

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

**Closing spatiotemporal gaps in hydrometeor
retrievals: exploiting geostationary infrared
observations via probabilistic deep learning**

ADRIÀ AMELL

Department of Environmental and Energy Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, 2026

Closing spatiotemporal gaps in hydrometeor retrievals: exploiting geostationary infrared observations via probabilistic deep learning

ADRIÀ AMELL

ISBN 978-91-8103-443-1

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

© 2026 Adrià Amell, except where otherwise stated.
All rights reserved.

Selected material from the author's Licentiate thesis, *Retrieving cloud ice masses from geostationary images with neural networks* (Amell, 2023), is republished in this PhD thesis.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5900
ISSN 0346-718X
<https://doi.org/10.63959/chalmers.dt/5900>

Division of Geoscience and Remote Sensing
Department of Environmental and Energy Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden, 2026

To my sister Anna,

for guiding me through my first days of university,
for helping me navigate life across borders,
and for sharing the unspoken language of a PhD journey.

Closing spatiotemporal gaps in hydrometeor retrievals: exploiting geostationary infrared observations via probabilistic deep learning

ADRIÀ AMELL

*Department of Environmental and Energy Sciences
Chalmers University of Technology*

Abstract

Comprehensive characterization of clouds and precipitation is fundamental to advancing our understanding of weather and climate systems. However, uncertainties persist in key quantities such as the mass of frozen hydrometeors and surface precipitation. Specialized spaceborne instruments provide high-quality measurements, but their orbital characteristics restrict their spatiotemporal sampling. Geostationary satellites, in contrast, offer continuous monitoring over continental scales, yet they have historically been underexploited. Relying solely on geostationary radiances, this thesis employs neural networks trained against gold-standard satellite products to retrieve frozen hydrometeor masses, cloud probabilities, and surface precipitation. Given the severely ill-posed nature of this inversion problem, the networks output case-specific probabilistic information targeting the irreducible uncertainty of these retrievals.

The appended papers challenge traditional paradigms by demonstrating the value of infrared (IR) observations beyond cloud top properties. Results show that retrievals of frozen hydrometeor masses using only thermal IR channels closely align in performance with those incorporating all visible and IR channels. This alignment overcomes the daytime-only limitation of existing physics-based methods. Extending this approach, machine learning delivers skilful 2D and 3D retrievals of frozen hydrometeor masses and cloud probabilities from only a single IR channel. This capability enables creating the *Chalmers Cloud Ice Climatology* from multi-decadal records of globally harmonized geostationary observations. Parallel to cloud properties, a multispectral IR-based retrieval developed for *Rain over Africa* provides rainfall estimates that outperform established products in accuracy, resolution, and latency.

These contributions unlock new applications for geostationary observations, providing unprecedented spatial and temporal continuity of hydrometeor data. The developed retrievals facilitate new insights into clouds, reduce observational uncertainties, and help validate climate models. Moreover, they carry clear societal value by allowing for timely estimates required for risk mitigation. Efforts to leverage these new capabilities are already underway.

Keywords

Machine learning, satellites, ice clouds, precipitation, Africa, remote sensing

Acknowledgements

Looking back on these years, I am moved by how many people have shaped this chapter of my life. Throughout this journey of learning, challenges, and growth, I have been fortunate to receive guidance and encouragement from so many. I have not come this far alone, and there are far more people to thank than I can name here.

First and foremost, I am profoundly grateful to my supervisor, Patrick Eriksson. His availability, his patience, and his detailed and constructive feedback have not only strengthened my research but also inspired me. The freedom he gave me and his finesse have been invaluable, both professionally and personally.

I extend my heartfelt thanks to Simon Pfreundschuh for the discussions that enriched my work. His willingness to meet at 7 a.m. his time, or even earlier, has been pivotal to our collaboration across time zones and stands as a testament to his genuine commitment to sharing knowledge.

My gratitude also goes to my examiner, Lars Ulander, to my co-supervisors, and to the administrative staff, especially Paulina Sjögren, who has been not only instrumental but also a cheerful presence throughout. A warm thanks to Santiago Arellano, my line manager, for his thoughtful support. I also appreciate my faculty opponent and grading committee for engaging with this thesis.

I gratefully acknowledge the Division of Geoscience and Remote Sensing and ultimately the taxpayers for funding my position and providing a working environment I have genuinely enjoyed. More generally, I am thankful to the Swedish system for the many ways it supports doctoral students, and to the city of Gothenburg for being a good place to live and work these past years.

To my office mates Franz, Hanna, and Fangyuan, who shared the day-to-day at different stages of this journey, thank you for the company, the conversations, and the small things that make a workplace feel like one. I owe a special thanks to my friend Raül, who shared that Chalmers was a great place to come for a master's and stay for a PhD. My appreciation also goes to my colleagues and fellow doctoral students, both at Chalmers and beyond, whose discussions have refined my questions and reminded me of the limits of my own knowledge. A quiet but sincere thanks to the anonymous reviewers of my papers, whose comments have improved both the work and the way I approach research.

A warm thanks to the four master's thesis students I supervised, Lilian, Axel, Filip, and Ludvig, as well as the undergraduates in the programming

course. Teaching at different levels has taught me as much as I hope I gave them in return.

Thanks also to Philippe Baron for hosting me at NICT in Tokyo. This short but memorable visit gave me a fantastic glimpse into the workings of a national research institute so far from home.

To the regulars and instructors in the group classes I join most evenings, thank you for pushing me to become stronger and more aware of my body, for the workouts, the chats, the friendships that grew alongside them, and for creating a space where I could disconnect after work and leave with a clearer head.

Many thanks to my friends, who have listened to both complaints and excitement over these years. I owe much to Rosa, Noe, and Cristina for taking such good care of my parents in difficult times. To my parents, and in loving memory of my father, whose absence I have felt this past year. To my siblings, Anna and Lluís, and to my extended family.

Finally, to Laura, whose love, support, and gentle pushes have meant more than I can express.

Ària Amell

Gothenburg, Sweden
May 2026

List of Publications

Appended papers

This thesis is based on the following papers:

- [I] **A. Amell**, P. Eriksson, and S. Pfreundschuh, *Ice water path retrievals from Meteosat-9 using quantile regression neural networks*
Atmos. Meas. Tech., 15, 5701–5717, 2022.
DOI: 10.5194/amt-15-5701-2022
- [II] **A. Amell***, S. Pfreundschuh*, and P. Eriksson, *The Chalmers Cloud Ice Climatology: retrieval implementation and validation*
Atmos. Meas. Tech., 17, 4337–4368, 2024.
DOI: 10.5194/amt-17-4337-2024
*Indicates equal contribution.
- [III] S. Pfreundschuh, J. Kukulies, **A. Amell**, H. Hallborn, E. May, and P. Eriksson, *The Chalmers Cloud Ice Climatology: A novel robust climate record of frozen cloud hydrometeor concentrations*
J. Geophys. Res. Atmos., 130, e2024JD042618, 2025.
DOI: 10.1029/2024JD042618
- [IV] **A. Amell**, S. Pfreundschuh, and P. Eriksson, *Fine-tuning a machine-learned 3D cloud climatology reveals aspects of cloud cover trends*
Under review at *Geophys. Res. Lett.*, 2026.
DOI: 10.22541/essoar.15001993/v1
- [V] **A. Amell**, L. Hee, S. Pfreundschuh, and P. Eriksson, *Probabilistic near-real-time retrievals of Rain over Africa using deep learning*
J. Geophys. Res. Atmos., 130, e2025JD044595, 2025.
DOI: 10.1029/2025JD044595

Summary of contributions

The contributions to the appended papers by the author of the thesis are:

- [I] Collected and prepared all data, performed the training and evaluation of the models, compared with CLAAS-2.1, and prepared the manuscript.
- [II] In close collaboration with S. Pfreundschuh, collected and prepared the data, developed the software package supporting the publication, evaluated the model, compared with other datasets, and prepared the manuscript. Additionally, supported the model training led by S. Pfreundschuh with technical discussions.
- [III] Collaborated with all authors on literature review and methodology, and contributed to the drafting of the manuscript. Processed and published the CCIC data record. Responsible for the collection and preparation of PATMOS-x and CCIC CPCIR time series data. Reviewed the codebase of the study, focusing on MODIS data processing and CCIC GridSat time series.
- [IV] Collected and prepared all data, including processing and managing the CCIC GridSat retrievals of 3D cloud probabilities. Performed the training and evaluation of the model. Analysed the impact of cloud overlap assumptions and spatial aggregation approaches, executed all comparisons, and prepared the manuscript.
- [V] Supervised the initial research phase, providing technical feedback on the training database construction and a first model version implemented by L. Hee. Optimized the software package supporting the publication. Trained and evaluated the published model. Performed the comparative analysis with all datasets in the publication and prepared the manuscript.

Other works

The following publications were published during my PhD studies or are currently under review but are not appended, as their focus differs from that of this thesis.

- P. Baron, **A. Amell**, S. Kawamura, S. Satoh, and T. Ushio, *3D precipitation nowcasting from phased array radar with uncertainty estimation using a quantile regression neural network*
2025 IEEE Radar Conference (RadarConf25), Krakow, Poland, 2025.
DOI: 10.1109/RadarConf2559087.2025.11204931
- S. Ageet, J.H. Marsham, D.J. Parker, A. Singh, C.M. Taylor, A.J. Roberts, J. Groves, E. Nkonde, and **A. Amell**, *Evaluation of satellite rainfall products for nowcasting over Africa using rain-gauges and lightning data*
Under review at *Meteorol. Appl.*, 2026.

Contents

Abstract	iii
Acknowledgements	v
List of Publications	vii
List of Acronyms	xv
I Introductory chapters	1
1 Introduction	3
1.1 Problem statement	4
1.2 Research objectives	5
1.3 Organization of the thesis	6
2 Spaceborne systems for retrievals of hydrometeors	7
2.1 Scientific context	7
2.1.1 Radiation physics	7
2.1.1.1 The electromagnetic spectrum	7
2.1.1.2 Intuitively interpreting thermal radiation . . .	8
2.1.1.3 Molecular absorption	8
2.1.1.4 Particle attenuation	10
2.1.1.5 Emission	10
2.1.2 Wavelength selection	11
2.1.2.1 IR versus MW	12
2.1.2.2 Angular resolution	13
2.1.3 Passive and active remote sensing	13
2.1.4 A note on the definition of clouds and precipitation . .	14
2.2 The critical role of satellites	16
2.3 Satellite orbits	16
2.4 Satellite data for this thesis	18
2.4.1 The CloudSat-CALIPSO tandem: profiling of clouds . .	19
2.4.2 GPM Core Observatory: the precipitation reference . .	22
2.4.3 Geostationary satellites: the input	23
2.4.3.1 Meteosat Second Generation	23
2.4.3.2 Geo-ring	24

3	Machine learning for retrievals of hydrometeors	27
3.1	Reshaping the retrieval landscape with machine learning	27
3.1.1	The retrieval problem	27
3.1.2	Traditional physics-based approaches	28
3.1.2.1	Accounting for retrieval uncertainty	29
3.1.3	The machine learning paradigm	29
3.2	Foundations of effective deep learning	30
3.2.1	Model capacity	30
3.2.2	Dataset design and quality	31
3.2.3	Experimental rigour	32
3.2.4	Computational considerations	33
3.3	Deep learning architectures	34
3.3.1	Fundamental building blocks	35
3.3.2	Design principles	36
3.3.3	Architecture selection	37
3.4	Uncertainty quantification	39
3.4.1	Types of uncertainty	39
3.4.2	Approaches to uncertainty estimation	39
4	Retrieval development	41
4.1	Constructing the datasets of collocations	41
4.2	Probabilistic retrievals	42
4.2.1	Quantile regression	42
4.2.2	Classification	44
4.2.2.1	Rationale for the loss function	44
4.2.2.2	Decision criteria for cloud classes	46
4.3	Training methodology	47
4.4	Evaluation	50
4.5	Computational environment	51
5	Summary of the papers	53
5.1	The Chalmers Cloud Ice Climatology	53
5.1.1	Precursor work	54
5.1.2	Retrieval development	54
5.2	Rain over Africa	56
5.3	Contributions to the research field	57
6	Downstream applications	59
6.1	Towards early warnings in Africa	59
6.2	Benchmark for atmospheric models	60
6.3	Baseline for new satellite products	62
7	Outlook	63
	References	65

II Appended Papers 79

Paper I – Ice water path retrievals from Meteosat-9 using quantile regression neural networks

Paper II – The Chalmers Cloud Ice Climatology: retrieval implementation and validation

Paper III – The Chalmers Cloud Ice Climatology: a novel robust climate record of frozen cloud hydrometeor concentrations

Paper IV – Fine-tuning a machine-learned 3D cloud climatology reveals aspects of cloud cover trends

Paper V – Probabilistic near-real-time retrievals of Rain over Africa using deep learning

List of Acronyms

CCIC	Chalmers Cloud Ice Climatology
CNN	Convolutional neural network
CPC	NOAA Climate Prediction Center
CPCIR	CPC globally merged IR product version 1
CPR	Cloud Profiling Radar
CPU	Central processing unit
DL	Deep learning
GEO	Geostationary orbit
GES DISC	NASA Goddard Earth Sciences Data and Information Services Center
GPM	Global Precipitation Measurement
GPM-CO	GPM Core Observatory
GPU	Graphics processing unit
GridSat	GridSat-B1 product version 2
IR	Infrared
IWC	Ice water content
IWP	Ice water path
LEO	Low Earth orbit
ML	Machine learning
MLP	Multilayer perceptron
MSG	Meteosat Second Generation
MW	Microwave
NCEI	NOAA National Centers for Environmental Information
NCEP	NOAA National Centers for Environmental Prediction
NN	Artificial neural network
OEM	Optimal estimation method
QPD	Quantile-parameterized distribution
QRNN	Quantile regression neural network
SEVIRI	Spinning Enhanced Visible Infra-Red Imager
T_B	Brightness temperature
TIR	Thermal IR
VIS	Visible

Part I

Introductory chapters

1 Introduction

Advances in machine learning (ML) are currently a key driver of progress in remote sensing and geoscience. The wealth of observational datasets and developments in computational systems have catalysed the exploration and implementation of data-driven algorithms. It suffices to attend any conference, engage in scientific discussions or follow remote sensing journals to notice that ML serves not only as a powerful tool for statistical data processing but also increasingly for learning, understanding, and extracting knowledge. This paradigm shift in modern remote sensing is also the basis for this thesis.

Acquiring atmospheric information is one of the applications of remote sensing. Clouds are an integral and dynamic part of the atmosphere. These suspended masses of particulate matter play a crucial role in the Earth’s weather and climate systems. Clouds are critical for the Earth’s radiation budget by reflecting incoming sunlight, resulting in a cooling effect, and by retaining outgoing heat radiated from the Earth’s surface, leading to a warming effect. Therefore, changes in surface temperature are sensitive to the distribution of clouds (Romano, 2020), including their shape, size, altitude, and concentrations of hydrometeors—liquid water and ice particles. According to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (2023), ‘the cloud feedback remains the largest contribution to overall uncertainty’ in global warming.

Much of this uncertainty is caused by the presence of ice clouds, which dominate the upper layers of the troposphere. Unlike the cooling effect of liquid clouds at lower altitudes, changes in the concentration of cloud ice masses can shift their radiative forcing, as thin ice clouds are relatively transparent to the incoming radiation but are highly effective at trapping outgoing heat (e.g., Wallace and Hobbs, 2006). Furthermore, the latent heat released during the formation of atmospheric ice fuels atmospheric circulation and intensifies storm systems (e.g., Houze Jr., 2014). Moreover, precipitation often originates from ice crystals that grow until they have a sufficient mass to leave the cloud (e.g., Lamb and Verlinde, 2011).

As a direct product of the cloud systems, falling hydrometeors, i.e. precipitation, represent another essential atmospheric variable. Given their fundamental importance for society and ecosystems, obtaining accurate estimates of surface precipitation is a primary objective of atmospheric remote sensing, a necessity for a robust risk management system. Ultimately, precipitation shapes the availability of freshwater resources and food systems.

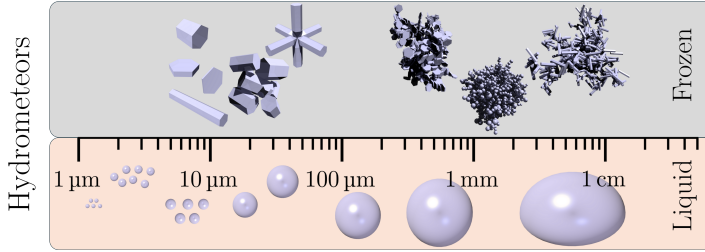


Figure 1.1: Examples of hydrometeors, categorized into frozen and liquid forms, ranging from pristine ice crystals and cloud droplets to large aggregates and raindrops. Shapes and renders adapted from Ekelund et al. (2020, CC BY-SA 4.0). Positions along the axis approximate particle sizes; not drawn to scale.

Regardless of the specific hydrometeor quantity of interest, remote sensing relies on indirect measurements of electromagnetic radiation that must be quantitatively inverted. This algorithmic process of mapping measured radiances back to physical parameters, such as hydrometeor mass, is known as a retrieval.

Satellites are indispensable for the remote sensing of clouds and precipitation: only with retrievals from spaceborne observations can the limited spatial coverage of ground-based instruments be overcome. Meteorological satellites are particularly valuable for monitoring remote, mountainous, or oceanic regions. Moreover, when observations from several missions are combined, they can deliver continuous and near-real-time estimates across the globe. Even a single geostationary satellite measures one side of the Earth within minutes, providing observations suitable for resolving full diurnal cycles and tracking atmospheric processes, a capability that a single platform in low Earth orbit cannot achieve due to its intermittent sampling over any given region. Nevertheless, spaceborne instruments face the inherent uncertainty that characterizes atmospheric retrievals. This difficulty is intensified by the broad variety of hydrometeor types and sizes, as illustrated in fig. 1.1.

1.1 Problem statement

Unsurprisingly, the scientific community has already addressed the challenge of retrieving frozen hydrometeors, surface precipitation, and other cloud properties. For geostationary observations, however, these efforts have been constrained by the nature of the measurements themselves, limited to reflected sunlight and thermal infrared radiances. Traditional retrieval approaches carry additional shortcomings of their own, summarized in section 3.1.2 and further elaborated in the appended papers.

Concerning cloud ice masses, the standard technique suffers from a major drawback: it is restricted to daytime observations. This implies that about half of the observations are not used and downstream analyses, such as diurnal cycles, are incomplete. Furthermore, resolving the vertical distribution of cloud

ice masses and cloud types has historically been deemed infeasible using only geostationary radiances, as these measurements mainly capture the information from the cloud top (section 2.1.1). Similarly, it has been widely assumed that geostationary data alone is insufficient for an accurate description of surface rainfall, necessitating the use of complementary data sources.

A further limitation exists in the way traditional methods quantify uncertainty. Many existing algorithms rely on Gaussian assumptions. However, such assumptions fail to capture the inherently skewed distributions and non-linear relationships between cloud ice, precipitation, and observed radiances. Consequently, the resulting error metrics often lack practical utility, highlighting the need for retrieval methods that can provide a more physically consistent description of the retrieval uncertainty.

1.2 Research objectives

The primary objective of this thesis is to enhance the retrieval of hydrometeor masses from geostationary satellites. Their continuous observations spanning multiple decades are uniquely suited to closing the spatiotemporal gaps that limit our understanding of weather and climate systems. To achieve this, the research departs from traditional physically-based inversions used in the field, adopting a radically data-driven ML methodology. Deep artificial neural networks (NNs) are trained to reproduce retrievals from specialized satellites considered a gold standard. Although the training datasets contain microphysical assumptions which are inherited, the algorithms developed in this thesis make no such assumptions, relying instead on purely empirical relationships.

A key methodological choice is that only thermal infrared radiances are selected as input to the models. Not only does this enable continuous retrievals throughout the day, allowing for a consistent diurnal description of the quantities, but it also leverages geostationary records of the 11 μm channel dating back to 1980 (Knapp et al., 2011). The information supplied to the NNs is chosen through exploration of the available geostationary data and the targets at stake. The choices are guided by black-box experimental results rather than physical models. The NNs used also vary in size and predefined capabilities, for example, the ability to leverage spatial information. Furthermore, these models integrate a probabilistic framework that provides retrieval uncertainty estimates that are not constrained by Gaussian statistics. This enables the deep learning (DL) models to generate probabilistic estimates suitable for sophisticated downstream applications.

Regarding the targeted variables, the research focuses primarily on the ice water path (IWP), i.e. the vertically-integrated atmospheric ice mass, followed by instantaneous surface precipitation. Besides IWP and surface precipitation, the research further challenges the traditional limitations of infrared retrievals by estimating the ice water content (IWC), i.e. the vertically-resolved IWP, alongside cloud cover and cloud types across different altitude levels. Geographically, IWP and rainfall retrievals are first developed and validated over Africa. Subsequently, retrievals of cloud ice and cloud types

are developed, validated, and studied on a quasi-global domain for multiple decades.

These objectives also set the limitations of this thesis. Exploring the explainability of the NNs is not considered and, therefore, the retrievals are simply treated as the output of a black-box. Although different NNs have been used, no exhaustive survey is conducted to find an optimal architecture; it can be argued that such an optimization would consume plenty of resources for slight improvements. The lack of suitable probabilistic reference data hinders the validation of retrieval uncertainties. Therefore, the probabilistic nature of the retrievals is limited to the evaluation of the internal consistency with the training data, in the sense of comparing the distribution of random samples with the training data distribution. The use of infrared radiances as input constrains the retrieval skill primarily to cloud top properties (section 2.1.2.1), making them well-suited for studying, e.g., feedbacks from high clouds. Nevertheless, this input choice also limits the vertical sensitivity of the retrievals, which remain inherently more uncertain and dependent on the training statistics. Finally, the research in this thesis does not address the forecasting of future scenarios at any scale, remaining strictly constrained to the inversion of the input data.

Overall, this thesis work provides a more comprehensive picture of frozen hydrometeor masses and cloud systems, together with more accurate and timely rainfall estimates. The retrievals implemented offer new information on the distribution and evolution of clouds, facilitating a better understanding of the changing climate and its repercussions on human activity. Furthermore, such datasets help support disaster risk management. A final core objective of this thesis is, therefore, to provide highly accessible datasets generated with the developed retrieval algorithms.

1.3 Organization of the thesis

This work is structured as a compilation thesis, comprising seven introductory chapters that provide context for the five appended papers. The papers are either published or under review at the time of writing the thesis. These introductory chapters also acknowledge aspects only briefly covered in the papers and offer broader context on the machine learning principles at the focus of this thesis. Chapter 1 through 4 establish the scientific background and research framework. Main contributions of the papers are summarized in chapter 5,¹ while chapter 6 is dedicated to summarizing specific applications that either directly result from this research or where this thesis work played a significant role. Finally, chapter 7 provides an outlook on potential research directions.

¹A detailed statement of the individual contributions of the author of the thesis to the appended papers can be found in the summary of contributions on page viii.

2 Spaceborne systems for retrievals of hydrometeors

2.1 Scientific context

A basic understanding of the governing physics of the atmosphere is essential for recognizing the possibilities and limitations of remote sensing systems. Similarly, a general notion of what is theoretically possible allows us to judge the performance of retrievals constructed with machine learning and set realistic expectations for the results. The appended papers focus on exploiting infrared measurements in the $6.2\text{ }\mu\text{m}$ to $13.4\text{ }\mu\text{m}$ range as input. However, the training targets consist of established microwave-derived data products. This section therefore provides an overview of the physical constraints governing both the input and reference data, which also motivate the rationale behind previous approaches. The following summary is based on standard textbooks, such as Rees (2012) or Wallace and Hobbs (2006).

2.1.1 Radiation physics

2.1.1.1 The electromagnetic spectrum

The electromagnetic spectrum is divided into ranges based on the wavelength λ or, equivalently, the frequency of its waves. The ranges are defined by practical applications. The specific thresholds used to limit each range can vary depending on the application or source.

Our eyes are sensitive to the $\lambda = 400\text{ nm}$ to 700 nm range, the visible region. Wavelengths shorter than the optical region consist of ultraviolet radiation, X-rays, and gamma radiation. The longest waves of the electromagnetic spectrum are radio waves with $\lambda \geq 1\text{ m}$. Microwaves (MWs) are short radio waves of $\lambda = 1\text{ mm}$ to 1 m . Radio and MWs are usually defined in terms of their frequency f instead of the wavelength. As a result, they can be defined by $f \leq 300\text{ MHz}$ to 300 GHz , respectively. The remaining region is the infrared (IR) region, covering 700 nm to 1 mm .

These ranges of the electromagnetic spectrum can be further subdivided. For instance, MWs used in radar applications are often designated by letters. Similarly, the IR can be divided into near, thermal, and far IR, with the latter acting as the bridge until the MW domain. Furthermore, instruments

measuring radiation are usually characterized by their spectral channels or bands, which refer to specific, discrete intervals of the electromagnetic spectrum that the sensor is designed to detect.

2.1.1.2 Intuitively interpreting thermal radiation

Any physical body with a temperature above 0 K emits electromagnetic waves, i.e. radiation. Its radiance is the quantity that describes the radiation strength per area and solid angle. Because objects in nature generally emit over a range of wavelengths, the spectral radiance L_λ ($\text{W m}^{-2} \text{sr}^{-1} \mu\text{m}^{-1}$) defines the variation of the radiance as a function of wavelength. When an object absorbs all incident radiation it is called a black-body; at thermodynamic equilibrium, the radiation it emits is known as black-body radiation.

In the early twentieth century, Max Planck derived the relationship between a black-body at temperature T and the spectral radiance, given by

$$L_\lambda = \frac{2hc^2}{\lambda^5 \left[\exp\left(\frac{hc}{\lambda k_B T}\right) - 1 \right]} \quad (2.1)$$

where h is the Planck constant, k_B the Boltzmann constant, and c the speed of light. From this spectral radiance, other radiometric quantities can be derived. In particular, by integrating L_λ over all directions from which radiation is received, we obtain the spectral irradiance E_λ , which reveals how the energy of the source is distributed across wavelengths. Integrating E_λ over all wavelengths yields the total irradiance; the fraction within a given spectral interval follows from the ratio of the interval's integral to this total.

Not all objects perfectly absorb and emit radiation and, thus, are not black-bodies. To relate the radiance measured by a sensor to the physical state of a body, the emissivity $\varepsilon(\lambda) \in [0, 1]$ describes how efficiently the body emits compared to a black-body at the same temperature. The brightness temperature T_B is then defined as the physical temperature a black-body would need to have to emit the same radiance as that observed; it is obtained by inverting Planck's law for the measured radiance. This conversion enables translating radiometric data into a more intuitive physical quantity that relates to the thermal state of the object, particularly when handling measurements across various wavelengths.

2.1.1.3 Molecular absorption

Electromagnetic waves travelling through a medium can interact with the elements that constitute it. As a consequence, the radiation experiences attenuation through absorption and scattering processes. In turn, the medium also emits radiation, as explained in section 2.1.1.5. The electromagnetic waves carry energy that is directly proportional to their frequency. When this energy matches one of the internal energy levels of a molecule, radiation can be absorbed.

There are three main absorption mechanisms. The higher energy radiation promotes electrons to higher orbitals. Infrared waves are not strong enough to

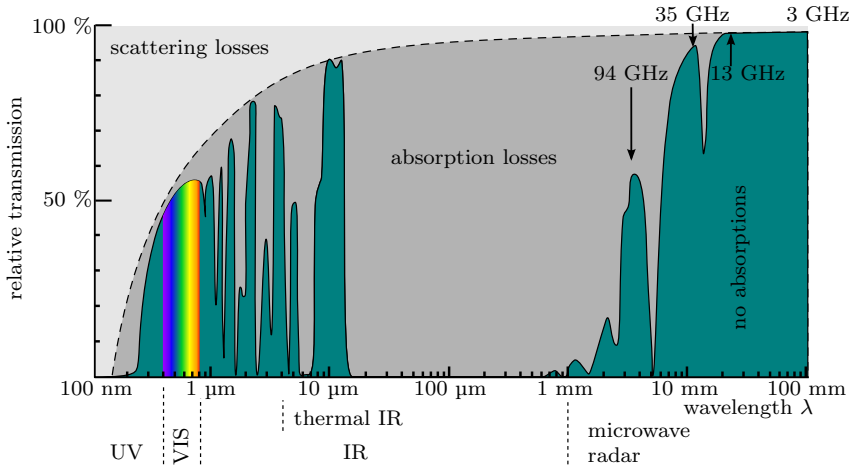


Figure 2.1: Schematic representation of the Earth's atmospheric transmissivity under clear-sky conditions. The diagram qualitatively illustrates regions of high transparency, such as the visible spectrum and the thermal IR window centred near $11\ \mu\text{m}$, as well as generally opaque regions, such as the IR wavelengths above $20\ \mu\text{m}$. The scattering and absorption curves are merely indicative. Adapted from *Atmospheric window* at Wikimedia Commons (public domain).

produce such electronic transitions, but they can make the molecular bonds between atoms vibrate, similar to a spring attaching two masses. Microwaves, which have less energy, cause rotations as molecules with an asymmetric distribution of their electric charge try to align with the travelling wave. Combinations of these mechanisms can also happen simultaneously, i.e. a molecule can rotate and vibrate at the same time.

Molecules have discrete energy levels. Combined with their unique structure and composition, molecules can be characterized by their absorption lines, which are akin to a molecular fingerprint. However, a given absorption process does not take place at a single, infinitesimal frequency. Instead, the absorption frequency experiences a phenomenon known as line broadening. While generally negligible for spaceborne remote sensing, the Heisenberg uncertainty principle already imposes a minimum line width. More relevant for atmospheric remote sensing are two other sources of line broadening.

Because of their thermal energy, molecules are moving. When the molecules move towards the sensor, the absorption appears at a higher frequency from the perspective of the sensor. Similarly, when the molecules move away from the sensor, the received frequency is lower. Molecules in the atmosphere will move in random directions, thus the sensor receives a range of slightly shifted frequencies, centred at the absorption line.

Another important line broadening mechanism is pressure broadening. In the atmosphere, molecules are constantly colliding with other molecules. These collisions shorten the lifetime of the involved energy states and, by the uncertainty principle, increase the uncertainty in the energy of the transition. This results in a widening of the absorption line, particularly in the dense lower

atmosphere.

The gases that constitute our atmosphere significantly attenuate radiation across a wide range of wavelengths. For instance, the upper levels of our atmosphere absorb the higher frequency—dangerous—incident radiation from space, the ozone layer significantly reduces the ultraviolet radiation, and water vapour absorbs most of the IR wavelengths. Ranges that are not blocked by the atmosphere are called atmospheric windows. For example, between $\lambda = 8\ \mu\text{m}$ and $15\ \mu\text{m}$, water vapour and other gases poorly absorb radiation, constituting an IR atmospheric window, as shown in fig. 2.1. The optical window is perhaps the most evident example, as our eyes take advantage of the high atmospheric transparency in this range. The MW region generally lacks gas attenuation and, thus, it plays an important role in remote sensing.

2.1.1.4 Particle attenuation

Scattering, the other attenuation process, deflects electromagnetic waves through their interaction with particles. The combined effects of absorption and scattering are also referred to as extinction in this context. To describe how much radiation is attenuated as it traverses a medium, the term optical thickness (also called optical depth) is used; optical thickness $\tau \geq 0$ is defined as the logarithm of the fraction of incident to transmitted radiant power.

The morphology of the particles and the particle size distribution, among other factors, determine the severity of scattering effects. Complex particle shapes, such as aggregates of ice crystals, lack a simple analytical description of the scattering efficiency. However, for spherical particles of radius r , the problem can sometimes be analysed in terms of the size parameter $x = 2\pi r/\lambda$. Therefore, x is instructive to characterize the scattering regime.

Particles with $x \ll 1$ are relatively inefficient at scattering radiation. Consequently, MW and IR scattering due to gaseous compounds is negligible. The situation is different for larger particles, such as aerosols or hydrometeors, where x approaches or exceeds unity, and scattering becomes a significant factor. This sensitivity to the size parameter is exploited by remote sensing instruments: MWs can penetrate thick clouds that are opaque to IR radiation, as the lower x values in the MW range result in significantly reduced scattering. Furthermore, the complex, asymmetric geometry of frozen hydrometeors implies that radiation is scattered differently depending on its polarization. This characteristic can be exploited to gain insights not only into the particle orientation, but also into its phase: spherical liquid droplets exhibit an isotropic signature, whereas the non-spherical geometry of mixed-phase and ice particles produces a distinctly anisotropic response.

2.1.1.5 Emission

To measure the atmosphere from space, we require a physical body to emit radiation. Both the Earth and the Sun emit electromagnetic radiation, though neither is a perfect black-body. However, the spectral radiance at the surface of the Earth approximately follows Planck's law, after scaling with the emissivity and using appropriate temperature values. As seen from space, the effective

Wavelength range	Region	Fraction of total irradiance	
		5772 K	255 K
10 nm to 400 nm	Ultraviolet	12.2 %	0.0000 %
400 nm to 700 nm	Visible	36.6 %	0.0000 %
700 nm to 1 mm	Infrared	51.2 %	99.9991 %
1 mm to 1 m	Microwave	0.0 %	0.0009 %
$\geq 3.9 \mu\text{m}$		1.1 %	99.9522 %
$\geq 6.2 \mu\text{m}$		0.3 %	98.0679 %

Table 2.1: Fractional irradiances for black-bodies at the effective temperatures of the Sun and the Earth, computed following section 2.1.1.2.

temperatures (the temperature a black-body would need to have to emit the same total amount of radiated power) of the Sun and the Earth are, respectively, 5772 K and 255 K. While the actual numbers are more complicated due to the presence of the atmosphere, table 2.1 serves to illustrate how each source can contribute to the radiation observed from a satellite observing the Earth: the Sun irradiates most strongly in the visible and the shorter IR wavelengths, while radiation from the Earth takes place in the IR, with a residual percentage in the MW region.

Remote sensing instruments measuring in the IR can be conditioned on the presence of sunlight. Particularly for shortwave IR radiation, e.g., $1.6 \mu\text{m}$, measured IR radiances originate from the reflection of solar radiances on the Earth. For this reason, this type of radiation can also be referred to as reflected IR, whereas thermal IR (TIR) describes the radiation that does not depend on the sunlight. As with the divisions of the electromagnetic spectrum, the exact boundaries between reflected and thermal IR radiation are not universal. The $3.9 \mu\text{m}$ wavelength, a channel measured by satellites used in this thesis, is at the transition between reflected and thermal IR, implying different information for day and night conditions. For this reason, this thesis does not include $3.9 \mu\text{m}$ or shorter wavelengths in the TIR and considers only wavelengths with negligible solar contamination as TIR, which for the data used implies wavelengths of at least $6.2 \mu\text{m}$.

2.1.2 Wavelength selection

The opacity of the atmosphere generally restricts the wavelengths suitable for remote sensing, though this limitation can also be turned into an advantage. For instance, emissions from the surface of the Earth are largely blocked at $6.2 \mu\text{m}$ (fig. 2.1), as this wavelength falls within a strong absorption band. Consequently, spaceborne sensors measuring at this wavelength primarily detect the thermal emission from higher altitudes, specifically from water vapour, thus enabling the monitoring of atmospheric moisture levels. Similarly, wavelengths that are transmitted through the gaseous atmosphere but are attenuated by hydrometeors are essential for retrieving cloud and precipitation properties.

2.1.2.1 IR versus MW

Both IR and MW radiation interact with hydrometeors, but they do so through different physical mechanisms and magnitudes. On the one hand, the liquid water and ice particles will absorb IR radiation that matches their molecular vibrational modes. On the other hand, because hydrometeors are about the same size or larger than IR wavelengths, scattering cannot be neglected, and its strength inevitably depends on the particle size and geometry. The combination of strong absorption and scattering makes hydrometeors highly opaque in the IR. Spaceborne instruments measuring in this spectral range therefore mainly receive radiation emitted from the cloud top.

The hydrometeor MW attenuation is significantly weaker than in the IR region. Concerning frozen hydrometeors, molecules are fixed in a lattice where they cannot freely rotate. As a result, ice crystals absorb little MW radiation (cf. section 2.1.1.3). In the case of liquid water, water molecules collide with each other as they try to rotate, absorbing some MW energy. Furthermore, the longer wavelengths of MWs render scattering highly inefficient for small hydrometeors, as the particles are small relative to such radiation wavelengths. As a consequence, the low absorption and inefficient scattering allows MWs to easily penetrate thick cloud layers. However, larger hydrometeors become more effective scatterers. By targeting specific frequencies, MW retrievals can thus provide vertical profiles of hydrometeor distributions.

The selection of MW frequencies is, therefore, a strategic compromise between the sensitivity to specific particles and the attenuation of the signal by particles of another kind. As an example, 94 GHz (3.19 mm) has been established as a frequency to quantify cloud ice from spaceborne radars, due to the sensitivity of ice particles to this frequency. The frequencies of 35.55 GHz (8.43 mm) and 13.6 GHz (22.04 mm), for their part, are better suited for precipitation retrievals. At these frequencies, a radar signal can better penetrate the overlying cloud ice to reach the larger hydrometeors associated with precipitation.

Surface emissivity also behaves differently between IR and MW. In the IR, $\varepsilon(\lambda)$ is close to 1 for most of the Earth's surface, though it decreases over desert areas (e.g., Zhou et al., 2011). This implies that the background IR signal is dominated by the physical temperature of the surface. In contrast, MW emissivity is more complicated. This complexity arises partly because the MW region spans a wider range of frequencies than the IR, over which emissivity varies significantly. In addition, MW emissivity is highly sensitive not only to the surface type but also to the polarization of the wave (e.g., Prigent et al., 2006). Over ocean, the low but uniform MW emissivity results in a radiometrically cold background as measured from space. Over land, however, the high variability of MW emissivity makes it difficult to separate the surface background signal from cloud signals, complicating MW retrievals. Addressing these challenges in the MW requires accurate modelling of the surface emissivity or the use of multiple spectral channels to disentangle cloud properties from the surface contribution. As evidenced in paper V, if this background emissivity is not handled correctly, retrievals that theoretically benefit from MWs can

actually underperform in practice.

To summarize, the traditional choice between IR and MW sensing is dictated by the desired depth of the observation subject to modelling constraints. Infrared sensors provide a precise characterization of the cloud top environment due to the strong IR attenuation, which enables the detection of tenuous clouds. Conversely, MWs inherently enable a better description of the vertical distribution of hydrometeors.

2.1.2.2 Angular resolution

The diffraction of electromagnetic waves limits the maximum spatial detail an instrument can resolve, known as the diffraction limit. Assuming a circular aperture, the angular resolution θ is defined by

$$\sin \theta \approx 1.22 \frac{\lambda}{D} \quad (2.2)$$

where D is the diameter of the aperture sensitive to the radiation, which can be approximated by the size of the antenna receiving the radiation. Because the angles θ are typically small in satellite remote sensing, eq. (2.2) is often expressed with the small-angle approximation, $\sin \theta \approx \theta$.

Equation (2.2) quickly reveals that larger wavelengths λ , such as MWs, will result in worse angular resolutions than IR for the same antenna size D . Moreover, the angular resolution is directly related to the ground footprint, that is, the spatial resolution on the Earth's surface resolved by the sensor, with

$$d \approx \theta H \quad (2.3)$$

where H is the altitude of the satellite. While MWs can be advantageous for vertically resolving the atmosphere, their poor resolution compared to IR limit their use. As an example, if we require a ground footprint of 3 km with a 94 GHz MW from a satellite orbiting at $H = 35\,786$ km (the altitude of geostationary satellites, see section 2.4.3), eq. (2.3) suggest an antenna size of 46 m—roughly the height of a 13-floor residential building. In contrast, we obtain an antenna size of only 16 cm for an IR wavelength of 11 μm , which is significantly more feasible to integrate into a satellite payload.

2.1.3 Passive and active remote sensing

The most natural way of measuring hydrometeors is to use a passive system. When we look at the sky and observe a cloud, our eyes act as passive instruments, detecting reflected visible radiation. In a similar manner, we can build passive sensors to measure the electromagnetic spectrum reflected by the Sun or emitted by the Earth's surface and atmosphere. Alternatively, we can employ active instruments, which generate their own signals and measure the portion of the energy returned after interacting with a target. This is the case of radars (radio detection and ranging) and lidars (light detection and ranging), which measure the backscattered radiation.

The power requirements for generating electromagnetic signals limit active instruments to a few specialized channels. In contrast, passive systems can easily incorporate a wide range of spectral bands. This energy efficiency, combined with broader scanning geometries, enable passive sensors to provide measurements over larger areas. However, since radars and lidars measure the time of flight—the duration for a signal to return to the sensor—active instruments provide a more direct and precise determination of the vertical altitude from which a signal originates. For this reason, active remote sensing is often considered the gold standard for vertical characterization of the atmosphere.

While passive remote sensing instruments can be used alone for estimating the state of the atmosphere at synoptic scales, retrievals designed to exploit synergies between active and passive instruments outperform those based on either alone. For example, the precipitation training labels used in paper V originate from a data product that combines radar and 13-channel MW radiometer data. In this specific case, the radar mitigates the ambiguity inherent in the passive signal by providing explicit vertical constraints necessary to reduce uncertainties in the retrieval. The radiometer, in turn, contributes its multiple frequencies and polarized channels to constrain the retrieval. Furthermore, the radiometer provides independent estimates of the total vertical attenuation and helps constrain the near-surface clutter zone, a region where surface backscatter makes radar measurements unreliable. This synergistic information enables disentangling whether the signal comes from suspended liquid or ice particles, collections of small droplets, or large raindrops: polarized signals are sensitive to particle shape, while the multi-frequency information captures size-dependent scattering characteristics.

2.1.4 A note on the definition of clouds and precipitation

Clouds are systematically identified and categorized, generally following a grouping based on their altitude and physical form as shown in fig. 2.2. For instance, cirrus clouds are at high altitudes and are essentially ice clouds, whereas cumulus are close to the surface and primarily composed of liquid particles.

Despite this established cloud nomenclature, there is no clear definition of what a cloud is. Spänkuch et al. (2022) have highlighted differences in the definition of clouds, including differences based on the different disciplines in atmospheric science, such as remote sensing or numerical modelling. For instance, the American Meteorological Society defines clouds as visible objects, while the World Meteorological Organization uses the broader term ‘perceivable’. The perceivable threshold is generally chosen to be a visible optical thickness of $\tau \geq 0.3$, a limit that excludes groups of subvisible, high-altitude ice masses as clouds. Furthermore, Koren et al. (2008) demonstrate that the transition between a cloud and a cloud-free atmosphere is not a sharp boundary; a significant fraction of cloud mass often falls below the spatial resolution of modern sensors. Thus, cloud properties such as cloud cover—the fraction of cloud contained in a grid cell of the retrieval—are inherently dependent on the resolution. In addition, the actual radiative signature of a cloud can fall below the sensitivity threshold of the measuring instrument or the retrieval

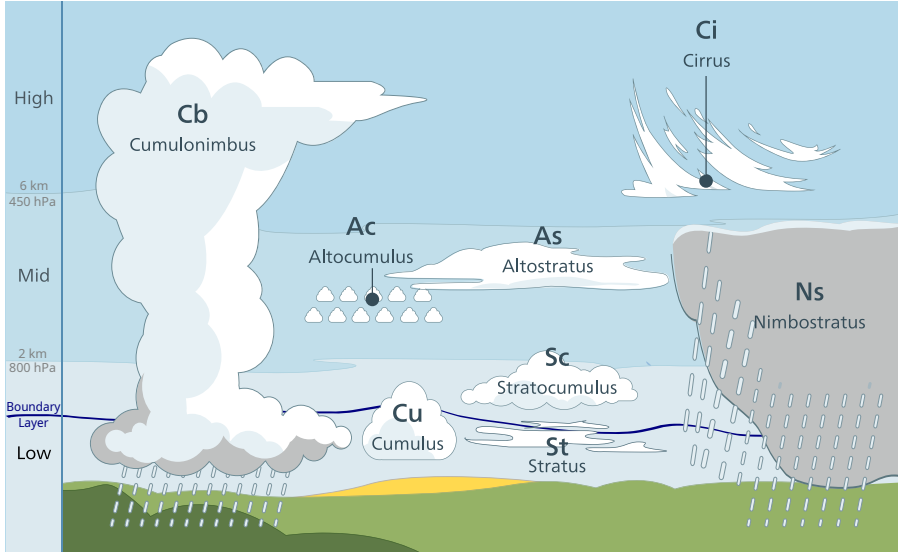


Figure 2.2: The cloud genera used in papers II and IV (therein, the term *deep convection* is used instead of *cumulonimbus*). The boundary layer is the lowermost atmospheric layer, which rapidly responds to surface fluxes of heat, moisture, or wind in timescales of an hour or less. Adapted from *Cloud types* at Wikimedia Commons (original by Valentin de Bruyn, CC BY-SA 3.0).

algorithm; differences in this threshold are evidenced in paper IV.

In any case, clouds and precipitation are not uniform blocks of hydrometeors but rather consist of distributions of such. Consequently, retrievals that model their radiative properties require making microphysical assumptions, such as the particle size distribution, shape, and phase (liquid, ice or mixed). This also holds for retrievals based on empirical relationships, such as the ones used in this thesis. However, in this case, the assumptions are not explicit but rather inherited from the training labels.

Beyond microphysical properties, the vertical distribution of hydrometeors is a critical factor in determining the observed radiative signal. For frozen water, this is characterized by the ice water content (IWC), which represents the mass of ice per unit volume at a specific atmospheric level, and is a quantity targeted in paper II. Integrating this quantity over the entire vertical column yields the ice water path (IWP), a quantity that is the primary target of paper I, but also of papers II and III. How the frozen hydrometeors are vertically distributed can also be summarized with column quantities, such as the mean mass height used in paper I.

As with clouds, the definition of IWP is not universal. Depending on the field of application, IWP will only account for the relatively small ice crystals inside clouds. In this way, frozen precipitation, such as snow or hail, is not included in IWP. Nevertheless, this thesis takes a pragmatic approach when retrieving the cloud properties, with IWP representing any frozen hydrometeor

throughout the vertical column in papers I to III. This also motivated using the term total IWP (TIWP) instead of IWP in papers II and III, although other authors also employ the term frozen water path. A pragmatic approach is also employed in paper V when referring to precipitation; the term is used as a synonym for surface precipitation.¹

2.2 The critical role of satellites

Localized rainfall can be directly quantified with rain gauges. However, such in-situ measurements suffer from limited spatial representativeness and geographical coverage. In fact, Kidd et al. (2017) estimated that less than half of a football field would be covered by combining all rain gauges from the Global Precipitation Climatology Project (GPCP; e.g., Gruber et al., 2007), which provides the basis for further precipitation datasets. The actual coverage of in-situ measurements is likely higher when accounting for rain gauges not included in international datasets, such as those maintained by commercial or private entities. Nonetheless, the characteristics of these measurements do not resolve the issue of global coverage, as they are unevenly distributed and only sparsely available over land. Moreover, maintaining a rain gauge network presents a variety of challenges, including consistent calibration across instruments, accounting for various systematic measurement biases, guaranteeing reliable reporting from the different sources, and achieving the low latency and high temporal resolution required to support decision-making. While rain gauges remain the most accurate and representative way to estimate precipitation at a local scale, more is needed to better describe precipitation at larger scales.

Consequently, for routinely covering regional scales we need to turn to remote sensing. Ground-based weather radars offer the best spatiotemporal resolution. However, a dense network is needed to cover large areas with sufficient precision. Similar to the rain gauges, ground-based weather radars are not evenly distributed over the globe, as illustrated in fig. 2.3. Moreover, ground-based radars can be obstructed in urban environments or by mountains, and lack the ocean coverage necessary to fully characterize the hydrological cycle.

Therefore, ground-based instruments alone are insufficient. Satellites are the only way to cover remote, mountainous, or oceanic regions inaccessible to ground-based networks. Although spaceborne data also requires calibration and distribution, even a single satellite can routinely observe at large scales with sufficient spatial and temporal resolution to address a multitude of questions.

2.3 Satellite orbits

Satellite orbits can be classified, among others, based on their altitude or function. Most scientific and meteorological satellites complete one orbit around the

¹The term refers to precipitation above the ground clutter of the reference data, not strictly at the surface. Because of the area covered in paper V, this precipitation is predominantly rain.

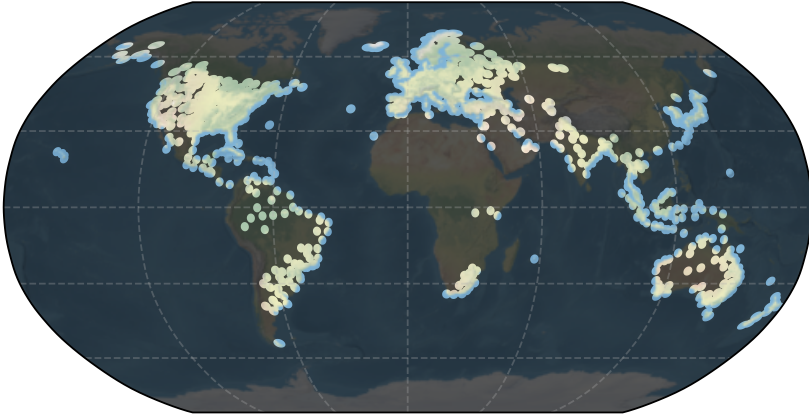


Figure 2.3: Coverage of ground-based operational weather radars in May 2025, assuming a range of 200 km for each radar. Data from the World Meteorological Organization radar database. Figure inspired by Saltikoff et al. (2019).

globe in ~ 100 min. This is possible by orbiting at an altitude of ~ 800 km. Such satellites are in low Earth orbit (LEO), which spans up to an altitude of 2000 km.

Moreover, most scientific and meteorological satellites are also placed in a sun-synchronous orbit. This orbit is characterized by an inclination angle beyond 90° , usually around $\sim 98^\circ$, enabling a global coverage; consequently, in this context, they are often referred to as (nearly) polar orbit satellites. While the sun-synchronous orbits used by such satellites have an orbital period of ~ 100 min, their revisit period—the time between two observations of the same point—is usually larger than that. The characteristic of this orbit is that the satellite maintains the same relationship with the Sun throughout one revolution, i.e. the satellite passes any point on the globe in the same direction at the same mean local solar time. Consequently, they are often defined by their equatorial crossing time in the ascent or descent of the orbit. The primary limitation of these types of orbits is that they do not allow for resolving the diurnal cycle. Nonetheless, the temporal consistency of these orbits allows for a direct comparison of seasonal data over multiple years without the interference of varying solar conditions, such as the solar zenith angle.

In contrast, the non-sun-synchronous orbit is also important. At an inclination of, for example, $\sim 65^\circ$, satellites on this orbit provide observations that drift in local solar time over several weeks. Given a sufficiently long record, non-sun-synchronous satellites sample the diurnal cycle of any point within their latitudinal range. Although such missions are rare, this unique characteristic is essential for understanding the temporal variability of the atmosphere measured from LEO platforms.

The geostationary orbit (GEO) is another type of orbit. At an inclination of 0° and an altitude of $\sim 35\,786$ km over the equator, geostationary satellites maintain a fixed position relative to the Earth, appearing static when viewed from the ground. In this case, the revisit period is determined by the scanning mechanism of the instrument rather than by orbital constraints. For

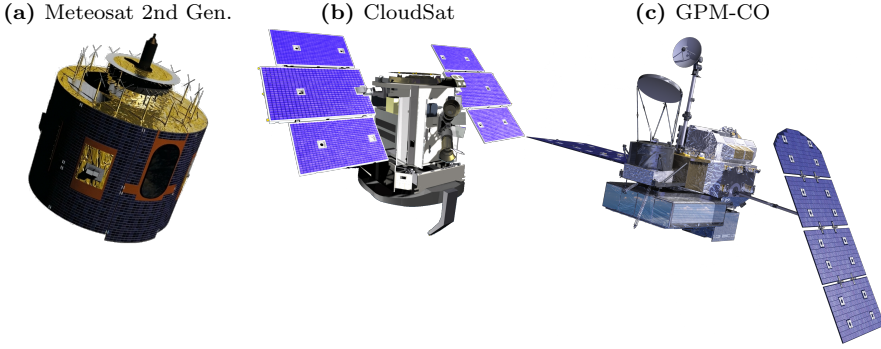


Figure 2.4: Artist’s renderings of satellites employed in this thesis (not to scale). Image credit: © EUMETSAT (a), NASA (b and c).

instance, the geostationary satellites used in papers I and V scan one side of the Earth every 15 min. Consequently, geostationary satellites allow continuously monitoring the same locations throughout the day.

A substantial limitation of the geostationary orbit is which region of the electromagnetic spectrum is feasible to measure. As shown in section 2.1.2.2, MW remote sensing necessitates significantly larger apertures to resolve the signal than those measuring in the IR to achieve the same ground resolution. Designs for synthetic aperture radiometers have been proposed to circumvent the aperture size required for geostationary MW observations (e.g., Lambrigtsen et al., 2022). However, no such instrument has yet been deployed, and MW measurements currently require LEO satellites.

All these orbits are complementary for several reasons. It is only LEO platforms that can have MW instruments, thanks to their lower altitude. Although constellations of sun-synchronous satellites can provide global coverage, they are temporally limited and have a constrained field of view, i.e. they can see only a small portion of the surface at any given time. Therefore, even such constellations face long revisit times, which in turn are a function of latitude, with a minimum at the poles where the orbits converge. The continuous monitoring from geostationary satellites can fill the gaps between LEO observations, depending on the application. In addition, non-sun-synchronous orbits enable collocating the paths of sun-synchronous satellites for, e.g., calibration. Collectively, a constellation of geostationary satellites can cover the Earth in longitude, a concept known as the ‘geo-ring’. However, because geostationary satellites cannot observe the highest latitudes, this configuration still requires polar orbiting satellites for global coverage.

2.4 Satellite data for this thesis

All appended papers employ geostationary radiances as the input to the neural networks. For their training, however, high-quality data derived from specialized active sensors serve as the target labels. Table 2.2 lists all satellites used in

Satellite	Orbit	Time span ^a	Instruments	In paper
<i>Dataset</i>				
Meteosat 2nd Gen. <i>Level 1.5 image data</i>	GEO, at 0° (81°W–81°E)	2002–Present	VIS+IR imager	I, V
Geo-ring <i>GridSat, CPCIR</i>	GEO (180°W–180°E)	1980–Present ^b	VIS+IR imagers	II–IV
CloudSat CALIPSO <i>DARDAR-CLOUD, 2C-ICE, 2B-CLDCLASS[-LIDAR]</i>	Sun-sync.	2006–2023 ^c	Radar (94 GHz) Lidar	I–IV
GPM-CO <i>GPM 2BCMB, IMERG^d</i>	Non-sun-sync.	2014–Present	Radar and MW radiometer	V

^a The range indicates only the period observed by the satellite, not necessarily the interval used in the papers. ^b There is no global coverage in 1980, see Knapp et al. (2011).

^c CloudSat operations were interrupted in April 2011 after the battery anomaly. The radar remained offline until October 2011, when daylight-only operations began.

^d The IMERG dataset merges observations from the GPM constellation, not only from GPM-CO. Additionally, it does not leverage the DPR from GPM-CO.

Table 2.2: Main satellites and observational datasets employed in this thesis. Three of the platforms are depicted in fig. 2.4.

this thesis together with their principal characteristics and the associated data products, whereas the subsequent sections provide a more detailed description of the scientific context and technical specifications of each satellite.

Using a scenario of clouds and precipitation as an example, fig. 2.5 illustrates both the capabilities and inherent challenges of the various data products employed in this work. As elaborated in section 2.1, IR signals are primarily sensitive to the cloud top, with lidar retrievals providing a precise detection of these upper boundaries. In contrast, radar signals penetrate thick clouds, interacting differently with the internal cloud structure based on their wavelength. Furthermore, the figure highlights the significant influence of the surface on the MW measurements, as opposed to the IR signal.

2.4.1 The CloudSat-CALIPSO tandem: profiling of clouds

CloudSat (Stephens et al., 2002) and CALIPSO (Winker et al., 2003), both now decommissioned, were launched together in 2006. They joined the afternoon constellation (Schoeberl, 2002), often abbreviated to “A-Train”, a series of sun-synchronous satellites with a nominal ascending crossing equatorial time of 13:30. The A-Train offered an unprecedented opportunity to make atmospheric cloud observations on a global level. CloudSat was equipped with the Cloud Profiling Radar (CPR), a nadir-looking radar with a tilt of 0.16° (Tanelli et al.,

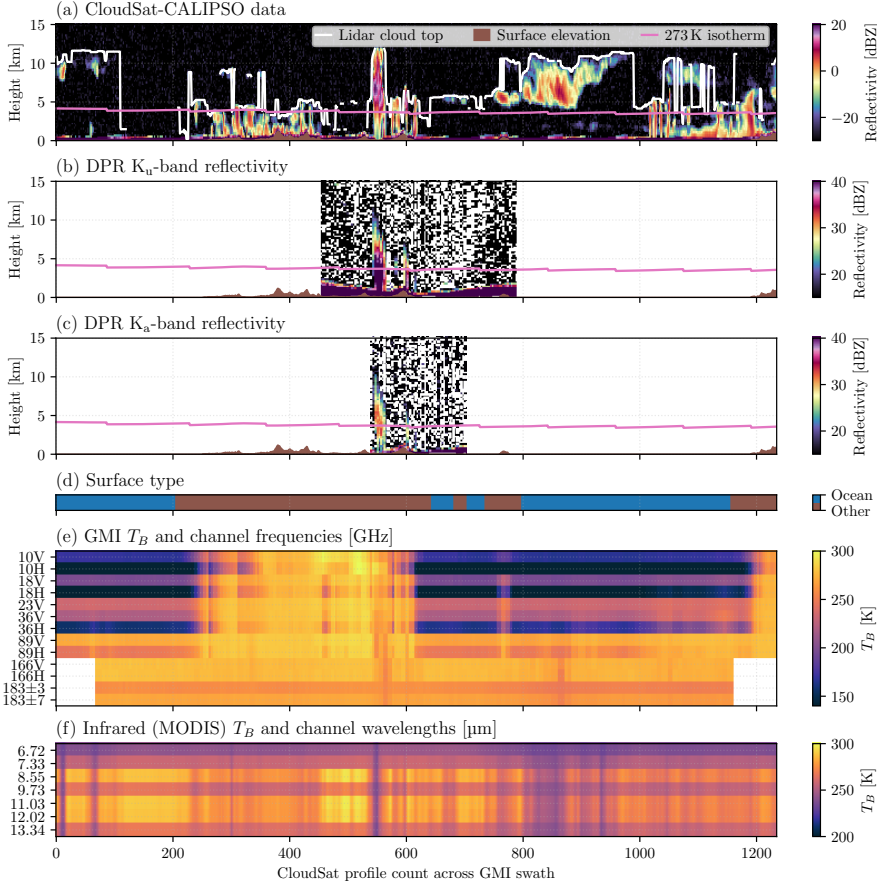


Figure 2.5: Collocated observations of a cloud scene from the 2B-CSATGPM CloudSat-GPM coincidence dataset (Turk et al., 2021). White indicates missing or invalid data. MODIS, also flying in the A-Train, is used here as a proxy for SEVIRI, as the scene falls outside its field of view. The MW channels show a clear dependency on surface type, whereas IR channels are largely insensitive to it. The scene was acquired on 2014-06-09 around 04:00 UTC over Japan.

2008). CALIPSO, in turn, carried the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), a nadir-looking lidar tilted at 3° (initially 0.3° , Hunt et al., 2009). These slight tilts improved the quality of the measurements by preventing the sensors from being saturated by specular reflection, i.e. mirror-like reflections from the surface or flat ice crystals.

The CPR provided cloud and precipitation information at a 500 m vertical resolution. Its frequency of 94 GHz represents an optimum compromise between cloud reflectivity and atmospheric attenuation, enabling the CPR to measure the vertical structure of clouds. With a sensitivity of almost -30 dBZ,² the

²dBZ is the radar reflectivity factor Z [$\text{mm}^6 \text{m}^{-3}$] on a relative logarithmic scale.

radar can profile most of the clouds affecting the radiation budget. Shallow boundary layer clouds (fig. 2.2) and cirrus clouds are two of the cloud types that CloudSat’s instrument struggled to detect for different reasons. While the small water particles from low-level clouds can return signals weaker than the CPR’s sensitivity, the low density of thin cirrus layers can give a lack of backscattered energy to the CPR. This underscores the need for additional measurements available from its companion satellite.

CALIPSO was designed to fly in close proximity to CloudSat. With CALIOP, it compensated for CloudSat’s lack of sensitivity to thin cirrus clouds. The 532 nm and 1064 nm wavelengths of CALIOP are more than 3000 times smaller than the size of a 94 GHz CPR wave. These shorter wavelengths are more adequate for detecting low concentrations found in tenuous cirrus clouds or the small particles from the water clouds at low altitudes, as well as for distinguishing between large and small particles. However, the lidar signal is heavily attenuated in dense clouds compared to the radar signal.

The CALIOP lidar also accurately determined the cloud phase, via the ratio of its two orthogonal channels; the non-sphericity of frozen hydrometeors alters the polarization. CALIOP additionally provided finer horizontal and vertical resolutions than the CPR. While data above 8.2 km was averaged before downlink to reduce the bandwidth, CALIOP achieved a vertical resolution down to 30 m at the surface. Compared to the 500 m resolution of CPR, the lidar provided greater spatial discrimination near the ground, enabling the detection of cloud particles that would be indistinguishable from the Earth’s surface from the radar echo.

Consequently, the specialized active instrumentation of CloudSat and CALIPSO established a synergy in the sense that vertical cloud properties missed by CPR, particularly the shallow lower boundary clouds and the subvisible cirrus, were captured by CALIOP. Thus, their combined retrievals provide a more complete description of the atmosphere. Readily available synergistic products include 2C-ICE (Deng et al., 2010, 2013, 2015) and 2B-CLDCLASS-LIDAR (Sassen et al., 2008) from the CloudSat Data Processing Center (Cooperative Institute for Research in the Atmosphere, Colorado State University, USA), or DARDAR-CLOUD (Delanoë and Hogan, 2010; Ceccaldi et al., 2013; Cazenave et al., 2019) from the AERIS/ICARE Data and Services Center (University of Lille, France), employing different microphysical assumptions.

Synergistic retrievals from CloudSat and CALIPSO have been the gold standard for global cloud profiling, have updated the interpretation of radiance observations, and have provided new insights and understanding of cloud processes (Stephens et al., 2018). However, these retrievals have three characteristics that limit the study of cloud distributions on a global scale: their spatial sampling, their sun-synchronous orbit, and their temporal coverage.

Being nadir-only profiling instruments, CPR and CALIOP traced a single, narrow line when projected on the ground. The cross-track effective field of view of CloudSat was approximately 1.4 km, which was also its cross-track coverage. CALIOP offered an even finer horizontal resolution. The narrow coverage of the synergistic retrievals is akin to taking a slice of the atmosphere along the orbit of the satellites. Hence, it is difficult to completely assess any

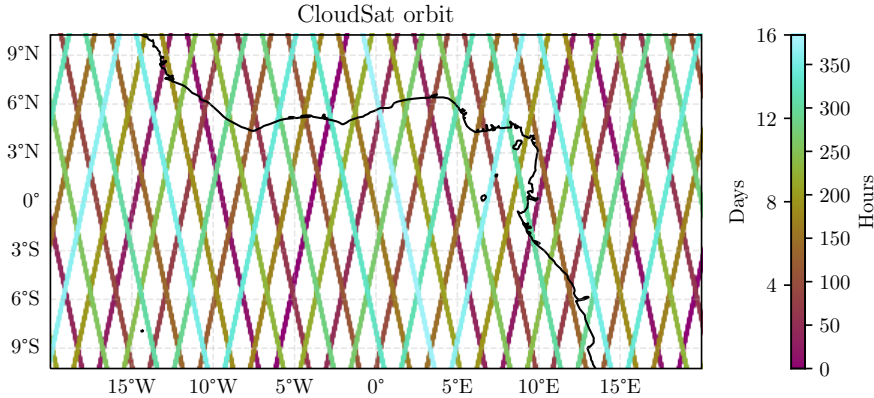


Figure 2.6: One full CloudSat repeat cycle for a region centred at 0° latitude and longitude. The cross-track field of view is not to scale.

spatial variability with the backscattered signals alone.

Their sun-synchronous orbit implies three additional limitations. Firstly, observations are discontinuous for any location on Earth, and therefore unable to track clouds in time; measurements are rather snapshots. Secondly, their repeat cycle is 16 days. These two limitations, jointly with the narrow field of view, cause a third limitation, which is its spatial coverage and revisit time. Figure 2.6 illustrates the issue: a considerable portion of the atmosphere is unobserved by the CloudSat-CALIPSO tandem; there is an equatorial track gap of ~ 170 km. For most locations observed, the revisit time is 16 days. However, for fortunate locations it is 8 days, consisting of a day and a night observation, always at the same local mean solar time.

The CloudSat and CALIPSO missions ended in 2023, exceeding by far their design requirement of two years of lifespan. However, their temporal coverage is interrupted due to a series of unfortunate events. The most limiting is that CloudSat suffered a battery anomaly in April 2011, resorting to only acquiring data during the sunlit portions of its orbit (Nayak, 2012). CloudSat also suffered an anomaly in one of the mechanisms that controlled its stability in the A-Train, threatening a possible collision with other satellites in the constellation, and was lowered 16 km in February 2018, followed by CALIPSO months after, establishing the C-Train constellation in October 2018 (Braun et al., 2019). Accordingly, there is only complete data until the first quarter of 2011, conditioned on operational downtimes.

2.4.2 GPM Core Observatory: the precipitation reference

The Global Precipitation Measurement (GPM) mission is an international satellite mission with the frequent sampling of precipitation as a primary objective (Hou et al., 2014). To achieve this goal, a constellation of LEO satellites delivering MW measurements is employed, with the GPM Core Observatory (GPM-CO) as its central reference. Launched in early 2014, GPM-CO serves

two purposes: calibrating the satellites in the GPM constellation and providing high-accuracy precipitation retrievals. Its 65° non-sun-synchronous orbit enables GPM-CO to frequently intersect the paths of other sun-synchronous satellites within the GPM constellation, acting as the satellite of opportunity for cross-calibration.

The GPM-CO payload features two instruments designed to measure precipitation: the multifrequency GPM Microwave Imager (GMI; Draper et al., 2015), and the Dual-frequency Precipitation Radar (DPR; Kojima et al., 2012), measuring at 13.6 GHz (K_u band) and 35.55 GHz (K_a band). The 13 channels of the radiometer enable measurements over a wide range of frequencies, capturing light to heavy precipitation, while the radar provides three-dimensional precipitation characteristics that allow for more accurate discrimination between rain and snow during the retrieval process. Compared to DPR-only or GMI-only methods, the combined GPM retrieval (Grecu et al., 2016) significantly reduces uncertainties (cf. section 2.1.3). In terms of spatial coverage, GMI covers a 885 km swath, whereas DPR measures over a narrower overlapping area of more than 245 km.³

2.4.3 Geostationary satellites: the input

Beyond its predominant use for telecommunication platforms, the geostationary belt hosts a series of meteorological satellites. These accommodate instruments that measure in the visible and IR regions of the spectrum, motivated by the aperture size, as MW radiometers necessitate significantly larger antennas (cf. section 2.1.2.2). Additionally, geostationary satellites face geometric constraints: as shown in fig. 2.7, distortion increases with latitude as larger viewing angles enlarge pixel footprints, leaving polar regions effectively unobservable.

2.4.3.1 Meteosat Second Generation

Meteosat is the name of the series of geostationary satellites from EUMETSAT. The first Meteosat was launched in 1977 (WMO, 2024), positioned at 0° , therefore covering Europe and Africa. In 2002, the debut of Meteosat-8 inaugurated the Meteosat Second Generation (MSG) series, improving the spatiotemporal resolution and expanding the number of channels. While the MSG series remains operational, the first satellite of the third generation was launched in 2022. However, this new platform only became operational at the end of 2024 after a complicated calibration process (Mousivand et al., 2025) and thus currently offers a limited data record.

All four MSG satellites carry replicas of the Spinning Enhanced Visible and InfraRed Imager (SEVIRI; Aminou et al., 1997), the specifications of which are summarized in table 2.3. SEVIRI operates by leveraging the east-to-west 100 rpm spin of the satellite while using a mirror to scan from south to north, thereby composing a full Earth disc image. Consequently, the imagery

³Originally, the K_a -band offered the narrowest swath. In May 2018, the DPR scanning pattern was reconfigured to expand the K_a -band coverage to match the K_u -band swath (Iguchi et al., 2024; Olson and the GPM Combined Radar-Radiometer Algorithm Team, 2022), while a boost in orbit altitude in 2023 increased the swath width (Kubota et al., 2024).

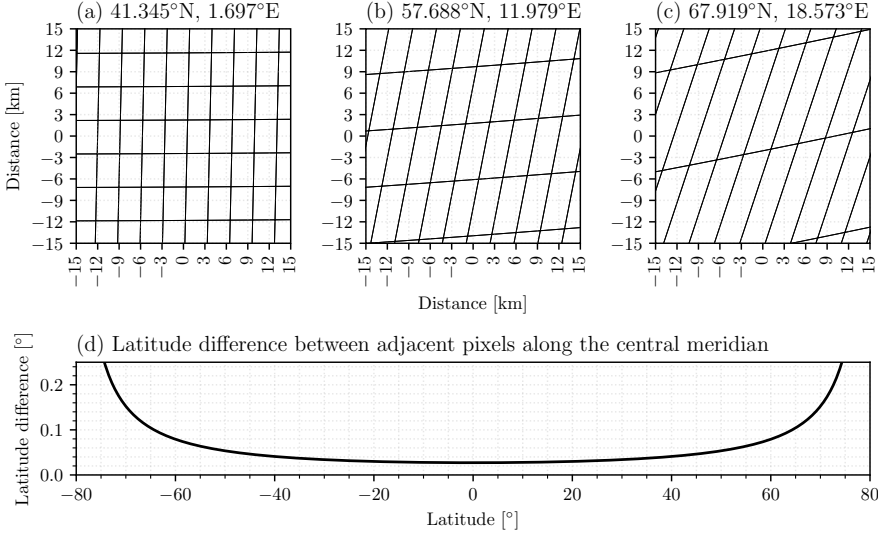


Figure 2.7: In the top row, pixel projections for MSG imagery for three different locations, using a transverse Mercator projection. The bottom plot shows the latitudinal variation in the native MSG projection at its centre, where projection distortion is minimized.

exhibits a temporal gradient, which may be critical for real-time applications. EUMETSAT publicly provides SEVIRI level 1.5 data (EUMETSAT, 2009, 2017) with minimal latency, making it suitable for such applications.

2.4.3.2 Geo-ring

Combining different geostationary satellites in a constellation offers the possibility to have quasi-global continuous observations in the IR region, only limited in latitude by the field of view of these satellites. The imagers on-board each satellite can differ in spatial resolution, revisit time, or measured wavelengths, among others. On the one hand, and as an example, the imager on-board Meteosat-9, source of the input data in paper I, located near the prime meridian during the time of the analysis (EUMETSAT, 2026), provided a 3 km resolution with a full-disc revisit time of 15 min for 11 spectral channels. On the other hand, the imager on-board the coexisting operational satellite GOES-11 (Menzel and Purdom, 1994), located at 135°W, offered only five spectral channels, with resolutions between 1 and 8 km and a worse full-disc revisit time in the routine mode.

The characteristics of each instrument will depend on the agencies responsible for the satellite and, inevitably, the technological achievements during their design phase. It should come as no surprise that data products tailored for one family of imagers cannot be applied directly to others. Nevertheless, observations from a geo-ring can be merged onto a predefined grid through selecting shared spectral bands—such as the 11 μm channel in the IR window—and

Dataset	Responsible agency	Channels [μm] (not used in this thesis)	Spatiotemporal resolution
SEVIRI level 1.5	EUMETSAT	0.635, 0.81, 1.64, 3.92, 6.25, 7.35, 8.70, 9.66, 10.80, 12.00, 13.40	3 km, 15 min
GridSat	NOAA (NCEI)	Since 1980: ~ 0.6 , ~ 11 Since 1998: ~ 6.7	8 km, 3 h
CPCIR	NOAA (NCEP/CPC)	Since 1998: ~ 11	4 km, 30 min

Table 2.3: Datasets of geostationary imagery used in this thesis. The spatial resolution is at the sub-satellite point. Although CPCIR originates from NOAA NCEP/CPC, it is curated and distributed by NASA GES DISC.

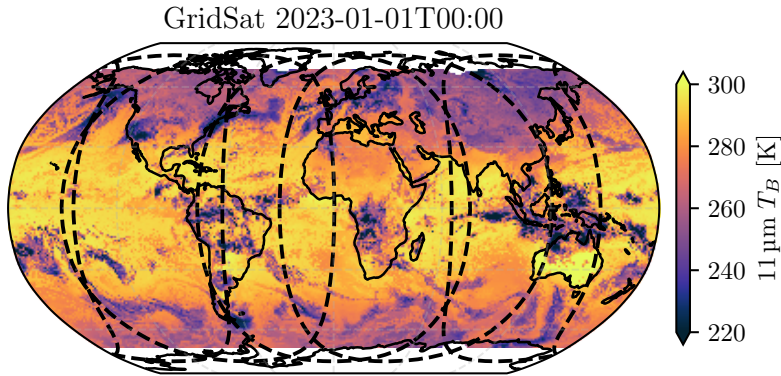


Figure 2.8: An $11 \mu\text{m}$ GridSat observation, combining GOES-17 (-137.2°), GOES-18 (-137.0°), GOES-16 (-75.2°), Meteosat-11 (0.0°), Meteosat-9 (45.5°), and Himawari-9 (140.7°) data. Dashed lines indicate the field of view of each satellite.

applying, for instance, viewing and parallax corrections.

Figure 2.8 illustrates a geo-ring observation from GridSat-B1 product version 2 (GridSat; Knapp et al., 2011; Knapp and NOAA CDR Program, 2014), a globally gridded product that merges IR observations from the geo-ring. The other geo-ring dataset used in this thesis is CPCIR—a shorthand used here for the NOAA Climate Prediction Center globally merged IR product (Janowiak et al., 2001; Janowiak et al., 2017). Such geo-ring datasets allow for developing data products with quasi-global coverage. In addition to the differences summarized in table 2.3, GridSat is designated a climate data record and employs a temporal calibration of the $11 \mu\text{m}$ channel for long-term stability. In contrast, CPCIR targets operational applications. Therefore, CPCIR and GridSat are complementary, offering distinct advantages.

3 Machine learning for retrievals of hydrometeors

As a form of machine learning, neural networks are central to the papers included in this thesis. This chapter motivates the use of deep learning for the retrieval problem. Afterwards, it outlines the core principles underlying the included work, rather than exhaustively justifying specific choices.

To set the context for this chapter, a neural network consists of layers of interconnected units called neurons. A deep neural network may be formally defined by the presence of more than one hidden layer, where each neuron k computes

$$h_k = g \left(\sum_j w_{kj} x_j + b_k \right) \quad (3.1)$$

with w_{kj} and b_k as trainable parameters, x_j the input, g an activation function, and h_k its output. Notice that eq. (3.1) essentially exploits differences among the data, learnt during training. Moreover, the argument of g can be expressed as a matrix operation, which modern hardware executes efficiently. Further familiarity with deep learning concepts is assumed throughout this chapter, as covered in standard texts such as Goodfellow et al. (2016) and Bishop and Bishop (2024).

3.1 Reshaping the retrieval landscape with machine learning

3.1.1 The retrieval problem

In the context of remote sensing, recall that retrieval refers to the acquisition of information about a geophysical parameter through indirect measurements. Mathematically, the measurements are functionally related to the geophysical parameter of interest, referred to as the state vector. Following ML convention, measurements are denoted as $\mathbf{x}$ and the state vector as $\mathbf{y}$.¹ A retrieval thus consists of inverting the forward function relating $\mathbf{y}$ to $\mathbf{x}$ or, equivalently, finding a function f such that $\mathbf{y} = f(\mathbf{x})$.

¹This is the opposite of the standard notation used in classical remote sensing literature.

The forward function F_{true} perfectly describes the physical relationship from the state vector to the measurements as

$$\mathbf{x} = F_{\text{true}}(\mathbf{y}, \mathbf{b}) \quad (3.2)$$

where $\mathbf{b}$ represents all other influencing parameters, such as atmospheric conditions, surface properties, or characteristics of the measurement instrument. In practice, however, F_{true} is either unknown or too complex to be represented exactly. Therefore, the classical approach to solving retrieval problems starts with a model for the true forward function, denoted as F . The relationship becomes

$$\mathbf{x} = F(\mathbf{y}, \mathbf{b}) + \varepsilon \quad (3.3)$$

where F , the forward model, represents the best approximation of the true forward function and ε accounts for both measurement errors and forward model errors, i.e. the discrepancy between F and F_{true} . Unsurprisingly, designing a forward model that accurately matches the forward function can be difficult: the real physics may be too complex, or the level of detail required can still be uncertain. In addition, a forward model needs to be numerically efficient, particularly when large amounts of data are processed.

Provided a forward model, one can design a retrieval method by mathematically inverting it. However, it is likely that this inversion returns a non-unique solution, yielding an ill-posed problem. Variational methods can solve the inversion problem by finding the minimizer of a cost function based on the forward model, but in their naive formulation they lack important information: the uncertainty of the retrieval. It is of little value to retrieve a quantity without a corresponding measure of error, particularly for sophisticated applications. Furthermore, external parameters or measurement noise can influence the retrieval and are a source of uncertainty.

3.1.2 Traditional physics-based approaches

Retrievals of cloud properties and precipitation have traditionally been grounded in physical principles complemented by empirical relationships, herein called physics-based approaches. For IWP retrievals from visible and IR data, the classical approach employs two spectral channels exploiting the fact that cloud reflectances in the visible region (e.g., at $0.6\ \mu\text{m}$) are strongly related to cloud optical thickness τ with little dependence on particle size, while IR reflectances (such as $1.6\ \mu\text{m}$ or $3.9\ \mu\text{m}$) are primarily sensitive to the effective particle radius r_e (Nakajima and King, 1990). Radiative transfer simulations are performed for a range of τ and r_e values to build lookup tables that map observed reflectances to τ and r_e . IWP is then taken to be proportional to τr_e (Stephens, 1978). Consequently, IWP retrievals from this bispectral method are limited to daytime. Paper I provides an overview of products employing this approach; more recent datasets continue to do so (e.g., Benas et al., 2023).

This method for retrieving IWP is, perhaps, the most established physics-based approach to determine one of the target variables covered in this thesis. There is no equivalent standard for providing vertically-resolved retrievals

of hydrometeors from visible and IR radiances. Concerning the detection of clouds, thresholds on the brightness temperatures have been used, as shown in paper II. Thresholds also on the cloud optical thickness and on estimates of the cloud top height have traditionally classified a cloud into one of the nine types defined by the International Satellite Cloud Climatology Project (ISCCP; Rossow and Schiffer, 1999). For precipitation, relationships between r_e and T_B (Rosenfeld and Lensky, 1998) or thresholds on T_B (e.g., Arkin and Meisner, 1987) have conventionally determined the presence of rain. Inevitably, such physics-based approaches are laden with assumptions.

3.1.2.1 Accounting for retrieval uncertainty

One solution to address the retrieval uncertainty in physics-based approaches is to consider the retrieved parameter as a random variable instead of a point estimate, and the measurement as a realization of another random variable. The Bayesian framework is well suited for this task, where the idea is that the measurement refines previous knowledge or assumptions of the parameter, yielding a posterior knowledge distribution for the parameter. The optimal estimation method (OEM; Rodgers, 2000) is widely used for solving atmospheric retrieval problems using the Bayesian framework. One characteristic of OEM is that it is an iterative method, since it repeatedly evaluates the Jacobian of the forward model. Therefore, one has to define the convergence criteria, which should include the possibility that there is no convergence. Furthermore, OEM is inherently limited to Gaussian statistics. In addition, this method requires initial conditions, the formal specification of prior knowledge, and that the linearization of the forward model is reasonable, which can all substantially impact the retrieval. Although OEM has a solid statistical foundation, its reliance on a forward model, iterations that can be expensive, and predefined convergence criteria can render the retrieval sub-optimal. More general methods such as Markov chain Monte Carlo are computationally heavy.

3.1.3 The machine learning paradigm

The proliferation of observational data has fundamentally reshaped the atmospheric retrieval landscape. In contrast to physics-based methods explicitly formulating the physical relationship between measurements and the state vector, modern approaches leverage empirical models derived from large datasets. By employing specialized remote sensing products as target labels, ML can implicitly approximate the inverse mapping of the forward function, as done in all appended papers detailing training results (papers I, II, IV, and V). When observational data is not available, synthetic datasets generated via radiative transfer models can serve as training data (e.g., May et al., 2024).

The ML paradigm reformulates the retrieval as a statistical optimization problem, employing tools that range from classical regression techniques to deep artificial neural networks, each offering varying degrees of model complexity and representational capacity. Because these models learn patterns directly from data rather than encoding explicit physical relationships, they offer

substantial flexibility in integrating heterogeneous data sources and capturing complex non-linear dependencies. Moreover, this approach circumvents the need for a complete analytical formulation, bypassing constraints imposed by an incomplete physical understanding or the inability to mathematically express complex phenomena. Paper II serves as a testament to this capability, in which a DL model produces vertically-resolved estimates of frozen hydrometeor masses and cloud probabilities from 11 μm geostationary radiances alone, representing a novel contribution to the field.

A critical limitation of such an empirical approach is that the resulting statistical relationships inevitably inherit both systematic biases and physical assumptions present in the training labels, as evidenced in paper III. Furthermore, ML models often lack transparency in their internal logic. While the field of explainable artificial intelligence is dedicated to providing insights into the decision-making process of these models (e.g., Molnar, 2025; Samek et al., 2019), they are frequently employed in atmospheric remote sensing as a black-box. This thesis adopts the same approach, focusing on model performance rather than the interpretability of the underlying structure.

From a computational perspective, ML methods effectively transfer the processing burden from the traditional retrieval scheme to an offline training phase. While this enables rapid inference once models are trained and facilitates the generation of near-real-time products with minimal computational footprint, as in paper V, it introduces a strong dependency on the representativeness of the training set. Machine learning models are inherently constrained by the training domain, with an expected degradation in performance when applied to conditions not adequately represented during training.

Probabilistic ML represents a principled framework to address the uncertainties inherent in the retrieval problem. Rather than producing a deterministic point estimate, such an approach can characterize the uncertainties associated with the state vector. This methodology combines the computational efficiency of the ML framework with the uncertainty quantification traditionally associated with, for instance, Bayesian schemes. By retrieving, e.g., a conditional probability distribution instead of a single value, the model can explicitly account for various sources of uncertainty and provide a quantifiable measure of credibility. A brief overview of probabilistic ML methods beyond those used in this thesis is given in section 3.4.2.

3.2 Foundations of effective deep learning

Although machine learning includes a broad range of methods, this thesis focuses exclusively on deep learning. Consequently, this section limits its scope to the foundational principles that are relevant for the appended papers.

3.2.1 Model capacity

The objective of DL is to learn the underlying structure of the training data such that the model can both fit known samples accurately and generalize to

unseen data. Its efficacy is governed, among other factors, by two architectural properties: expressivity, which defines the complexity of functions the model can represent, and capacity, which refers to the model's ability to store and represent information from a given dataset. When capacity is insufficient relative to the complexity of the data, the model experiences underfitting, failing to capture the underlying patterns adequately. In essence, expressivity is a product of the network's structural design, while capacity is primarily determined by the total number of trainable parameters.

Conventionally, an excessive capacity of the model has been linked to overfitting the training data. In this state, the model becomes overly sensitive to unique features and noise specific to the training dataset, rather than learning the underlying patterns. This high variance—the sensitivity to training-specific patterns—prevents the model from accurately generalizing on unseen data. Modern training techniques help mitigate this issue. Interestingly, research on overparameterized models suggests that extremely large networks can actually use their massive capacity to find patterns that improve, rather than hurt, generalization (e.g., Belkin et al., 2019). In the appended papers, however, model capacity was adjusted pragmatically based on available computational resources and dataset characteristics, rather than through, e.g., systematic hyperparameter optimization or scaling laws. Early stopping was employed in papers I and V as a precautionary measure against overfitting. In papers II and IV this issue was not observed, though likely because the computationally expensive training limited experimentation beyond the main training run.

3.2.2 Dataset design and quality

The quality of a DL model is fundamentally constrained by the quality of its training data. A well-constructed dataset must satisfy several critical properties to enable effective learning. First, it must be accurate: labels should be factually correct and consistent, as errors in the ground truth directly propagate to the learned model. Second, the dataset must be diverse, covering the full range of scenarios that the model is expected to encounter. A model trained on narrow or biased data will fail to generalize beyond those specific conditions. This issue can be alleviated with data augmentation techniques. Finally, the dataset should present sufficiently complex patterns that require learning non-trivial features; for simple datasets, classical methods often suffice, while DL carries a high risk of overfitting.

In practice, assembling such training datasets in atmospheric remote sensing is a resource-intensive aspect of a ML project, even when the source data is publicly accessible. The sheer volume of satellite data, combined with the need to collocate observations often from different instruments, creates substantial computational demands. Each data source can employ distinct projections, spatial grids, and temporal resolutions, requiring careful harmonization through regridding and resampling. Moreover, simplifications may be needed to make the problem tractable. For instance, in papers II and IV, high-resolution vertical profiles were reduced to a small number of levels. However, such simplifications can have clear consequences, as evidenced in paper IV: the vertical coarsening

of cloud masks omitted thin cloud layers from the 3D training labels, leading to inconsistencies between 2D and 3D retrieval outputs. The preprocessing pipeline required to transform raw satellite observations can be as demanding and impactful as the subsequent model development.

For probabilistic models, data quality has additional implications: reliable uncertainty quantification requires that the training data accurately reflects the true variability of the underlying distribution. Furthermore, insufficient data volumes can result in poor uncertainty estimates, particularly in sparsely sampled regions, as probabilistic models must learn the range of plausible outcomes across different scenarios. Throughout this thesis, the aim was to use the large data volumes that remained computationally manageable: where CloudSat data serves as training labels (papers I, II, and IV), most of the record with complete diurnal sampling was used for training and evaluation.²

However, even with large data volumes, the reference datasets used as training labels carry their own intrinsic uncertainties, which limit both the achievable accuracy and the reliability of the uncertainty estimates of the retrievals. In principle, these label uncertainties could be incorporated during training by, e.g., sampling from their distributions. In practice, however, the available error estimates may not reliably represent the true error structure. Consequently, this thesis did not explore such approaches, leaving their practical integration and evaluation as open questions.

3.2.3 Experimental rigour

Effective DL requires rigorous experimental practices to ensure that models learn generalizable patterns instead of training-specific artefacts. A critical concern is data leakage, where information from the evaluation data inadvertently influences model training. This can occur through improper splitting of training, validation, and test sets, leading to models that appear to perform well during development but fail on truly unseen data.

Data splitting strategies depend on the underlying data structure and the intended application, typically categorized into random and systematic approaches. Random splitting is suitable for independent and identically distributed samples. In satellite imagery, however, spatiotemporal autocorrelation can cause random splits to produce data leakage, artificially inflating performance metrics without reflecting the true generalization capability. Despite this limitation, a random split was used in paper I. Although this likely resulted in more optimistic performance values for test data, it provided a consistent internal baseline for comparing the different models. To more rigorously assess generalization, papers II, IV, and V employed systematic splits by assigning distinct time periods to the training, validation, and test sets. This systematic approach was essential for these papers, as each evaluated a single model where absolute performance accuracy mattered more than relative comparisons.

²Paper I is restricted to data from 6 May 2008 onwards to ensure homogeneity in the collocated input data; prior to this date, Meteosat data followed a different definition of radiance (EUMETSAT, 2012). Paper II and IV omit the 2011 record because of its limited volume, a decision that also serves to simplify the methodology.

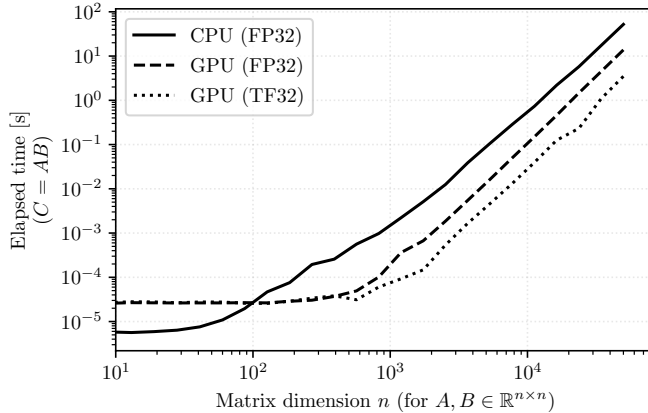


Figure 3.1: Elapsed times for matrix multiplication on a CPU and a GPU, using the same PyTorch code. Results averaged over 100 runs, with 5 warmups, and in an isolated computing environment. At small dimensions ($n \leq 100$), overhead causes the CPU to outperform this specific GPU. However, DL workloads typically operate at larger scales. FP32 and TF32 indicate single-precision floating point operations, but the latter employs the cores dedicated to matrix operations. GPU: NVIDIA A100 40 GiB; CPU: Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz, 64 cores. Data created directly on the CPU or GPU.

Finally, rigorous experimental design requires integrating domain knowledge to account for physical and measurement constraints. Without this context, performance metrics may obscure the true strengths or limitations of a model. For instance, paper V differentiates between land and ocean surfaces when comparing IR and MW precipitation retrievals. This is crucial because the radiometric complexity of land complicates MW-based retrievals, as explained in section 2.1.2.1. By acknowledging these physical constraints, the evaluation allows for a more accurate interpretation of the model performance.

3.2.4 Computational considerations

Deep learning architectures primarily consist of a multitude of linear algebra operations, such as that in eq. (3.1), which can be executed in parallel. Among the various hardware accelerators employed in DL (see, e.g., Xu et al., 2026), graphics processing units (GPUs) excel at such tasks, containing many specialized cores that perform simultaneous operations. Goodfellow et al. (2016, section 12.1.2) elaborates on this compelling argument. For instance, the NVIDIA A100 GPU used for training in paper IV—common in ML clusters—features, in addition to 6912 cores for general single-precision arithmetic, 432 cores dedicated to matrix operations (NVIDIA, 2020). In contrast, CPUs such as Intel Xeon or AMD EPYC—commonly found in high-performance computing clusters—offer at most 40–96 cores per socket favouring sequential execution, which makes them far less efficient for the highly parallelizable computations employed in DL. Even when individual network components are

small, ML software libraries recast the computational problem into large-scale matrices, further leveraging the capabilities of hardware accelerators. Figure 3.1 illustrates the time required for matrix multiplication for a GPU and a CPU. Since neural networks consist of many such operations, even a modest speedup translates to a substantial reduction in total computation time.

However, GPUs are not always advantageous. For small models, the overhead of transferring data to GPU memory can exceed the computational time saved, making CPUs more efficient. Models with highly sequential operations or conditional logic also benefit less from parallelization but more serial power. Additionally, training can become CPU-bound when on-the-fly preprocessing leaves the GPU idle awaiting the next data batch.

Efficient data pipelines are equally critical. Repeatedly reading training data from disk can become a bottleneck that limits throughput regardless of GPU capacity, making fast storage solutions essential. Beyond hardware selection, computational efficiency can be improved through strategies like transfer learning. In paper IV, updating only a subset of parameters through partial fine-tuning reduced memory requirements and enabled a four-fold increase in the batch size—the data volume per training iteration—within the same GPU constraints, thereby requiring less training time.

Ultimately, effective DL requires not only sufficient computational resources but also model architectures designed to exploit parallelization and the special features of the dedicated hardware, as well as data systems capable of sustaining high throughput. Even with optimal infrastructure, however, the physics of the retrieval problem imposes inherent limits on predictive performance.

3.3 Deep learning architectures

The design and selection of a neural network is a complex problem given its flexibility. Besides limitations in software and computing resources, this process relies heavily on the expertise and intuition of the researcher, and can even resemble an art more than a science (Erenshteyn et al., 1994). A primary decision concerns the network architecture, i.e. how connections and layers are organized and how data is transformed as it flows through the network. This includes, for example, whether the data flows strictly in one direction, as in feedforward networks, or through feedback loops. The architecture choice also shapes the inductive bias of the model, i.e. built-in assumptions about the structure of the data, such as the presence of local spatial features or the relevance of patterns at multiple scales.

An architecture can be described in terms of its building blocks. Complex architectures typically combine different layer types or nest entire networks within larger ones, introducing depth and inductive biases that better fit the model to the structure of the problem. Therefore, the terms block, layer, and network are used somewhat interchangeably to refer to different architectural components and should be understood from context.

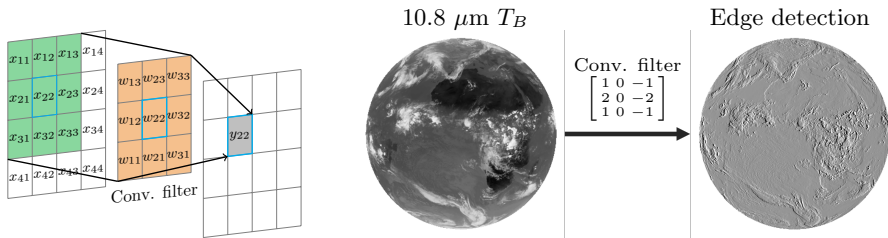


Figure 3.2: Left: schematic of a convolutional filter inside a CNN. The input $\mathbf{x}$ is convolved with the weights w_{ij} , producing an output value y_{mn} for each receptive field (in green) as the filter slides across the input; the weights are learned during training. Right: a filter that detects edges, applied to a 10.8 μm SEVIRI image; the output highlights cloud boundaries.

3.3.1 Fundamental building blocks

The multilayer perceptron (MLP) is the most fundamental type of modern neural network, in which every neuron in one layer is connected to every neuron in the next. Hence, MLPs are also referred to as fully-connected or dense networks. This type of network is agnostic to any structure in the data, which is both its strength and weakness. When features lack an inherent structural order, such as in tabular data, MLPs are well suited, as they can freely learn interactions between any pair of inputs regardless of their position. However, this global connectivity comes at a cost: the number of parameters grows rapidly with layer size. For high-dimensional structured data, such as images, this becomes problematic. A fully-connected network must learn a separate weight for every connection across all layers, leading to an explosion in parameters and high computational cost. More critically, MLPs do not efficiently exploit the spatial structure present in, e.g., images, as a pixel at the edge of a feature is treated no differently from one at its centre. Nevertheless, dense layers are commonly used as the final layers of larger networks, where earlier blocks have already extracted structured representations of the data, allowing global connectivity at this stage to capture relationships within the resulting feature space.

The inherent limitations of MLPs motivated the development of architectures with stronger inductive biases, such as convolutional neural networks (CNNs) for spatial data. Convolutional filters define CNNs. Rather than assigning an independent weight to each input element, the same filter is applied across the entire input, extracting local features within the receptive field of each filter, as illustrated in fig. 3.2. This parameter sharing makes CNNs far more efficient than fully-connected layers for structured data such as images. Sliding the filter across the input also grants CNNs translational equivariance: the resulting feature map shifts correspondingly when a feature moves in the input. This property is particularly desirable for dense prediction tasks, where a value is assigned to each input element. Furthermore, stacking convolutional layers allows CNNs to automatically learn hierarchical representations: early

layers capture simple features such as edges and textures, and deeper layers combine these into complex structures that inform the prediction.

Despite their advantages over MLPs, CNNs introduce their own challenges. Because each filter only sees a small region of the input, determined by the size of the filter, distant dependencies can only be captured indirectly through network depth. Moreover, because the output of convolutional filters is locally aggregated information, sharp spatial boundaries are difficult to achieve. This is particularly experienced with pixelwise averaging losses, such as the mean squared error or the mean absolute error, which reward smooth predictions over sharp but slightly misplaced ones. This difference is clearly evidenced in paper I, where MLP and CNN retrievals are directly compared.

These limitations can be addressed with improved training methodologies, more architectural components, or strategical decisions. For instance, while CNNs do not inherently possess rotation equivariance, this property can be approximated by applying physically meaningful rotations during training. To capture large-scale features more efficiently, pooling operations—such as taking the maximum value in a receptive field—or strided convolutions downsample feature maps, so that subsequent layers focus on coarser features, though at the cost of losing fine spatial structures required for dense prediction tasks. The lost details from pooling layers or multiple convolutional layers can be recovered through skip connections, which shortcut distant layers: by directly concatenating information from earlier layers to deeper ones, the network can reintroduce detailed structures, while additive skip connections primarily ease the training of deep networks (He et al., 2016). Finally, normalization layers inside the network, such as batch normalization (Ioffe and Szegedy, 2015) or layer normalization (Ba et al., 2016), improve the conditioning of the optimization problem by smoothing the loss landscape (Santurkar et al., 2018).

More architectural considerations exist beyond those summarized here, including the choice of non-linear activation functions and the logic to follow when arranging the different blocks. The principles outlined here, however, reflect the rationale for the networks used in this thesis.

3.3.2 Design principles

As illustrated by fig. 3.3, clouds and precipitation exhibit an inherent morphological structure. Consequently, incorporating inductive biases that exploit spatial context is highly beneficial for the retrievals in this thesis: without such spatial awareness, for instance, the convective squall line in fig. 3.3f would be difficult to retrieve from spectral data alone. In addition to complementing spectral information, spatial context can also compensate for its absence. For instance, the network in paper II relies solely on a single-channel image.

Convolutional filters provide an efficient mechanism for incorporating this context in the retrievals. Alongside the critical reduction in parameters compared to MLPs, their translational equivariance is particularly relevant, as meteorological features are not confined to specific locations within the image.

Furthermore, allowing for both network depth and width is advantageous. Atmospheric data is inherently hierarchical: individual hydrometeors give

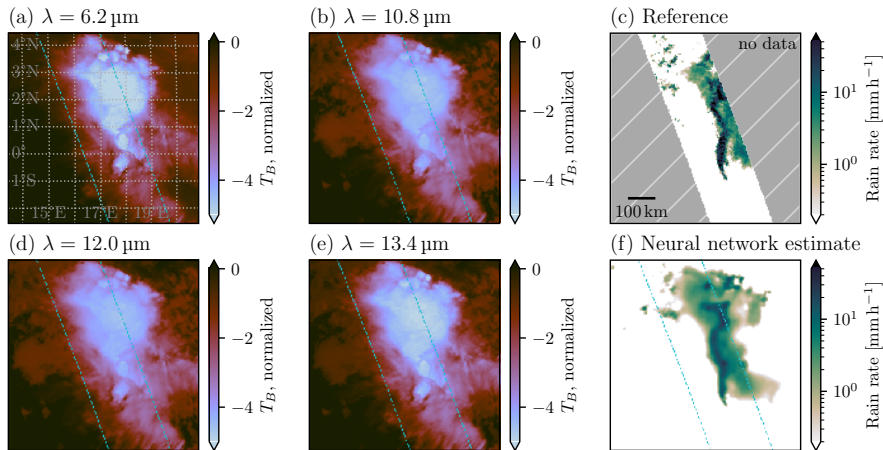


Figure 3.3: A cloud with a precipitation field over Africa, as shown in (c). The left and centre panels show four of the seven input channels used by the network in paper V to produce the retrieval shown in (f), each providing different features. The brightness temperatures are shown as seen by the network, i.e. standardized using the training set statistics.

rise to local cloud textures, which then organize into small-scale cells. These small features, in turn, aggregate to larger structures with clear hierarchical dependencies: anvils, which may display uniform radiative signatures across vast areas, are inextricably linked to the cloud cores. This is effectively exploited with increasing network depth, as successive layers build representations based on previous layers, thereby exploiting the natural hierarchy of the data. At the same time, increasing layer width—through more convolutional filters or wider fully-connected layers—enables the network to develop neurons that specialize in distinct but complementary features, while also allowing the network to form a consensus across neurons under ambiguous conditions.

Unfortunately, there are no principled answers to the question of how many layers, neurons, or what spatial context to incorporate: these all add to the large list of hyperparameters. In practice, the most pragmatic approach is to adapt architectures developed for similar problems to the hardware and data at hand, adjusting for the number of input features and the desired output. One guiding principle, however, is to have fully-connected layers at the end of the network, for the reasons outlined in the previous section.

3.3.3 Architecture selection

The use of neural networks for retrieving frozen hydrometeors from geostationary observations is relatively novel, and even for precipitation retrievals the literature offers limited guidance. This contrasts with other fields where well-tested, standard architectures have widely converged, leaving architecture choice here largely an empirical process.

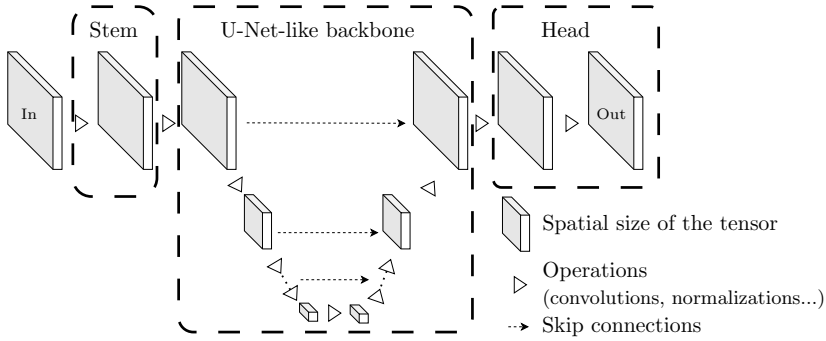


Figure 3.4: Schema of a U-Net-like architecture.

Paper I started with few assumptions about a suitable architecture. First, several principles were gathered to propose a MLP without using any spatial information. The hyperparameters of this MLP were tuned over a regular grid to obtain a satisfactory performance. Although it can be argued that a grid search is not as efficient as, for example, a Bayesian optimization, the search space was small and manageable.

The CNNs trained in this thesis consisted of many more parameters than the MLP in paper I.³ This posed a challenge: both hyperparameter tuning and designing architectures from scratch can be computationally prohibitive and practically infeasible within the scope of this thesis. For this reason, all appended papers employed CNN backbones that, based on previous publications, had been empirically observed to work reasonably well for hydrometeor retrievals, employing limited hyperparameter optimization. Only paper I included some hyperparameter optimization for the CNN, though this was modest compared to the optimization performed for the MLP in the same paper.

To provide one retrieval per input pixel, the CNNs in this thesis are also fully CNNs, which preserve the spatial dimensions of the input data. Furthermore, fully CNNs support different input sizes. Although the backbones differ slightly across papers due to new ML approaches adopted over time, they share a common idea: the use of a U-Net-like architecture (Ronneberger et al., 2015), illustrated in fig. 3.4. The U-Net architecture was specifically designed for semantic segmentation tasks, i.e. pixelwise classification. Nevertheless, its structure can easily be adapted for pixelwise regression tasks. In a U-Net-like architecture, the encoder path extracts high-level features, while the decoder path recovers the original spatial dimensions. By leveraging skip connections, the decoder also merges spatial details directly concatenated from the contracting path with the more abstract representations of the upsampling path. This type of architecture is, therefore, a natural fit for the retrieval of hydrometeors, as it exploits features across multiple spatial scales to provide pixelwise estimates.

³The MLP in paper I has $\sim 2.8 \times 10^5$ parameters, and the CNN in the same paper $\sim 5.3 \times 10^5$. The CNN in paper V is slightly larger, at $\sim 6.3 \times 10^5$, while the CNN introduced in paper II (and fine-tuned in paper IV) is almost two orders of magnitude larger, at $\sim 4.7 \times 10^7$.

3.4 Uncertainty quantification

Since retrievals generally map one-to-many possible states, employing a single point estimate as a retrieval discards information about the solution space. Knowing how much uncertainty a retrieval carries can matter as much as the retrieval itself, particularly for downstream applications where a range of plausible outcomes is required. Uncertainty, however, has different origins, not all of which are captured equally by the same method.

3.4.1 Types of uncertainty

If a retrieval comes with, for example, a 5% uncertainty, does this reflect a fundamental ambiguity in the problem itself or a lack of confidence in the model? This distinction exemplifies the two classical types of uncertainty.

Epistemic uncertainty arises from a knowledge gap due to limited training data, inadequate models, or a mismatch between the training data and the data the model encounters in practice.⁴ Also known as model uncertainty, epistemic uncertainty is in principle reducible given enough high-quality training data and sufficiently expressive retrieval models.

In contrast, aleatoric uncertainty arises from variance intrinsic to the data itself. For instance, the same IR radiance can correspond to fundamentally different cloud or precipitation states, implying that not even a perfect model can resolve such ambiguity from IR data alone. It is therefore generally considered irreducible.

In practice, however, these two types are neither strictly orthogonal nor have a universal mathematical definition (e.g., Hüllermeier and Waegeman, 2021; Gruber et al., 2025; Bickford Smith et al., 2025): their distinction depends on the school of thought (Kirchhof et al., 2025). Furthermore, they cannot always be disentangled (e.g., Mucsányi et al., 2024; de Jong et al., 2026). Nonetheless, the distinction serves a useful communicative purpose: it clarifies what one targets to solve with a specific uncertainty estimator.

3.4.2 Approaches to uncertainty estimation

One can construct an uncertainty estimator by designing a model to predict an error quantity as the target variable; for example, a standard deviation in regression or class probabilities in classification. However, there are methods that inherently quantify uncertainty. This section briefly presents five of them. For a more comprehensive survey, see, e.g., Murphy (2023) and Izbicki (2025).

Three common examples for quantifying epistemic uncertainty are ensemble methods, Bayesian neural networks, and Monte Carlo Dropout. In an ensemble, the disagreement among the outputs of each predictor serves as a measure of

⁴This mismatch is also referred to as distributional uncertainty (e.g., Sullivan, 2015). It is a practically relevant concern: for instance, models trained on data from one season may face substantial differences when applied to another. Although dataset shift can be treated as an uncertainty class of its own (e.g., Malinin and Gales, 2018), here it is subsumed under epistemic uncertainty, as it reflects an incomplete knowledge of the true data-generating process.

model uncertainty. In Bayesian neural networks, model parameters are modelled by probability distributions: by repeatedly sampling these weights, the resulting predictive distribution reflects parameter uncertainty. Similarly, in Monte Carlo Dropout, neurons are randomly deactivated at each forward pass during inference, and the distribution obtained across a number of such predictions approximates the parameter uncertainty (Gal and Ghahramani, 2016).

A distinct approach is conformal prediction, which directly combines epistemic and aleatoric uncertainty by producing prediction regions (intervals for regression, sets of labels for classification) calibrated to contain the true value with a chosen probability (Vovk et al., 2022; Angelopoulos and Bates, 2023). Although this framework ensures valid uncertainty estimates provided the calibration data and the unseen samples are exchangeable,⁵ the guarantees lack validity for conditional estimates: they hold only on average and may yield uninformative intervals.

Quantile regression, on the other hand, directly estimates conditional quantiles of a continuous variable. This approach relies on correct model specification, requiring sufficient model capacity and representative data to capture the true underlying distribution. When these conditions are met, the model characterizes the irreducible uncertainty present in the data. It also admits a Bayesian interpretation, as estimating conditional quantiles is equivalent to characterizing the posterior predictive distribution of the target given the input. Moreover, quantile regression is straightforward to implement and is computationally efficient.

The ill-posed nature of the inverse problems addressed in this thesis suggests that aleatoric uncertainty is expected to dominate. Consequently, the appended papers focus on capturing this irreducible uncertainty. For regression, quantile regression is chosen to estimate multiple conditional quantiles, with its mathematical formulation and implementation given in section 4.2.1. For classification, the selected approach is cross-entropy minimization, which naturally captures aleatoric uncertainty through class probabilities, further detailed in section 4.2.2.

⁵The property by which the joint distribution of the calibration points and any test points is invariant under permutation.

4 Retrieval development

The development of retrievals through supervised learning is shaped by the quality of the training data, the statistical techniques employed, and the overarching methodology followed. Considerations regarding these aspects, including practical ones that further shaped the retrievals in the appended papers, are detailed here.

4.1 Constructing the datasets of collocations

The retrievals developed in this thesis aim to reproduce the estimates of specialized spaceborne sensors using geostationary observations, treating those estimates as absolute ground truth. Therefore, it is natural to frame the machine learning tasks in a supervised learning process. Formally, supervised learning consists of maximizing the performance of an algorithm when mapping input data points $\mathbf{x}$ to known associated outcomes $\mathbf{y}$. Consequently, any supervised learning task requires a training dataset $\mathcal{D}_{\text{training}} = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^m$ generated from a distribution $P_{\mathbf{X}, \mathbf{Y}}$ on $\mathcal{X} \times \mathcal{Y}$. Note that the dimensionality of $\mathbf{x}$ and $\mathbf{y}$ will depend on the problem addressed. In this thesis, $\mathcal{X}$ are geostationary observations and $\mathcal{Y}$ retrievals from CloudSat-CALIPSO (frozen hydrometeor masses, cloud masks, and cloud types) or GPM-CO (surface precipitation).

Despite recent efforts towards benchmark datasets (Pfreundschuh et al., 2026), readily available collocations $\mathcal{D}$ between geostationary observations and the specific variables in the reference products targeted in this thesis do not exist. This necessitated producing large datasets of collocations tailored to the machine learning tasks, and stored in an analysis-ready format suitable for iterative model training, while keeping the storage footprint low. Generating these datasets carries a non-negligible computational expense, and design choices made at this stage are difficult to revisit after model development, making early decisions important.

Constructing $\mathcal{D}$ is inherently challenging, as GEO and LEO satellites differ in spatial sampling geometries, temporal sampling, and data product projections. Papers I, II, and IV used more advanced approaches than a nearest neighbour search for the spatial collocation. This is hypothesized to improve the quality of $\mathcal{D}$, and consequently the quality of $\mathcal{D}_{\text{training}} \subset \mathcal{D}$ used to guide the algorithms. Temporal collocation was determined by the temporal resolution of the geostationary observations.

A further complication arises specifically for frozen hydrometeors and cloud masks. The availability of two different sets of equivalent data products (from AERIS/ICARE or from the CloudSat Data Processing Center, cf. section 2.4.1) targeting the same physical variables, presents a dilemma regarding which to employ. Choosing between them is not only a scientific question but also has practical aspects. Paper I trains and evaluates on DARDAR-CLOUD, and also compares with physics-based retrievals of IWP from the CLAAS edition 2.1 dataset (Benas et al., 2017; Finkensieper et al., 2020), which is also validated against DARDAR-CLOUD by the dataset authors. In contrast, paper II employs 2C-ICE for training and evaluation of IWP, mainly motivated by data attribution requirements and further practical considerations. As evidenced in paper III, however, DARDAR-CLOUD and 2C-ICE can yield different estimates, raising the question of which constitutes the best reference, if such a distinction can be made at all.

4.2 Probabilistic retrievals

At the centre of each of the retrievals developed in this thesis is a neural network f_{θ} , which serves as a numerical approximation of f , the inverse of the forward function. The goal of f_{θ} is to learn optimal values for the learnable parameters θ such that f_{θ} accurately maps inputs from the observation space $\mathcal{X}$ to the corresponding outputs in the state space $\mathcal{Y}$, according to the underlying distribution of the collocations. As argued in section 3.1.1, the inherent non-uniqueness of this mapping benefits from a probabilistic approach. Consequently, f_{θ} is formulated to estimate the conditional probability distribution $P_{Y|\mathbf{X}}$, rather than a deterministic point estimate.

The specific method to learn $P_{Y|\mathbf{X}}$ depends on the nature of the target variable. For continuous variables (such as IWP, IWC, or precipitation), the appended papers employ quantile regression (section 4.2.1). For the categorical targets in papers II and IV, i.e. cloud types and masks, f_{θ} is trained to minimize the cross-entropy (section 4.2.2.1). With this cost function for the classification task, the model yields a probability distribution over all possible classes. Although the class with the highest probability is often used to indicate the state, the full distribution can address classification ambiguities. Rather than simply selecting the most probable class, paper IV highlights the influence of probability thresholds to determine cloud occurrences, which facilitates reconciling estimates not only from different retrieval configurations, but also with estimates from other datasets. The theoretical basis and implementation for these two probabilistic approaches are described below. Despite being described separately here, a network f_{θ} can integrate both regression and classification tasks by employing multiple network heads, as demonstrated in, e.g., paper II.

4.2.1 Quantile regression

Koenker and Bassett (1978) introduced linear quantile regression. This method provides a more comprehensive characterization of the relationship between a

target variable y and predictors $\mathbf{x}$ than the common least squares regression. It achieves a better description by estimating conditional quantiles rather than the conditional mean. Given the cumulative distribution function $F_{Y|\mathbf{X}}$ for a distribution $P_{Y|\mathbf{X}}$, the conditional quantile $y_\tau = y_\tau(\mathbf{x})$ at level $\tau \in (0, 1)$ is defined to be the value such that $y_\tau = \inf\{y : F_{Y|\mathbf{X}}(y | \mathbf{x}) \geq \tau\}$. The quantile loss function is defined as

$$\mathcal{L}_\tau(\hat{y}_\tau, y) = \begin{cases} \tau(y - \hat{y}_\tau) & \text{if } \hat{y}_\tau < y \\ (1 - \tau)(\hat{y}_\tau - y) & \text{if } \hat{y}_\tau \geq y \end{cases}. \quad (4.1)$$

It can be proven that the expectation $\mathbb{E}[\mathcal{L}_\tau(\hat{y}_\tau, Y) | \mathbf{X} = \mathbf{x}]$ is minimized when $\hat{y}_\tau = y_\tau$ (Koenker, 2005, pp. 5–9).

A quantile regression neural network (QRNN) is a non-linear extension of quantile regression, suitable for the non-linear nature of remote sensing retrievals. Therefore, using the notation in this thesis and adopting a Bayesian perspective, a QRNN seeks an estimator f_θ able to minimize $\mathbb{E}[\mathcal{L}_\tau(f_\theta(\mathbf{x}), Y) | \mathbf{X} = \mathbf{x}]$, where $\mathbf{x}$ are realizations of the random variable $\mathbf{X}$.

Pfreundschuh et al. (2018) analysed the performance of QRNNs as an alternative to estimating the posterior distribution of Bayesian remote sensing retrievals. Their work can be considered foundational for this thesis, as QRNNs are the unique method used here for the retrieval of continuous variables. The actual implementation of QRNNs also follows the approach of Pfreundschuh et al.: f_θ is designed to estimate a set of quantiles at fixed levels $\mathcal{T}$. That is, the loss function of QRNNs for a sample $(\mathbf{x}, y)$ is given by

$$\mathcal{L}(f_\theta(\mathbf{x}), y) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathcal{L}_\tau(f_\theta(\mathbf{x}; \tau), y) \quad (4.2)$$

where $f_\theta(\mathbf{x}; \tau)$ indicates the estimated quantile at level τ . Since the network predicts quantiles for a finite set of levels $\mathcal{T}$, the output is a quantile-parameterized distribution (QPD), as illustrated in fig. 4.1. The target distribution $P_{Y|\mathbf{X}}$ can then be approximated by, e.g., a linear extension of the predicted QPD.

As the notation suggests, this formulation of quantile regression is restricted to scalar target variables. Consequently, for vector retrievals such as IWC profiles or precipitation fields, the model only provides marginal distributions for each variable. Estimating joint distributions is an active area of research and lies outside the scope of this thesis work.

In any case, the essence of QRNNs consists of training neural networks to estimate quantiles of the irreducible uncertainty, from which quantities such as the expected value of a retrieval can be derived. The QRNN approach does not require assumptions on the target variable, such as a plausible family of distributions, and can be completely data-driven. For hydrometeor retrievals, this is particularly relevant: in the hypothetical case that the retrieval reproduces the distribution of the training data, such distribution can be far from Gaussian, as illustrated in fig. 4.2. It is also reasonable to expect heteroskedasticity, i.e. that the uncertainties vary with the input. For instance, the uncertainty of an IWP retrieval for a cloud-free pixel should be minimal when compared with a cloudy

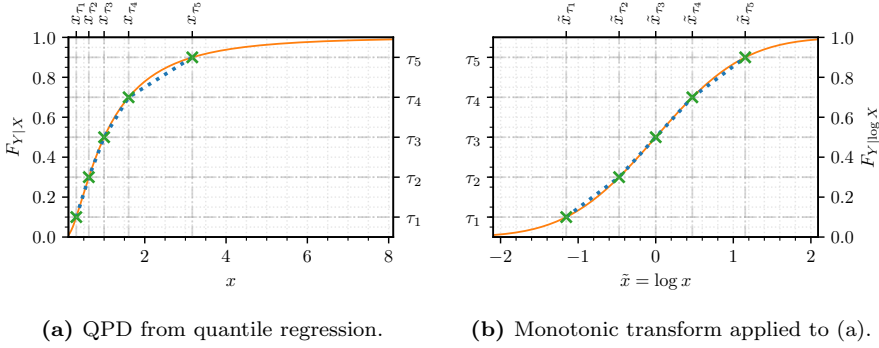


Figure 4.1: Optimal quantile regression examples for arbitrary random variables and quantile levels $\mathcal{T} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The solid orange curve is the true cumulative distribution function $F_{Y|\cdot}$. The green markers indicate quantiles estimated with quantile regression (a QPD), and the blue dotted curve their linear interpolation.

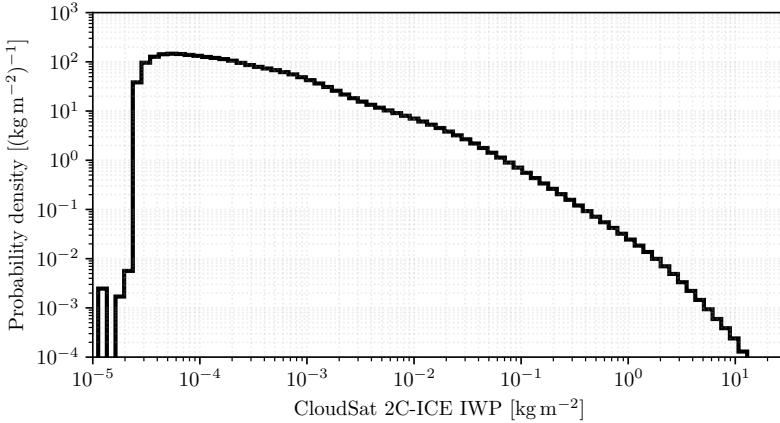


Figure 4.2: Probability density function from one year of CloudSat 2C-ICE IWP retrievals, where $\Pr(\text{IWP} \leq 10^{-5} \text{ kg m}^{-2}) = 49.7\%$. Note the logarithmic axes.

pixel. The non-linear nature of QRNNs is well suited for capturing this variability. In addition, quantile regression is invariant to monotonic transformations. Figure 4.1(b) provides a graphical proof of this property; formally, if h is a non-decreasing function, then $\tau = \Pr(Y \leq y_\tau \mid \mathbf{x}) = \Pr(h(Y) \leq h(y_\tau) \mid \mathbf{x}) = \tau$.

4.2.2 Classification

4.2.2.1 Rationale for the loss function

Although quantile regression can be adapted for retrieving categorical variables (e.g., Ghosh et al., 2017), retrieving cloud types and masks is more naturally addressed through a probabilistic classification approach. To motivate this, the following derivation provides a pedagogical description of this approach,

focusing on the cloud type retrieval, including clear-sky conditions. For cloud masks, this approach simply reduces to two classes: cloud and no-cloud.

In the appended papers, the training data consists of mutually exclusive cloud types: the training labels contain one and only one cloud class type. The objective of the model is to maximize $\Pr(Y_i = c^* \mid \mathbf{x}_i)$, i.e. the probability of predicting the cloud class $c = c^*$ of the training label $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \dots, y_{i,K}]^\top$ given the input $\mathbf{x}_i$. It can be rewritten as

$$\Pr(Y_i = c^* \mid \mathbf{x}_i) = \hat{y}_{i,c^*} = \hat{y}_{i,1}^0 \hat{y}_{i,2}^0 \cdots \hat{y}_{i,c^*}^1 \cdots \hat{y}_{i,K}^0 = \prod_{c=1}^K \hat{y}_{i,c}^{y_{i,c}} \quad (4.3)$$

where $\mathbf{y}_i$ is the one-hot encoding of the cloud type, and $\hat{\mathbf{y}}_i = f_\theta(\mathbf{x}_i) = [\hat{y}_{i,1}, \hat{y}_{i,2}, \dots, \hat{y}_{i,K}]^\top$ represent predicted probabilities for each of the K cloud types. Because the logarithm is a strictly increasing function, maximizing $\log \Pr(Y_i = c^* \mid \mathbf{x}_i)$ is equivalent to maximizing $\Pr(Y_i = c^* \mid \mathbf{x}_i)$. Machine learning usually seeks to minimize a loss function. Therefore, maximizing $\log \Pr(Y_i = c^* \mid \mathbf{x}_i)$ is equivalent to minimizing its negative. Through this trick we obtain

$$\mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = -\log \prod_{c=1}^K \hat{y}_{i,c}^{y_{i,c}} = -\sum_{c=1}^K y_{i,c} \log \hat{y}_{i,c} = -\sum_{c=1}^K y_{i,c} \log \hat{p}(c \mid \mathbf{x}_i) \quad (4.4)$$

for one training sample, where $\hat{y}_{i,c} = \hat{p}_\theta(c \mid \mathbf{x}_i)$ is used to emphasize that $\hat{y}_{i,c}$ is the probability of predicting class c conditioned on the input data $\mathbf{x}_i$. Equation (4.4) is also known as the negative log-likelihood.

In practice, the loss function in eq. (4.4) is minimized by computing the average loss across the M samples of the training dataset, as implemented in Papers II and IV. Using information theory, it can be shown¹ that

$$\frac{1}{M} \sum_{i=1}^M \mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \mathbb{E} \left[\underbrace{H(p_{\text{data}}(\cdot \mid \mathbf{x}))}_{\text{Data entropy, irreducible}} + \underbrace{D_{\text{KL}}(p_{\text{data}}(\cdot \mid \mathbf{x}) \parallel \hat{p}_\theta(\cdot \mid \mathbf{x}))}_{\text{Kullback-Leibler divergence}} \right] \quad (4.5)$$

(independent of network parameters θ)

where the expectation is computed over $p_{\text{data}}(\mathbf{x})$, which denotes the underlying marginal distribution that the training inputs approximate, $p_{\text{data}}(\cdot \mid \mathbf{x})$ is the conditional distribution over the K classes given $\mathbf{x}$, and $\hat{p}_\theta(\cdot \mid \mathbf{x})$ the corresponding model estimate. The expression inside the expectation of eq. (4.5) is also known as the cross-entropy. Therefore, minimizing the average negative log-likelihood is equivalent to minimizing the expected cross-entropy. Since the data entropy is independent of the model parameters, only the Kullback-Leibler divergence is minimized. Under ideal conditions, this divergence vanishes, yielding $\hat{p}_\theta(c \mid \mathbf{x}) = p_{\text{data}}(c \mid \mathbf{x})$, i.e. the true underlying conditional distribution. The probabilities $\hat{p}_\theta(c \mid \mathbf{x})$ therefore, under these conditions, quantify the aleatoric uncertainty of the retrievals.

¹The full derivation connecting the empirical loss to the information-theoretic quantities is omitted for brevity. Standard textbooks (e.g., Zhang et al., 2023, section 22.11) cover the underlying steps and concepts, including definitions and interpretations of $H(\cdot)$ and $D_{\text{KL}}(\cdot \parallel \cdot)$.

The derivation above implicitly assumed that the network outputs, $\hat{\mathbf{y}}_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i)$, are valid probabilities. In practice, however, raw network outputs are unnormalized and can take any real value: in the context of deep learning, these outputs are referred to as logits, here indicated with $\mathbf{z}_i$. Consequently, logits must be transformed into a valid probability distribution. The softmax function, defined for class c as

$$\text{softmax}(\mathbf{z}_i)_c = \frac{\exp z_{i,c}}{\sum_{j=1}^K \exp z_{i,j}} \quad (4.6)$$

maps logits to a categorical probability distribution over the K possible classes. Alternative mapping functions exist (e.g., Martins and Astudillo, 2016), and research has reported that the softmax can introduce overconfidence and miscalibration (Guo et al., 2017; Pearce et al., 2021). Nevertheless, the softmax is the standard function paired with the cross-entropy loss and naturally arises by reparameterizing the conditional probability.² This standard practice is evidenced by the implementation of the cross-entropy loss function in, e.g., PyTorch (Paszke et al., 2019), which applies a softmax operation directly in the loss.^{3,4} Following this standard, papers II and IV also employ the softmax to convert the network outputs to cloud type probabilities.

4.2.2.2 Decision criteria for cloud classes

Minimizing the cross-entropy loss does not directly produce a discrete classification, but instead outputs a probability distribution across the K cloud types. For simplicity, consider the case of retrieving a binary cloud mask ($K = 2$, i.e. no cloud or cloud): a probability decision threshold is needed to transform a probability into a class. The standard cross-entropy loss treats all misclassifications equally: the same penalty is given to a false positive (predicting a cloud when there is none) as to a false negative (missing an actual cloud). Because of the averaging, a model can achieve a low loss by optimizing for dominant patterns while performing poorly on ambiguous cases; this is often encountered when there is a large class imbalance, further discussed in section 4.3. Consequently, adjusting the probability decision threshold allows for weighting different types of errors. This adjustment determines the criteria used to flag a retrieval as cloudy.

The precision and recall⁵ metrics quantify the cost of false cloud detections and the cost of missing clouds, respectively. Using the cloudy conditions as the positive class, and TP , FP , and FN indicating true positives, false positives,

²For class c , the conditional probability can be rewritten as $\hat{y}_{i,c} = \hat{p}_{\boldsymbol{\theta}}(c | \mathbf{x}_i) = \frac{\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i, c)}{\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i)} = \frac{\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i | c) \hat{p}_{\boldsymbol{\theta}}(c)}{\sum_{j=1}^K \hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i | j) \hat{p}_{\boldsymbol{\theta}}(j)} = \frac{\exp \log(\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i | c) \hat{p}_{\boldsymbol{\theta}}(c))}{\sum_{j=1}^K \exp \log(\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i | j) \hat{p}_{\boldsymbol{\theta}}(j))} = \frac{\exp z_{i,c}}{\sum_{j=1}^K \exp z_{i,j}} = \text{softmax}(\mathbf{z}_i)_c$, with logits $z_{i,j}$ representing $\log(\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i | j) \hat{p}_{\boldsymbol{\theta}}(j))$.

³While PyTorch's `torch.nn.CrossEntropyLoss` automatically applies this operation, its negative log-likelihood loss class `torch.nn.NLLLoss` requires manually applying the softmax.

⁴Computing the softmax within the loss function allows for a more numerically stable implementation than applying it as a separate layer.

⁵Recall is also known as sensitivity.

and false negatives, these metrics are defined as

$$\text{Precision} = \frac{TP}{TP + FP} = \Pr(Y = \text{cloud} \mid \hat{Y} = \text{cloud}) \quad (4.7)$$

$$\text{Recall} = \frac{TP}{TP + FN} = \Pr(\hat{Y} = \text{cloud} \mid Y = \text{cloud}) \quad (4.8)$$

Given the irreducible uncertainty in the data, cloudy and clear conditions are not perfectly separable in all cases. This creates an inherent trade-off: eqs. (4.7) and (4.8) reveal that by flagging every case as cloudy a model achieves perfect recall but low precision ($FN \rightarrow 0$ while FP increases), whereas flagging only the most certain cloudy cases yields high precision at the expense of recall ($FP \rightarrow 0$ while FN increases). Consequently, these metrics are most informative when used in tandem. Through varying the probability decision threshold, papers II and IV show precision-recall curves to visualize this trade-off on the quality of cloud detections, ignoring the easier task of correctly identifying clear sky conditions (precision and recall do not involve true negatives).

To determine an optimal decision threshold that balances both metrics, the appended papers minimize $\sqrt{(1 - \text{precision})^2 + (1 - \text{recall})^2}$, i.e. the Euclidean distance to the point of perfect precision and recall. This approach is visually intuitive and, given that the curves in the papers are largely symmetric, it results in values similar to an alternative common approach: maximizing the harmonic mean of precision and recall (the F_1 score).⁶ Papers II and IV employ this optimal threshold as a formal criterion to separate clear from cloudy retrievals. Beyond its use as a binary cutoff, paper IV shows the impact of the sensitivity to this threshold. The resulting insights provide a more nuanced physical interpretation of the differences between 2D and 3D retrieval characteristics. Finally, this threshold is a necessary precursor for the 3D classification in the papers, as it defines the cloudy voxels that are then assigned one of eight distinct cloud types based on their maximum predicted probability.

4.3 Training methodology

Recall that a network f_{θ} is a parameterized function whose trainable parameters $\theta_i \in \theta$ are optimized during training to minimize the loss function $\mathcal{L}$. The methodology in this thesis follows the standard practice of updating these parameters using gradient descent as

$$\theta_i \leftarrow \theta_i - \eta \frac{\partial \mathcal{L}}{\partial \theta_i} \quad (4.9)$$

employing mini-batches of data, with the gradients computed via backpropagation, i.e. by recursively applying the chain rule through the computational graph

⁶As an example using the CPCIR data in paper IV, maximizing the F_1 score for the 2D cloud mask yields a threshold of 0.460 instead of 0.496, which is obtained by minimizing the Euclidean distance. This results in a precision and recall of 0.895 and 0.936, respectively, compared to 0.904 and 0.926 as presented in the manuscript.

to obtain $\partial\mathcal{L}/\partial\theta_i$. The scalar η denotes the learning rate, a hyperparameter of the optimization algorithm governing the step size of each update. In practice, rather than using a fixed global learning rate η , modern optimization algorithms modify the update rule in eq. (4.9) to dynamically adjust the update step for each parameter. Adam (Kingma and Ba, 2017) and AdamW (Loshchilov and Hutter, 2019) are two such optimizers, employed in the appended papers. Additionally, papers II, IV, and V employ a cosine annealing learning rate scheduler: this modifies η following a cosine function during training to further aid in avoiding poor local minima.

The practical implementation of these optimization strategies, along with considerations for numerical stability, is provided by modern deep learning frameworks. In this thesis, PyTorch (Paszke et al., 2019) serves as the primary framework. Two additional high-level interfaces for PyTorch are used: quantnn (Pfreundschuh, 2022), providing specialized functions for training and implementing QRNNs, and Lightning (Falcon and The PyTorch Lightning team, 2019), to reduce repetitive code. While these libraries and the architectural choices argued in section 3.3.3 streamline the retrieval development workflow, the methodology further necessitates a series of principled decisions regarding data representation, generalization capacity, and training efficiency.

In the appended papers, the dataset of collocations $\mathcal{D}$ was partitioned into training, validation, and test splits. Training was monitored using both the loss curve and auxiliary diagnostics, such as the mean squared error of the expected value of the retrieval and scatter plots at selected checkpoints, providing a physical interpretation of the training evolution. The validation set was used to assess generalization performance, from which a network state was selected as the final model, before overfitting where it occurred. Given the iterative nature of this process, the final models reflect a trade-off between retrieval performance and the time required for implementation, training, and evaluation. The held-out test set was then used to evaluate the retrieval performance of the final models.

To improve both the data representation and the generalization capacity of the CNNs, the training accounted for the sampling geometry of the reference satellites. The collocation of CloudSat-CALIPSO or GPM-CO data with geostationary imagery introduces a directional bias in the training labels: the loss is computed only along the swath, which crosses the image diagonally. To mitigate the risk of the network learning these directional artefacts and to enable generalization to features regardless of their orientation, data augmentation was applied at training time: specifically, random rotations and both vertical and horizontal data flips. This procedure encourages the CNNs to achieve rotational invariance, decoupling the retrieval performance from the relative orientation of the satellite path and the atmospheric features.

Concerning the loss function, three core concepts were applied across the appended papers. First, all retrieval models technically perform multi-task learning (Caruana, 1997), where multiple targets are simultaneously predicted for each input pixel. The underlying idea of multi-task learning is that some of the factors explaining the relationships in the data are shared across different tasks. Consequently, jointly optimizing for these tasks pressures the model to

learn more robust relationships, which benefits generalization. In the case of QRNNs, these tasks correspond to the estimation of multiple quantiles, while in paper II the model targets both multiple quantiles and class probabilities. Intuitively, quantifying frozen hydrometeors connects to identifying not only the presence of a cloud but also a cloud type. In all cases, the loss function minimized is the sum of the losses for each variable. In this configuration, the loss for one variable can dominate the optimization of the shared backbone at the expense of others. Nevertheless, more complex approaches, such as a scaling factor for each term, were not pursued, as the straightforward approach already provided satisfactory retrievals.

The idea that features are shared across tasks also motivated the fine-tuning strategy in paper IV. Instead of re-initializing the trainable parameters or retraining the entire model, only specific network heads were fine-tuned. By preserving the backbone weights from the radar-only cloud type retrievals learnt in paper II, the optimization focused on adapting the existing feature representations in the input data to the more informative radar-lidar training labels of paper IV.

The second concept involves applying log-based transformations to the training labels for the regression tasks. These transformations were empirically observed to improve retrieval performance: intuitively, the target variables (frozen hydrometeor masses and precipitation) are physically meaningful across several orders of magnitude and exhibit highly-skewed distributions. Applying such a transformation allows the model to learn relationships in a space where relative differences are preserved across scales: for example, in log-space and for precipitation retrievals, differences between 1 mm h^{-1} and 10 mm h^{-1} or 10 mm h^{-1} and 100 mm h^{-1} are equidistant, allowing the model to represent the target distribution more uniformly across the relevant range.⁷ Because the logarithmic function and its inverse, the exponential, are strictly increasing functions, and quantile regression is invariant to monotonic transformations (section 4.2.1), transforming the predictions back to the original physical space yields valid quantiles.

Third, regarding the inherent imbalance of the data, with highly-skewed distributions and approximately 90 % of the observed volume being cloud-free (table A1 in paper II or table S2 in paper IV), sample weighting presents a potential strategy to facilitate the learning of rare but physically significant cases. However, this technique was intentionally avoided for several reasons. From a practical standpoint, it introduces an arbitrary hyperparameter regarding the assignment of weights. Furthermore, it affects quantile regression and classification differently.⁸ More fundamentally, artificially weighting the training data permanently biases the model’s learnt priors. This structural bias directly conflicts with the statistical objective of the QRNNs as employed in this thesis, which is to characterize the aleatoric uncertainty of the retrieval: any weighting

⁷For quantile regression, only the factor $\partial\mathcal{L}/\partial\hat{y}_\tau = \tau$ or $1 - \tau$ (cf. eq. 4.1) is propagated during the weight update rule (eq. 4.9), i.e. the gradient is independent of the error magnitude.

⁸For quantile regression, sample weighting with weight w scales the constant factor of its gradient (cf. footnote 7). In contrast, for the cross-entropy loss, $\partial\mathcal{L}/\partial z_i = \hat{y}_i - y_i$: here, the weight w acts as a multiplier of the actual error, i.e. $w(\hat{y}_i - y_i)$.

would distort the physical validity of the retrieved distributions. Similarly, for classification, weighting distorts the output probabilities relative to the occurrence rates in the training data. Although the retrieval distribution can be adjusted after training by, for example, correcting for the imposed prior, such methods increase the complexity of the retrieval pipeline. Ultimately, the unweighted optimization proved sufficient for satisfactory retrieval performance.

One critical factor during the development of the retrievals was access to specialized hardware, as motivated in section 3.2.4. PyTorch leverages this hardware by providing efficient implementations of common deep learning operations. Moreover, its implementation of distributed data parallel (DDP) training (Li et al., 2020) was used to scale training across four GPUs simultaneously in papers II and IV. In this technique, each GPU maintains a local copy of the model and optimizer. During each training step, the data batch is partitioned such that distinct inputs are processed by each GPU to compute local gradients for the parameter updates. However, the resulting gradients vary across devices as the inputs differ. To maintain consistency, the DDP paradigm aggregates and synchronizes these gradients prior to the weights update. This ensures that the parameter updates are identical across all GPUs, simulating the processing of a larger batch on one more powerful device.

4.4 Evaluation

The evaluation of the retrievals developed in this thesis is constrained by the nature of the reference data products, which are themselves retrievals. Consequently, microphysical assumptions are present at every stage of the development, including the evaluation phase.

The terminology adopted throughout the appended papers reflects the fact that the trained models implicitly leverage the microphysical assumptions of their reference datasets. The terms *evaluation* or *characterization* are generally preferred over *validation*, as the latter conveys the existence of an error-free ground truth; however, validation is used when evaluating against estimates that constitute an independent benchmark of high quality. When contrasting estimates with other retrievals that rely on common input data, the term *comparison* is employed to acknowledge that neither retrieval is necessarily absolutely correct.

Although the data products used for some of the training labels provide their own uncertainty estimates, these were not used in the evaluation process. This choice was motivated by the same concerns outlined when discussing data quality in section 3.2.2: reference uncertainties depend on the specific statistical models and prior assumptions of the data product. Consequently, the uncertainties of the retrievals developed in this thesis were not evaluated for absolute accuracy. Instead, the evaluation focused on the model’s ability to reproduce expected values and statistical characteristics of the retrieval distribution.

Relying on a single metric for evaluation can be misleading. To obtain a better picture of model performance, the appended papers employ a diverse

suite of metrics. These include, among others, bias, root mean squared error (sensitive to large errors) and mean absolute error (a more robust indicator of average performance) for continuous variables, as well as precision and recall for categorical variables. However, such standard pixelwise metrics can fail to capture the performance regarding spatial features. To address this limitation, paper V employs a technique relying on Fourier transforms of the data to evaluate the spatial agreement of the retrievals across multiple scales (Guiloteau et al., 2021), yielding an interpretable metric of spatial consistency.

In addition to summary statistics, graphical tools inform the evaluation. In all developed retrievals, the conditional distributions $p(y_{\text{retrieved}} | y_{\text{reference}})$ are presented on log-log axes as a proxy for raw scatter plots, which suffer from overplotting. The logarithmic scale ensures that the wide range of values relevant to the target variables is adequately examined. Moreover, selected cases are examined qualitatively to complement the statistical analysis.

Finally, evaluating retrievals that offer spatiotemporal continuity against LEO satellite data raises specific challenges. For all papers except paper V, the test data was derived from polar-orbiting satellites, restricting the evaluation to fixed sampling times and leaving temporal gaps outside those specific overpasses. Furthermore, constructing a continuous time series to quantitatively evaluate performance at a specific location is impossible when all reference satellites are LEO. Temporal consistency was therefore assessed qualitatively, relying on physical intuition (e.g., in the video supplement of paper I or figs. 8, S4, and Movie S1 of paper V). Paper II partly mitigated these limitations by validating the developed retrievals against estimates from independent specialized instruments: a ground-based cloud radar, as well as an airborne cloud radar and in-situ measurements from flight campaigns. However, these independent estimates are not without caveats: flight campaigns are confined to specific geographical regions and cloud scenarios, while probes specialized for quantifying hydrometeors are themselves subject to measurement errors (Strapp et al., 2016).

4.5 Computational environment

The retrieval development relied on a technical ecosystem spanning satellite data handling, large-scale data management, and a deep learning framework. For reading and resampling satellite observations, the Pytroll framework (Raspaud et al., 2018) provided the necessary tools. Among the cross-domain libraries employed, several were critical: xarray (Hoyer and Hamman, 2017) for manipulating multidimensional data arrays, dask (Dask Development Team, 2016) not only for parallel computing but also for handling computations that exceed system memory, and the fsspec project (fsspec developers, 2019) for accessing data remotely. The latter was particularly leveraged in paper IV, where terabyte-scale data was processed on-the-fly under constrained computational resources. The Zarr format (Zarr developers, 2015) also played a critical role: its flexible compression capabilities reduced the storage footprint, which exceeded 10 TiB on disk for papers III and IV, without sacrificing efficiency in

data analysis.

Finally, the deep learning workflow relied on the PyTorch library (Paszke et al., 2019) and its ecosystem, with the machine learning tasks generally executed on a local high-performance computing cluster. Within this environment, the integration of more modern GPUs accelerated the research by reducing computation time. For example, while NVIDIA V100 GPUs were used for training in paper II, NVIDIA A100 GPUs became available for all subsequent papers; the latter offer more compute capability and provide a higher data transfer rate. This increased bandwidth was particularly advantageous when producing the extensive data records used in papers III and IV.

5 Summary of the papers

The papers appended to this thesis establish two sets of contributions, both grounded in deep learning applied to geostationary infrared observations but addressing distinct hydrometeor types. The first set targets frozen hydrometeors and clouds. Paper I motivated the development of the Chalmers Cloud Ice Climatology, which is introduced and characterized in papers II–IV. The second set targets precipitation: paper V introduces Rain over Africa, a precipitation retrieval algorithm for the African continent. This chapter summarizes each paper and closes with an overview of their combined contribution to the scientific field. Practical applications and further research enabled by these works are discussed in chapter 6.

5.1 The Chalmers Cloud Ice Climatology

Figure 5.1 illustrates the scope and main targets of papers I–IV. The four papers progress from a regional exploratory study to a quasi-global, multidecadal retrieval algorithm and dataset.

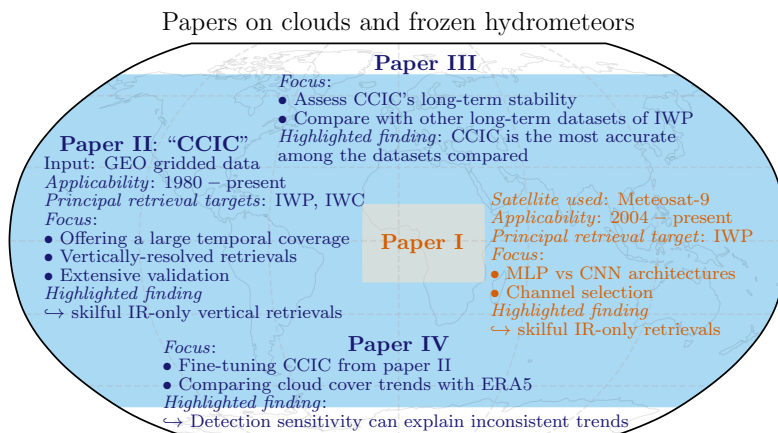


Figure 5.1: Summary of appended papers targeting retrievals of clouds or frozen hydrometeors. The colours indicate the regions covered by the papers.

5.1.1 Precursor work

Paper I proposes retrieving IWP from geostationary imagery using QRNNs trained against CloudSat-CALIPSO IWP retrievals. The study focuses on tropical Africa, using SEVIRI data as input. Due to the training and evaluation period chosen, only Meteosat-9 imagery was used, but since all MSG satellites carry SEVIRI replicas, the approach should generalize to any MSG satellite.

Five neural network configurations were compared by varying the input channels and architecture. Three pixelwise MLPs, differing only in their spectral inputs, were trained on daytime observations: a network using all 11 visible and IR channels performed best, though IR-only models followed closely. This strong IR-only performance motivated retraining on both daytime and nighttime data, with comparable results. Finally, a CNN trained on the same IR inputs generally produced more satisfactory retrievals.

Subsequently, the CNN retrievals were compared with physics-based retrievals from a publicly available dataset, also derived from Meteosat-9. Unsurprisingly, the CNN reproduced CloudSat-CALIPSO retrievals more closely. Diurnal cycles were also examined, and the patterns between the two approaches generally agreed, though they differed in IWP magnitude. Importantly, by relying solely on thermal infrared radiances, the CNN provides retrievals during nighttime, closing the gap left by the physics-based approach (section 3.1.2).

Paper I also explores retrieving the mean mass size and mean mass height of the frozen hydrometeor column, hinting at the possibility of resolving the vertical distribution of frozen hydrometeor mass. Overall, the paper demonstrates that deep learning can retrieve IWP from geostationary satellite infrared observations even with a limited number of channels.

5.1.2 Retrieval development

Figure 5.1 highlights a limitation of paper I: its geographically limited coverage. Paper II extends the scope to quasi-global coverage, constrained only by the combined field of view of geostationary satellites. This broader scope relies on two complementary datasets of globally gridded geostationary satellite observations, GridSat and CPCIR (section 2.4.3.2), which also span a longer time period, with one extending back to 1980.

Extending coverage introduces a challenge: harmonizing spectral information across different satellites. Since the $11\text{ }\mu\text{m}$ channel is the only one common to both input datasets, paper II uses it as the sole input to a CNN, building on the strong IR-only performance demonstrated in paper I. The outputs of this network constitute the Chalmers Cloud Ice Climatology (CCIC).

CCIC retrieves not only IWP, but also cloud probabilities. Moreover, CCIC retrieves IWC and vertically-resolved cloud classes, uncommon for geostationary imagery. All target variables in paper II were trained against CloudSat-derived products, using quantile regression for continuous variables and cross-entropy minimization for classification. The evaluation shows that CCIC-retrieved IWP correlates well with CloudSat-retrieved IWP, as do the corresponding zonal means. The vertically-resolved retrievals perform somewhat worse, though this

is unsurprising given the inherently greater complexity in the vertical dimension. Nevertheless, as the first long-term dataset to provide such vertically-resolved retrievals from 11 μm data, the results exceeded expectations. The CCIC characterization goes beyond the held-out test dataset to include independent measurements from flight campaigns and a ground-based cloud radar outside the training period, showing reasonable agreement.

Paper III extends the characterization of CCIC by evaluating its IWP estimates against those from 2C-ICE and DARDAR (section 2.4.1) over the entire CloudSat-CALIPSO mission lifetime. Moreover, it compares the full CCIC IWP record through 2023 with three other long-term observational datasets, namely ISCCP (H-series dataset Knapp et al., 2016), PATMOS-x (Foster et al., 2021), and MODIS (daily gridded level 3 product Platnick, 2019), as well as ERA5 reanalysis data (Hersbach et al., 2020). The results show that CCIC generally provides the most accurate IWP estimates relative to the CloudSat-CALIPSO reference data, thus demonstrating its long-term stability. Although absolute global IWP estimates differ substantially across datasets, they agree on key structural features. Trends over 20 years reveal consistent interannual variability between CCIC and most other datasets, though this consistency weakens over 40-year trends: the competing datasets exhibit clear artefacts or apparent biases, and while CCIC also shows some artefacts, they are directly traceable to the input data.

A limitation of papers II and III is that cloud probability retrievals were not a primary focus. Moreover, the original cloud probability outputs were trained on radar-only data which did not capture the thinnest clouds (cf. section 2.1), resulting in an incomplete cloud description. Figure 5.2 illustrates this by comparing radar-only with combined radar-lidar retrievals. Paper IV addresses these limitations by updating the cloud mask and cloud class training labels originally extracted from the 2B-CLDCLASS product with 2B-CLDCLASS-LIDAR data, and fine-tuning only the network heads corresponding to cloud probabilities. This update also underscores that CCIC is applicable to all cloud types, not only ice clouds; the name reflects its principal target, not its full scope. The evaluation in paper IV follows the same approach as paper II, with results

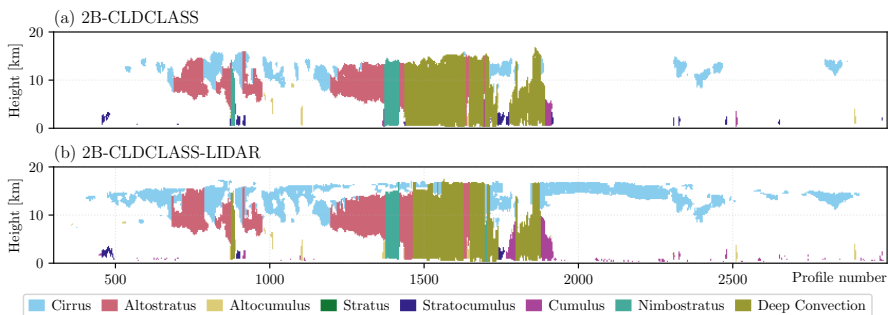


Figure 5.2: Comparison of cloud classification products for a fraction of a CloudSat orbit. The comparison illustrates the enhanced sensitivity of the radar-lidar product (2B-CLDCLASS-LIDAR) in capturing thin cloud layers, particularly cirrus, which are often missed by radar-only retrievals (2B-CLDCLASS).

consistent with the revised training labels, while also expanding on the vertical characterization of CCIC for cloud probabilities.

Paper IV further analyses cloud cover trends stratified by altitude. Existing long-term observational datasets providing stratified cloud cover, such as CLARA-A3 (Karlsson et al., 2023), report only the topmost cloud layer, missing clouds hidden beneath. Addressing this gap, the paper compares CCIC, the first long-term observational dataset providing retrievals in multilayer cloud scenarios, with ERA5 at different levels across more than four decades. Both datasets agree on a general decline in total cloud cover over the past two decades, though ERA5 exhibits an abrupt increase around the year 2000. CCIC reproduces this pattern under a specific retrieval configuration, and its flexibility allows this discrepancy to be reconciled by adjusting the cloud-detection sensitivity. This suggests that thin clouds can explain the discrepancies, at least in CCIC trends. This reconciliation is also evident in local trends computed over a 30-year period.

5.2 Rain over Africa

Accurate estimation of precipitation is a critical societal challenge, as it directly impacts water resources, agriculture, and more generally, disaster preparedness. Despite its importance, timely and reliable rain rate data remains scarce across Africa. Existing datasets, such as IMERG (Huffman et al., 2023), are well established but suffer from high latency. The Rain over Africa (RoA) project aims to fill these gaps by providing timely and reliable precipitation data.

Paper V introduces the RoA retrieval, a QRNN that provides instantaneous rain rate estimates for Africa using all SEVIRI thermal infrared channels as input to a CNN. The network was trained against retrievals from the GPM 2BCMB data product (Olson, 2022), which integrates passive MW and radar measurements (section 2.4.2). Training was conducted with Meteosat-11 data between 2018 and 2021, but the evaluation used observations from 2023 predominantly from Meteosat-10, thus providing a more rigorous assessment of RoA’s generalization ability.

The paper also compares RoA with both the final and IR-only versions of IMERG. This comparison demonstrates that IMERG IR-only performs considerably worse than RoA, even though both products rely on the same geostationary satellite. Although RoA and IMERG are comparable across several metrics, RoA exhibits higher fidelity to GPM 2BCMB in terms of effective spatial resolution over land, while IMERG maintains an advantage over the ocean. This disparity is attributed to the complex microwave emissivity of different surface types (section 2.1.2.1). Further comparisons of annual mean distributions show that RoA maintains closer agreement with the GPM 2BCMB mean.

A key advantage of RoA over established products is its timeliness. While the final version of IMERG requires several months to be produced, paper V demonstrates that RoA estimates can be generated in less than a minute after the dissemination of a Meteosat input image, using the accompanying open-source RoA code. Furthermore, since the RoA retrieval consists of a

QRNN, it enables probabilistic statements regarding extreme rainfall events, which are critical for flood and drought risk management.

5.3 Contributions to the research field

This thesis advances the application of machine learning in atmospheric remote sensing, specifically by leveraging geostationary satellite data to close existing gaps in hydrometeor observations. While geostationary platforms provide decades of continuous spatial and temporal coverage, their spectral characteristics have historically limited the accuracy or availability of cloud and precipitation retrievals. This thesis work challenges these limitations by demonstrating that deep learning architectures can enable IR-only retrievals that are competitive with, and in specific contexts exceed, the performance of established products.

A primary contribution of this research is CCIC, which, by exploiting the full geostationary record, provides the first quasi-global, 3D estimates of frozen hydrometeors and cloud probabilities continuous in time over more than four decades. Expanding the same methodological framework to precipitation, this thesis also introduces RoA, a retrieval algorithm for Africa that remains highly competitive with established products such as IMERG. Crucially, RoA provides probabilistic estimates that are timely, adding another dimension for risk mitigation at minimal computational cost.

This thesis is also committed to open and accessible science. Both CCIC and RoA are available as open-source software. High-performance computing resources are beneficial for processing the vast volumes of satellite data involved. Nevertheless, the software can be executed on any modern workstation. The corresponding datasets have likewise been made publicly available: at the time of writing (May 2026), the Registry of Open Data on Amazon Web Services hosts both CCIC (<https://registry.opendata.aws/ccic>) and RoA (<https://registry.opendata.aws/roa>) with cloud-optimized access. A more complete CCIC record, updated monthly, is additionally available at <https://clouds-and-precip.group/datasets/ccic>, while RoA is also distributed through third parties, as discussed later in section 6.1. Together, these efforts lower the barrier to adoption and aim to make the contributions of this thesis directly usable by the scientific community.

Ultimately, the developments presented here stand to influence established retrieval frameworks, support new research through the released datasets, and complement retrievals from upcoming mission-specific satellite instruments. Chapter 6 illustrates several such applications.

6 Downstream applications

The value of the open-access tools and datasets developed in this thesis lies in the new scientific capabilities they enable. As demonstrated in this chapter, both local and external research groups are already utilizing RoA and CCIC to facilitate downstream applications that were previously difficult or impossible. This ongoing work illustrates how the outputs of this thesis contribute to closing spatiotemporal gaps in hydrometeor retrievals.

6.1 Towards early warnings in Africa

A notable use of RoA has come through its adoption in the Weather and Climate Information Services programme’s Early Warnings for Southern Africa project (WISER-EWSA, wiser-ews.org), which aims to ensure that urban communities in southern Africa receive early weather warnings and know how to act on them. To this end, the project works with regional meteorological agencies and non-governmental organizations to co-produce nowcasting tools with community feedback.

As part of its activities, WISER-EWSA conducted two testbeds (e.g., Ageet et al., 2024; San Roman Pineda et al., 2025): short, intensive campaigns that brought together key stakeholders to evaluate new tools in real time. The testbeds evaluated RoA alongside other nowcasting and forecasting products, with the project team eventually running their own implementation of the RoA code; a catalogue of RoA retrievals produced by WISER-EWSA is available at <https://science.ncas.ac.uk/wiserews>. Using data shared under the WISER-EWSA project, Ageet et al. (2026) evaluate RoA and other precipitation products against sub-hourly rain gauge measurements, finding that RoA gener-

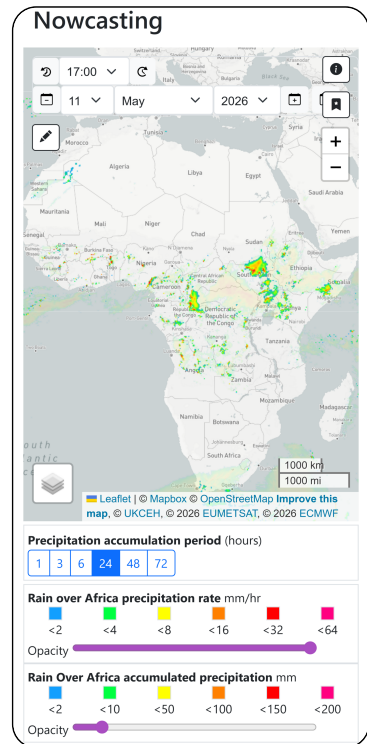


Figure 6.1: Screenshot of <https://africa-hydrology.ceh.ac.uk/nowcasting> (May 2026).

ally outperforms the alternatives across most metrics, even for accumulations over eight hours. Demonstrating further adoption, the UK Centre for Ecology & Hydrology also provides RoA estimates in near real time, as shown in fig. 6.1.

In a separate effort, RoA has been included in ongoing research on soil moisture assessment. Ahmad et al. (2026, in preparation) evaluate medium-resolution (0.1°) soil moisture datasets developed under ESA’s Climate Change Initiative, using daytime land surface temperature as a proxy for spatial variability in soil moisture. Among the datasets evaluated, only the soil moisture derived from ASCAT (a spaceborne 5.255 GHz radar) significantly captured spatial variability in land surface temperature at medium resolution. ASCAT soil moisture at 0.1° is then used to rank precipitation products based on the correlation between soil moisture and accumulated rainfall patterns at sub-daily timescales. In this ranking, RoA outperformed the other products, including IMERG and ERA5-Land.

6.2 Benchmark for atmospheric models

CCIC’s combination of multidecadal coverage with half-hourly, quasi-global coverage has made it a reference dataset for evaluating not only reanalysis data but also atmospheric models down to the km-scale. The works below illustrate a range of applications CCIC can cover.

Deutloff et al. (2026a) identified a previously unrecognized positive climate feedback using km-scale simulations: surface warming causes a partial shift of tropical deep convective clouds from daytime to nighttime, reducing sunlight reflection and amplifying the warming. Deutloff et al. (2026b) recently tested this prediction using CCIC’s multi-decadal, diurnally continuous IWP record, and derived a global feedback of $+0.12 \pm 0.02 \text{ W m}^{-2} \text{ K}^{-1}$, which is comparable in magnitude to other cloud feedbacks and implies a higher climate sensitivity.

Building on Leko (2025), Leko et al. (2026, in preparation) analyse the agreement of CCIC, ERA5, IMERG, and the km-scale models from the DYAMOND intercomparison project (Stevens et al., 2019). As a novel application of CCIC, Leko et al. not only compare IWP estimates across the datasets, but also

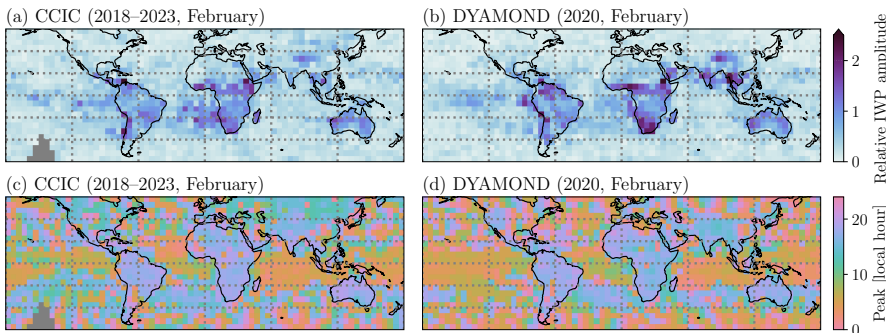


Figure 6.2: Harmonic-fitted diurnal amplitude and peak time of IWP for CCIC and the DYAMOND multi-model average. Adapted from Leko (2025), with permission.

the high cloud cover following the update in paper IV. Specifically, the study examines how these quantities vary geographically and covary, together with precipitation, throughout the day across six tropical regions.

The covariance between IWP and precipitation further allows for the estimation of precipitation efficiency. While strictly defined as the ratio of surface precipitation to the rate of new hydrometeor mass forming from water vapour, an approximation can be derived by replacing the denominator with the temporal rate of change of IWP (Kukulies et al., 2024). Kukulies et al. (2026) exploit the high temporal resolution of the CCIC IWP alongside this formulation to assess how well a km-scale model simulation estimates precipitation efficiency for six storm types over the contiguous United States. CCIC IWP estimates have also been used over the Asian monsoon region as a robustness check on precipitation efficiency derived from IMERG and ERA5 (Makgoale et al., 2026).

In a different setting, CCIC has provided the observational reference for evaluating simulated IWP in a km-scale model simulation of an Arctic marine cold-air outbreak (Lin et al., 2026). For this application, the spatiotemporally continuous coverage of CCIC complemented the limited sampling of the CloudSat-based reference.

Finally, Eriksson et al. (2026) leverage CCIC in three distinct roles. First, it is used in a cross-comparison of CloudSat-derived estimates given by 2C-ICE, DARDAR, and inversions with the Atmospheric Radiative Transfer Simulator (Buehler et al., 2025), extending the evaluation work presented in this thesis. The two other roles are oriented towards model evaluation. On the one hand, CCIC’s multidecadal record serves as the observational benchmark against which climate models are evaluated, specifically for the annual cycle of IWP. On the other hand, CCIC is used to evaluate km-scale models from DYAMOND, with an emphasis on characterizing the spatial organization of convection, as well as the diurnal cycle, number, and size of convective cores, with the latter two shown in fig. 6.3.

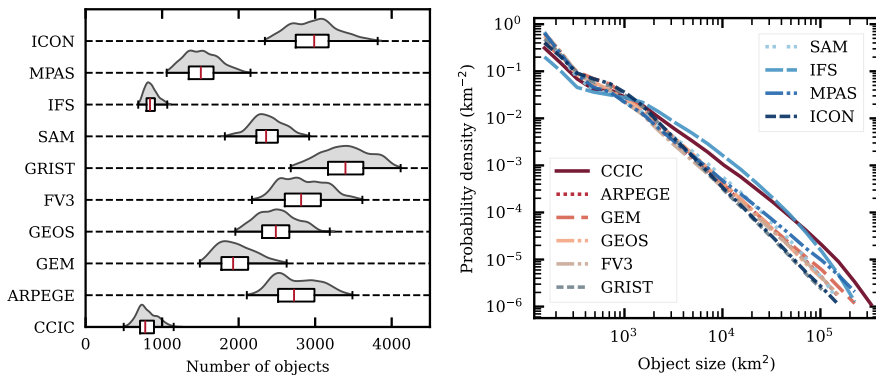


Figure 6.3: Number (left) and size (right) of tropical convective cores from CCIC and DYAMOND models; figs. 13 and 14 from Eriksson et al. (2026, CC BY 4.0).

6.3 Baseline for new satellite products

In a parallel role, CCIC provides the necessary data to support the evaluation of new IWP retrievals. The following two examples demonstrate this application.

Operational since 2025, the Arctic Weather Satellite (AWS, Eriksson et al., 2025) is the prototype for the upcoming EPS Sterna constellation. For the first time in an operational setting, AWS provides sub-millimetre (above 300 GHz) observations in addition to MW information. Building on these new observations, McEvoy et al. (2026) present the Chalmers Hydrometeor Inversion Product from AWS (CHIP-AWS), which currently retrieves IWP using a QRNN. CCIC, in turn, provides an IWP dataset independent of CHIP-AWS, since the latter is trained against radiative transfer simulations. Given its sub-millimetre channels, CHIP-AWS can achieve higher accuracy at larger IWP values. Furthermore, it resolves finer-scale features than CCIC, due to its instrument characteristics and retrieval design. Figure 6.4 illustrates both this difference in resolution and a strength of CCIC: the limited swath inherent to LEO platforms leaves gaps in the scene that CCIC’s geostationary-based retrieval naturally fills.

CCIC’s quasi-global continuous coverage, ensuring overlap with LEO platforms, is also leveraged in the evaluation of FYAI (Yang et al., 2026), a new dataset of IWP estimates derived from MW observations acquired by the Fengyun-3 satellites. Given their shared reliance on a QRNN trained against 2C-ICE, the value of CCIC here lies in providing an additional observational dataset for assessing the performance of FYAI during a tropical cyclone overpass.

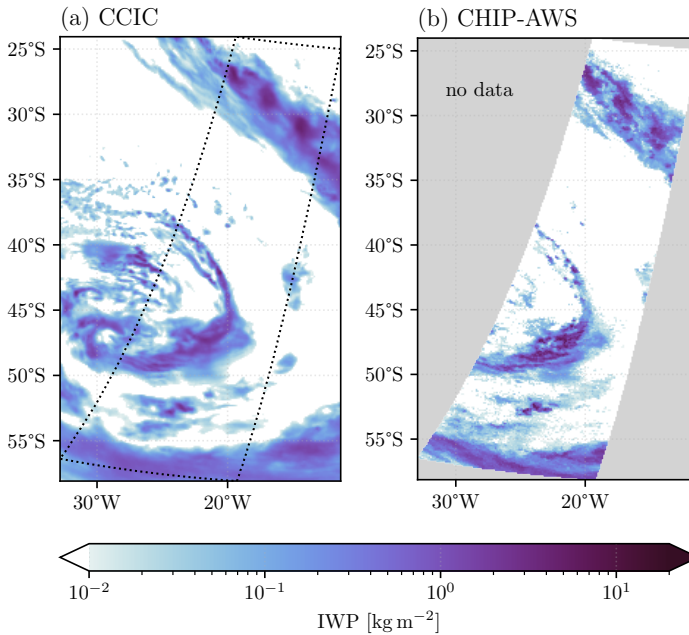


Figure 6.4: IWP estimates for the same scene. The dotted outline in (a) indicates the AWS swath footprint; regions outside the swath are not retrieved by CHIP-AWS but are covered by CCIC. Adapted from McEvoy et al. (2026), with permission.

7 Outlook

Building on the retrievals and datasets presented here, this chapter outlines several research directions to further develop them, together with challenges that remain open. These topics span more advanced algorithms, the downstream use of the probabilistic outputs, and the maintenance of the large data records.

A fundamental question is whether more advanced ML approaches will substantially improve the retrievals. Although the models are ultimately constrained by physics and cannot be expected to overcome irreducible uncertainties, extensions of the methods suggest the answer may be affirmative. Two directions are particularly relevant.

First, the temporal continuity that distinguishes geostationary observations is unexploited by the networks developed here, which process each image independently. Informing the models of the preceding state of the atmosphere would provide a basis for the networks to capture dynamics of atmospheric systems, such as the evolution and trajectory of deep convective systems, potentially reducing retrieval uncertainties.

Second, as discussed in section 4.2.1, the quantile regression approach adopted here describes only the marginal distributions of the output variables, not the correlations between them. One line of research is therefore to explore statistical learning methods capable of approximating their joint distribution. Capturing the dependency between spatially adjacent outputs would, moreover, make it possible to determine a case-specific effective resolution of the retrievals.

Even without further algorithmic development, the practical application of the current probabilistic outputs presents an unexploited opportunity. While the retrievals presented here output full probability distributions, in practice they are reduced to their expected values, both within this thesis and in downstream applications. Making use of this information requires addressing research questions where retrieval uncertainty itself is not only a statistic, but constitutes the core of the answer.

For the datasets produced in this thesis to deliver lasting value to the scientific community, they must be both maintained and extended as new observations become available. This poses processing and long-term storage challenges; these terabyte-scale datasets exceed what standard repositories such as Zenodo offer. To date, the distribution of CCIC has relied on a time-limited industry sponsorship, complemented by a self-managed solution. A truly sustainable solution requires a robust and durable infrastructure, ideally managed by organizations with the capacity to keep such large data records

openly accessible over the long term. Furthermore, it is unrealistic to leave the processing burden to researchers simply because the code and input data are available, particularly given the relatively small community working on such problems. Consequently, the extension of the datasets calls for an operational pipeline, ideally hosted by those with the capacity to do so.

The extension of the datasets also matters in light of new and upcoming missions, such as EarthCARE (Illingworth et al., 2015) or the imminent Ice Cloud Imager on MetOp-SG (Eriksson et al., 2020). These satellites will narrow existing observational gaps, but will not entirely close them. Retrievals from geostationary radiances will thus remain complementary, providing the spatiotemporal continuity and long record that these LEO platforms cannot offer.

References

- Ageet, A., A. H. Marsham, A. J. Parker, A. Singh, A. M. Taylor, A. J. Roberts, A. Groves, A. Nkonde and A. Amell (2026). ‘Evaluation of satellite rainfall products for nowcasting over Africa using rain-gauges and lightning data’. Manuscript under review at *Meteorol. Appl.* (cit. on p. 59).
- Ageet, S., D. J. Parker, H. Msemo et al. (2024). *Weather and Climate Information Services Early Warning for Southern Africa (WISER-EWSA) testbed 1*. Report. The University of Leeds. DOI: 10.48785/100/267 (cit. on p. 59).
- Ahmad, J. A., C. M. Taylor, B. Harris, D. D.Kovács, S. Anderson, A. Amell, C. Moldenhauer, S. Wells and W. Dorigo (2026). ‘ESA CCI 0.1° soil moisture: evaluation and applications for precipitation verification and storm nowcasting’. In preparation (cit. on p. 60).
- Amell, A. (2023). *Retrieving cloud ice masses from geostationary images with neural networks*. Chalmers University of Technology (Sweden). URL: <https://research.chalmers.se/publication/537675> (visited on 2026-05-16) (cit. on p. ii).
- Aminou, D. M. A., B. Jacquet and F. Pasternak (1997). ‘Characteristics of the Meteosat Second Generation (MSG) radiometer/imager: SEVIRI’. In: *Sensors, Systems, and Next-Generation Satellites*. Ed. by H. Fujisada. Vol. 3221. International Society for Optics and Photonics. SPIE, pp. 19–31. DOI: 10.1117/12.298084 (cit. on p. 23).
- Angelopoulos, A. N. and S. Bates (2023). ‘Conformal Prediction: A Gentle Introduction’. In: *Foundations and Trends in Machine Learning* 16.4, pp. 494–591. ISSN: 1935-8237. DOI: 10.1561/2200000101 (cit. on p. 40).
- Arkin, P. A. and B. N. Meisner (1987). ‘The Relationship between Large-Scale Convective Rainfall and Cold Cloud over the Western Hemisphere during 1982-84’. In: *Monthly Weather Review* 115.1, pp. 51–74. DOI: 10.1175/1520-0493(1987)115<0051:TRBLSC>2.0.CO;2 (cit. on p. 29).
- Ba, J. L., J. R. Kiros and G. E. Hinton (2016). *Layer Normalization*. DOI: 10.48550/arXiv.1607.06450 (cit. on p. 36).
- Belkin, M., D. Hsu, S. Ma and S. Mandal (2019). ‘Reconciling modern machine-learning practice and the classical bias–variance trade-off’. In: *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854. DOI: 10.1073/pnas.1903070116 (cit. on p. 31).
- 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 (cit. on p. 42).
- Benas, N., I. Solodovnik, M. Stengel et al. (2023). 'CLAAS-3: The Third Edition of the CM SAF Cloud Data Record Based on SEVIRI Observations'. In: *Earth System Science Data* 15.11, pp. 5153–5170. DOI: 10.5194/essd-15-5153-2023 (cit. on p. 28).
- Bickford Smith, F., J. Kossen, E. Trollope, M. van der Wilk, A. Foster and T. Rainforth (2025). *Rethinking Aleatoric and Epistemic Uncertainty*. DOI: 10.48550/arXiv.2412.20892. arXiv: 2412.20892 (cit. on p. 39).
- Bishop, C. M. and H. Bishop (2024). *Deep Learning: Foundations and Concepts*. Springer Cham. ISBN: 978-3-031-45468-4. URL: <https://www.bishopbook.com> (visited on 2026-05-16) (cit. on p. 27).
- Braun, B. M., T. H. Sweetser, C. Graham and J. Bartsch (2019). 'CloudSat's A-Train Exit and the Formation of the C-Train: An Orbital Dynamics Perspective'. In: *2019 IEEE Aerospace Conference*. 2019 IEEE Aerospace Conference. Big Sky, MT, USA: IEEE, pp. 1–10. ISBN: 978-1-5386-6854-2. DOI: 10.1109/AERO.2019.8741958 (cit. on p. 22).
- Buehler, S. A., R. Larsson, O. Lemke, S. Pfreundschuh, M. Brath, I. Adams, S. Fox, F. E. Roemer, P. Czarnecki and P. Eriksson (2025). 'The atmospheric radiative transfer simulator ARTS, version 2.6 — Deep python integration'. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 341, p. 109443. ISSN: 0022-4073. DOI: 10.1016/j.jqsrt.2025.109443 (cit. on p. 61).
- Caruana, R. (1997). 'Multitask learning'. In: *Machine Learning* 28.1, pp. 41–75. ISSN: 1573-0565. DOI: 10.1023/A:1007379606734 (cit. on p. 48).
- Cazenave, Q., M. Ceccaldi, J. Delanoë, J. Pelon, S. Groß and A. Heymsfield (2019). 'Evolution of DARDAR-CLOUD Ice Cloud Retrievals: New Parameters and Impacts on the Retrieved Microphysical Properties'. In: *Atmospheric Measurement Techniques* 12.5, pp. 2819–2835. ISSN: 1867-8548. DOI: 10.5194/amt-12-2819-2019 (cit. on p. 21).
- Ceccaldi, M., J. Delanoë, R. J. Hogan, N. L. Pounder, A. Protat and J. Pelon (2013). 'From CloudSat-CALIPSO to EarthCare: Evolution of the DARDAR Cloud Classification and Its Comparison to Airborne Radar-Lidar Observations'. In: *Journal of Geophysical Research: Atmospheres* 118.14, pp. 7962–7981. DOI: 10.1002/jgrd.50579 (cit. on p. 21).
- Dask Development Team (2016). *Dask: Library for dynamic task scheduling*. URL: <http://dask.pydata.org> (visited on 2026-04-17) (cit. on p. 51).
- de Jong, I. P., A. I. Sburlea, M. Sabatelli and M. Valdenegro-Toro (2026). *Measuring Orthogonality as the Blind-Spot of Uncertainty Disentanglement*. DOI: 10.48550/arXiv.2408.12175. arXiv: 2408.12175 (cit. on p. 39).
- 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 (cit. on p. 21).
- Deng, M., G. G. Mace, Z. Wang and R. P. Lawson (2013). 'Evaluation of Several A-Train Ice Cloud Retrieval Products with in Situ Measurements Collected during the SPARTICUS Campaign'. In: *Journal of Applied Meteorology and*

- Climatology* 52.4, pp. 1014–1030. DOI: 10.1175/JAMC-D-12-054.1 (cit. on p. 21).
- Deng, M., G. G. Mace, Z. Wang and H. Okamoto (2010). ‘Tropical Composition, Cloud and Climate Coupling Experiment Validation for Cirrus Cloud Profiling Retrieval Using CloudSat Radar and CALIPSO Lidar’. In: *Journal of Geophysical Research* 115, D00J15. ISSN: 0148-0227. DOI: 10.1029/2009JD013104 (cit. on p. 21).
- Deng, M., Gerald. G. Mace, Z. Wang and E. Berry (2015). ‘CloudSat 2C-ICE Product Update with a New Ze Parameterization in Lidar-Only Region’. In: *Journal of Geophysical Research: Atmospheres* 120.23, pp. 12, 198–12, 208. DOI: 10.1002/2015JD023600 (cit. on p. 21).
- Deutloff, J., S. Buehler, J. Windmiller and A. K. Naumann (2026a). *Facets of the Tropical High-Cloud Feedback in a Global Storm-Resolving Model*. [Preprint]. DOI: 10.21203/rs.3.rs-9133974/v1 (cit. on p. 60).
- (2026b). *Observed Diurnal Cycle Change of Tropical Deep Convection Amplifies Surface Warming*. [Preprint]. DOI: 10.21203/rs.3.rs-9293749/v1 (cit. on p. 60).
- Draper, D. W., D. A. Newell, F. J. Wentz, S. Krimchansky and G. M. Skofronick-Jackson (2015). ‘The Global Precipitation Measurement (GPM) Microwave Imager (GMI): Instrument Overview and Early on-Orbit Performance’. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.7, pp. 3452–3462. DOI: 10.1109/JSTARS.2015.2403303 (cit. on p. 23).
- Ekelund, R., M. Brath, J. Mendrok and P. Eriksson (2020). *ARTS Microwave Single Scattering Properties Database*. Version 1.1.0. Zenodo. DOI: 10.5281/zenodo.4646605 (cit. on p. 4).
- Erenshateyn, R., R. Foulds and S. Galuska (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 (cit. on p. 34).
- Eriksson, P., A. Baró Pérez, N. Müller, H. Hallborn, E. May, M. Brath, S. A. Buehler and L. Ickes (2026). ‘Advancements and continued challenges in observations and global modelling of atmospheric ice mass’. In: *Atmospheric Chemistry and Physics* 26.4, pp. 2741–2768. DOI: 10.5194/acp-26-2741-2026 (cit. on p. 61).
- Eriksson, P., A. Emrich, K. Kempe et al. (2025). ‘The Arctic Weather Satellite radiometer’. In: *Atmospheric Measurement Techniques* 18.18, pp. 4709–4729. DOI: 10.5194/amt-18-4709-2025 (cit. on p. 62).
- Eriksson, P., B. Rydberg, V. Mattioli, A. Thoss, C. Accadia, U. Klein and S. A. Buehler (2020). ‘Towards an operational Ice Cloud Imager (ICI) retrieval product’. In: *Atmospheric Measurement Techniques* 13.1, pp. 53–71. DOI: 10.5194/amt-13-53-2020 (cit. on p. 64).
- EUMETSAT (2009). *High Rate SEVIRI Level 1.5 Image Data - MSG - 0 degree*. URL: <https://data.eumetsat.int/product/E0:EUM:DAT:MSG:HRSEVIRI> (visited on 2026-02-02) (cit. on p. 24).
- (2012). *Effective Radiance and Brightness Temperature Relation Tables for Meteosat Second Generation*. Tech. rep. EUM/OPS-MSG/TEN/08/0024. EUMETSAT. URL: <https://user.eumetsat.int/s3/eup-strap>

- media/pdf_ten_080024_rad_bright_temp_f754be79ca.pdf (visited on 2026-05-17) (cit. on p. 32).
- EUMETSAT (2017). *MSG Level 1.5 Image Data Format Description*. Tech. rep. EUM/MSG/ICD/105. EUMETSAT. URL: https://user.eumetsat.int/s3/eup-strap-media/pdf_ten_05105_msg_img_data_e7c8b315e6.pdf (visited on 2026-05-17) (cit. on p. 24).
- (2026). *Meteosat orbital parameters*. URL: <https://user.eumetsat.int/resources/service-statuses/meteosat-orbital-parameters> (visited on 2026-03-13) (cit. on p. 24).
- Falcon, W. and The PyTorch Lightning team (2019). *PyTorch Lightning*. Version 1.4. DOI: 10.5281/zenodo.3828935. URL: <https://github.com/Lightning-AI/lightning> (visited on 2026-05-17) (cit. on p. 48).
- Finkensieper, S., J. F. Meirink, G.-J. van Zadelhoff, T. Hanschmann, N. Benas, M. Stengel, P. Fuchs, R. Hollmann, J. Kaiser and M. Werscheck (2020). *CLAAS-2.1: CM SAF CLOUD property dAtaset using SEVIRI - Edition 2.1*. DOI: 10.5676/EUM_SAF_CM/CLAAS/V002_01 (cit. on p. 42).
- Foster, M. J., C. Phillips, A. K. Heidinger and NOAA CDR PROGRAM (2021). *NOAA Climate Data Record (CDR) of Advanced Very High Resolution Radiometer (AVHRR) and High-resolution Infra-Red Sounder (HIRS) Reflectance, Brightness Temperature, and Cloud Products from Pathfinder Atmospheres - Extended (PATMOS-x), Version 6.0*. DOI: 10.7289/V5X9287S (cit. on p. 55).
- fsspec developers (2019). *fsspec: Filesystem interfaces for Python*. URL: <https://github.com/fsspec> (visited on 2026-04-17) (cit. on p. 51).
- Gal, Y. and Z. Ghahramani (2016). ‘Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning’. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1050–1059. URL: <https://proceedings.mlr.press/v48/gal16.html> (visited on 2026-05-17) (cit. on p. 40).
- Ghosh, A., H. Kumar and P. S. Sastry (2017). ‘Robust Loss Functions under Label Noise for Deep Neural Networks’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1. DOI: 10.1609/aaai.v31i1.10894 (cit. on p. 44).
- Goodfellow, I., Y. Bengio and A. Courville (2016). *Deep Learning*. MIT Press. URL: <http://www.deeplearningbook.org> (visited on 2026-05-17) (cit. on pp. 27, 33).
- Greco, M., W. S. Olson, S. J. Munchak, S. Ringerud, L. Liao, Z. Haddad, B. L. Kelley and S. F. McLaughlin (2016). ‘The GPM Combined Algorithm’. In: *Journal of Atmospheric and Oceanic Technology* 33.10, pp. 2225–2245. DOI: 10.1175/JTECH-D-16-0019.1 (cit. on p. 23).
- Gruber, A., B. Rudolf, M. M. Morrissey, T. Kurino, J. E. Janowiak, R. R. Ferraro, R. Francis, A. Chang and R. F. Adler (2007). ‘The Global Precipitation Climatology Project’. In: *Measuring Precipitation From Space: EURAINSAT and the Future*. Ed. by V. Levizzani, P. Bauer and F. J. Turk.

- Dordrecht: Springer Netherlands, pp. 25–35. ISBN: 978-1-4020-5835-6. DOI: 10.1007/978-1-4020-5835-6_3 (cit. on p. 16).
- Gruber, C., P. O. Schenk, M. Schierholz, F. Kreuter and G. Kauermann (2025). *Sources of Uncertainty in Supervised Machine Learning – a Statisticians’ View*. DOI: 10.48550/arXiv.2305.16703 (cit. on p. 39).
- Guiloteau, C., E. Foufoula-Georgiou, P. Kirstetter, J. Tan and G. J. Huffman (2021). ‘How Well Do Multisatellite Products Capture the Space–Time Dynamics of Precipitation? Part I: Five Products Assessed via a Wavenumber–Frequency Decomposition’. In: *Journal of Hydrometeorology* 22.11, pp. 2805–2823. DOI: 10.1175/JHM-D-21-0075.1 (cit. on p. 51).
- Guo, C., G. Pleiss, Y. Sun and K. Q. Weinberger (2017). ‘On Calibration of Modern Neural Networks’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330. URL: <https://proceedings.mlr.press/v70/guo17a.html> (visited on 2026-05-17) (cit. on p. 46).
- He, K., X. Zhang, S. Ren and J. Sun (2016). ‘Deep Residual Learning for Image Recognition’. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 36).
- Hersbach, H., B. Bell, P. Berrisford et al. (2020). ‘The ERA5 global reanalysis’. In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: <https://doi.org/10.1002/qj.3803> (cit. on p. 55).
- Hou, A. Y., R. K. Kakar, S. Neeck, A. A. Azarbarzin, C. D. Kummerow, M. Kojima, R. Oki, K. Nakamura and T. Iguchi (2014). ‘The Global Precipitation Measurement Mission’. In: *Bulletin of the American Meteorological Society* 95.5, pp. 701–722. DOI: 10.1175/BAMS-D-13-00164.1 (cit. on p. 22).
- Houze Jr., R. A. (2014). *Cloud Dynamics*. 2nd ed. Vol. 104. Academic Press. ISBN: 9780080921464 (cit. on p. 3).
- Hoyer, S. and J. Hamman (2017). ‘xarray: N-D labeled arrays and datasets in Python’. In: *Journal of Open Research Software* 5.1. DOI: 10.5334/jors.148 (cit. on p. 51).
- Huffman, G. J., D. T. Bolvin, R. Joyce et al. (2023). *NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Version 07*. Algorithm Theoretical Basis Document (ATBD) V07. NASA/Goddard Space Flight Center. URL: https://arthurhou.pps.eosdis.nasa.gov/Documents/IMERG_V07_ATBD_final.pdf (visited on 2026-04-22) (cit. on p. 56).
- Hüllermeier, E. and W. Waegeman (2021). ‘Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods’. In: *Machine Learning* 110.3, pp. 457–506. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-05946-3 (cit. on p. 39).
- Hunt, W. H., D. M. Winker, M. A. Vaughan, K. A. Powell, P. L. Luckner and C. Weimer (2009). ‘CALIPSO Lidar Description and Performance Assessment’. In: *Journal of Atmospheric and Oceanic Technology* 26.7, pp. 1214–1228. DOI: 10.1175/2009JTECHA1223.1 (cit. on p. 20).

- Iguchi, T., S. Seto, R. Meneghini, N. Yoshida, J. Awaka, M. Le, V. Chandrasekar, S. Brodzik, T. Kubota and N. Takahashi (2024). *GPM/DPR Level-2 Algorithm Theoretical Basis Document*. URL: https://www.eorc.jaxa.jp/GPM/doc/algorithm/ATBD_DPR_L2.pdf (visited on 2026-02-02) (cit. on p. 23).
- Illingworth, A. J., H. W. Barker, A. Beljaars et al. (2015). ‘The EarthCARE Satellite: The Next Step Forward in Global Measurements of Clouds, Aerosols, Precipitation, and Radiation’. In: *Bulletin of the American Meteorological Society* 96.8, pp. 1311–1332. DOI: 10.1175/BAMS-D-12-00227.1 (cit. on p. 64).
- Intergovernmental Panel on Climate Change (2023). ‘The Earth’s Energy Budget, Climate Feedbacks and Climate Sensitivity’. In: *Climate Change 2021 – the Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, pp. 923–1054. DOI: 10.1017/9781009157896.009 (cit. on p. 3).
- Ioffe, S. and C. Szegedy (2015). ‘Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift’. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456. URL: <https://proceedings.mlr.press/v37/ioffe15.html> (visited on 2026-05-17) (cit. on p. 36).
- Izbicki, R. (2025). *Machine Learning Beyond Point Predictions: Uncertainty Quantification*. 1st, p. 260. ISBN: 978-65-01-20272-3. URL: <https://rafaelizbicki.com/uq4ml> (visited on 2026-05-17) (cit. on p. 39).
- Janowiak, J., B. Joyce and P. Xie (2017). *NCEP/CPC L3 Half Hourly 4km Global (60S - 60N) Merged IR V1*. Ed. by A. Savtchenko. Accessed: 31 January 2023. Greenbelt, MD. DOI: 10.5067/P4HQB9N27EKU (cit. on p. 25).
- Janowiak, J. E., R. J. Joyce and Y. Yarosh (2001). ‘A Real-Time Global Half-Hourly Pixel-Resolution Infrared Dataset and Its Applications’. In: *Bulletin of the American Meteorological Society* 82.2, pp. 205–218. DOI: 10.1175/1520-0477(2001)082<0205:ARTGHH>2.3.CO;2 (cit. on p. 25).
- Karlsson, K.-G., M. Stengel, J. F. Meirink et al. (2023). ‘CLARA-A3: The third edition of the AVHRR-based CM SAF climate data record on clouds, radiation and surface albedo covering the period 1979 to 2023’. In: *Earth System Science Data* 15.11, pp. 4901–4926. DOI: 10.5194/essd-15-4901-2023 (cit. on p. 56).
- Kidd, C., A. Becker, G. J. Huffman, C. L. Muller, P. Joe, G. Skofronick-Jackson and D. B. Kirschbaum (2017). ‘So, How Much of the Earth’s Surface Is Covered by Rain Gauges?’ In: *Bulletin of the American Meteorological Society* 98.1, pp. 69–78. DOI: 10.1175/BAMS-D-14-00283.1 (cit. on p. 16).
- Kingma, D. P. and J. Ba (2017). *Adam: A Method for Stochastic Optimization*. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 (cit. on p. 48).
- Kirchhof, M., G. Kasneci and E. Kasneci (2025). ‘Reexamining the Aleatoric and Epistemic Uncertainty Dichotomy’. In: *ICLR Blogposts 2025*. URL: <https://iclr-blogposts.github.io/2025/blog/reexamining-the->

- aleatoric-and-epistemic-uncertainty-dichotomy (visited on 2026-05-17) (cit. on p. 39).
- Knapp, K. R., S. Ansari, C. L. Bain et al. (2011). ‘Globally Gridded Satellite Observations for Climate Studies’. In: *Bulletin of the American Meteorological Society* 92.7, pp. 893–907. DOI: 10.1175/2011BAMS3039.1 (cit. on pp. 5, 19, 25).
- Knapp, K. R. and NOAA CDR Program (2014). *NOAA Climate Data Record (CDR) of Gridded Satellite Data from ISCCP B1 (GridSat-B1) Infrared Channel Brightness Temperature, Version 2*. NOAA National Centers for Environmental Information. DOI: 10.7289/V59P2ZKR (cit. on p. 25).
- Knapp, W. B., W. A., V. Golea, K. R. Knapp, A. Young, A. Inamdar, B. Hankins and NOAA’s Climate Data Record Program (2016). *International Satellite Cloud Climatology Project Climate Data Record, H-Series [HGM, HGG]*. NOAA National Centers for Environmental Information. DOI: 10.7289/V5QZ281S (cit. on p. 55).
- Koenker, R. (2005). *Quantile regression*. Cambridge university press. ISBN: 9780511754098. DOI: 10.1017/CB09780511754098 (cit. on p. 43).
- Koenker, R. and G. Bassett (1978). ‘Regression Quantiles’. In: *Econometrica* 46.1, p. 33. ISSN: 00129682. DOI: 10.2307/1913643 (cit. on p. 42).
- Kojima, M., T. Miura, K. Furukawa, Y. Hyakusoku, T. Ishikiri, H. Kai, T. Iguchi, H. Hanado and K. Nakagawa (2012). ‘Dual-Frequency Precipitation Radar (DPR) Development on the Global Precipitation Measurement (GPM) Core Observatory’. In: *Earth Observing Missions and Sensors: Development, Implementation, and Characterization II*. Ed. by H. Shimoda, X. Xiong, C. Cao, X. Gu, C. Kim and A. S. K. Kumar. Vol. 8528. SPIE / International Society for Optics and Photonics, 85281A. DOI: 10.1117/12.976823 (cit. on p. 23).
- Koren, I., L. Oreopoulos, G. Feingold, L. A. Remer and O. Altaratz (2008). ‘How Small Is a Small Cloud?’ In: *Atmospheric Chemistry and Physics* 8.14, pp. 3855–3864. DOI: 10.5194/acp-8-3855-2008 (cit. on p. 14).
- Kubota, T., T. Masaki, G. Kikuchi, M. Ito, T. Higashiuwatoko, K. Kanemaru, N. Takahashi, K. Yamamoto, K. Furukawa and T. Nio (2024). ‘Evaluation of Effects on Dual-Frequency Precipitation Radar Observations Due to the Orbit Boost of the GPM Core Observatory’. In: *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 709–712. DOI: 10.1109/IGARSS53475.2024.10641066 (cit. on p. 23).
- Kukulies, J., A. F. Prein, J. M. Done, A. M. Stansfield and R. Rios-Berrios (2026). ‘Precipitation Efficiency by Storm Type in km-Scale Climate Simulations and Satellite Observations’. In: *Journal of Geophysical Research: Atmospheres* 131.5. e2025JD044800 2025JD044800, e2025JD044800. DOI: <https://doi.org/10.1029/2025JD044800> (cit. on p. 61).
- Kukulies, J., A. F. Prein and H. Morrison (2024). ‘Simulating Precipitation Efficiency Across the Deep Convective Gray Zone’. In: *Journal of Geophysical Research: Atmospheres* 129.24. e2024JD041924 2024JD041924, e2024JD041924. DOI: <https://doi.org/10.1029/2024JD041924> (cit. on p. 61).

- Lamb, D. and J. Verlinde (2011). *Physics and Chemistry of Clouds*. Cambridge University Press. ISBN: 9780511976377. DOI: 10.1017/CB09780511976377 (cit. on p. 3).
- Lambrigtsen, B., P. Kangaslahti, O. Montes, N. Niamsuwan, D. Posselt, J. Roman, M. Schreier, A. Tanner, L. Wu and I. Yanovsky (2022). ‘A Geostationary Microwave Sounder: Design, Implementation and Performance’. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15, pp. 623–640. DOI: 10.1109/JSTARS.2021.3132238 (cit. on p. 18).
- Leko, L., G. Behrens, N. Müller, A. Amell, A. Lauer and P. Eriksson (2026). ‘Diurnal cycles of convective processes in satellite measurements, reanalysis and km-scale models’. In preparation (cit. on p. 60).
- Leko, L. (2025). *The diurnal cycle of cloud ice water path*. URL: <http://hdl.handle.net/20.500.12380/309563> (visited on 2026-06-17) (cit. on p. 60).
- Li, S., Y. Zhao, R. Varma et al. (2020). ‘PyTorch Distributed: Experiences on Accelerating Data Parallel Training’. In: *Proceedings of the VLDB Endowment* 13.12, pp. 3005–3018. ISSN: 2150-8097. DOI: 10.14778/3415478.3415530 (cit. on p. 50).
- Lin, L., Y. Zhang, H. Beydoun, X. Zheng, M. Zhang, P. Bogenschutz, P. Wu and P. M. Caldwell (2026). ‘Exposing and Reducing Biases of Simulating Mixed-Phase Clouds in the Convection-Permitting E3SM Atmosphere Model: Lessons From an Arctic Cold-Air Outbreak’. In: *Journal of Geophysical Research: Atmospheres* 131.3. e2025JD044660 2025JD044660, e2025JD044660. DOI: 10.1029/2025JD044660 (cit. on p. 61).
- Loshchilov, I. and F. Hutter (2019). *Decoupled Weight Decay Regularization*. DOI: 10.48550/arXiv.1711.05101. arXiv: 1711.05101 (cit. on p. 48).
- Makgoale, T. E., S. C. Sullivan and J. Kukulies (2026). ‘An Observational Perspective on Precipitation Efficiency of Mesoscale Convective Systems over the Asian Monsoon Region’. In: *EGUsphere* 2026, pp. 1–23. DOI: 10.5194/egusphere-2026-1395 (cit. on p. 61).
- Malinin, A. and M. Gales (2018). ‘Predictive Uncertainty Estimation via Prior Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/3ea2db50e62ceefceaf70a9d9a56a6f4-Paper.pdf (visited on 2026-05-17) (cit. on p. 39).
- Martins, A. and R. Astudillo (2016). ‘From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification’. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1614–1623. URL: <https://proceedings.mlr.press/v48/martins16.html> (visited on 2026-05-17) (cit. on p. 46).
- May, E., B. Rydberg, I. Kaur, V. Mattioli, H. Hallborn and P. Eriksson (2024). ‘The Ice Cloud Imager: retrieval of frozen water column properties’. In: *Atmospheric Measurement Techniques* 17.19, pp. 5957–5987. DOI: 10.5194/amt-17-5957-2024 (cit. on p. 29).

- McEvoy, P., E. May and P. Eriksson (2026). ‘The Arctic Weather Satellite, introducing a new wavelength range for ice hydrometer retrievals’. Manuscript submitted for publication. URL: <https://clouds-and-precip.group/datasets/chip-aws> (visited on 2026-05-17) (cit. on p. 62).
- Menzel, W. P. and J. F. W. Purdom (1994). ‘Introducing GOES-I: The First of a New Generation of Geostationary Operational Environmental Satellites’. In: *Bulletin of the American Meteorological Society* 75.5, pp. 757–781. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(1994)075<0757:IGITF0>2.0.CO;2 (cit. on p. 24).
- Molnar, C. (2025). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3rd ed. ISBN: 978-3-911578-03-5. URL: <https://christophm.github.io/interpretable-ml-book> (visited on 2026-05-17) (cit. on p. 30).
- Mousivand, A., C. Straif, A. Burini, M. Lekouara, V. Debaecker, T. Hewison, S. Stock and B. Bojkov (2025). ‘In-Flight Calibration of Geostationary Meteorological Imagers Using Alternative Methods: MTG-I1 FCI Case Study’. In: *Remote Sensing* 17.2369 (14). ISSN: 2072-4292. DOI: 10.3390/rs17142369 (cit. on p. 23).
- Mucsányi, B., M. Kirchhof and S. J. Oh (2024). ‘Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks’. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak and C. Zhang. Vol. 37. Curran Associates, Inc., pp. 50972–51038. DOI: 10.52202/079017-1614. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/5afa9cb1e917b898ad418216dc726fbd-Paper-Datasets_and_Benchmarks_Track.pdf (visited on 2026-05-17) (cit. on p. 39).
- Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press. URL: <http://probml.github.io/book2> (visited on 2026-05-17) (cit. on p. 39).
- 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 the Atmospheric Sciences* 47.15, pp. 1878–1893. ISSN: 0022-4928, 1520-0469. DOI: 10.1175/1520-0469(1990)047<1878:DOTOTA>2.0.CO;2 (cit. on p. 28).
- Nayak, M. (2012). ‘CloudSat Anomaly Recovery and Operational Lessons Learned’. In: *SpaceOps 2012 Conference*. SpaceOps 2012. Stockholm, Sweden: American Institute of Aeronautics and Astronautics. DOI: 10.2514/6.2012-1295798 (cit. on p. 22).
- NVIDIA (2020). *NVIDIA A100 Tensor Core GPU Architecture*. URL: <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf> (visited on 2026-02-12) (cit. on p. 33).
- Olson, W. (2022). *GPM DPR and GMI Combined Precipitation L2B 1.5 hours 5 km V07*. Greenbelt, MD: Goddard Earth Sciences Data and Information Services Center (GES DISC). DOI: 10.5067/GPM/DPRGMI/CMB/2B/07 (cit. on p. 56).

- Olson, W. S. and the GPM Combined Radar-Radiometer Algorithm Team (2022). *GPM Combined Radar-Radiometer Precipitation Algorithm Theoretical Basis Document (Version 7)*. URL: https://pps.gsfc.nasa.gov/Documents/Combined_algorithm_ATBD.V07.pdf (visited on 2026-02-02) (cit. on p. 23).
- Paszke, A., S. Gross, F. Massa et al. (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 2026-03-27) (cit. on pp. 46, 48, 52).
- Pearce, T., A. Brintrup and J. Zhu (2021). *Understanding Softmax Confidence and Uncertainty*. DOI: 10.48550/arXiv.2106.04972. arXiv: 2106.04972 (cit. on p. 46).
- Pfreundschuh, S. (2022). *quantnn*. Version v0.0.4dev. DOI: 10.5281/zenodo.7127652. URL: <https://github.com/simonpf/quantnn> (visited on 2026-05-17) (cit. on p. 48).
- Pfreundschuh, S., M. Arulraj, A. Behrangi et al. (2026). ‘A Benchmark Dataset for Satellite-Based Estimation and Detection of Rain’. In: *Scientific Data* 13.1, p. 244. ISSN: 2052-4463. DOI: 10.1038/s41597-026-06565-0 (cit. on p. 41).
- 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. ISSN: 1867-8548. DOI: 10.5194/amt-11-4627-2018 (cit. on p. 43).
- Platnick, S. (2019). *MODIS Atmosphere L3 Cloud Properties Product. Version-1.1*. DOI: 10.5067/MODIS/CLDPROP_M3_MODIS_Aqua.011 (cit. on p. 55).
- Prigent, C., F. Aires and W. B. Rossow (2006). ‘Land Surface Microwave Emissivities over the Globe for a Decade’. In: *Bulletin of the American Meteorological Society* 87.11, pp. 1573–1584. DOI: 10.1175/BAMS-87-11-1573 (cit. on p. 12).
- Raspaud, M., D. Hoese, A. Dybbroe et al. (2018). ‘PyTroll: An Open-Source, Community-Driven Python Framework to Process Earth Observation Satellite Data’. In: *Bulletin of the American Meteorological Society* 99.7, pp. 1329–1336. DOI: 10.1175/BAMS-D-17-0277.1 (cit. on p. 51).
- Rees, W. G. (2012). *Physical Principles of Remote Sensing*. 3rd ed. Cambridge University Press. ISBN: 9781139017411. DOI: 10.1017/CB09781139017411 (cit. on p. 7).
- Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding*. World Scientific. DOI: 10.1142/3171 (cit. on p. 29).
- Romano, F. (2020). ‘Editorial for the Special Issue “Remote Sensing of Clouds”’. In: *Remote Sensing* 12.24, p. 4085. ISSN: 2072-4292. DOI: 10.3390/rs12244085 (cit. on p. 3).
- Ronneberger, O., P. Fischer and T. Brox (2015). ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’. In: *Medical Image Computing*

- and *Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4 (cit. on p. 38).
- Rosenfeld, D. and I. M. Lensky (1998). ‘Satellite-Based Insights into Precipitation Formation Processes in Continental and Maritime Convective Clouds’. In: *Bulletin of the American Meteorological Society* 79.11, pp. 2457–2476. DOI: 10.1175/1520-0477(1998)079<2457:SBIIPF>2.0.CO;2 (cit. on p. 29).
- Rossow, W. B. and R. A. Schiffer (1999). ‘Advances in Understanding Clouds from ISCCP’. In: *Bulletin of the American Meteorological Society* 80.11, pp. 2261–2288. DOI: 10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2 (cit. on p. 29).
- Saltikoff, E., K. Friedrich, J. Soderholm, K. Lengfeld, B. Nelson, A. Becker, R. Hollmann, B. Urban, M. Heistermann and C. Tassone (2019). ‘An Overview of Using Weather Radar for Climatological Studies: Successes, Challenges, and Potential’. In: *Bulletin of the American Meteorological Society* 100.9, pp. 1739–1752. DOI: 10.1175/BAMS-D-18-0166.1 (cit. on p. 17).
- Samek, W., G. Montavon, A. Vedaldi, L. K. Hansen and K.-R. Müller (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 1st ed. Springer Cham. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6 (cit. on p. 30).
- San Roman Pineda, I., K. Vincent, W. Colenbrander et al. (2025). *WISER EWSA King-Size Testbed Report*. Report. The University of Leeds. DOI: 10.48785/100/394 (cit. on p. 59).
- Santurkar, S., D. Tsipras, A. Ilyas and A. Madry (2018). ‘How Does Batch Normalization Help Optimization?’ In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf (visited on 2026-05-17) (cit. on p. 36).
- Sassen, K., Z. Wang and D. Liu (2008). ‘Global Distribution of Cirrus Clouds from CloudSat/Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) Measurements’. In: *Journal of Geophysical Research: Atmospheres* 113.D8. DOI: 10.1029/2008JD009972 (cit. on p. 21).
- Schoeberl, M. (2002). ‘The afternoon constellation: a formation of Earth observing systems for the atmosphere and hydrosphere’. In: *IEEE International Geoscience and Remote Sensing Symposium*. Vol. 1, 354–356 vol.1. DOI: 10.1109/IGARSS.2002.1025038 (cit. on p. 19).
- Spänkuch, D., O. Hellmuth and U. Görsdorf (2022). ‘What Is a Cloud? Toward a More Precise Definition’. In: *Bulletin of the American Meteorological Society* 103.8, E1894–E1929. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-21-0032.1 (cit. on p. 14).
- Stephens, G. L. (1978). ‘Radiation Profiles in Extended Water Clouds. II: Parameterization Schemes’. In: *Journal of the Atmospheric Sciences* 35.11, pp. 2123–2132. ISSN: 0022-4928, 1520-0469. DOI: 10.1175/1520-0469(1978)035<2123:RPIEWC>2.0.CO;2 (cit. on p. 28).

- Stephens, G., D. Winker, J. Pelon, C. Trepte, D. Vane, C. Yuhas, T. L'Ecuyer and M. Lebsock (2018). 'CloudSat and CALIPSO within the A-Train: Ten Years of Actively Observing the Earth System'. In: *Bulletin of the American Meteorological Society* 99.3, pp. 569–581. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-16-0324.1 (cit. on p. 21).
- Stephens, G. L., D. G. Vane, R. J. Boain et al. (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. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-83-12-1771 (cit. on p. 19).
- Stevens, B., M. Satoh, L. Auger et al. (2019). 'DYAMOND: The Dynamics of the Atmospheric General Circulation Modeled On Non-hydrostatic Domains'. In: *Progress in Earth and Planetary Science* 6.1, p. 61. ISSN: 2197-4284. DOI: 10.1186/s40645-019-0304-z (cit. on p. 60).
- Strapp, J. W., L. E. Lilie, T. P. Ratvasky, C. R. Davison and C. Dumont (2016). 'Isokinetic TWC Evaporator Probe: Development of the IKP2 and Performance Testing for the HAIC-HIWC Darwin 2014 and Cayenne Field Campaigns'. In: *8th AIAA Atmospheric and Space Environments Conference*. 8th AIAA Atmospheric and Space Environments Conference. Washington, D.C.: American Institute of Aeronautics and Astronautics. ISBN: 978-1-62410-433-6. DOI: 10.2514/6.2016-4059 (cit. on p. 51).
- Sullivan, T. J. (2015). 'Distributional Uncertainty'. In: *Introduction to Uncertainty Quantification*. Cham: Springer International Publishing, pp. 295–318. ISBN: 978-3-319-23395-6. DOI: 10.1007/978-3-319-23395-6_14 (cit. on p. 39).
- Tanelli, S., S. L. Durden, E. Im, K. S. Pak, D. G. Reinke, P. Partain, J. M. Haynes and R. T. Marchand (2008). 'CloudSat's Cloud Profiling Radar After Two Years in Orbit: Performance, Calibration, and Processing'. In: *IEEE Transactions on Geoscience and Remote Sensing* 46.11, pp. 3560–3573. DOI: 10.1109/TGRS.2008.2002030 (cit. on p. 19).
- Turk, F. J., S. E. Ringerud, A. Camplani et al. (2021). 'Applications of a CloudSat-TRMM and CloudSat-GPM Satellite Coincidence Dataset'. In: *Remote Sensing* 13.2264 (12). ISSN: 2072-4292. DOI: 10.3390/rs13122264 (cit. on p. 20).
- Vovk, V., A. Gammerman and G. Shafer (2022). *Algorithmic Learning in a Random World*. 2nd ed. Springer. ISBN: 978-3-031-06649-8. DOI: 10.1007/978-3-031-06649-8 (cit. on p. 40).
- Wallace, J. M. and P. V. Hobbs (2006). *Atmospheric Science. An Introductory Survey*. 2nd ed. Vol. 92. Elsevier. ISBN: 978-0-12-732951-2. DOI: 10.1016/C2009-0-00034-8 (cit. on pp. 3, 7).
- 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 (cit. on p. 19).
- WMO (2024). *OSCAR*. URL: <https://space.oscar.wmo.int> (visited on 2026-02-02) (cit. on p. 23).

- Xu, B., A. Banerjee and S. Gupta (2026). *Hardware Acceleration for Neural Networks: A Comprehensive Survey*. DOI: 10.48550/arXiv.2512.23914. arXiv: 2512.23914 (cit. on p. 33).
- Yang, Y., T. Dou, G. Xu, R. Zhou, B. Li, L. Husi, W. Wang and C. Xiao (2026). ‘FYAI: a Fengyun satellite-based dataset for atmospheric ice water path’. In: *Earth System Science Data* 18.2, pp. 1287–1305. DOI: 10.5194/essd-18-1287-2026 (cit. on p. 62).
- Zarr developers (2015). *Zarr: An implementation of chunked, compressed, N-dimensional arrays for Python*. URL: <https://zarr.dev> (visited on 2026-04-17) (cit. on p. 51).
- Zhang, A., Z. C. Lipton, M. Li and A. J. Smola (2023). *Dive into Deep Learning*. Cambridge University Press. URL: <https://D2L.ai> (visited on 2026-05-17) (cit. on p. 45).
- Zhou, D. K., A. M. Larar, X. Liu, W. L. Smith, L. L. Strow, P. Yang, P. Schlüssel and X. Calbet (2011). ‘Global Land Surface Emissivity Retrieved From Satellite Ultraspectral IR Measurements’. In: *IEEE Transactions on Geoscience and Remote Sensing* 49.4, pp. 1277–1290. DOI: 10.1109/TGRS.2010.2051036 (cit. on p. 12).

