)}$ indicates differentiation with respect to u evaluated at $u^{(t)}$. Repeated iterations with a suitable choice of learning rate will lead to local minima.

In Equation (3.2), the gradient for each parameter will generally need to make use of the chain rule. An efficient implementation consists of computing the gradients one layer at a time, propagating them backwards from the output layer. This method of computing the gradients is generally known as backpropagation and makes training efficient.

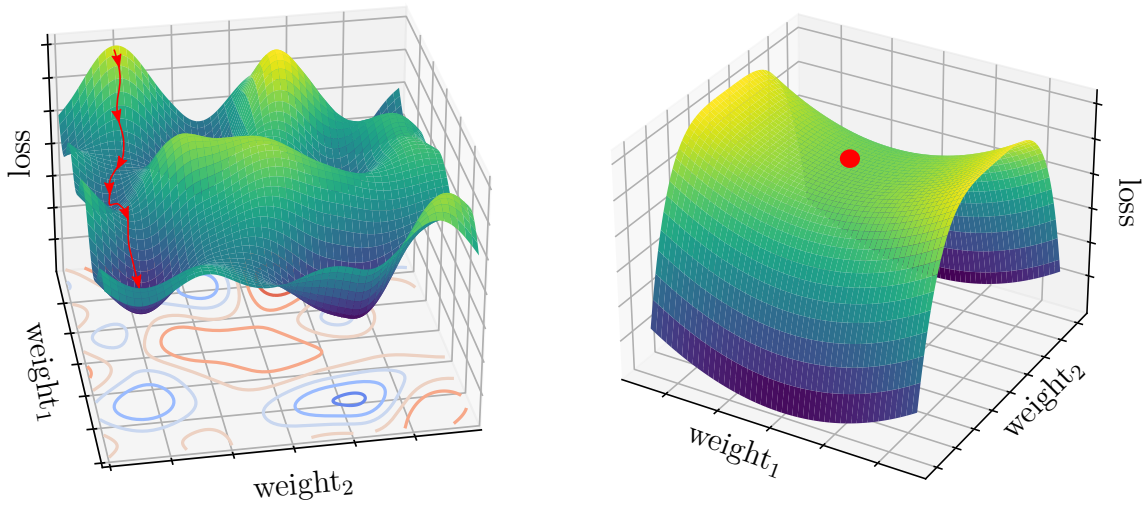
Neural networks are essentially large compositions of functions with a large number of learnable parameters with multiple local minima. This poses at least three problems. Firstly, avoiding bad local minima, that is, minima far from optimal minima. Secondly, the choice of a learning rate. Thirdly, numerical stability.

Stochastic gradient descent (SGD) is an optimisation method based on gradient descent. It replaces the gradient of the loss function with an estimate computed on a subset of the data samples. This reduces the computational burden, particularly for large datasets. Usually, the training data is shuffled and randomly split into batches of the same size, called mini-batches. After propagating a mini-batch through the network, SGD computes the gradient of the loss function evaluated with the mini-batch and updates the learnable parameters. A training epoch is completed when all training data has been observed by the network and the learnable parameters have been updated. Afterwards, the data is usually shuffled again and new mini-batches are defined. SGD also facilitates avoiding bad local minima. The idea is that not taking a step in the direction of steepest descent for all training data can lead to better local minima. This is implicitly achieved by using mini-batches.

However, if a bad local minimum is reached, neither gradient descent nor SGD can escape from it without further improvements. The momentum method is an improvement to the optimisation method, or optimiser in short, that linearly combines the current value of the gradient with the previous update, so Equation (3.2) at iteration t becomes

$$u^{(t+1)} = u^{(t)} + \Delta u^{(t)} + \frac{\alpha}{\eta}\Delta u^{(t-1)}, \quad \Delta u^{(t)} = -\eta\frac{\partial \mathcal{L}}{\partial u^{(t)}}, \quad \text{where } u \in \mathcal{P} \tag{3.3}$$

where $\alpha \in [0, 1]$ is the momentum parameter that determines the contribution of earlier gradients to the current gradient. This is particularly useful for escaping from shallow minima coming from a steep descent, as in Figure 3.1, or when the loss reaches a saddle point or is located on a plateau, as in Figure 3.1b.



(a) Reaching a local minimum.

(b) Non-strict saddle point.

Figure 3.1: Schematics of a loss function landscape. The lowest minimum is not reached but convergence to a shallow minimum is avoided, represented in (a), and navigating the loss landscape can be slow if it has plateaus and saddle points, indicated in (b) with a red point.

The updates on the network parameters depend on the value of the learning rate. If it is set too large, the network may not converge, and if it is too small convergence

can be too slow. Thus, it needs to be carefully chosen. Moreover, there can be learnable parameters that change more rapidly the loss function value than others. Adam (Kingma & Ba, 2015) is an optimisation algorithm that combines an adaptive learning rate with momentum (the name is derived from adaptive moment estimation). The learning rate is adapted for each weight and bias by dividing it with an average of recent gradients. As a consequence, this optimiser is more forgivable to a bad initialisation of the parameters or choice of the learning rate, and can also achieve convergence much faster.

Additionally, a learning rate scheduler is often chosen. The scheduler modifies the learning rate as the training proceeds. A typical scheduler reduces the learning rate at prescribed epochs. Another common scheduler consists of reducing the learning rate when a metric stops improving. Cosine annealing and cyclic schedulers decrease and increase the learning rate. This can allow escaping bad local minima, but also escaping from good local minima if the maximum learning rate is too large. Warm restarts can also be used, that is, restarting the scheduler but not the values of the learnable parameters.

Regardless of the choice of optimiser and learning rate scheduler, the network requires an initialisation of its learnable parameters. The initialisation usually consists in assigning random values to each of the learnable parameters. Different trainings of the same model can reach different local minima with different initalisations and mini-batch permutations, and while good minima they may not be the best or satisfactory, as in Figure 3.1. If possible, it is recommended to repeat a training with different initialisations.

The training can be made easier and faster by normalisation of the inputs, as it can be expected that the parameters will have the same range of values. This can produce a more symmetric loss function. One of the most common normalisation methods is standardisation, which applies a precomputed and fixed pair of parameters, a sample mean μ and a sample standard deviation σ, to any input x as

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{3.4}$$

so the standardised inputs $\tilde{x}$ have zero mean and unit variance.

For networks of several layers, the normalisation of the input features will have little effect on the output layer. Batch normalisation (Ioffe & Szegedy, 2015) is a method that normalises the input values in an activation function[1] achieving faster convergences. When training and for a given hidden neuron, the mean $\mu_{\mathcal{B}}$ and standard deviation $\sigma_{\mathcal{B}}$ are computed from the values $v_i \in \mathcal{B}$, where $\mathcal{B}$ is a mini-batch of size m. After standardising the values[2], the normalised value $\tilde{v}_i$ is obtained after scaling and shifting it with additional parameters γ and β which become learnable parameters:

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i}^{m} v_i, \quad \sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i}^{m}(v_i - \mu_{\mathcal{B}})^2 \tag{3.5}$$

$$\hat{v}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \tag{3.6}$$

$$\tilde{v}_i = \gamma \hat{v}_i + \beta \tag{3.7}$$

[1]There is some debate about whether the normalisation should be applied before or after the activation function. The text describes the approach used in this work.

[2]A small value ϵ is added for numerical stability to avoid cases in which the variance is very small.

When evaluating, Equation (3.5) is replaced with the population statistics.

Normalisation of the data can help achieve faster convergence and numerical stability, since the values are expected to be in the same range. In addition, fast convergence and in particular numerical stability are highly influenced by the choice of activation functions. Given that the network is a composition of functions, the chain rule is applied in the majority of the cases when differentiating the loss function. Activation functions, such as the sigmoid function, can have nearly zero derivatives if the point at which they are evaluated is large in absolute value. If in an application of the chain rule a few values are close to zero, the numerical value can result nearly zero. Then it is said that the gradient vanishes. This makes training slow, as the network will barely update the learnable parameters, even if required, as it cannot see their effect at the output layer. Layers far away from the output can be heavily affected by the multitude of factors in the chain rule.

The rectified linear unit (ReLU) activation function and its derivative are defined by

$$g(x) = \max(0, x), \quad \frac{\mathrm{d}g(x)}{\mathrm{d}x} = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}. \tag{3.8}$$

It has the main advantages of reducing the vanishing gradient problem by having a constant derivative and that it is efficient to compute. However, the derivative is undefined at zero, although implementations use *ad-hoc* solutions, and it can also suffer the dying ReLU problem. Neurons might be set in a state where they become inactive for all inputs. In the inactive state the ReLU gradient is zero, so gradients are not backpropagated. Then the neurons get in a permanent inactive state, as the network cannot update them. Furthermore, ReLU is unbounded. If large gradients are accumulated, then it is said that there is an exploding gradient and the loss function can diverge.

3.1.3 Quantile regression neural networks

Least squares regression finds the average response of a dependent variable y_i, $i = 1, \ldots, n$, from a set of independent variables x_i by finding a function f such that it minimises the sum of squared residuals S, that is,

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2. \tag{3.9}$$

Therefore, it only provides a partial view of the relationship between the dependent and independent variables. On the other hand, quantile regression models the relationship between one or more quantiles of the dependent variable and the independent variables. In other words, it can give the relationship at different points of the conditional distribution of the dependent variable on the independent ones, which makes it more robust to outliers. Quantile regression is then well suited to model a context with heteroskedasticity, providing a way to explore heterogeneity in the relationships[3]. If several quantiles for a dependent variable are predicted, then its conditional distribution from the data can

[3]In a situation where the true distribution of the data has, for example, a non-constant variance (heteroskedastic variance) but a constant mean, the usual least squares regression would only provide an estimation of the constant mean. That is, it would not explain the non-constant variance. However, the regression for different quantiles could provide a good picture of the non-constant variance.

be reconstructed. The reconstructed distribution can be used to make predictions with associated uncertainties.

For a cumulative distribution function $F(x)$, the quantile x_τ at level $\tau \in [0, 1]$ is the value such that

$$x_\tau = \inf\{x : F(x) \geq \tau\}. \tag{3.10}$$

The expectation with respect to x of the loss function

$$\mathcal{L}_\tau(x_\tau, x) = \begin{cases} \tau|x - x_\tau| & \text{if } x_\tau < x \\ (1-\tau)|x - x_\tau| & \text{otherwise} \end{cases} \tag{3.11}$$

is minimised by the quantile x_τ (Koenker, 2005, pp. 5–6). The loss function $\mathcal{L}_\tau$, also called pinball loss, is the function minimised in quantile regression. The neural networks that seek to minimise the pinball loss are called quantile regression neural networks (QRNNs). They have the advantage that they can easily capture nonlinearities. For a set of quantile levels $\mathcal{Q}$ and mini-batch B, and an indicator function $\mathbb{1}_\mathcal{V}(x) = \{1 \text{ if } x \in \mathcal{V}; 0 \text{ otherwise}\}$, where $\mathcal{V}$ is the set of valid points, a QRNN minimises the loss

$$\mathcal{L}^{(B)} = \frac{\sum_{i \in B} \sum_{\tau \in \mathcal{Q}} \mathbb{1}_\mathcal{V}(x_i) \mathcal{L}_\tau(\hat{x}_{i,\tau}, x_i)}{\sum_{i \in B} \sum_{\tau \in \mathcal{Q}} \mathbb{1}_\mathcal{V}(x_i)} \tag{3.12}$$

at the end of each mini-batch step, and approximates the total loss $\mathcal{L}$ from the set of all mini-batches $\mathcal{B}$ with a weighted sum

$$\mathcal{L} = \frac{\sum_{B \in \mathcal{B}} w_B \mathcal{L}^{(L)}}{\sum_{B \in \mathcal{B}} w_B}, \quad w_B = \sum_{i \in B} \mathbb{1}_\mathcal{V}(x_i) \tag{3.13}$$

where $i \in B$ indicates the element i of B with an abuse of notation, and x_i and $\hat{x}_{i,\tau}$ indicate the reference value and predicted quantile at level τ, respectively.

It is important to remark that when a QRNN minimises the pinball loss it does not take into account the error[4] values, which would happen with a least squares regression. This comes from that minimisation of the pinball loss is achieved with some form of gradient descent, where the derivative of the loss does not include the error value; underprediction is penalised with τ and overprediction with $(1 - \tau)$.

3.1.4 Neural network architectures

A neural network architecture is the arrangement of the different units that constitute a neural network model. This includes determining the number of neurons and layers the model should include, and how these are connected. Generally, the choice of a satisfactory architecture cannot be done before any empirical evaluation, and their design is restricted to the intuition of the investigator and their choice of paradigms, as mentioned in Section 3.1.1.

3.1.4.1 Multilayer perceptron

The multilayer perceptron (MLP) is arguably the simplest feedforward network with at least one hidden layer. Its architecture consists of an input layer, hidden layers and the

[4]Throughout this text, the word error is used as a synonym for residual for consistency.

output layer with nodes in adjacent layers fully connected, that is, each node j in one layer is connected with a certain weight w_{ij} to every node i in the adjacent layer. The number of hidden neurons in each hidden layer does not need to be the same for each layer and is a parameter to be tuned together with the number of hidden layers.

Each IWP input image to the network consists of arrays of size $128 \times 128 \times \mathrm{C}$, where C is the depth (channels) dimension. Therefore, the ground truth used is of size $128 \times 128 \times 1$. That is, the pixel from the input image at position (i, j), which has C channels, corresponds to pixel (i, j) from the ground truth. In the MLP architecture, each pixel of the image $(1 \times 1 \times \mathrm{C})$ is propagated individually through the network to predict a vector of quantiles, as shown in Figure 3.2. The MLP architectures in this work had the same number of hidden neurons for each hidden layer. No batch normalisation was applied in the MLP models, as it did not benefit premature results.

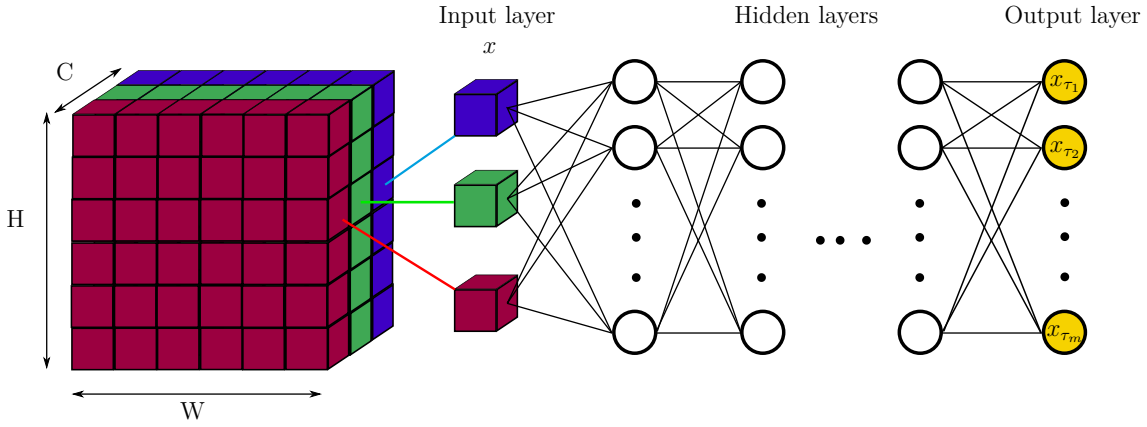


Figure 3.2: Schematics of a multilayer perceptron, where the input to the network is a single element x in the image of dimension $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$. The connections are in the forward direction, which predict the vector of quantiles x_{τ_i} at level τ_i, $i = 1, \ldots, m$, for x.

3.1.4.2 Convolutional neural network

Any spatial feature in the spatial dimensions is not considered in the MLP architecture presented. Convolutional neural networks (CNN) are architectures that can take advantage of any spatial information by having at least one convolutional layer which performs translation-invariant convolutions with a set of filters that each produce a feature map. These filters contain a set of learnable parameters which aim to learn spatial features from the inputs, as opposed to pooling layers which act exactly like the filters but compute a prescribed statistic.

A convolutional filter can have different sizes, that is, the receptive field, as shown in Figure 3.3. Three examples are filters that span in spatial dimensions (height and width) and are shared across the depth dimension (channels), filters that span in spatial dimensions which are not shared across the depth dimension, called depthwise convolutions, or filters that span only throughout the depth dimension, which is known as a pointwise convolutions or 1×1 filters. Pointwise convolutions are particularly useful for depth dimensionality reduction. Note that a layer containing only pointwise convolutions is mathematically equivalent to a hidden layer in a MLP, where each filter corresponds to a hidden neuron. Convolutional filters also incorporate another number of hyperparameters, for example, a stride indicating how the filter should move on the

inputs or the padding which defines how the boundaries should be treated. If padding is not used and the filter has a size greater than one in any spatial dimension, the resulting feature map will have different spatial dimensions than the previous layer.

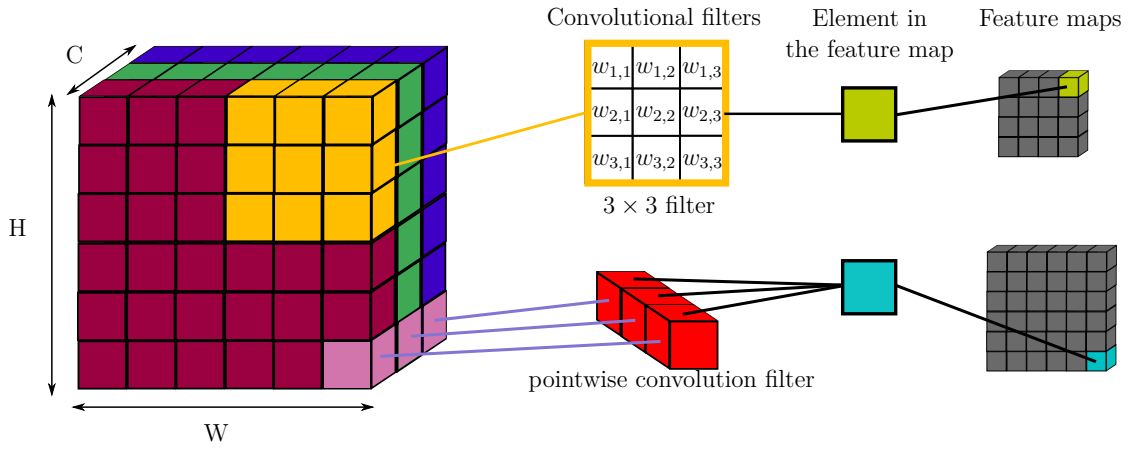


Figure 3.3: Two convolution filters acting on an input image. The yellow and pink colours indicate the receptive field of each filter. The 3×3 filter with weights $w_{i,j}$ embeds the information from a pixel and its direct eight neighbours from the same channel into a single element in the feature map while the pointwise convolution filter does it across the channels of a given pixel. The corresponding feature maps result of different size.

All these parameters make the design of a CNN much more complicated than a MLP design where the number of hidden layers and neurons are the main parameters to be chosen. A common practice for choosing a CNN architecture is to select architectures designed for similar problems and slightly adapt them, rather than designing an architecture from scratch. CNN architectures have been mainly developed for classification, which makes the choice of an architecture for regression problems more challenging.

3.1.4.2.1 XceptionFPN

It can be difficult to decide *a priori* the filter size when designing a convolutional layer. One approach is to not make one choice for the filter size but instead use several filter sizes for each layer. This delegates the task of determining which type of filters provide more meaningful information, technically referred as strong semantics, to the network training. However, there is one problem with this approach which is computational efficiency. When applying filters to a tensor with a large amount of channels the computations can be computationally prohibitive for a large network. If correlation across channels can be considered sufficiently decoupled from spatial correlations, pointwise convolutions can be used to reduce the channels dimensionality before applying different filters. These two concepts can be recognised in the Inception modules, introduced with the Inception network (Szegedy *et al.*, 2015), which essentially consists in many connected Inception modules, and which have been refined in subsequent papers.

The Xception network (Chollet, 2017), which stands for Extreme Inception, builds on the last concept by strongly assuming that correlation across channels can be entirely decoupled from spatial correlations. The paper reports significant performance gains with respect to Inception V3 (Szegedy *et al.*, 2016) for some given datasets, where it is claimed that the parameters are more efficiently used as both networks have the same number of parameters. The Xception block is formed by a linear stack of depthwise

separable convolution layers, which are depthwise convolutions followed by pointwise convolutions, with residual connections.

Recognising patterns at different scales can be achieved with feature pyramids that are connected to image pyramids. The scale of the pattern changes in each level of these pyramids applying scale invariant transformations. This makes possible detecting patterns across a number of pyramid levels. Another possibility for pattern recognition is extracting features from the pattern which are then scaled at different levels, avoiding the image pyramid.[5] The Feature Pyramid Network (FPN) (Lin *et al.*, 2017) leverages this approach with strong semantics at each pyramid level, by introducing a bottom-up feature pathway, a top-down feature pathway, and residual connections for each pyramid level. Each level may consist of several layers of the same size, which are said to be in the same stage. The FPN defines one pyramid level for each stage.

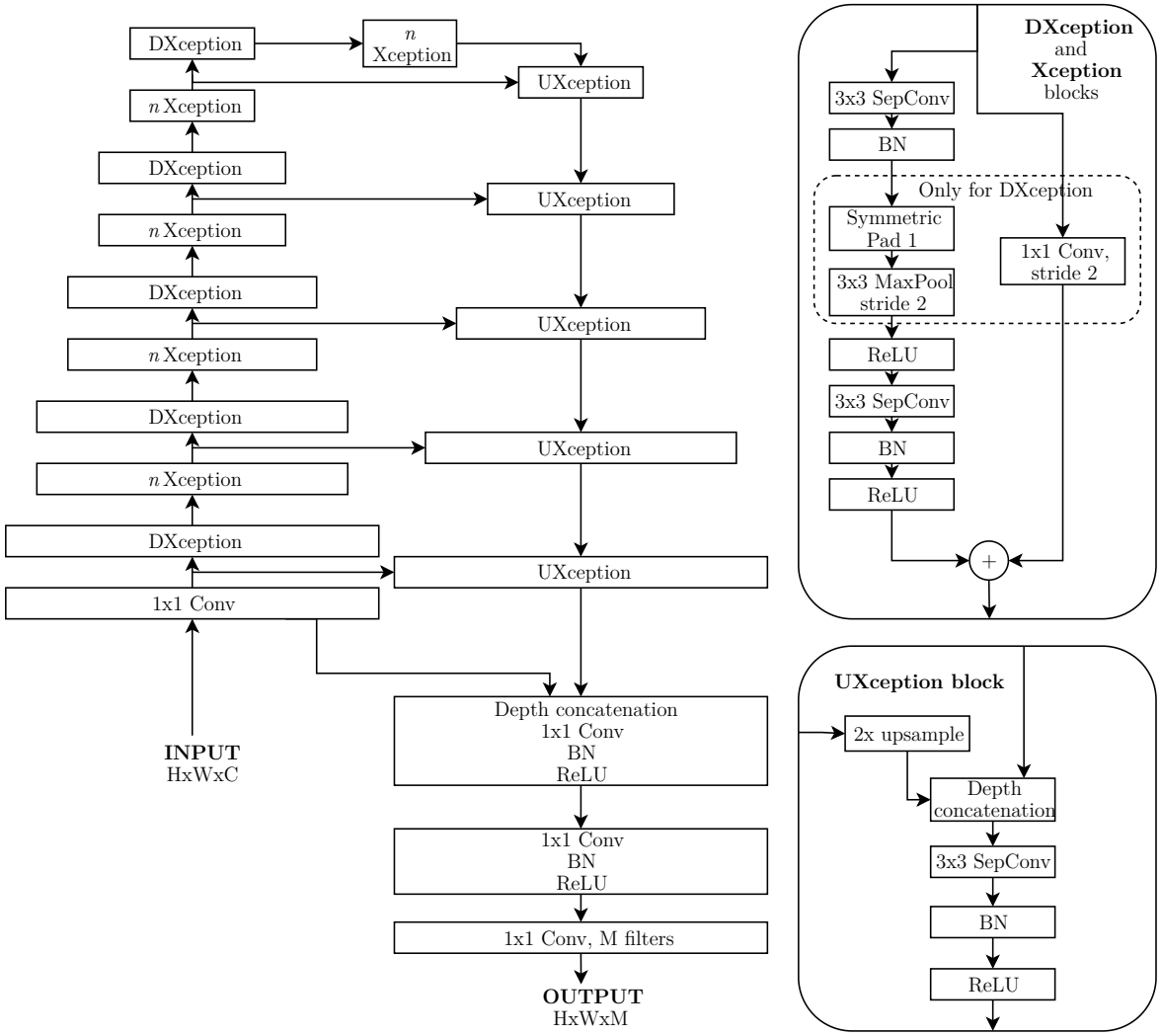


Figure 3.4: The XceptionFPN architecture. Block widths relate to the spatial sizes (not to scale). The feature pyramid halves the spatial sizes at each level. All convolutional layers use the same amount of filters except for the output layer which uses M filters. n Xception indicates n blocks connected at the same stage. The depthwise separable convolutions (SepConv) preserve the spatial size using a replicated padding of 1. Strides of 1, otherwise indicated.

[5]For an illustrative description of these approaches refer to Figure 1 in (Lin *et al.*, 2017).

The XceptionFPN architecture from the quantnn library (Pfreundschuh, 2021) combines the Xception and FPN concepts producing an architecture prepared for a regression problem. It defines a FPN with an asymmetric bottom-up and top-down feature pathway by using multiple Xception blocks at each stage of the bottom-up pathway. The specifications of the network are given in Figure 3.4. This was the CNN model chosen in this work.

3.2 Other performance metrics

The value of the pinball loss function can be difficult to interpret. There are other widely used performance metrics that can provide easy interpretations of the retrievals, although they do not necessarily work with quantiles. Given a reference value x and a predicted value $\hat{x}$, the root mean square error (RMSE), mean absolute error (MAE), and mean error (ME), defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}}, \quad \text{MAE} = \frac{\sum_{i=1}^{n}|\hat{x}_i - x_i|}{n}, \quad \text{ME} = \frac{\sum_{i=1}^{n}\hat{x}_i - x_i}{n} \tag{3.14}$$

are three frequent measures of performance. MAE can be easily interpreted, and RMSE and ME are reminiscent of the standard deviation and estimator bias. This makes interpretation simple as they also are computed in the same units as the estimated quantity. They require a single value as an estimate. The expected value of the predicted distribution, that is, its mean value, was used for this purpose. The mean value was also used for any other tool that does not work with distributions and requires a single value.

Evaluating predicted distributions is nontrivial. Ideally, a prediction should be:

- Of high resolution;
- Sharp, that is, the probability concentrated around a value for small uncertainty;
- Reliable, that is, the distribution should be well calibrated.

To have high resolution with QRNN it is sufficient to have a large amount of quantiles, ideally equally distributed. However, the more quantiles, the higher is the risk for quantile crossing, which consists in not respecting the level order of the predicted quantiles, for example, the quantile at level 0.6 being smaller than that at level 0.4. Improvements could be made to avoid this problem (Cannon, 2018), but it was not seen to be a problem for the final models chosen in this work for the quantile levels chosen.

A generalisation of the MAE for distribution prediction is the continuous ranked probability score (CRPS), defined as

$$\text{CRPS} = \int_{-\infty}^{+\infty}\left[\hat{F}(x) - F(x)\right]^2 \mathrm{d}x \tag{3.15}$$

for a predicted and true cumulative distribution function, $\hat{F}(x)$ and $F(x)$, respectively. As the ground truth IWP values are points, $F(x)$ results the indicator function $\mathbb{1}_y(x) = \{1 \text{ if } x \geq y; 0 \text{ otherwise}\}$, where y is the observed IWP. The CRPS can assess the overall prediction performance. Nonetheless, it is nontrivial how to interpret the CRPS value, but the lower the better. In the ideal case of small uncertainty, the predicted

distribution would be extremely sharp, concentrating the probability around a point. Sharpness around the central tendency of predicted distributions can be assessed with the interquartile range (IQR), the difference between the quantiles at level 0.25 and 0.75, also called first and third quartiles, respectively, and general sharpness with the difference between the quantiles at levels 0.05 and 0.95. Therefore, a distribution results sharper the lower these differences are.

For a quantile x_τ at level τ from a predicted distribution, it is expected that the ratio of samples from this distribution less than x_τ will be τ. In addition, given an interval that covers, for example, 95% of the distribution, it is expected that samples will be in this interval 95% of the times. A model that is well calibrated has these properties. Reliability diagrams, illustrated in Figure 3.5, summarise the calibration of a model for a given set of distributions by estimating the observed frequency of each predicted quantile level with respect to the nominal level. They only provide the information on a global level, which is a drawback. That is, they may show perfect calibration at the global level but may mask a bad calibration for a subset of reference values. To evaluate both situations, reliability diagrams with and without including the zero IWP values were computed, as zero values are roughly 45% of the IWP values. Poorly calibrated models can be calibrated with a post-processing step at the global level (Kuleshov, Fenner & Ermon, 2018) or at the distribution level (Song *et al.*, 2019). However, this introduces another modelling problem which was regarded as unnecessary for this work.

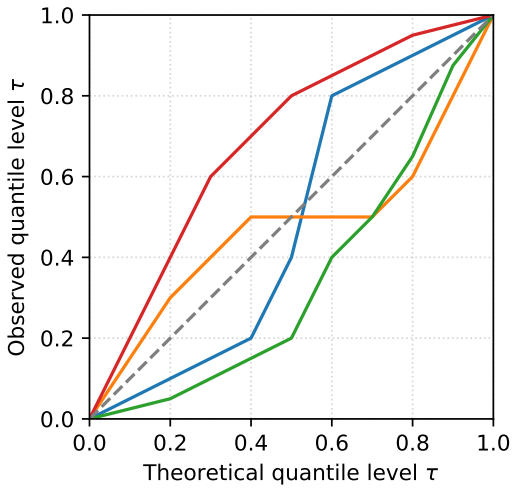


Figure 3.5: Examples of reliability diagrams. The ideal curve follows the diagonal dashed line. The green line indicates underestimation, the red line overestimation. The blue (orange) line underestimates (overestimates) small quantiles and overestimates (underestimates) large quantiles, therefore it indicates overconfidence (underconfidence).

All these tools, including the loss function used in the regression, have one disadvantage which is that they condense much information from values between different orders of magnitude in a single value or a curve. The RMSE, MAE, ME and the differences between quantiles were also computed for ranges of true IWP values for a better interpretation. However, this still condenses much information in a single value. One way to avoid such condensation of information which can hide valuable information is by visualising every prediction with scatter plots, where the mean of the predicted IWP distribution is plotted against the ground truth. Because of the large number of points in the IWP

dataset, it was required to visualise binned scatter plots, which consists of assigning each prediction to the closest bin from a precomputed set of bins. These plots result very informative, though evaluations on scatter plots are prone to subjectivity. To mitigate this problem when comparing the plots, the median and the first and third quartiles were used as a measure of central tendency and dispersion.

3.3 Methodology

The best performing MLP and CNN models were determined in an iterative approach. This approach sought to find answers to the following:

- Appropriate hyperparameters for each model: hidden layers and hidden neurons for the MLP and number of filters and number of Xception blocks at each stage (n in Figure 3.4);
- Data transformations: in particular whether to predict in a transformed space and data augmentation;
- Efficient computation according to the training infrastructure and data;
- Quick and good convergence with the optimiser and scheduler;
- Evaluation of the performance with other tools besides the loss function.

The quantile levels to predict were fixed to be all the percentile levels, that is the quantile levels $1\%, 2\%, \ldots, 98\%$, and 99%. To reduce the spectrum of possibilities, the approach used a survey of the possibilities and the most promising ones based on a compromise between observed preliminary results and practicalities were chosen for performing the more thorough analysis. In addition, all decisions except the hyperparameters of the architectures were taken based on the results of using all SEVIRI channels.

The appropriate values for the hyperparameters in the MLP and CNN architectures were searched over a fixed set of values, presented in Table 3.1, as opposed to a random search. This was done to simplify comparisons between different configurations. The models were trained at least once for each combination, and those presenting the lowest loss value for the test set were determined suitable.

The IWP values range across several orders of magnitude, and their distribution is approximately exponential, with zero values being the most common. While a QRNN does not use the value of the prediction error when updating the learnable parameters, as remarked in Section 3.1.3, it is intuitive that a logarithmic transform can facilitate the learning to the network. To cope with the problem that the zero is undefined for the logarithm, zero IWP values were replaced with a small random value. The smallest non-zero value is larger than $10^{-5}\,\mathrm{kg\,m^{-2}}$, and the data transform for an IWP value x applied was

$$\tilde{x} = \begin{cases} \log x & \text{if } x > 10^{-6} \\ \log \epsilon & \text{otherwise} \end{cases} \qquad \text{with } \epsilon \sim \text{Uniform}(10^{-4}, 10^{-6}) \tag{3.16}$$

where ϵ is sampled every time x is accessed and all units are in $\mathrm{kg\,m^{-2}}$. Since the loss functions computed with the untransformed and transformed data are not directly

comparable unless all predictions in one space are transformed to the other, which was not practical, and because the network may learn the data differently, determining if the data transform was beneficial was done with scatter plots on the test set.

Arguably, the performance of a model will increase with the amount of data used during training and it will generalise better the more diverse the training data is. Data augmentation modifies the training dataset with random but reasonable transformations. In this work, rotations of 0°, 90°, 180° or 270°, as well as vertical and horizontal flips, all applied randomly to each image read, were used as a data augmentation method. Thus, it was only significant for the CNN models. The data augmentation performance was not only evaluated on the test set loss value but also on the scatter plots.

Because the DARDAR swath results less than 3 pixels wide on the collocated 128×128 pixels data samples, most of the resampled IWP pixels are NaN, which poses a computational problem. For efficiency, the MLP was trained only on pixels with ground truth. The CNN could not avoid the NaNs, so they were masked. The batch size, that is, the number of data instances that can be loaded at once in the memory for each mini-batch, was adjusted for the CNN in order to maximise the GPU usage for two parallel trainings and set to 32. To reduce any bias effect by different batch sizes, the batch size for the MLP was similar to the average number of valid ground truth pixels for each CNN mini-batch and set to 4096. The batch sizes were the same for any training configuration.

The optimisers chosen were SGD with momentum and Adam, and the schedulers consisted in reducing the learning rate when the loss stopped improving on the validation set (reduce on plateau) and cosine annealing. Their parameters are also presented in Table 3.1. The networks were trained for 100 epochs.

The training, validation, and test sets were fixed and shared for any training and evaluation. All input data to the models, that is, the SEVIRI data, was normalised with standardisation. The mean and standard deviation were computed from the training set. No other information, such as a time reference or the geographical coordinates of each pixel, was provided to the network.

Finally, those models with the corresponding training configurations that showed not only a small loss function but also good performance on other metrics were retrained with different initialisations to potentially find better results, as reasoned in Section 3.1.

Table 3.1: Sets of architecture parameters considered, and optimisers and schedulers used.

	MLP	**Optimiser**	**Learning rate**
Hidden layers	1, 2, 4, 6, 8, 16, 32	Adam	10^{-3}
Hidden neurons	8, 16, 32, 64, 128, 256	SGD (0.9 momentum)	10^{-2}, 5×10^{-3}, 2.5×10^{-3}, 10^{-3}
	CNN	**Scheduler**	**Parameters**
Xception blocks	0, 2, 4, 6	Reduce on plateau	0.1 reduce factor, 5 epochs patience
Filters	64, 128	Cosine annealing	10, 20 epochs period

3.3.1 Training infrastructure

All networks were executed with the quantnn library (Pfreundschuh, 2021) with Pytorch backend (Paszke *et al.*, 2019). They were trained on NVIDIA Tesla V100 32 GB SMX2 GPUs. Each MLP experiment could take up to 3 hours and each CNN experiment up to 12 hours, with large fluctuations depending on the parameters chosen and if any training was being executed in parallel.

4

Results

The MLP and CNN models with more satisfactory retrieval performance are reported in Section 4.1, which also compares the influence of spatial information. Section 4.2 remarks several relevant findings found during the training process. Regarding the SEVIRI channels included in the models, two approaches are presented: including all channels (VISIR) and including all infrared channels without a solar contribution (IR), which excludes the 3.9 µm channel. Section 4.3 comments on the inclusion of the 3.9 µm channel with the IR channels. The comparison between VISIR and IR channels facilitates interpreting how the networks react to each setting for daytime and nighttime retrievals. All results are reported for the test set, though the remarks are also valid for the training set (Appendix A.1). Finally, Section 4.4 compares the monthly means and mean diurnal variations obtained with the QRNNs models with those from the CLAAS dataset.

4.1 Best performing models

One model per set of channels considered, referred to as channels setting, and architecture was selected as the best performing model. The parameters for these models are presented in Table 4.1. Both data augmentation and the log-transform of the data were beneficial. The loss value for the whole test set is lower when the VISIR channels are considered, as seen in Figure 4.1, with good agreement between different initialisations. The models can be considered to be well calibrated at the global level as shown in Figure 4.2a, though not necessarily for any predicted distribution, as seen in Figure 4.2b. This implies that some distributions can underestimate or overestimate the IWP values.

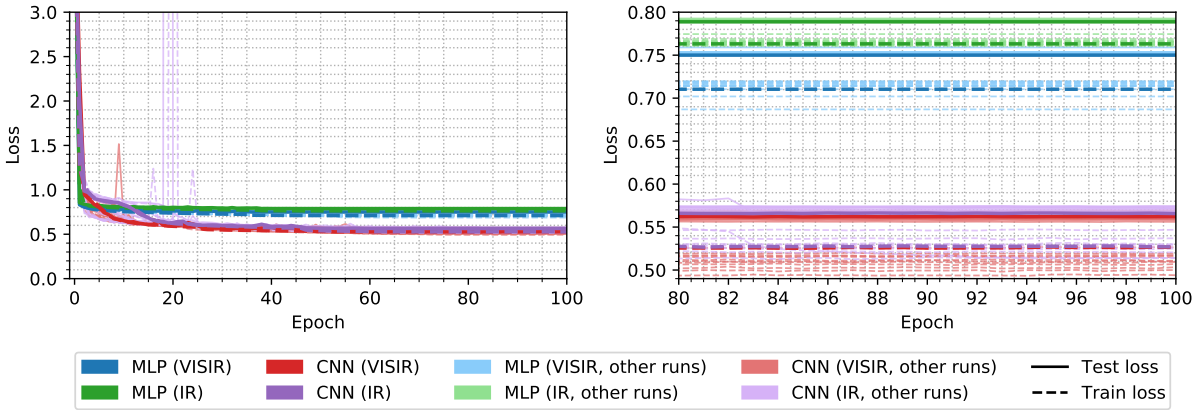


Figure 4.1: The loss value for the test set as a function of training epoch, where other runs indicate the results of running with different initialisations.

Table 4.1: Parameters and configurations for the best performing networks.

	MLP		CNN	
	VISIR	IR	VISIR	IR
Hidden layers/ Xception blocks	4	4	2	4
Hidden neurons/ Filters	128	128	128	128
Transformed space	Yes			
Optimiser	Adam			
Scheduler	Reduce on plateau			
Data augmentation	Not applicable		Yes	

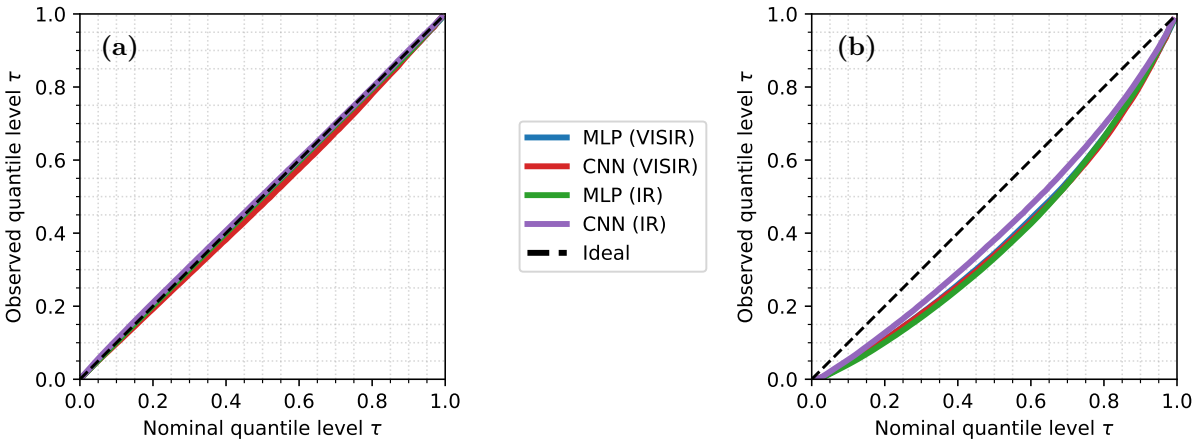


Figure 4.2: Reliability diagrams for the test set where in (a) all IWP values are considered to compute the observed level and in (b) zero DARDAR IWP values were discarded for the calculation. The MLP (VISIR) line is hardly visible as it is completely overlapped by the other curves. CNN (IR) does not hide another curve in (b).

Figure 4.3 presents the statistics derived from the binned scatter plots explained in Section 3.2. The MLP models have more dispersion when compared with the CNN models. For IWP between $10^{-3}\,\mathrm{kg\,m^{-2}}$ to $10^{1}\,\mathrm{kg\,m^{-2}}$, both models show a central tendency close to the ideal prediction, particularly at daytime. Comparing VISIR with IR, there is less dispersion for nighttime when only the IR channels are considered, while it is the opposite for daytime. The analogous is also true for the central tendency: it is closer to the ideal prediction for the IR channels for nighttime and the opposite for daytime.

To determine overestimation and underestimation at a global level, the probability density functions (PDFs) for each architecture and channels setting were computed and are presented in Figure 4.4. Firstly, it is seen that both architectures and channels settings have slightly higher PDF at small values, but smaller PDF from roughly $4.5\,\mathrm{kg\,m^{-2}}$. Secondly, only the PDF for the VISIR setting is notably different between daytime and nighttime retrievals. Thirdly, in the IR setting the PDF for the CNN is closer to the DARDAR PDF for both daytime and nighttime, while in the VISIR setting the CNN is closer to the DARDAR PDF for nighttime but they result similar for daytime.

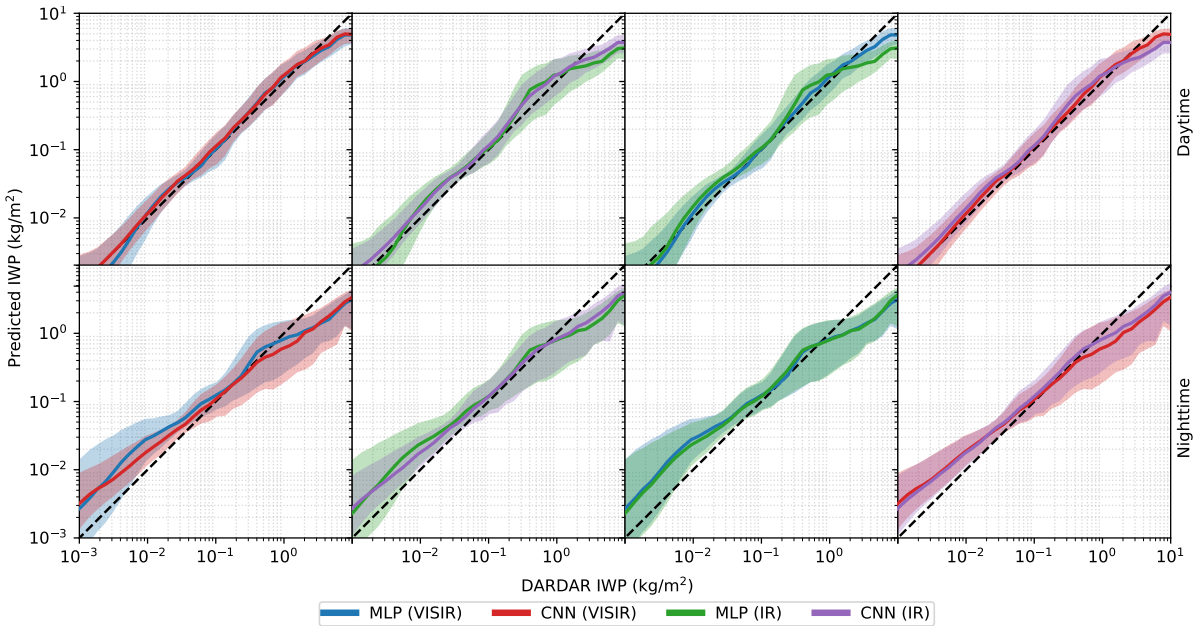


Figure 4.3: Median (solid line) bounded by the first and third quartiles (filled area) indicating the central tendency and dispersion of the data points in the test set for the binned scatter plots explained in Section 3.2, where each data point is the predicted distribution mean. The same channels settings in the two different architectures are compared, and vice versa, for daytime (top row) and nighttime (bottom row) retrievals.

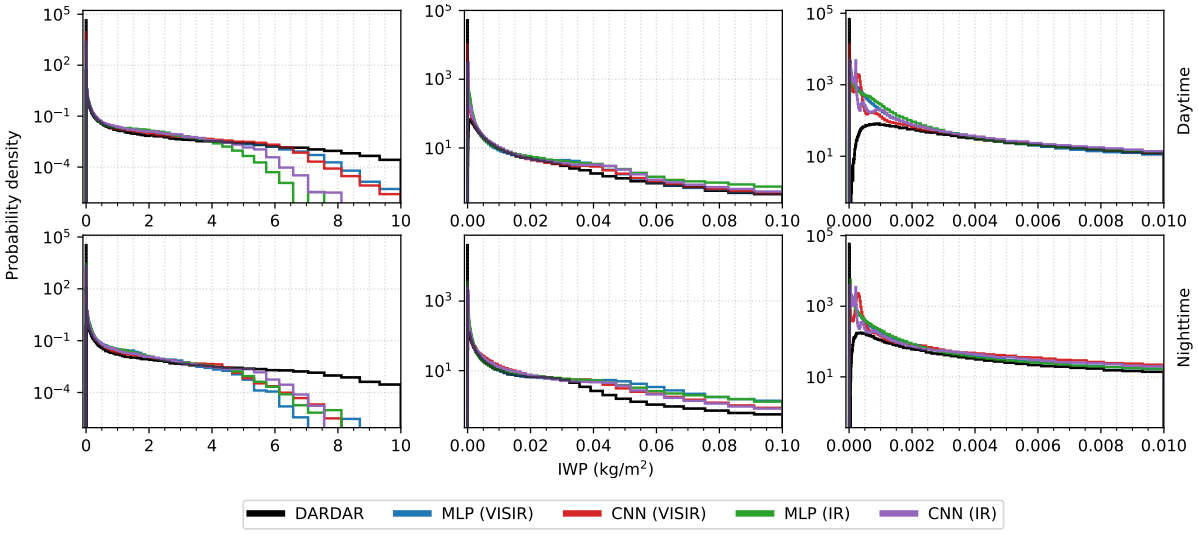


Figure 4.4: Probability density functions for the test set for daytime and nighttime retrievals.

To better understand how the MLP and the CNN models retrieve the IWP, video animations were produced for a given area showing the IWP retrieval as a time series. Figure 4.5a aims to help summarise the observed patterns in these animations. The main observation is that the MLP can have large differences between neighbouring pixels for a given retrieval. This finding is also seen over the time series, where the same pixel can have a rough evolution of the retrieved IWP for the MLP architecture. The CNN instead presents retrievals that are smoother both for a given time and through time. Unfortunately, this analysis of the retrieval with animations lacked ground truth data to compare against due to the long revisit period of CloudSat.

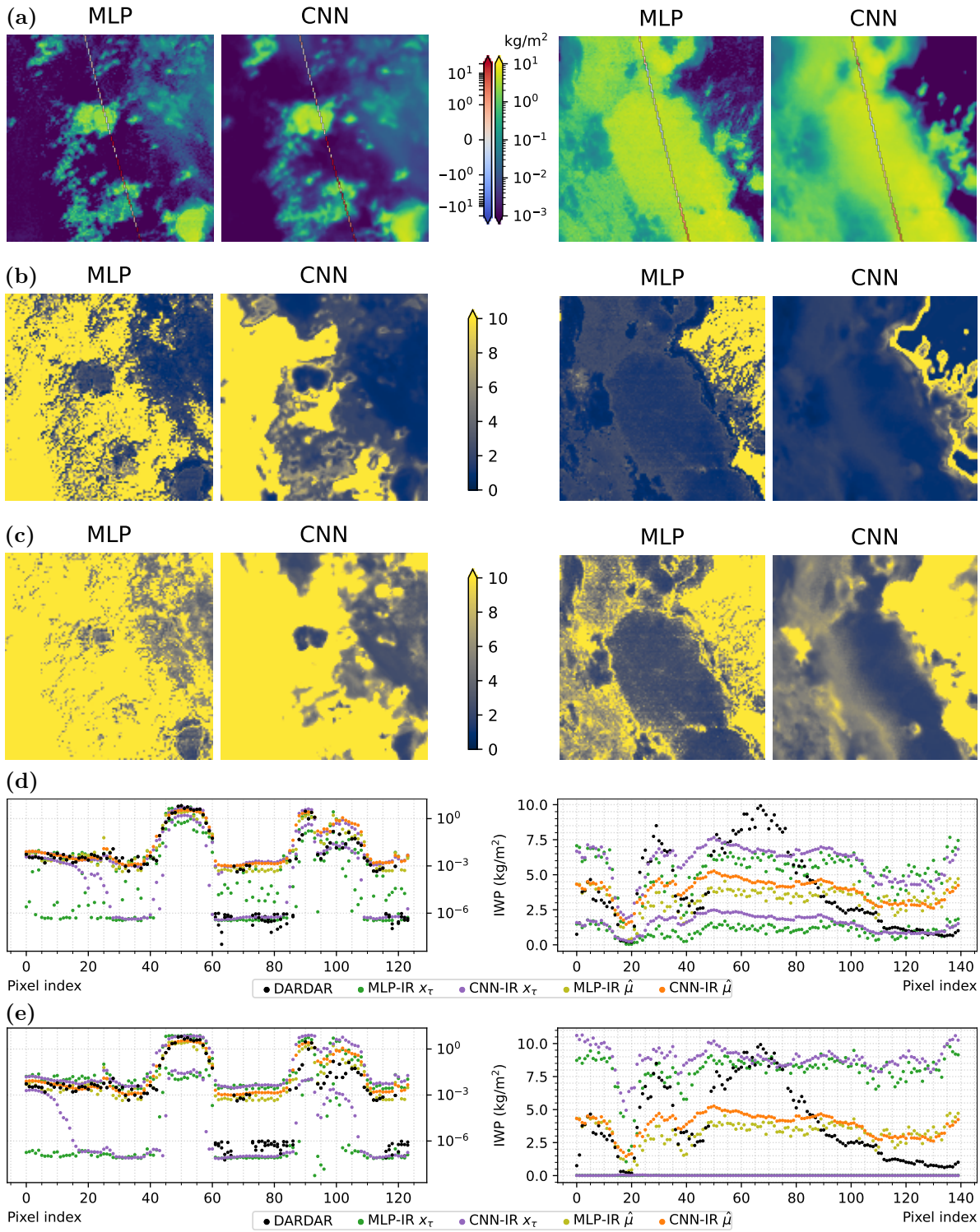


Figure 4.5: In (a), two IWP daytime retrievals using the IR channels setting with the relative error $(\hat{\mu}-y)/y$ overlaid, where $\hat{\mu}$ is the predicted distribution mean and y is the DARDAR IWP. Relative sharpnesses in (b) and (c), computed as $(x_{0.75}-x_{0.25})/x_{0.50}$ and $(x_{0.95}-x_{0.05})/x_{0.50}$, where x_τ indicates the quantiles at level τ. The quantiles at levels 0.25 and 0.75 and means along the DARDAR swath in (d), where the index increases top-to-bottom of the image. In (e) same as (d) but for levels 0.05 and 0.95.

The smooth output of the CNN models is also seen in the relative sharpnesses shown in Figures 4.5b and 4.5c: the CNNs have relative sharpnesses with less variation for a given area than MLPs, to some extent. It is highlighted that the CNN can clearly have a change in the sharpnesses at the boundaries of the clouds, where there are large IWP gradients predicted. These observations are also valid for the same results using the VISIR channels setting, presented in Figure A.4. Because of the large DARDAR IWP values on the left panels of Figures 4.5 and A.4, it is difficult to say which model provides a closer retrieval to DARDAR. On the other hand, on the right panels it can be claimed that the CNN has a retrieval value closer to the DARDAR IWP according to Figures 4.5d and 4.5e. However, Figures A.4d and A.4e can give the impression that it is the MLP that does better.

This example presented only the outputs for two input images, but this is not necessarily representative of the general situation. The analysis of Table 4.2 offers a better general picture of the retrievals. As a general trend and for the same architecture, the VISIR setting provides smaller RMSE, MAE and unsigned ME values than the IR setting. Also as a general trend, the CNN model provides smaller MAE and unsigned ME values than the MLP model in the same channels setting, though the opposite for the RMSE. The negative ME for non-zero ground truths agrees well with the underestimation seen in Figure 4.2b. The computed metrics values decrease as the ground truth values decrease, though the metrics show a relatively worse performance as the IWP magnitude decreases.

Regarding the sharpness of the predicted distributions and as a general trend, it is seen that the VISIR setting provides sharper predictions for both the central part, assessed with the IQR, as well as overall, assessed with the difference between quantiles at level 0.05 and 0.95. The CNN provides sharper predictions than the MLP in the same channels settings. Concerning the CRPS, the mean value results higher for the VISIR setting in the same architecture, but the median smaller, which indicates occasionally higher CRPS as the mean is sensible to outliers. In the same channels setting but different architectures, the CRPS mean results smaller for the MLP.

Accidentally, one of the samples of the test set was found to have the SEVIRI data mainly corrupted. It was not considered worth finding the cause of the corruption due to this rare occurrence, but it is seen that the CNN can recover more information than the MLP, as shown in Figure 4.6. Data points found to be corrupted with NaNs in at least one channel, such as data points in Figure 4.6, were excluded when computing the values in Tables 4.2 and A.1.

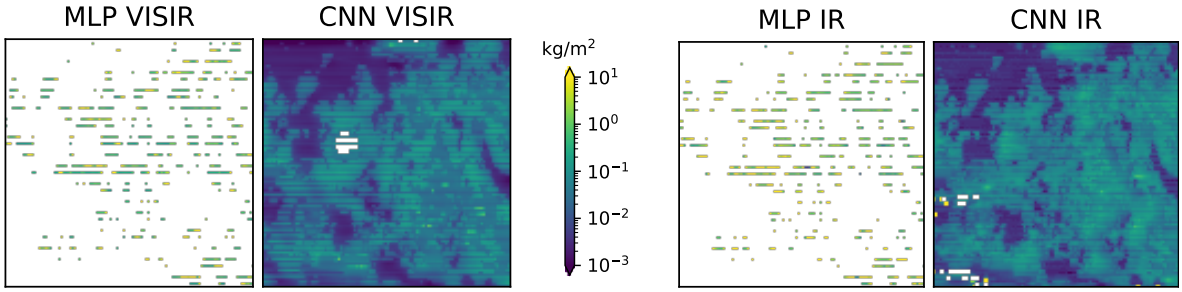


Figure 4.6: Retrieved IWP from a corrupted input file to the model.

Table 4.2: Other performance metrics computed for the test set. All values in $\mathrm{kg\,m^{-2}}$. The intervals indicate that only data points with ground truth values in the given interval are considered to compute the metric, and x_τ the quantiles at level τ. Values closest to zero are highlighted: architecture-wise in light grey, and row-wise in dark grey (considering all decimals).

Metric	Ground truth interval	MLP		CNN	
		VISIR	IR	VISIR	IR
RMSE	All values	5.02×10^{-1}	5.41×10^{-1}	5.05×10^{-1}	5.13×10^{-1}
	> 0	6.78×10^{-1}	7.30×10^{-1}	6.81×10^{-1}	6.93×10^{-1}
	1 to 10^{1}	2.30	2.49	2.30	2.32
	10^{-1} to 1	6.17×10^{-1}	7.37×10^{-1}	6.41×10^{-1}	6.69×10^{-1}
	10^{-2} to 10^{-1}	2.75×10^{-1}	2.86×10^{-1}	2.82×10^{-1}	2.97×10^{-1}
	10^{-3} to 10^{-2}	1.74×10^{-1}	1.87×10^{-1}	1.82×10^{-1}	1.90×10^{-1}
	10^{-5} to 10^{-3}	1.35×10^{-1}	1.36×10^{-1}	1.35×10^{-1}	1.44×10^{-1}
	$< 10^{-5}$	6.48×10^{-2}	6.73×10^{-2}	7.23×10^{-2}	7.16×10^{-2}
MAE	All values	9.61×10^{-2}	1.09×10^{-1}	9.14×10^{-2}	9.70×10^{-2}
	> 0	1.72×10^{-1}	1.95×10^{-1}	1.64×10^{-1}	1.74×10^{-1}
	1 to 10^{1}	1.70	1.90	1.72	1.73
	10^{-1} to 1	3.45×10^{-1}	4.36×10^{-1}	3.33×10^{-1}	3.76×10^{-1}
	10^{-2} to 10^{-1}	7.40×10^{-2}	7.89×10^{-2}	6.24×10^{-2}	7.10×10^{-2}
	10^{-3} to 10^{-2}	3.21×10^{-2}	3.39×10^{-2}	2.46×10^{-2}	2.79×10^{-2}
	10^{-5} to 10^{-3}	2.04×10^{-2}	2.12×10^{-2}	1.51×10^{-2}	1.77×10^{-2}
	$< 10^{-5}$	5.32×10^{-3}	6.19×10^{-3}	4.18×10^{-3}	4.67×10^{-3}
ME	All values	-1.48×10^{-3}	-5.00×10^{-3}	-5.08×10^{-3}	-3.48×10^{-3}
	> 0	-7.16×10^{-3}	-1.44×10^{-2}	-1.28×10^{-2}	-1.03×10^{-2}
	1 to 10^{1}	-1.17	-1.52	-1.08	-1.24
	10^{-1} to 1	1.99×10^{-1}	2.81×10^{-1}	1.72×10^{-1}	2.33×10^{-1}
	10^{-2} to 10^{-1}	6.59×10^{-2}	7.11×10^{-2}	5.32×10^{-2}	6.28×10^{-2}
	10^{-3} to 10^{-2}	3.04×10^{-2}	3.22×10^{-2}	2.35×10^{-2}	2.70×10^{-2}
	10^{-5} to 10^{-3}	2.02×10^{-2}	2.10×10^{-2}	1.50×10^{-2}	1.75×10^{-2}
	$< 10^{-5}$	5.32×10^{-3}	6.19×10^{-3}	4.18×10^{-3}	4.67×10^{-3}
$\boldsymbol{x_{0.75} - x_{0.25}}$ **(median value)**	All values	2.07×10^{-3}	2.17×10^{-3}	1.35×10^{-3}	1.57×10^{-3}
	> 0	1.18×10^{-2}	1.25×10^{-2}	8.92×10^{-3}	9.86×10^{-3}
	1 to 10^{1}	1.84	1.83	1.93	2.06
	10^{-1} to 1	1.54×10^{-1}	1.53×10^{-1}	1.52×10^{-1}	1.72×10^{-1}
	10^{-2} to 10^{-1}	1.74×10^{-2}	1.77×10^{-2}	1.55×10^{-2}	1.56×10^{-2}
	10^{-3} to 10^{-2}	4.43×10^{-3}	4.51×10^{-3}	3.15×10^{-3}	3.30×10^{-3}
	10^{-5} to 10^{-3}	1.11×10^{-3}	1.23×10^{-3}	9.32×10^{-4}	1.19×10^{-3}
	$< 10^{-5}$	1.15×10^{-6}	2.43×10^{-6}	5.74×10^{-7}	5.51×10^{-7}
$\boldsymbol{x_{0.95} - x_{0.05}}$ **(median value)**	All values	9.11×10^{-3}	1.02×10^{-2}	5.90×10^{-3}	6.14×10^{-3}
	> 0	5.98×10^{-2}	6.51×10^{-2}	4.87×10^{-2}	5.05×10^{-2}
	1 to 10^{1}	5.06	5.32	5.09	5.21
	10^{-1} to 1	6.64×10^{-1}	9.75×10^{-1}	6.15×10^{-1}	8.87×10^{-1}
	10^{-2} to 10^{-1}	9.05×10^{-2}	9.60×10^{-2}	8.23×10^{-2}	7.95×10^{-2}
	10^{-3} to 10^{-2}	1.60×10^{-2}	1.76×10^{-2}	1.33×10^{-2}	1.38×10^{-2}
	10^{-5} to 10^{-3}	5.89×10^{-3}	6.22×10^{-3}	4.44×10^{-3}	4.31×10^{-3}
	$< 10^{-5}$	2.05×10^{-3}	2.69×10^{-3}	1.32×10^{-3}	1.20×10^{-3}
CRPS mean	All values	5.13×10^{-2}	4.61×10^{-2}	5.40×10^{-2}	4.88×10^{-2}
median		4.23×10^{-4}	4.72×10^{-4}	4.04×10^{-4}	5.59×10^{-4}

4.2 Observed trends in trainings

Several observations were made in the process of determining the best performing models. The Adam optimiser occasionally produced bursts in the loss, although it quickly corrected that. This can be seen in the left panel of Figure 4.1 from training a model not determined to have the best performance. The SGD optimiser did not present these bursts, but it showed a much slower decrease in the loss function. However, it did not achieve the same loss level with the parameters chosen. Moreover, SGD experienced exploding gradients when the learning rate was large, as well as substantially larger convergence times when using small learning rates. For the set of hyperparameters chosen for the two learning rate schedulers used, both provided similar results.

Concerning the number of hidden layers and neurons in the MLP, it was seen that deeper networks had a higher loss value for the test set than shallower networks. What made a beneficial impact on the loss of the test set was increasing the width of the network. Nonetheless, the narrowest networks presented smaller loss of the test set when increasing their depth. The training loss was smaller for deeper and wider networks.

As expected, the test set benefited from using data augmentation in the training. Particularly, it achieved smaller loss values. On the other hand, the loss value for the training set resulted higher when using data augmentation. This is reasonable as at each training epoch the training data experiences a random transformation, as explained in Section 3.3. That is, at each epoch the network uses slightly different training data and thus struggles to perfectly fit it, but it learns to generalise better as it has to adapt to different data at each epoch.

Finally, the reliability diagrams were always close to the ideal curve when using different initialisations for the MLP. This is not observed to the same degree for the CNN, especially in the VISIR setting. In this case, models often presented curves deviating from the ideal curve, and several trainings were needed to find models that closely matched it. However, in the IR setting, this undesired outcome was not observed with the same frequency.

4.3 About the 3.9 µm channel

The results presented in the previous sections come from models that use either the VISIR setting, where all available channels in SEVIRI are considered, or the IR setting, where all channels without a solar contribution are considered. The same analyses were performed where the IR setting included the 3.9 µm channel.

The inclusion of this channel produced results that are in between the VISIR and IR settings, and are not reported to interpret better the comparisons between the VISIR and IR settings. The loss values including the 3.9 µm channel result in between the VISIR and IR values for each architecture, and the central tendency and dispersion in the scatter plots are in good agreement with the IR setting, as well as with the probability density functions.

The exclusion of this channel was decided after observing a possible artefact in the diurnal variations. This possible artefact is also seen for the VISIR setting, and is

presented in the next section. This can be correlated with the fact that the 3.9 µm channel has a solar reflectance contribution during daytime. Therefore, it may experience the same issues as the visible and near infrared channels that depend on the solar reflectance.

4.4 Comparison with CLAAS

The monthly mean IWP and monthly mean diurnal cycle were computed for the QRNN models in the same way as specified in the CLAAS Product User Manual (Finkensieper *et al.*, 2020b). The monthly mean was obtained by averaging daily means, where these were obtained from all available time slots. All time slots within one hour were averaged to a daily mean diurnal cycle. Afterwards, all diurnal cycles of a month were averaged to obtain the monthly mean diurnal cycle. In both cases, all retrievals for each area are averaged to provide a mean IWP for the area. The results are presented in Figures 4.7 to 4.9, where the monthly means for the CLAAS dataset correspond to the all-sky data variables from the dataset. Three out of 35 110 retrievals for the land area with the CNN VISIR model presented mean values above $1.2 \times 10^3\,\mathrm{kg\,m^{-2}}$ and were excluded, but no other anomalies were detected.

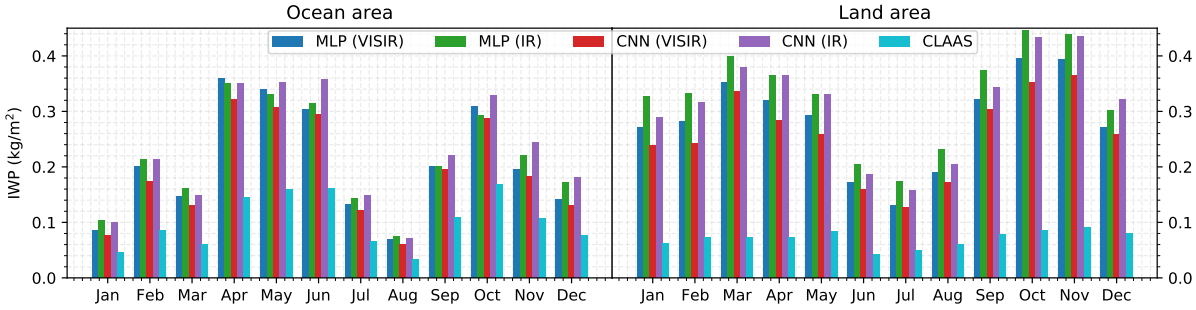


Figure 4.7: Monthly means for 2012 obtained with the QRNN models and CLAAS, computed according to CLAAS Product User Manual (Finkensieper *et al.*, 2020b).

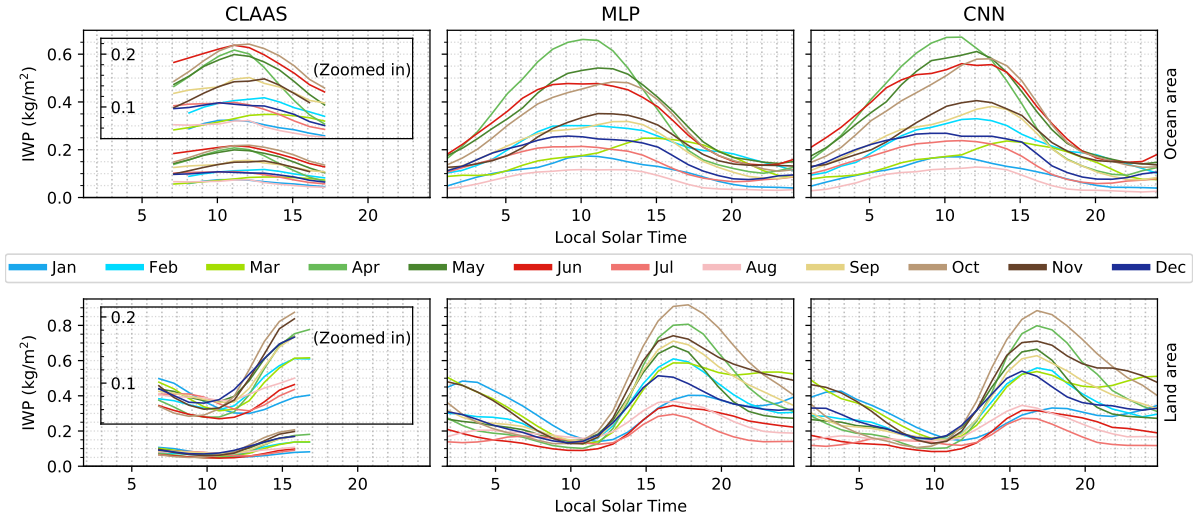


Figure 4.8: Mean IWP diurnal cycle for each month in 2012 considering only IR channels.

The principal observation is that QRNN estimates higher quantities of IWP for these two areas than CLAAS, and in particular for the land area which covers a tropical

rainforest. If the monthly IWP median is computed instead of the monthly IWP mean, then the monthly IWP for both areas decreases, but this is not comparable with the datasets considered from CLAAS. Besides differences in the mean value for each month, roughly the same pattern of mean IWP as a function of the months is observed for the QRNNs and CLAAS, with good agreement in months with higher and lower mean values. These observations are also seen in the mean diurnal variations in the IR channels setting in Figure 4.8, where there is generally good agreement between months but the magnitude is different.

Comparing the QRNN models, it is seen that the VISIR setting tends to estimate a lower IWP mean per area in the monthly means for the same architecture. Comparing the MLP and CNN with the same channels setting, in the land area the CNN tends to estimate a lower IWP mean in the monthly means than the MLP for both VISIR and IR settings. This is also observed for the VISIR setting in the ocean area, but in the IR setting this trend is only observed for the first quarter of the year with the contrary pattern for the remaining months.

When Figure 4.8 is repeated but for the VISIR channels setting, it is observed a peculiar pattern taking place in the early morning as well as later in the afternoon. It is suspected that it is an artefact correlated with phenomena related to sunrise and sunset. This pattern for the VISIR setting is observed in Figure 4.9, which presents the annually averaged monthly mean IWP diurnal cycles for each model considered, together with the time coverage of the observations in the training dataset.

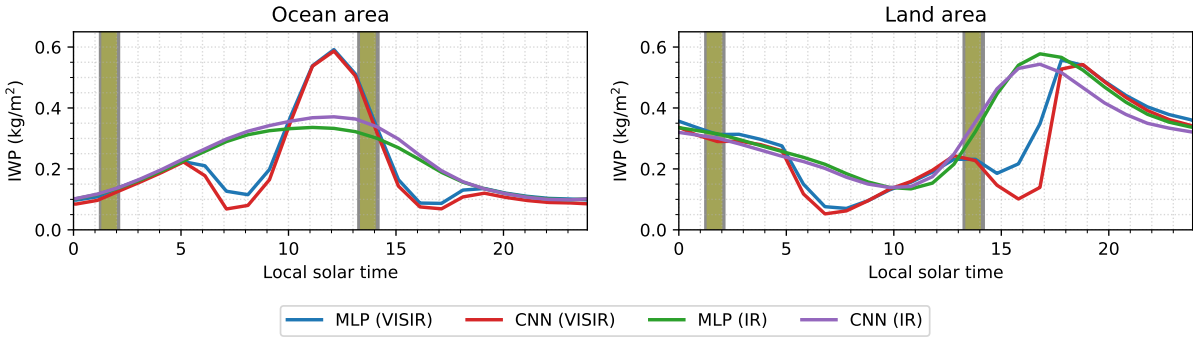


Figure 4.9: Mean IWP diurnal cycle for 2012, averaged from the monthly mean IWP diurnal cycle. The yellow areas indicate the local solar time coverage of the ground truth data points in the training set, and the grey areas behind them the local solar time coverage of all data points used in training, where not all of them have ground truth but the spatial information can be used in the CNN architecture.

5

Discussion

The results support the hypothesis that quantile regression neural networks calibrated against DARDAR can provide a satisfactory retrieval of IWP with a high spatial and temporal coverage from SEVIRI measurements. This machine learning approach not only provides a predicted distribution mean that tends to be close to the IWP from DARDAR as seen on the scatter plots, in particular for the daytime retrievals, but also an associated uncertainty, which can be calculated, for example, from an interval of the predicted distribution related to its mean.

Benas *et al.* (2017) provide a comparison of CLAAS against DARDAR. This referenced work provides a figure similar to Figure 4.3, where IWP predicted with CLAAS is plotted versus IWP from DARDAR, and where the median is also used to indicate the central tendency. It is seen there that the IWP from CLAAS has a larger deviation from the ideal prediction: CLAAS overestimates and underestimates to greater degrees smaller and larger IWP values, respectively, than IWP retrieved with QRNNs. Benas *et al.* argument that these results should be attributed to differences in the particle effective radius. Thus, an area with high IWP should be more underestimated in CLAAS than with the QRNN models. On the other hand, and for the land area, the QRNNs with IR channels settings clearly indicate that the IWP is higher for retrievals that CLAAS does not provide, that is, nighttime retrievals, and therefore CLAAS should also underestimate the mean IWP for this area. These two remarks correlate well with the differences in magnitude in the monthly means and mean diurnal variations obtained with the QRNN models.

As can be expected, the QRNN models struggle to retrieve the larger IWP values. This is reasonable as SEVIRI, which can only retrieve the IWP indirectly, will not be able to discern accurately the characteristics of thick clouds, which contain a large quantity of atmospheric ice, as it will receive roughly the same signal. In addition, the retrieval of the largest values can also be further penalised by collocation errors. These difficulties in large IWP retrievals can be observed in the probability density functions, where the IWP PDFs for the QRNNs diverge from the DARDAR PDF from roughly $4.5\,\mathrm{kg\,m^{-2}}$.

Regardless of the network architecture, daytime retrievals using all SEVIRI channels provide better retrievals than using only channels without a solar contribution. This supports the physical approaches in which only channels with solar contributions are used to retrieve IWP at daytime mentioned in Section 1.2, although all QRNN models learn to leverage the infrared channels for nighttime retrievals. The physical approaches referenced in Section 1.2 cannot perform nighttime retrievals, except for SatCORPS. A

comparison with SatCORPS was not possible due to data availability[1] and a comparison with CiPS was considered inappropriate as the latter targets a specific type of cloud. Bearing this in mind, this work has the advantage that can perform satisfactory daytime and nighttime retrievals.

Nevertheless, the models in the VISIR setting perform worse than in the IR setting for nighttime retrievals as seen in the scatter plots, particularly for the CNN model. This might be resolved by providing additional information to the networks to make them more robust, for example, the non-approximated local solar time of each pixel, and can be a continuation of this work. The differences in daytime and nighttime retrievals for the IR setting are likely caused by differences in the observations in each group, for instance, the nighttime observations containing cases more difficult to predict.

The improvement of the retrieval using spatial information is particularly seen in the scatter plots where the MLP and CNN architectures are compared, where the latter deviates less from the ideal prediction (Figure 4.3). The CNN architecture also presents smaller values of loss, MAE, ME, and median values of sharpness. The latter indicates that the probability is more concentrated and thus leads to smaller uncertainties, although these can still be relatively large. Nevertheless, the CNN presents two debatable results. Firstly, it has a general trend to have higher RMSE than the MLP in the same channels setting for different intervals of ground truth values considered. This can be considered a drawback and can be interpreted as occasionally obtaining larger errors, since the RMSE penalises larger errors more due to the squared term. Secondly, since for a given predicted IWP over an area, such as in Figure 4.5, the CNN provides a smoother variation of the IWP it might reduce true large IWP gradients among neighbouring predictions, which might be correlated with the larger RMSE values. Nonetheless, this reduction of large IWP gradients can hardly be assessed with the ground truth dataset considered.

Complementing the DARDAR data with data from other instruments that provide ice measurements, for example, a hypothetical network of ground-based radars, could help assess the last statement. Besides, the CNN training data is constrained by the narrow width of the DARDAR swath. If a data source that has larger spatial coverage per observation with the same sensitivity is used, then the CNN retrieval might leverage this larger spatial information in the reference data. However, such data sources are not available for the region used to train the models, to the best knowledge of the author. The use of another data source with IWP measurements at different local solar times should also improve the retrievals, and in particular if this data source has observations relatively close to sunrise and sunset. Despite it can only be speculated as the DARDAR data does not cover these time windows and that a characterisation of the models was not conducted, it is reasonable to assume that the models in the VISIR setting experience an artefact in these time ranges; models in the IR setting do not present suspicious dips in the diurnal variations. Reference data covering these time windows may resolve the issue. Complementing DARDAR with other data sources can be an area of further research, but it should be borne in mind that there can be large discrepancies between satellite observations (Duncan & Eriksson, 2018). In addition, it could be investigated if the artefacts are also observed in other areas or only in the two areas used for the comparison.

[1]The SatCORPS website (NASA LaRC, 2017) provides cloud products for a set of satellites, including the IWP for Meteosat satellites. However, it was not found sufficient Meteosat-9 data for a comparison with SatCORPS.

It is possible to increase the amount of DARDAR data used in training to further improve the models developed, as it can be assumed that this will improve the retrievals. Data before 6 May 2008 was discarded because some SEVIRI data was only available from a different processing algorithm than data from this date. An analysis of the implications of potential differences resulting from these two algorithms might reveal how to treat the data from the different algorithms so that it can be used for training the models. After the CloudSat battery malfunction in April 2011, the DARDAR data contains only daytime measurements. These measurements can be included, although the change in the distribution of daytime and nighttime measurements should be taken into account. Additionally, the training data can be expanded to a larger geographical area, in particular a larger latitude range. This expansion should increase the range of local solar times of the data with which the models are trained. Nonetheless, the larger the deviation from the nadir point of the satellite, the larger the distortions experienced in the signal measured by SEVIRI, which should be borne in mind in any further work in this direction. There exists a fourth way to increase the training data, which consists in including the validation data in the training set. The reduce on plateau learning rate scheduler used the validation data to reduce the learning rate, but if other schedulers that do not make use of it are used then this set can be merged into the training set.

A characterisation of the differences among the SEVIRI instruments aboard the different Meteosat satellites can reveal if the models trained on Meteosat-9 observations are also applicable to observations from the other Meteosat satellites. This would not only allow to expand the use of these models beyond the start and end of life of Meteosat-9, but also make it possible to increase the temporal resolution if the region of interest is between the latitudes +15° and +70°; the Rapid Scanning Service (RSS) (EUMETSAT, 2021b) provides scans covering this latitude range every 5 minutes instead of 15 minutes, the same as current weather radars, and the disseminated images result compatible with the data used to train the models. If a region of interest is between these latitudes and the period of study expands beyond April 2013, this characterisation is a requirement: then Meteosat-9 stopped the full disc scans and took over the RSS from Meteosat-8. Additionally, the applicability of the models for Meteosat-8 data can be regarded as particularly important: this satellite was re-located from roughly the prime meridian to +41.5° in 2016, and therefore covers another area. In this case, validation of the models for another geographical area with DARDAR data can only be assessed with daytime measurements.

Transfer learning is a machine learning method that can help when not much data is available. This method leverages a previously trained model to adapt it for a similar problem. In the hypothetical case where there are small but non-negligible differences between Meteosat observations, transfer learning could be considered to adapt the models for the Meteosat satellites for which there is only daytime DARDAR data. Additionally, transfer learning could be used as a possible starting point for investigating the applicability of the models, trained with Meteosat-9 observations, on observations from other imagers which have similar spectral channels but that cover other geographical regions and started operating later than Meteosat-9.

Concerning the machine learning methodology, it was chosen to focus on two particular architectures tuned over a fixed set of parameters. An alternative approach would have been to not deeply focus on two architectures but rather explore shallowly many models,

for instance simpler CNNs or MLPs with different regularisation methods to combat overfitting when going deeper, or to tune over a random set of parameters using prior knowledge on their influence for each model, such as that reported in Section 4.2. Furthermore, the choice of the optimiser and its hyperparameters including the learning rate can be critical to improve the results, though empirical studies can indicate opposite directions in this choice, that is, SGD generalising better than adaptive methods such as Adam (Wilson *et al.*, 2017), or that adaptive methods should never underperform the ones they approximate when carefully tuned (Choi *et al.*, 2020).

6

Conclusion

The primary goal of this work was to evaluate how well quantile regression neural networks can retrieve atmospheric ice mass, quantified as the ice water path, from geostationary observations which have a higher spatial and temporal coverage than CloudSat. The results support that they perform a satisfactory retrieval, subject to the limitations imposed by the physics of remote sensors operating at the visible and infrared spectrum, albeit with large uncertainties, which seem to be unavoidable in this context.

Daytime retrievals are more satisfactory when the models use all SEVIRI channels instead of only thermal channels, but for nighttime retrievals the situation is reversed. It is reasoned that including the local solar time of each image pixel or expanding the latitude of the geographical region used for training may make the models using all channels more robust. This can be explored in further research. The QRNN models, which work with publicly available data of easy access, overcome the limitation of physical-based models which perform only daytime retrievals. This facilitates the study of diurnal variations. Comparing with the CLAAS-2.1 dataset, which relies on a physical-based approach, there is a correlation that favours using the QRNN models for the two areas studied. Diurnal variations obtained using all channels show a potential artefact not observed when only thermal channels are used, which favours the use of the latter.

The use of spatial information in the retrieval, evaluated with the CNN retrievals, appears to be beneficial. Besides a higher tendency to be close to the DARDAR IWP, the CNN also predicts sharper distributions than the MLP, which indicates that smaller uncertainties can be obtained with the former. Only one CNN architecture, although with different hyperparameters, has been considered. The endless possibilities to choose from for both CNN architectures and hyperparameters suggest that better results can be achieved. Nevertheless, it is argumented from visual inspection that the CNN may mask true large IWP gradients as retrievals are smoother both in space and time when compared to MLP retrievals. This may be a limitation imposed by the narrow swath of the DARDAR data, where the CNN models do not have much reference data with spatial information to learn from.

The models were trained with data from the SEVIRI instrument aboard Meteosat-9. This satellite belongs to the Meteosat Second Generation series, where all these satellites have a SEVIRI instrument aboard. If the differences between their SEVIRI instruments are small, the QRNN models could be adapted to be used with the other Meteosat satellites, expanding their usage beyond the lifetime of Meteosat-9.

References

AERIS/ICARE Data and Services Center (2021). *ICARE Online Data Archive*. URL: https://www.icare.univ-lille.fr/data-access/data-archive-access/ (visited on 03/05/2021).

Benas, N., S. Finkensieper, M. Stengel, G.-J. van Zadelhoff, T. Hanschmann, R. Hollmann and J. F. Meirink (2017). 'The MSG-SEVIRI-based cloud property data record CLAAS-2'. In: *Earth System Science Data* 9.2, pp. 415–434. DOI: 10.5194/essd-9-415-2017.

Boucher, O., D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao, U. Lohmann, P. Rasch, S. Satheesh, S. Sherwood, B. Stevens and X. Zhang (2013). 'Clouds and Aerosols Supplementary Material'. 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*. Ed. by T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. Midgley. Cambridge University Press. DOI: 10.1017/CBO9781107415324.016.

Bugliaro, L., T. Zinner, C. Keil, B. Mayer, R. Hollmann, M. Reuter and W. Thomas (2011). 'Validation of cloud property retrievals with simulated satellite radiances: a case study for SEVIRI'. In: *Atmospheric Chemistry and Physics* 11.12, pp. 5603–5624. DOI: 10.5194/acp-11-5603-2011.

Cannon, A. J. (2018). 'Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes'. In: *Stochastic environmental research and risk assessment* 32.11, pp. 3207–3225. DOI: 10.1007/s00477-018-1573-6.

Choi, D., C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison and G. E. Dahl (2020). *On Empirical Comparisons of Optimizers for Deep Learning*. arXiv: 1910.05446.

Chollet, F. (2017). 'Xception: Deep Learning with Depthwise Separable Convolutions'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

Delanoë, J. and R. J. Hogan (2010). 'Combined CloudSat-CALIPSO-MODIS retrievals of the properties of ice clouds'. In: *Journal of Geophysical Research: Atmospheres* 115.D4. DOI: 10.1029/2009JD012346.

Duncan, D. I. and P. Eriksson (2018). 'An update on global atmospheric ice estimates from satellite observations and reanalyses'. In: *Atmospheric Chemistry and Physics* 18.15, pp. 11205–11219. DOI: 10.5194/acp-18-11205-2018.

Erenshteyn, R., R. Foulds and S. Galuska (July 1994). 'Is Designing a Neural Network Application an Art or a Science?' In: *SIGCHI Bull.* 26.3, pp. 23–29. ISSN: 0736-6906. DOI: 10.1145/181518.181522.

EUMETSAT (Sept. 2017). *MSG Level 1.5 Image Data Format Description*. Tech. rep. URL: https://www-cdn.eumetsat.int/files/2020-05/pdf_ten_05105_msg_img_data.pdf (visited on 24/05/2021).

EUMETSAT (23rd Mar. 2019). *High Rate SEVIRI Level 1.5 Image Data – MSG – 0 degree*. URL: https://navigator.eumetsat.int/product/EO:EUM:DAT:MSG:HRSEVIRI (visited on 27/05/2021).

EUMETSAT (2021a). *Data Store*. URL: https://data.eumetsat.int/ (visited on 24/05/2021).

EUMETSAT (2021b). *Rapid Scanning Service*. URL: https://www.eumetsat.int/rapid-scanning-service (visited on 20/05/2021).

Field, P. R. and A. J. Heymsfield (2015). 'Importance of snow to global precipitation'. In: *Geophysical Research Letters* 42.21, pp. 9512–9520. DOI: 10.1002/2015GL065497.

Finkensieper, S., S. Meirink, G.-J. van Zadelhoff, T. Hanschmann, N. Benas, M. Stengel, P. Fuchs, R. Hollmann, J. Kaiser and M. Werscheck (2020a). *CLAAS-2.1: CM SAF CLoud property dAtAset using SEVIRI - Edition 2.1*. Satellite Application Facility on Climate Monitoring. DOI: 10.5676/EUM_SAF_CM/CLAAS/V002_01.

Finkensieper, S., M. Stengel, N. Benas and J. F. Meirink (Apr. 2020b). *CLAAS Edition 2.1, Product User Manual*. Ed. by R. Hollman. Reference number SAF/CM/KNMI/PUM/SEV/CLD, Issue 2.5. DOI: 10.5676/EUM_SAF_CM/CLAAS/V002_01.

Ioffe, S. and C. Szegedy (2015). 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *International conference on machine learning*. PMLR, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html (visited on 24/05/2021).

Kerkmann, J. (30th June 2004). *Applications of Meteosat Second Generation (MSG): Meteorological use of the SEVIRI IR3.9 channel*. EUMETSAT. URL: http://eumetrain.org/IntGuide/PowerPoints/Channels/Channel_IR39.ppt (visited on 12/05/2021).

Kingma, D. P. and J. L. Ba (2015). 'Adam: A method for stochastic gradient descent'. In: *ICLR: International Conference on Learning Representations*, pp. 1–15. arXiv: 1412.6980.

Koenker, R. (2005). *Quantile regression*. Cambridge New York: Cambridge University Press. ISBN: 978-0-521-60827-5.

Kuleshov, V., N. Fenner and S. Ermon (July 2018). 'Accurate Uncertainties for Deep Learning Using Calibrated Regression'. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2796–2804. URL: http://proceedings.mlr.press/v80/kuleshov18a.html (visited on 24/05/2021).

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie (2017). 'Feature pyramid networks for object detection'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125. arXiv: 1612.03144.

Minnis, P., G. Hong, S. Sun-Mack, W. L. Smith Jr., Y. Chen and S. D. Miller (2016). 'Estimating nocturnal opaque ice cloud optical depth from MODIS multispectral infrared radiances using a neural network method'. In: *Journal of Geophysical Research: Atmospheres* 121.9, pp. 4907–4932. DOI: 10.1002/2015JD024456.

Minnis, P., L. Nguyen, R. Palikonda, P. W. Heck, D. A. Spangenberg, D. R. Doelling, J. K. Ayers, W. L. S. Jr., M. M. Khaiyer, Q. Z. Trepte, L. A. Avey, F.-L. Chang, C. R. Yost, T. L. Chee and S.-M. Szedung (2008). 'Near-real time cloud retrievals from operational and research meteorological satellites'. In: *Remote Sensing of Clouds*

and the Atmosphere XIII. Ed. by R. H. Picard, A. Comeron, K. Schäfer, A. Amodeo and M. van Weele. Vol. 7107. International Society for Optics and Photonics. SPIE, pp. 19–26. DOI: 10.1117/12.800344.

Nakajima, T. and M. D. King (1990). 'Determination of the Optical Thickness and Effective Particle Radius of Clouds from Reflected Solar Radiation Measurements. Part I: Theory'. In: *Journal of Atmospheric Sciences* 47.15, pp. 1878–1893. DOI: 10.1175/1520-0469(1990)047<1878:DOTOTA>2.0.CO;2.

NASA LaRC (30th Aug. 2017). *Satellite Imagery and Cloud Products*. URL: https://satcorps.larc.nasa.gov/ (visited on 10/06/2021).

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala (2019). 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (visited on 24/05/2021).

Pfreundschuh, S., P. Eriksson, D. Duncan, B. Rydberg, N. Håkansson and A. Thoss (2018). 'A neural network approach to estimating a posteriori distributions of Bayesian retrieval problems'. In: *Atmospheric Measurement Techniques* 11.8, pp. 4627–4643. DOI: 10.5194/amt-11-4627-2018.

Pfreundschuh, S. (2021). *Quantile regression neural networks on top of Keras and Pytorch*. URL: https://github.com/simonpf/quantnn (visited on 24/05/2021).

Raspaud, M., D. Hoese, P. Lahtinen, S. Finkensieper, G. Holl, A. Dybbroe, S. Proud, A. Meraner, X. Zhang, S. Joro, joleenf, W. Roberts, L. Ø. Rasmussen, J. H. B. Méndez, Y. Zhu, R. Daruwala, strandgren, BENR0, T. Jasmin, T. Barnie, E. Sigurðsson, R. K. Garcia, T. Leppelt, ColinDuff, U. Egede, LTMeyer, M. Itkin, R. Goodson, jkotro and peters77 (Mar. 2021). *Pytroll/Satpy*. Version v0.27.0. DOI: 10.5281/zenodo.4638572.

Roebeling, R. A., A. J. Feijt and P. Stammes (2006). 'Cloud property retrievals for climate monitoring: Implications of differences between Spinning Enhanced Visible and Infrared Imager (SEVIRI) on METEOSAT-8 and Advanced Very High Resolution Radiometer (AVHRR) on NOAA-17'. In: *Journal of Geophysical Research: Atmospheres* 111.D20. DOI: https://doi.org/10.1029/2005JD006990.

Schmid, J. (2000). 'The SEVIRI instrument'. In: *Proceedings of the 2000 EUMETSAT meteorological satellite data user's conference, Bologna, Italy*. Vol. 29.

Song, H., T. Diethe, M. Kull and P. Flach (June 2019). 'Distribution calibration for regression'. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5897–5906. URL: http://proceedings.mlr.press/v97/song19a.html (visited on 24/05/2021).

Stephens, G. L., D. G. Vane, R. J. Boain, G. G. Mace, K. Sassen, Z. Wang, A. J. Illingworth, E. J. O'connor, W. B. Rossow, S. L. Durden, S. D. Miller, R. T. Austin, A. Benedetti, C. Mitrescu and the CloudSat Science Team (2002). 'The CloudSat Mission and the A-Train: A New Dimension of Space-Based Observations of Clouds and Precipitation'. In: *Bulletin of the American Meteorological Society* 83.12, pp. 1771–1790. DOI: 10.1175/BAMS-83-12-1771.

Strandgren, J., L. Bugliaro, F. Sehnke and L. Schröder (2017). 'Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks'. In: *Atmospheric Measurement Techniques* 10.9, pp. 3547–3573. DOI: 10.5194/amt-10-3547-2017.

Stubenrauch, C. J., W. B. Rossow, S. Kinne, S. Ackerman, G. Cesana, H. Chepfer, L. D. Girolamo, B. Getzewich, A. Guignard, A. Heidinger, B. C. Maddux, W. P. Menzel, P. Minnis, C. Pearl, S. Platnick, C. Poulsen, J. Riedi, S. Sun-Mack, A. Walther, D. Winker, S. Zeng and G. Zhao (2013). 'Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel'. In: *Bulletin of the American Meteorological Society* 94.7, pp. 1031–1049. DOI: 10.1175/BAMS-D-12-00117.1.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich (2015). 'Going deeper with convolutions'. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna (2016). 'Rethinking the Inception Architecture for Computer Vision'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

Trepte, Q. Z., P. Minnis, S. Sun-Mack, C. R. Yost, Y. Chen, Z. Jin, G. Hong, F.-L. Chang, W. L. Smith, K. M. Bedka and T. L. Chee (2019). 'Global Cloud Detection for CERES Edition 4 Using Terra and Aqua MODIS Data'. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.11, pp. 9410–9449. DOI: 10.1109/TGRS.2019.2926620.

Waliser, D. E., J.-L. F. Li, C. P. Woods, R. T. Austin, J. Bacmeister, J. Chern, A. Del Genio, J. H. Jiang, Z. Kuang, H. Meng, P. Minnis, S. Platnick, W. B. Rossow, G. L. Stephens, S. Sun-Mack, W.-K. Tao, A. M. Tompkins, D. G. Vane, C. Walker and D. Wu (2009). 'Cloud ice: A climate model challenge with signs and expectations of progress'. In: *Journal of Geophysical Research: Atmospheres* 114.D8. DOI: 10.1029/2008JD010015.

Wilson, A. C., R. Roelofs, M. Stern, N. Srebro and B. Recht (2017). 'The marginal value of adaptive gradient methods in machine learning'. In: arXiv: 1705.08292.

Winker, D. M., J. R. Pelon and M. P. McCormick (2003). 'CALIPSO mission: spaceborne lidar for observation of aerosols and clouds'. In: *Lidar Remote Sensing for Industry and Environment Monitoring III*. Ed. by U. N. Singh, T. Itabe and Z. Liu. Vol. 4893. International Society for Optics and Photonics. SPIE, pp. 1–11. DOI: 10.1117/12.466539.

Winker, D. M., M. A. Vaughan, A. Omar, Y. Hu, K. A. Powell, Z. Liu, W. H. Hunt and S. A. Young (2009). 'Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms'. In: *Journal of Atmospheric and Oceanic Technology* 26.11, pp. 2310–2323. DOI: 10.1175/2009JTECHA1281.1.

Yost, C. R., P. Minnis, S. Sun-Mack, Y. Chen and W. L. Smith (2021). 'CERES MODIS Cloud Product Retrievals for Edition 4—Part II: Comparisons to CloudSat and CALIPSO'. In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5, pp. 3695–3724. DOI: 10.1109/TGRS.2020.3015155.

A

Supplementary material

A.1 Training set visuals

This Section presents the analogous Figures and Table in Section 4.1 for the training set data. The values in Table A.1 were computed as for the test set, although the means from the predicted distributions which resulted above $100\,\mathrm{kg\,m^{-2}}$, a situation which was never observed in the test set, were not considered to get a better bigger picture, as they completely biased the metrics. They represented less than 0.001 % of the data points (34/3 475 649), and this only occurred for the MLP models. The comments to be made are the same as for the test set. The results are slightly better, which can be expected since it is the data the models fit.

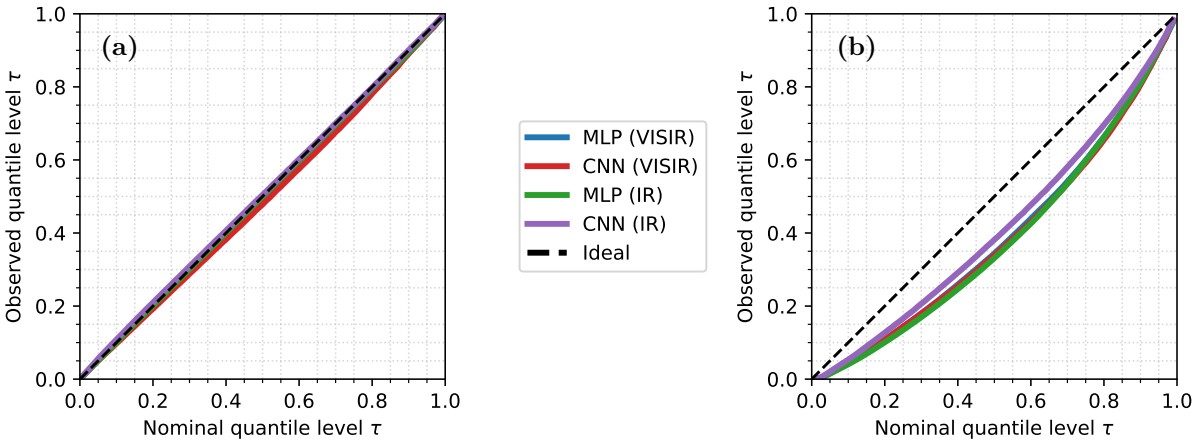


Figure A.1: Reliability diagrams for the training set where in (a) all IWP values are considered to compute the observed level and in (b) zero DARDAR IWP values were discarded for the calculation. The MLP (VISIR) line is hardly visible as it is completely overlapped by the other curves. CNN (IR) does not hide another curve in (b).

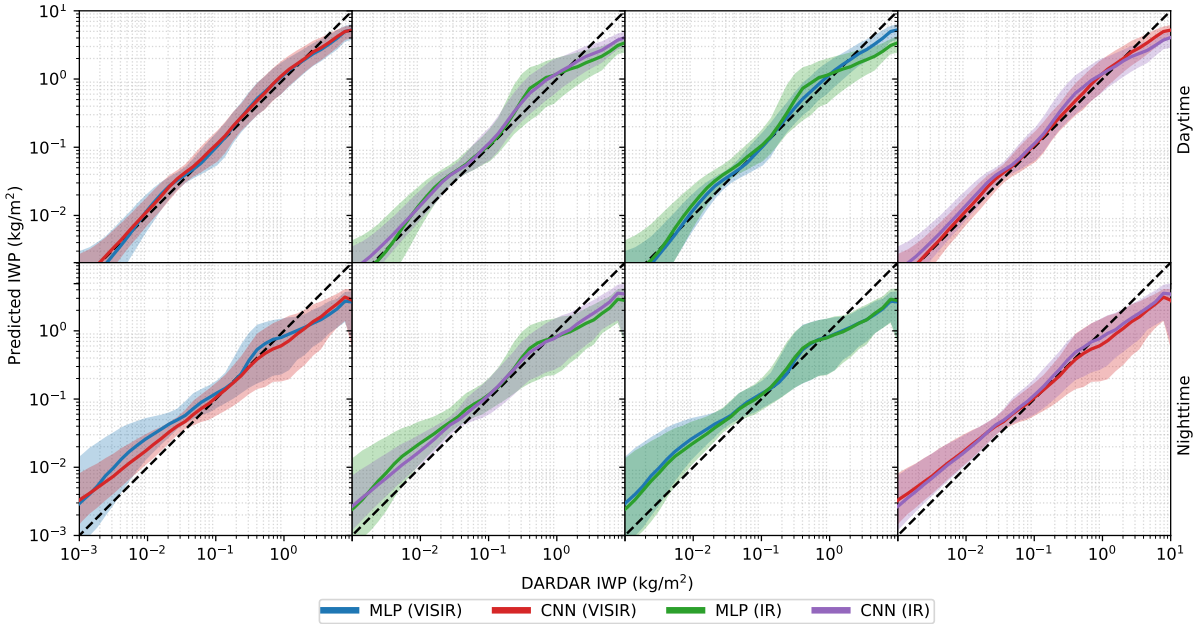


Figure A.2: Median (solid line) bounded by the first and third quartiles (filled area) indicating the central tendency and dispersion of the data points in the training set for the binned scatter plots explained in Section 3.2, where each data point is the mean of the predicted distribution. The same channels settings in the two different architectures are compared, and vice versa, for daytime (top row) and nighttime (bottom row) retrievals.

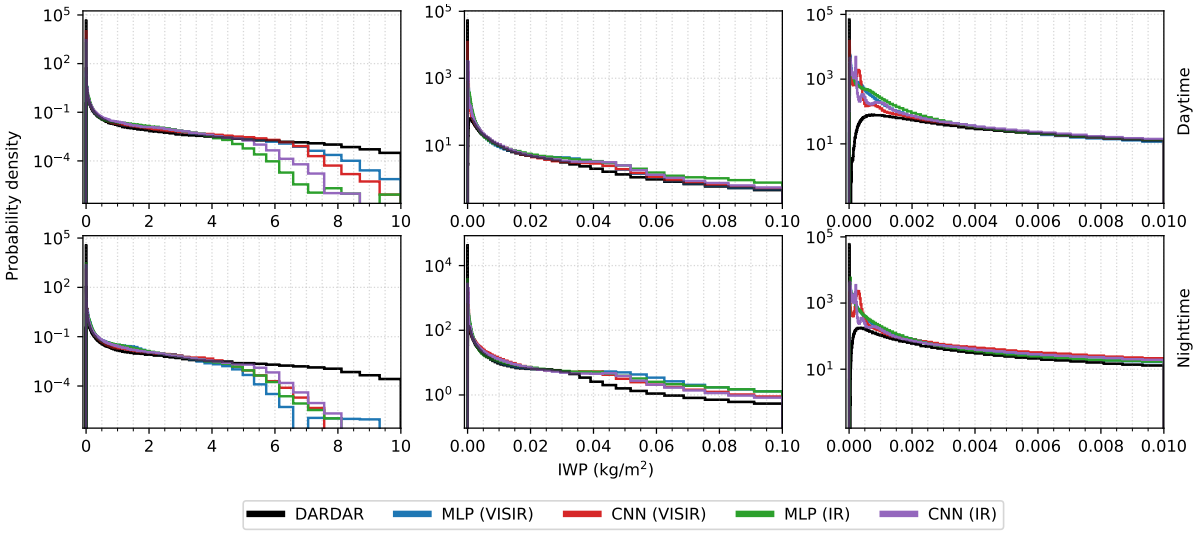


Figure A.3: Probability density functions for the training set for daytime and nighttime retrievals.

Table A.1: Other performance metrics computed for the training set. All values in $\mathrm{kg\,m^{-2}}$. The intervals indicate that only data points with ground truth values in the given interval are considered to compute the metric, and x_τ the quantiles at level τ. Values closest to zero are highlighted: architecture-wise in light grey, and row-wise in dark grey (considering all decimals).

Metric	Ground truth interval	MLP		CNN	
		VISIR	IR	VISIR	IR
RMSE	All values	4.83×10^{-1}	5.25×10^{-1}	4.81×10^{-1}	4.91×10^{-1}
	> 0	6.57×10^{-1}	7.13×10^{-1}	6.52×10^{-1}	6.65×10^{-1}
	1 to 10^{1}	2.30	2.49	2.28	2.32
	10^{-1} to 1	6.27×10^{-1}	7.38×10^{-1}	6.48×10^{-1}	6.62×10^{-1}
	10^{-2} to 10^{-1}	2.73×10^{-1}	2.82×10^{-1}	2.60×10^{-1}	2.77×10^{-1}
	10^{-3} to 10^{-2}	1.77×10^{-1}	2.09×10^{-1}	1.88×10^{-1}	1.87×10^{-1}
	10^{-5} to 10^{-3}	1.80×10^{-1}	1.70×10^{-1}	1.82×10^{-1}	1.84×10^{-1}
	$< 10^{-5}$	6.36×10^{-2}	6.76×10^{-2}	7.00×10^{-2}	7.75×10^{-2}
MAE	All values	9.55×10^{-2}	1.09×10^{-1}	9.06×10^{-2}	9.61×10^{-2}
	> 0	1.74×10^{-1}	1.98×10^{-1}	1.65×10^{-1}	1.75×10^{-1}
	1 to 10^{1}	1.70	1.89	1.70	1.72
	10^{-1} to 1	3.51×10^{-1}	4.39×10^{-1}	3.40×10^{-1}	3.76×10^{-1}
	10^{-2} to 10^{-1}	7.00×10^{-2}	7.69×10^{-2}	5.84×10^{-2}	6.79×10^{-2}
	10^{-3} to 10^{-2}	3.19×10^{-2}	3.39×10^{-2}	2.50×10^{-2}	2.77×10^{-2}
	10^{-5} to 10^{-3}	2.32×10^{-2}	2.48×10^{-2}	1.95×10^{-2}	2.08×10^{-2}
	$< 10^{-5}$	4.81×10^{-3}	5.86×10^{-3}	3.85×10^{-3}	4.45×10^{-3}
ME	All values	-2.57×10^{-3}	-4.99×10^{-3}	-4.83×10^{-3}	-3.48×10^{-3}
	> 0	-8.94×10^{-3}	-1.43×10^{-2}	-1.23×10^{-2}	-9.89×10^{-3}
	1 to 10^{1}	-1.18	-1.49	-1.06	-1.21
	10^{-1} to 1	2.06×10^{-1}	2.85×10^{-1}	1.82×10^{-1}	2.34×10^{-1}
	10^{-2} to 10^{-1}	6.21×10^{-2}	6.92×10^{-2}	4.92×10^{-2}	5.99×10^{-2}
	10^{-3} to 10^{-2}	3.03×10^{-2}	3.23×10^{-2}	2.38×10^{-2}	2.70×10^{-2}
	10^{-5} to 10^{-3}	2.30×10^{-2}	2.47×10^{-2}	1.93×10^{-2}	2.09×10^{-2}
	$< 10^{-5}$	4.81×10^{-3}	5.86×10^{-3}	3.87×10^{-3}	4.40×10^{-3}
$\boldsymbol{x_{0.75} - x_{0.25}}$ **(median value)**	All values	2.00×10^{-3}	2.10×10^{-3}	1.28×10^{-3}	1.49×10^{-3}
	> 0	1.20×10^{-2}	1.26×10^{-2}	9.23×10^{-3}	1.01×10^{-2}
	1 to 10^{1}	1.87	1.89	1.99	2.12
	10^{-1} to 1	1.59×10^{-1}	1.62×10^{-1}	1.57×10^{-1}	1.80×10^{-1}
	10^{-2} to 10^{-1}	1.76×10^{-2}	1.78×10^{-2}	1.57×10^{-2}	1.58×10^{-2}
	10^{-3} to 10^{-2}	4.70×10^{-3}	4.71×10^{-3}	3.21×10^{-3}	3.37×10^{-3}
	10^{-5} to 10^{-3}	1.14×10^{-3}	1.24×10^{-3}	9.29×10^{-4}	1.16×10^{-3}
	$< 10^{-5}$	9.20×10^{-7}	1.47×10^{-6}	5.73×10^{-7}	5.50×10^{-7}
$\boldsymbol{x_{0.95} - x_{0.05}}$ **(median value)**	All values	8.87×10^{-3}	9.89×10^{-3}	5.60×10^{-3}	5.84×10^{-3}
	> 0	6.03×10^{-2}	6.61×10^{-2}	4.97×10^{-2}	5.19×10^{-2}
	1 to 10^{1}	5.08	5.46	5.24	5.25
	10^{-1} to 1	6.82×10^{-1}	1.02	6.36×10^{-1}	9.28×10^{-1}
	10^{-2} to 10^{-1}	9.09×10^{-2}	9.68×10^{-2}	8.31×10^{-2}	8.09×10^{-2}
	10^{-3} to 10^{-2}	1.69×10^{-2}	1.82×10^{-2}	1.33×10^{-2}	1.40×10^{-2}
	10^{-5} to 10^{-3}	6.02×10^{-3}	6.33×10^{-3}	4.41×10^{-3}	4.25×10^{-3}
	$< 10^{-5}$	1.95×10^{-3}	2.52×10^{-3}	1.28×10^{-3}	1.15×10^{-3}
CRPS mean	All values	5.14×10^{-2}	4.61×10^{-2}	5.47×10^{-2}	4.92×10^{-2}
median		4.04×10^{-4}	4.49×10^{-4}	3.64×10^{-4}	4.95×10^{-4}

A.2 Retrieval example using all channels

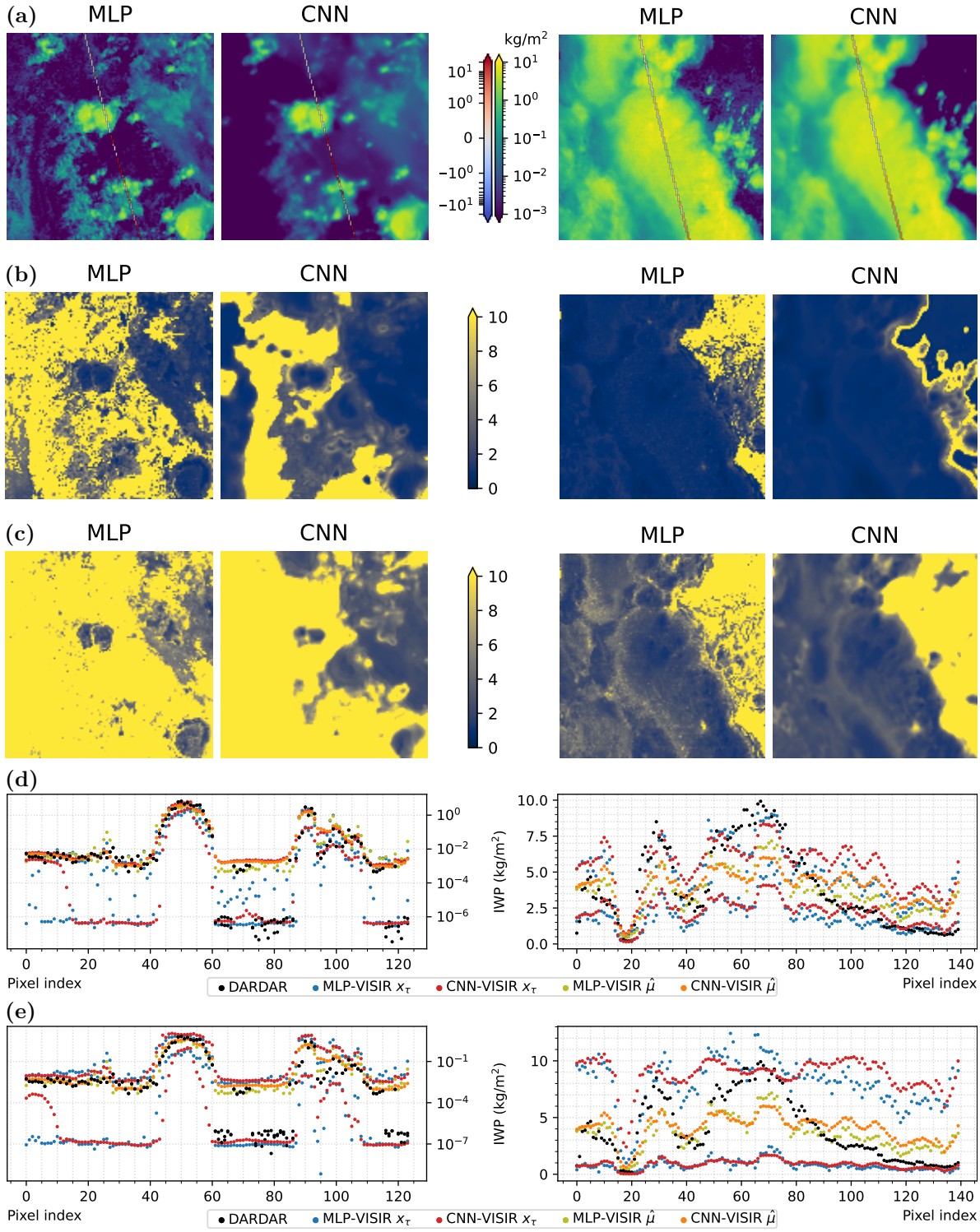


Figure A.4: In (a), two IWP daytime retrievals using the VISIR channels setting with the relative error $(\hat{\mu} - y)/y$ overlaid, where $\hat{\mu}$ is the predicted distribution mean and y is the DARDAR IWP. Relative sharpnesses in (b) and (c), omputed as $(x_{0.75} - x_{0.25})/x_{0.50}$ and $(x_{0.95} - x_{0.05})/x_{0.50}$, where x_τ indicates the quantiles at level τ. The quantiles at levels 0.25 and 0.75 and means along the DARDAR swath in (d), where the index increases top-to-bottom of the image. In (e) same as (d) but for levels 0.05 and 0.95.

DEPARTMENT OF SPACE, EARTH AND ENVIRONMENT
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY