Reliability of Neural Networks as Diagnostic Tools for Information Extraction

Ann-Christin Wörl

28.04.2026

DOI: 10.25358/openscience-15770

Institute of Computer Science

JG|U

Reliability of Neural Networks as Diagnostic Tools for Information Extraction

Dissertation

submitted for the award of the title

“Doctor of Natural Sciences”

to the Faculty of Physics, Mathematics and Computer Science
of Johannes Gutenberg University Mainz
in Mainz

Ann-Christin Wörl

Born in Groß-Gerau

Mainz, 28.04.2026

Ann-Christin Wörl
Reliability of Neural Networks as Diagnostic Tools for Information Extraction
Dissertation, 28.04.2026
Reviewers: [name removed] and [name removed]

Date of the oral examination: 02.07.2026

Johannes Gutenberg University Mainz
Computational Geometry
Institute of Computer Science
FB08
Staudingerweg 9
55128 Mainz

Abstract

Deep neural networks have become powerful tools for modeling complex systems in many scientific fields. Beyond their predictive capabilities, they enable the extraction of meaningful information from high-dimensional data and thus support scientific hypothesis generation. In this context, two complementary approaches can be distinguished. First, interpretability methods such as saliency maps attribute model predictions to input features and identify which variables are relevant for the task. Second, information can be probed implicitly by systematically restricting the input space. If predictive performance remains stable under such constraints, this indicates that the underlying information is robustly represented in the data.

This thesis investigates under which conditions neural networks can be used as reliable diagnostic tools for information extraction. We analyze the robustness of gradient-based saliency methods with respect to random model initialization, which can be identified as a significant source of variability in attribution maps, even when predictive performance remains unchanged. To quantify the variability, we introduce a signal to noise ratio framework that measures the stability of explanations across independently trained models and propose marginalization over the initialization distribution as a practical strategy to reduce attribution variability.

As an application, we study the ability of deep neural networks to reconstruct cloud structures based on atmospheric state variables and analyze how cloud-relevant information is encoded in the data. Our results show that this information is distributed in a redundant and complementary manner across physical variables and remains partially preserved under reduced input representations and generalization across regions, time shifts, and data sources. Applying the robustness framework in this setting further reveals that saliency patterns for the most relevant variables are considerably more stable than in classification tasks, enabling more reliable identification of physically meaningful structures.

Overall, this work clarifies under which conditions neural networks can serve as reliable tools for scientific information extraction. It shows that interpretability is inherently non-deterministic and must be evaluated in terms of robustness and uncertainty. Beyond attribution-based methods, it demonstrates that systematically probing the input space enables complementary and often more stable insights into how relevant information is structured within the data. Consequently, insights derived from data-driven models should not be interpreted as causal, but as evidence of structured information in the data that requires careful validation and uncertainty assessment.

Zusammenfassung

Tiefe neuronale Netze haben sich als leistungsstarkes Werkzeug zur Modellierung komplexer Systeme in vielen wissenschaftlichen Bereichen etabliert. Neben ihrer Vorhersagefähigkeit ermöglichen sie die Extraktion aussagekräftiger Informationen aus komplexen Daten und können so die Hypothesenentwicklung unterstützen. Dabei lassen sich zwei Ansätze unterscheiden: Zum einen untersuchen Methoden der Interpretierbarkeit den Zusammenhang zwischen Modellvorhersage und Eingabemerkmalen, etwa durch Salienzkarten. Zum anderen können Informationen implizit analysiert werden, indem der Eingaberaum systematisch eingeschränkt und die Güte der Vorhersage bewertet wird. Bleibt diese unter solchen Einschränkungen erhalten, deutet dies auf robust in den Daten enthaltene Informationen hin.

In dieser Arbeit untersuchen wir, unter welchen Bedingungen neuronale Netze als zuverlässige Diagnosewerkzeuge zur Informationsextraktion genutzt werden können. Dazu analysieren wir die Robustheit gradientbasierter Salienzmethoden gegenüber zufälliger Modellinitialisierung, die sich als wesentliche Quelle für Variabilität herausstellt – selbst bei gleicher Vorhersagegüte. Zur Quantifizierung nutzen wir das Signal-Rausch-Verhältnis über unabhängig trainierte Modelle hinweg und schlagen die Marginalisierung über die Initialisierungsverteilung als Strategie zur Reduzierung der Attributionsvariabilität vor.

Als Anwendungsfall dient die Rekonstruktion von Wolkenstrukturen aus atmosphärischen Zustandsvariablen mithilfe von neuronalen Netzen. Dies ermöglicht eine Analyse, wie Informationen über Wolkenbildung in den Daten kodiert sind. Die Ergebnisse zeigen eine redundante und komplementäre Verteilung dieser Information über verschiedene physikalische Variablen. Selbst reduzierte Eingaben oder eine Generalisierung über Regionen und Datenquellen beeinträchtigen die Vorhersagequalität nur geringfügig. Zudem zeigt sich, dass Salienzmuster in diesem Regressionssetting deutlich stabiler sind als in Klassifikationsaufgaben, wodurch eine zuverlässigere Interpretation physikalischer Zusammenhänge möglich wird.

Insgesamt zeigt diese Arbeit, unter welchen Bedingungen neuronale Netze als zuverlässige Werkzeuge für die wissenschaftliche Informationsextraktion dienen können. Sie macht deutlich, dass Interpretierbarkeit nicht deterministisch ist und hinsichtlich Robustheit und Unsicherheit bewertet werden muss. Über attributionsbasierte Methoden hinaus liefert die systematische Untersuchung des Eingaberaums ergänzende und oft stabilere Einblicke. Erkenntnisse aus datengesteuerten Modellen sollten nicht kausal interpretiert werden, sondern als Hinweise auf strukturierte Informationen, die sorgfältiger Validierung und Unsicherheitsbewertung bedürfen.

Contents

1 Introduction

1.1 Motivation and Problem Statement

Deep neural networks have demonstrated a remarkable ability to solve problems that were previously considered unsolvable. Their ability to approximate highly complex functions allows them to discover patterns and relationships in data beyond the scope of traditional statistical or physically motivated methods. As a consequence, neural networks have become powerful analytical tools in many scientific disciplines. Their applications span from histopathology, where morphological features in tissue samples provide cues to the presence or absence of cancer, to atmospheric physics, where they infer cloud structures from coarse-scale physical fields.

However, the remarkable predictive abilities of neural networks have an important drawback: they are difficult to interpret. While these models demonstrate superior accuracy, their internal decision-making processes remain unclear. This is particularly challenging in fields where the objective is not only to make accurate predictions, but also to gain a deeper understanding of underlying physical or biological processes. As a result, interpretability has emerged as a central research topic. Instead of asking how a classifier implements its decision, researchers are interested in identifying which input features carry relevant information about the phenomenon itself. Saliency methods are one possible approach to this problem, attempting to attribute model output to specific regions or variables of the input. Nevertheless, the usefulness of such attribution methods depends on their reliability. When the goal is to draw scientific conclusion, such as identifying the atmospheric variables that drive cloud formations, one must assume that saliency maps are stable and robust.

This dissertation investigates when and under which conditions neural networks can be used as reliable diagnostic instruments for extracting information from complex, high-dimensional data. Here, information is understood in an operational sense: as structured input-output dependencies that enables accurate prediction while generalizing across spatial and temporal

shifts and yielding stable attributions. Under this definition, information is not merely predictive performance, but reproducible and transferable structure captured by the model.

We address random initialization of neural networks as a potential source of attribution fluctuation and provide a marginalization framework to reduce this source of instability. Cloud physics serves as a challenging real-world application to stress-test interpretability methods under complex, structured data. We employ deep neural networks to predict cloud structures from atmospheric state variables and use this setup to explore physically meaningful dependencies relevant to cloud formations.

First, we investigate how accurately neural networks can reconstruct cloud patterns from large-scale reanalysis and forecast data, and which input variables contribute most consistently to this task. Second, we assess the robustness of these predictions across regional and temporal shifts to distinguish genuine physical dependencies from potential memorization effects. In this context, generalization performance functions as an empirical proxy for transferable information content. Finally, we apply the proposed reliability framework to our cloud-prediction setting to evaluate the stability of attribution-based scientific conclusions under model perturbation.

Together, these steps clarify both the potential and the limitations of neural networks as instruments for scientific information extraction. This work thereby provides methodological guidance on when and how neural network explanations can be interpreted as insights into underlying physical processes rather than artifacts of model instability.

1.2 Thesis Structure

This thesis investigates how neural networks can be used as tools for generating scientifically meaningful insights into physical processes. To this end, we combine gradient-based attribution methods with the use of neural networks as information extraction models. Since scientific interpretation requires reliability, a central focus of this work is the systematic assessment of the robustness and stability of these interpretability approaches.

Chapter 2 introduces the theoretical and methodological background of the thesis. It provides an overview of neural network interpretability methods,

outlines the foundations of atmospheric modeling, and reviews the role of machine learning in this domain.

The main contribution of this thesis is structured into three parts:

The first part (Chapter 3 to 6) takes a methodological perspective. It investigates how random initialization affects gradient-based attributions and evaluates the resulting saliency maps. This part establishes a framework for assessing interpretability robustness independent of a specific physical application.

The second part (Chapter 7 to 10) turns to an applied setting in atmospheric cloud modeling. Here, neural networks are trained to predict cloud structures from atmospheric state variables. The analysis focuses on identifying which input variables contain transferable information about cloud formation and how this information is spatially and across variables distributed.

The third part (Chapter 11 to 14) connects the methodological and applied components of this thesis. The robustness framework developed in Part I is applied to the cloud prediction setup from Part II in order to evaluate the stability of attribution-based scientific conclusions. The thesis concludes with a general discussion of the scientific implications and an overall synthesis of the findings.

1.3 Publications

This thesis is based on the following publications, from which parts of the content are directly incorporated.

“Initialization Noise in Image Gradients and Saliency Maps”
A.-C. Woerl, J. Disselhoff, and M. Wand
In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1766–1775.

This publication [62] investigates the influence of random initialization on a range of commonly used saliency methods, with the aim of quantifying the robustness of input attributions. Furthermore, a framework based on stochastic integration is proposed to mitigate initialization-induced artifacts. This work was published at the Conference on Computer Vision and Pattern Recognition (CVPR).

The work was developed in collaboration with Michael Wand. The conceptualization of the research idea was carried out jointly. My contribution includes the design of the experiments, the complete implementation, the execution and analysis of all experiments, and the writing of the main parts of the manuscript. Contributions by Jan Disselhoff that are not directly relevant to the present dissertation have been omitted.

"What Atmospheric Data Knows About Clouds: An Information-Theoretic Approach"
A.-C. Woerl, M. Wand, and P. Spichtinger
In preparation

This publication [65] applies neural networks to analyze the information content of atmospheric data with respect to cloud formation, focusing on the distribution and redundancy of predictive information across physical variables. In addition, a framework for the analysis of high-dimensional attribution data is developed. This work is currently in preparation, but preliminary results were published in [64].

The development of the work was a collaboration with Michael Wand and Peter Spichtinger. The conceptual design was developed jointly, while I was responsible for the methodological design, implemented the models and analysis pipeline, conducted all experiments, and wrote the main parts of the manuscript.

In both publications, I was the main contributor with responsibility for the core methodological development, implementation, and data analysis.

In addition, I contributed to several publications in the field of medical imaging [63, 44, 15]. In these works, I was involved in software implementation and data analysis. As these contributions are not directly related to the topic of this dissertation, they are not discussed further.

2 Foundations

This chapter introduces the theoretical and methodological foundations of this thesis. It first reviews central concepts of neural network interpretability and how these can be applied to cloud prediction and weather forecasting. Subsequently, it discusses the use of machine learning in atmospheric physics and weather forecasting, followed by a summary of established knowledge about clouds from atmospheric models. Together, these foundations provide the conceptual basis for the methodological and application-oriented parts that follow.

2.1 Interpretability of Neural Networks

Deep Neural Networks achieve high predictive performance, but their internal decision process is often difficult to understand. This lack of transparency is particularly problematic in scientific applications, where a model is not only expected to produce accurate predictions, but also to support the extraction of meaningful information from the data. In such settings, interpretability can increase confidence in model outputs and provide hints about which structures in the input are associated with the prediction [29, 11, 70, 32].

At the same time, interpretability is not a single, universally agreed concept. The literature distinguishes between different notions such as transparency of the model itself, post-hoc explanations of an already trained black box, local explanations for individual predictions, and global explanations of overall model behavior [29, 11]. In this work, we focus on *post-hoc local interpretability*.

More specifically, we consider feature-attribution methods, which assign relevance scores to input features in order to explain an individual model prediction. For image-like or spatially structured data, these scores are commonly visualized as saliency maps. Such approaches are especially attractive

in scientific applications because they promise to connect model predictions back to physically meaningful structures in the input [32, 70].

A useful distinction can be made within feature-attribution methods between *gradient-based* and *perturbation-based* approaches. Gradient-based methods estimate relevance from derivatives of the output with respect to the input or intermediate representations and therefore provide a local view of model sensitivity. Perturbation-based methods, in contrast, estimate feature importance by systematically modifying, masking, or removing parts of the input and measuring the resulting change in prediction [2, 70].

The following sections introduce these two methodological classes of saliency methods and discuss their conceptual foundations as well as their practical limitations.

2.1.1 Gradient-based methods

Gradient-based explanations start from the observation that, for a differentiable model, the gradient of an output score with respect to the input describes how sensitively the score changes under infinitesimal changes of each input feature. In this sense, the gradient provides a first-order local approximation of feature importance around the current sample. The most direct method is the “Vanilla” Gradient (also called Input Gradient), where the saliency map is simply given by the derivative of the output with respect to the input [49]. Large absolute gradient values indicate directions in input space along which the prediction would change most strongly.

Closely related variants differ mainly in terms of how gradients propagate through nonlinearities during backpropagation. Deconvolution [68] and “guided” backpropagation [53] modify the backward pass by suppressing negative signals at ReLU units, which often produces visually sharper maps than raw gradients. However, these methods remain based in local derivatives and therefore inherit the same basic limitation: they characterize only the immediate neighborhood of the current input rather than the full contribution of a feature over a larger region of input space.

A well-known practical problem is that raw gradient maps are often noisy and difficult to interpret visually. SmoothGrad [51, 35] addresses this by averaging saliency maps over multiple noisy copies of the same input. The underlying idea is not to change the explanation target, but to smooth the lo-

cal sensitivity estimate in input space: features that are consistently important across small perturbations remain visible, while high-frequency fluctuations are reduced. This often improves visual readability substantially. Related work has argued that part of the apparent noise in saliency maps originates from the propagation of irrelevant signals through ReLU activations, and proposes modified backpropagation rules such as Rectified Gradient to suppress these effects [23].

Another important line of gradient-based attribution works with higher-level feature maps rather than the raw input. GradCAM [45] computes the gradient of the target output with respect to convolutional feature maps and aggregates these gradients into a coarse localization heatmap. Compared with pixel-level gradients, this yields explanations that are often easier to interpret semantically, because they are expressed at a representation level where the network has already combined low-level patterns into more meaningful structures. The trade-off is that such maps are spatially coarser, especially in architectures with downsampling or pooling.

Despite their popularity, gradient-based methods must be interpreted carefully. By construction, they measure local sensitivity, not necessarily actual feature contribution over a finite range. In particular, they can underestimate the importance of features in saturated regions where the output is already insensitive to small local changes. This is one of the main motivations for non-local attribution methods such as integrated gradients or reference-based propagation rule [56, 48].

2.1.2 Perturbation-based methods

Perturbation-based methods estimate feature importance more directly: instead of relying on an infinitesimal derivative, they modify or remove parts of the input and observe how strongly the prediction changes. This makes them conceptually appealing, because the explanation is tied to an explicit intervention on the model input. More generally, many of these methods can be understood as non-local approaches, since they compare the prediction at the observed sample with predictions at altered or reference inputs [70].

A central representative is integrated gradients [56]. Rather than evaluating the gradient only at the input itself, the method integrates gradients along a path from a baseline input to the actual sample. This addresses the saturation problem of local gradients by accumulating contributions over a larger

portion of input space. Closely related is DeepLIFT [48], which compares neuron activations to a reference activation and propagates contribution scores based on activation differences instead of local derivatives. Both methods therefore depend on a chosen baseline or reference state, which can substantially influence the resulting attribution.

A second branch of perturbation methods explains predictions through explicit feature masking or replacement. In computer vision, this includes occluding image regions, blurring them, or replacing them with a background pattern, and then measuring how strongly the target score drops [68, 16, 22]. Such approaches are attractive because they make the notion of "importance" operational: a region is important if removing it substantially changes the prediction. However, the result can depend strongly on how the perturbation is implemented, since artificial masks may shift the input away from the data distribution and create model responses that are not solely attributable to the removed feature itself.

Local surrogate approaches follow a related philosophy. LIME [39] explains an individual prediction by sampling perturbed versions of the input around the instance of interest and fitting an interpretable surrogate model locally. SHAP [31] places this idea into a game-theoretic framework and computes additive feature attributions based on Shapley values, thereby providing a principled notion of local contribution under explicit assumptions. These methods are widely used because they are flexible and can be applied to many model classes. At the same time, the explanation is not obtained directly from the original network, but from a local approximation or feature-coalition analysis, so the result depends on the sampling strategy, feature representation, and assumptions about missingness or feature independence.

For image models, perturbation methods tend to be more intuitive than purely local gradients, as they explicitly determine necessary or sufficient evidence for a prediction. This motivates more advanced masking formulations such as meaningful perturbation [16] and optimization-based saliency methods like iGOS++ [22], which search for masks that either maximally remove or preserve predictive evidence under regularization constraints.

2.1.3 Limitations

Although attribution maps are widely used, their reliability remains an active research question. A central conceptual limitation is that post-hoc ex-

planations do not automatically reveal the true internal reasoning of a deep model. As emphasized in the interpretability literature, explanations can be useful for debugging, hypothesis generation, or qualitative inspection, but they should not be mistaken for a full causal account of model behavior [29, 11, 42].

For saliency methods specifically, several studies have identified severe robustness and validity issues. Adebayo et al. [1] showed that some popular saliency methods are surprisingly insensitive to learned model parameters and even to the data-generating process: after randomizing network weights or labels, the resulting maps can remain visually similar. This finding is particularly important because it demonstrates that visually plausible heatmaps are not necessarily faithful explanations of learned behavior. Similarly, Kindermans et al. [24] showed that some attribution methods violate input invariance, meaning that explanations can change under transformations that leave the model output unchanged.

Another major concern is vulnerability to adversarial or otherwise unnoticeable changes. Explanations can be significantly altered while maintaining the original prediction. This means that saliency maps may be far less stable than they appear [10, 18]. More broadly, the statistical nature of deep learning itself must be taken into account: trained models depend on random initialization, stochastic optimization, data order, and other sources of variation. Recent work in interpretable machine learning has therefore argued that explanations should not be treated as deterministic objects, but analyzed with respect to uncertainty, variability, and robustness across model instances [34, 33].

These concerns are highly relevant for the present work. Our focus is not to argue that saliency methods are generally unusable, but rather to examine one specific and practically important source of uncertainty: variability induced by initialization noise. This question is motivated by the observation that many methods produce visually cleaner maps through smoothing, integration, or higher-level aggregation, which can easily create the impression of stability. However, visual smoothness does not imply robustness across independently trained models. In Part I, we therefore study to what extent common gradient- and perturbation-based methods vary under repeated training runs and show that initialization noise can remain substantial even for methods such as SmoothGrad, GradCAM, integrated gradients, and SHAP/LIME-style explanations.

Finally, a growing body of work suggests that interpretability also depends on the features learned during training. In particular, models trained under robust, generative, or explicitly interpretable regimes often yield explanations that are more semantically aligned with human expectations [4, 58]. This motivates applying attribution methods not only as a visualization tool, but also as a means of studying how training objectives and model classes shape the features that are ultimately used for prediction.

2.2 Machine Learning in Cloud Prediction and Weather Forecasting

Numerical Weather Prediction (NWP) involves using physical models based on fluid dynamics and thermodynamics equations to simulate the atmospheric evolution. These models take initial states from reanalysis data and use time-stepping schemes to forecast future conditions. While NWP has led to major advances in forecasting skills, it struggles with sub-grid-scale processes, especially those involving cloud microphysics. Traditional approaches rely on bulk microphysical parameterization to estimate quantities such as optical thickness or cloud cover. However, these models tend to fail to resolve the fine-scale spatial structure. This has motivated the use of machine learning (ML) both as a supplement and alternative for traditional NWP in tasks such as cloud estimation, parameterization and nowcasting.

Recent advances in ML have opened new pathways for modeling cloud processes and weather forecasting. Convolutional neural networks (CNN), in particular, have been applied to cloud classification [67], satellite-based cloud segmentation [8] and short-term prediction of cloud evolution using satellite data [36]. However, most of this work either treats the problem as image translation or classification, without grounding their model in physical variables like reanalysis data.

Beyond CNNs, generative models are emerging as powerful forecasting tools. *SATcast* applies a conditional diffusion model to predict vapor channel evolution from satellite data and FuXi NWP forecasts, achieving skillful cloud predictions up to 8 hours ahead [7]. Chase et al. [6] use score-based diffusion models to predict clouds and precipitation up to three hours using satellite data. These methods generate sharper, more realistic outputs compared to deterministic CNNs and outperform optical flow extrapolation in both RMSE and structural similarity.

Hybrid approaches try to blend physical simulations with neural components [38] or learned parameterization of cloud processes [5]. Meanwhile, large-scale ML forecasting systems like *Pangu-Weather* [3], *FuXi* [55] or *GraphCast* [27] demonstrate that deep models can rival NWP skill in broad atmospheric fields, but their treatment of cloud microphysics remains limited, often relying on learned correlations rather than physical drivers.

Purely neural approaches combine satellite images and weather forecast data obtained from deep-learning weather [55] predict future cloud fields by using diffusion models [7]. While promising, these approaches typically lack interpretability.

Most prior work focuses on end-to-end forecasting or cloud classification, often treating the problem as image translation without physical grounding; our approach reframes cloud reconstruction as an information extraction task. Rather than replacing physical models, we investigate how much predictive signal is embedded in coarse atmospheric fields and which physical variables drive cloud structures over space and time. By evaluating attribution consistency across space and time, we aim to link learned representations back to physically meaningful drivers, bridging data-driven modeling and scientific insights. The results can be found in Part II.

2.3 (Cloud) Information in Physical Models of Atmospheric Conditions

Classical numerical weather prediction and climate models operate on comparatively coarse spatial grids and therefore explicitly represent only motions and processes on the resolved scales. In current large-scale atmospheric models, the effective horizontal resolution is typically on the order of several tens of kilometers. On these scales, the fundamental state variables of the atmosphere, such as horizontal and vertical wind components (u, v, w), temperature (T), pressure (p), and water vapor (q_v), are represented explicitly. Cloud processes, however, largely occur on smaller scales and are therefore not resolved directly. Instead, their effects on the resolved atmospheric state must be represented through parameterizations.

For non-convective clouds, the formulation is usually based on rate equations in terms of ordinary differential equations. The resulting model variables typically include mean mass concentrations of water vapor (q_v), liquid

water (q_l), and ice (q_i). For practical interpretation and output, the condensed phases are often expressed as cloud liquid water content ($clwc$) and cloud ice water content ($ciwc$), respectively. While the details of cloud processes may dramatically differ between parameterizations [see, e.g., 41], the basic principles are included in all formulations.

At a basic level, cloud formation is controlled by the large-scale thermodynamic state of the atmosphere. The resolved variables p, T, and q_v determine whether air is close to saturation and thus whether phase transitions between water vapor, liquid water, and ice are favored. Vertical motion is particularly important in this context: upward motion leads to adiabatic cooling, which increases relative humidity and can initiate condensation and cloud formation. Conversely, large-scale subsidence tends to warm and dry the air and therefore suppress cloud formation. From this large-scale perspective, one would therefore expect clouds to be preferentially associated with ascending motion and to be less likely in regions dominated by subsidence.

However, cloud evolution also depends on a wide range of small-scale microphysical processes (as e.g. nucleation, collision, breakup). These processes can produce complex cloud structures and organization patterns [e.g. 54]. As a result, the relationship between resolved atmospheric conditions and the resulting cloud field is neither simple nor deterministic at the local scale. Even if the large-scale environment contains relevant information about cloud formation, this information is indirect, distributed, and filtered through unresolved dynamics and microphysics.

This creates a central challenge for physically based cloud prediction. On the one hand, clouds are not explicitly resolved and their detailed structure cannot be inferred directly from coarse atmospheric fields alone. On the other hand, those same coarse-scale fields do contain physically meaningful information about the thermodynamic and dynamical conditions under which clouds form and evolve. Different cloud regimes and characteristic cloud patterns are indeed associated with different large-scale environments across the globe. The key question is therefore not whether coarse-scale atmospheric variables contain cloud-relevant information, but how much of this information is present, how it is distributed across variables and spatial context, and to what extent it can be extracted for prediction.

Part I

Initialization Noise in Saliency Maps

3 Overview

Chapter 3 to 6 are based on the paper:

“Initialization Noise in Image Gradients and Saliency Maps”
A.-C. Woerl, J. Disselhoff, and M. Wand
In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1766–1775.

Several text passages and images are taken directly from there. The experimental setting was designed in collaboration with Micheal Wand and Jan Disselhoff. I implemented the whole software pipeline, performed and evaluated the experiments, and wrote the main parts of the manuscript. Due to duplications in related work, the foundations are transferred to Chapter 2 and can be reviewed there.

This part provides a methodological analysis of neural networks as instruments for scientific information extraction. We focus in particular on gradient-based saliency methods as a class of attribution techniques and examine how variability induced by model initialization affects their reliability.

The results demonstrate that attribution maps can exhibit substantial instability across randomly initialized but otherwise equivalent models. Consequently, a systematic treatment of initialization-induced variability is required if neural networks are to be used as robust tools for extracting scientifically meaningful information.

3.1 Abstract

In this work, we examine gradients of logits of image classification CNNs by input pixel values. We observe that these fluctuate considerably with the initial random initialization of the networks (after training). We extend our study to gradients of intermediate layers, obtained via GradCAM, as well as

popular network saliency estimators such as DeepLIFT, SHAP, LIME, Integrated Gradients, and SmoothGrad. While empirical noise levels vary, qualitatively different attributions to image features are still possible with all of these, which comes with implications for interpreting such attributions, in particular when seeking data-driven explanations of the phenomenon generating the data. Finally, we demonstrate that the observed artifacts can be removed by marginalization over the initialization distribution by simple stochastic integration.

3.2 Introduction

Deep neural networks have revolutionized pattern recognition, detecting complex structures at accuracies unheard of just a few years back. Unsurprisingly, the newly gained ability to model complex phenomena comes at costs in terms of interpretability — it is usually not obvious how nonlinear, multi-layer networks reach their conclusions. Correspondingly, a lot of research has focused on developing *interpretation* methods for explaining how deep networks make decisions [21], and this often takes the form of *attributing* decisions to subsets of the data. In the case of image classification, this usually leads to *saliency maps* highlighting the image area containing decisive information [49, 39, 48, 45, 56, 51].

Strong classifiers trained from example data combined with suitable attribution methods have opened up a new approach to empirical research: *understanding phenomena* by *interpreting learned models* [17]. We often know of posterior outcomes (for example, tumor growth rates or treatability with certain medication) but do not understand how these are related to prior data (say, findings from histological tissue samples). If we are able to train a strong classifier that can predict posterior outcomes from prior data, an attribution method could potentially explain which aspects of the data predict this outcome (for example, which visual features in the histology indicate a negative or positive therapeutic prognosis [63]), thereby providing new insight into the phenomenon at hand.

For these kinds of research approaches, the classifier might only be an auxiliary tool: In terms of attribution, we are not primarily interested in explaining how *the classifier* reaches its decision (which, of course, would be highly relevant when studying potential data leakage or the fairness of decisions [46, 9]), but our actual goal is to accurately characterize which *features*

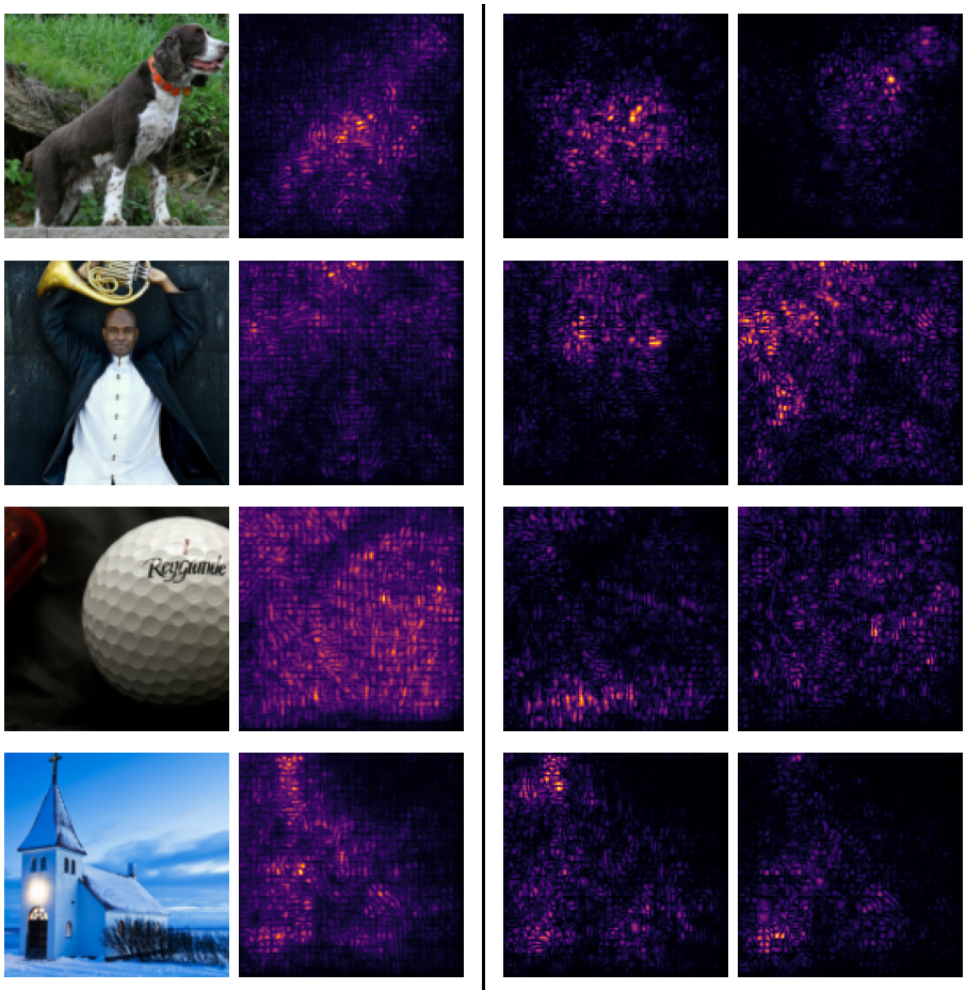


Fig. 3.1.: Logit-by-image gradients (ResNet18 on "Imagenette" [@2]). First column: reference image; second column: mean over 50 models, column 3-4: single models with random initialization.

in the data are related to *the phenomenon* to be explained. Ultimately, it is, of course, impossible to be sure whether an ad-hoc classifier (even with great statistical performance and hypothetical perfect attribution) actually does exploit all relevant information (and only this), but we would, of course, in such cases, make an effort to avoid wrong or incomplete information or misattributions that we are already aware of.

The main insight and contribution of this chapter is to point out one such source of fluctuations in attributions, the impact of which, to the best of our knowledge, has not yet been documented in literature so far: *In nonlinear CNNs, image gradients of network outputs can contain significant initialization noise* (Fig. 3.1). The level of such noise, i.e., information unrelated to

the data itself, *often exceeds the level of the attribution signal*, and variability includes coarse-scale variations that could suggest qualitatively varying attributions. Surprisingly, this still holds (and can even be worse) for more sophisticated attribution techniques, including popular approaches such as SHAP [31], LIME [39], DeepLIFT [48], or Integrated Gradients [56]. Even class activation maps (including top- and intermediate-level GradCAMs [45], the former to a lesser degree) can be affected by noise to an extent that could alter coarse-scale attribution.

Exploring the phenomenon further, we observe that gradient initialization noise grows with depth, is rather weak in simple convex architectures (linear softmax regression), and dampened stochastically for wide networks (as suggested by the known convexity in the infinite-width limit [14, 30]). This indicates that nonlinearity and nonconvexity might play an important role in causing the problem by either amplifying numerical noise (for example, from SGD) or convergence to different local minima.

We further show that initialization noise artifacts can be removed by marginalization over the initialization distribution [61], which in practice can be implemented with simple stochastic integration: By averaging the results of tens of independently initialized and trained networks, signal to noise levels can be brought to acceptable levels.

4 Method

This chapter presents the methodological framework used to analyze the robustness of saliency methods. It introduces the attribution techniques considered in this work, describes the marginalization approach used to reduce initialization-induced artifacts, and defines the signal to noise framework that serves as the main quantitative measure of attribution stability.

4.1 Saliency Maps

We begin by giving some formal definitions. We denote the *input* of the network as vector $x \in \mathbb{R}^d$. A neural network computes a map $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$, where C is the dimension of the output; in our experiments we consider only classification, so C also denotes the number of classes. $\theta \in \mathbb{R}^{d_p}$ describes the parameters of the network. In practical settings, the values for θ are the result of a training process. Formally, the result is dependent on the initialization weights θ_0, the data set D, and training hyperparameters, such as *optimizer, batchsize, learning rate*, as well as the random choices such as the ordering of batches. We use T to denote the function that transforms initial into final parameters using data D, $\theta = T(\theta_0, D)$. In this view, both θ_0 and T are random variables, drawn from distributions $\theta_0 \sim p(\theta_0), T \sim p(T)$.

We now define a *saliency method* as a deterministic map

$$S_{f_\theta} : \mathbb{R}^d \to \mathbb{R}^d, \tag{4.1}$$

that takes an input vector x and returns an output of the same shape, with the goal of "explaining" the behavior of f_θ by highlighting salient entries in x [1].

Taking the influence of training and initialization into account, the salience mapping becomes

$$S_{f_\theta}(x) = S_{f_{T(\theta_0, D)}}(x), \tag{4.2}$$

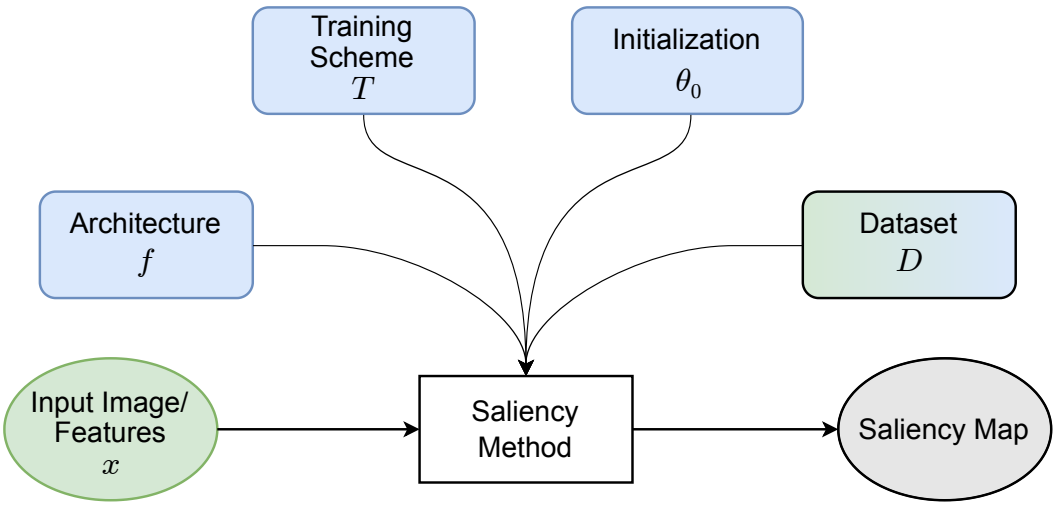


Fig. 4.1.: Typical influences on saliency methods. When such methods are used to explain phenomena, the influence of factors extrinsic to the problem (here in blue) should be minimized. While biases in a dataset can be seen as extrinsic factors, the underlying true distribution is intrinsic to the phenomenon.

i.e., f_θ is completely determined given training randomness $T \sim p(T)$, initialization $\theta_0 \sim p(\theta_0)$, and dataset D. In this formulation, the result of the saliency map depends not only on the input x, but on the training scheme, initialization, dataset, and architecture encoded in f. (see Fig. 4.1)

Most of these values describe unwanted influences on the saliency map if we are looking for patterns belonging to the problem that is modeled. In other words, the choice of architecture or initialization does not influence the underlying patterns of the phenomenon of interest, but it still has an influence on the saliency map. In the following, we will call factors that can influence the saliency map but are independent of the phenomenon **extrinsic**, while factors that belong solely to the phenomenon are **intrinsic**. Depending on the specific problem, even the dataset and its sampling can contain extrinsic information we might want to remove; however, in this work, we treat D as fixed, non-random information.

Note that saliency methods that are implementation agnostic – or even treat the model as a blackbox – can still be influenced by such extrinsic factors, as architecture and initialization fundamentally influence which function is learned by the model. In general, we want to minimize the effect of these extrinsic factors in order to produce a clear signal of the phenomenon of interest.

4.2 Bayesian Marginalization

In a Bayesian view, irrelevant random factors can be removed by marginalization [61, 17], i.e., integrating the joint distribution over all variants. For noise introduced from initialization and random training choices, this is simple. We just estimate the mean saliency map as

$$\mu = \mathbb{E}_{\theta_0 \sim p(\theta_0), T \sim p(T)} \left[S(f_{T(\theta_0)}, x) \right] \tag{4.3}$$

Note that $\mu \in \mathbb{R}^d$ is again vector in input shape (i.e., in our experiments, an image). In order to quantify the level of noise, we also compute the pixel-wise variance $\sigma \in \mathbb{R}^d$ of the saliency maps:

$$\sigma^2 = \mathbb{E}_{\theta_0 \sim p(\theta_0), T \sim p(T)} \left[S(f_{T(\theta_0)}, x) - \mu \right]^2 \tag{4.4}$$

Remark: An ideal approach should also remove the extrinsic influence of the choice of architecture. However, marginalization over architecture is difficult conceptually (what would be an appropriate general set of architectures?) and computationally. For this reason, we restrict ourselves to removing initialization noise (and gain the insight that this already has substantial influence). Our computation does, however, capture training randomness due to stochastic batch gradient descent. In our discussion, we will simply subsume this under the notion of "initialization noise" for simplicity. Hyperparameters are also still treated as non-random constants.

4.3 Stochastic Integration

The integrals in Eq. (4.3) and Eq. (4.4) are extremely high-dimensional; thus we resort to stochastic integration. As most saliency methods provide bounded results (either implicitly or explicitly) the per-pixel variance is also bounded [47]. Thus, the central limit theorem applies and we obtain stochastic convergence of the mean (Eq. (4.3)) at a rate of $\mathcal{O}(1/\sqrt{n})$ when averaging over n independently initialized and trained models. Experiments (Fig. 5.1) are also visually consistent with estimates stabilizing quickly.

4.4 Signal to Noise Ratio

In order to quantify variability due to initialization noise, we compute a signal to noise ratio (SNR), given by.

$$\text{SNR} = \frac{||\mu||_2}{||\sigma||_2}, \tag{4.5}$$

where $||x||_2 := \sqrt{\sum_i x_i^2}$ denotes the standard ℓ_2-norm for vectors, i.e., we divide the norm of μ by the total standard deviation over all pixels, respectively.

The SNR does not reveal the structure of the noise; a high SNR could hypothetically also be caused by high-frequency fluctuations that do not lead to qualitatively different judgments. We therefore also perform a visual examination to confirm the potential for variability in structure. We do not try to formalize this, as perceptual image comparison metrics are still an open field of research [69] and an objective quantification of structural differences might, if at all, only be possible with specific knowledge of the application domain.

5 Experiments

We now assess the influence of initialization on saliency maps experimentally. All experiments were conducted on a single commodity PC equipped with an AMD Ryzen 9 5900X CPU, 128GB of RAM and an Nvidia GeForce RTX 3090 GPU. Deep networks were implemented using PyTorch.

As *data sets*, we consider *CIFAR10*[26], *FashionMNIST* [66] and "*Imagenette*" [@2] at a resolution of 128×128. The latter is a subset of 10 distinct classes from the Imagenet dataset. As our experiments require a large number of models to be trained, we chose Imagenette as a compromise between complexity and computational cost.

In terms of networks, we restrict ourselves to the *VGG19* architecture [50] and the *ResNet18* and *ResNet101* architectures [20] as popular feed-forward CNN architectures, as well as a simple custom CNN-design for testing the influence of width and depth. In all three cases, we use the standard implementation from torchvision; for VGG19 on Imagenette we replace ReLUs with softplus to examine the effect of alternative activation functions.

5.1 Analysis of Input Gradients

First, we investigate the behavior of input gradients, i.e.,

$$S_{f_\theta}(x) := \nabla_x f_\theta(x) \tag{5.1}$$

As function f, we consider the full network up to the logits; we omit the final softmax-layer, as it can be easily seen that the input gradients vanish at class-label output probabilities of zero or one, which renders the probability gradients even less suitable for saliency detection. To prove this, let $\sigma(z)_i = {e^{z_i}}/{\sum_j e^{z_j}}$ be the softmax-output with $z_i := f_\theta(x)_i$ denoting the i-th logit.

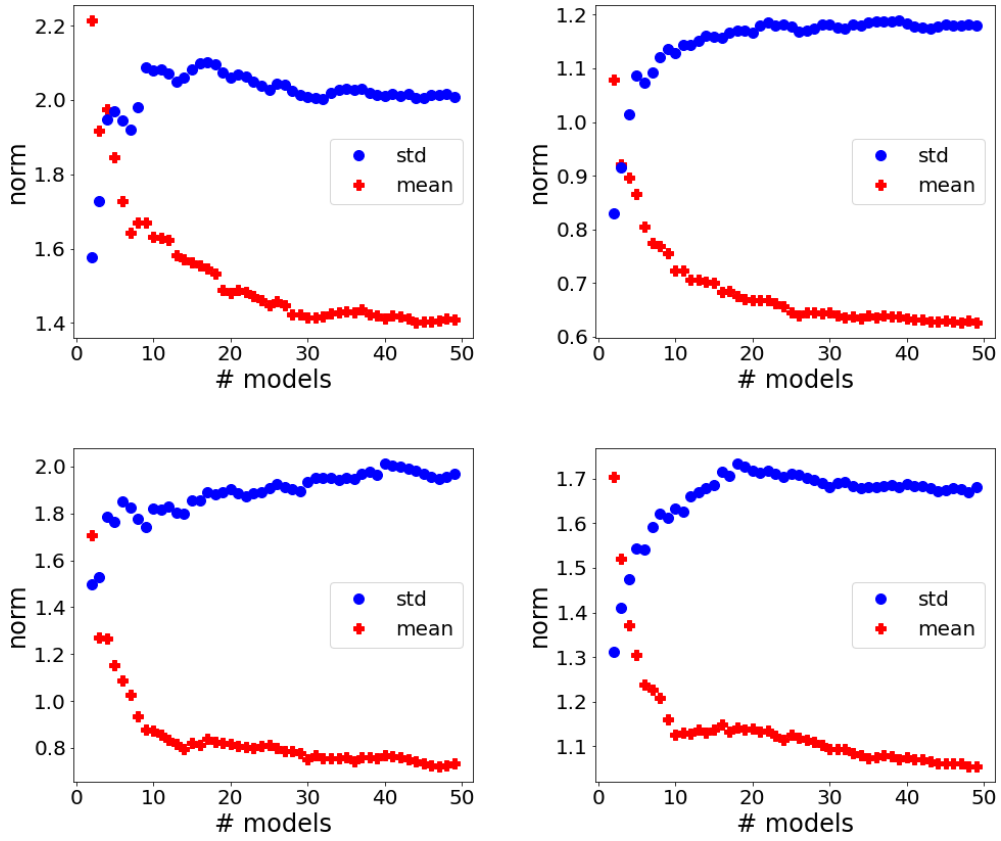


Fig. 5.1.: Examples for experimentally calculated approximations for μ and σ (gradient maps from Fig. 3.1). Increasing the number of models quickly leads to a stable estimate for both.

Following the chain rule, the input gradient for the class-label output is then given by

$$\nabla_x \sigma(z)_i = \nabla_x \left[\frac{e^{z_i}}{\sum_j e^{z_j}} \right] \tag{5.2}$$

$$= \sigma(z)_i \left[\nabla_x z_i - \sum_j \sigma(z)_j \nabla_x z_j \right]. \tag{5.3}$$

As the class-label output probability approaches zero or one, Eq. (5.3) converges towards zero. We compute the logit-gradient analytically using the built-in automatic differentiation of pytorch.

Gradients are not only an early saliency method but are also often used for empirical studies of deep networks as well as as a basis for more elaborate methods [2]. Therefore, their behavior is of particular interest.

Fig. 5.2.: signal to noise ratio (SNR) of Imagenette classes. The values are produced using ResNet18. Each colored dot corresponds to a single sample. The dotted line describes the overall average SNR.

We train 30 ResNet18 models with random He initialization up to a validation accuracy of 87.4 - 89.1 %. For training, we use ADAM-Optimizer [25] and One-Cycle-LR-Scheduler [52] with a maximum learning rate of 0.01. We chose this combination for its fast convergence and good generalization performance, exceeding for example simple step-decay+SGD by a substantial margin at roughly one-tenth of the cost.

Signal to noise ratios: Table 5.1 provides mean SNR-values over the validation part of the datasets based on 30 models. We consistently obtain values below one (in the range of $0.44...0.59$), i.e., the noise is more prominent than the actual saliency signal, roughly a factor of two.

Fig. 5.1 displays the norm of the mean ($||\mu||_2$, red/lower curve) and standard deviation ($||\sigma||_2$, blue/upper curve) over an increasing number of models. The four plots correspond to the samples in Fig. 3.1. Both the norms of mean and standard deviation appear to converge already after averaging a small number of models. The characteristic drop in the norm of the mean indicates the presence of uncorrelated noise.

Fig. 5.2 illustrates the SNR for all images of the validation set of Imagenette over 30 ResNet18-models. The 10 different classes are plotted in different colors. The mean SNR over all images is approximately 0.59, with noise being 70% larger than the signal (indicated by the black dashed line). The SNR varies for different classes; however, an analysis of the softmax predictions shows that there is no correlation between the quality of the prediction and the SNR. Repeating this experiment with VGG19 and ResNet101 shows a similar pattern of the SNR for different classes but at an even worse (lower) level, which might already suggest that deeper architectures tend to be noisier.

Qualitative comparison: Fig. 3.1 shows input gradients for four random images. The first column displays the input image, the second column shows the mean gradient, and the last two columns show the gradients of two single models, picked manually for variability. As one can see, the specific input gradients for different models differ significantly from each other and from the mean; in some cases (e.g., golf ball in row three), the attributed area has almost no overlap and differs drastically from the mean.

Since there is no generic metric to capture the human perception, we decide to use a second metric for verifying our results. We rely on the established structural similarity index measure (SSIM), which is (in particular) less susceptible to purely high-frequency noise (unlike SNR). Fig. 5.3 compare SSIM and SNR for different saliency methods calculated on the four examples of Fig. 3.1. The ranking by SNR and SSIM are qualitatively similar (with a deviation to the worse for LIME) so we choose SNR as a metric.

Discussion: Marginalization removes the variability, but it is important to stress that the value of input gradient as a saliency method is still limited due to its well-known conceptual issue of often not detecting key features of a class [56, 2]). Our results, however, caution against the treatment of raw gradients as properties of the data examined. A substantial amount of the information depicted originates from the initialization, not the data.

Training protocol: Having trained the networks with “superconvergent” one-cycle might raise the concern that gradient noise might be an artifact of accelerated training. To eliminate the concern of the results being caused by numerical issues of the chosen training scheme, we repeat the experiment, training additional 30 instances of a ResNet18 on Imagenette with stochastic gradient descent with a learning rate of 0.01 for 30 epochs and a step decay of 0.5 after epochs 15 and 25. The resulting networks perform a little worse than networks trained with “superconvergence” but we obtain a similar SNR

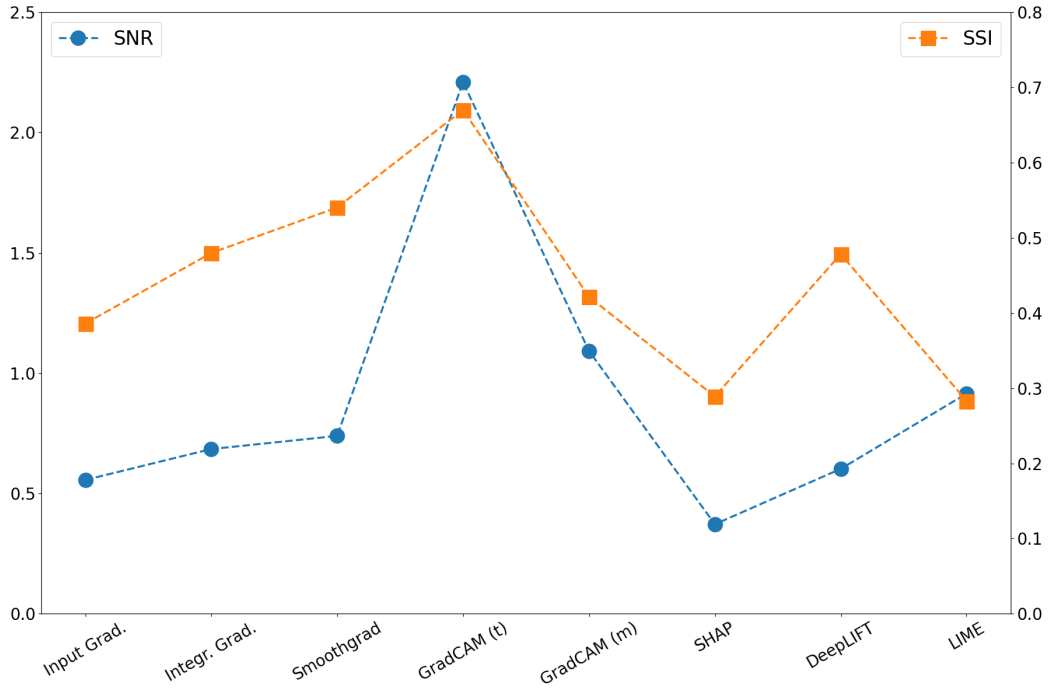


Fig. 5.3.: SSIM values show a qualitatively similar behavior to SNR, with LIME deviating a bit to the worse (computed for the four example images from Fig. 3.1.)

value (0.79), so it is unlikely that the seen results are due to numerical issues specific to ADAM or one-cycle.

Optimization vs. initialization: Noise can be categorized into initialization noise and optimization noise. In order to analyze their effects, we trained an ensemble of ResNet18 models with fixed initialization for each instance. This resulted in an improved SNR compared to varying initialization (average SNR varying: 0.694 vs. average SNR fixed: 0.830). However, a visual examination of the saliency maps revealed that even with fixed initialization, there were variations in feature localization.

5.2 Influence of Architectural Parameters

Next, we study the influence of depth, width and amount of nonlinearity on the noise level. Previous experiments have already suggested that depth increases initialization noise. For a systematic study, we run experiments on the FashionMNIST dataset as its low complexity allows comparing many architectures at an acceptable computational cost. The dataset contains 48.000

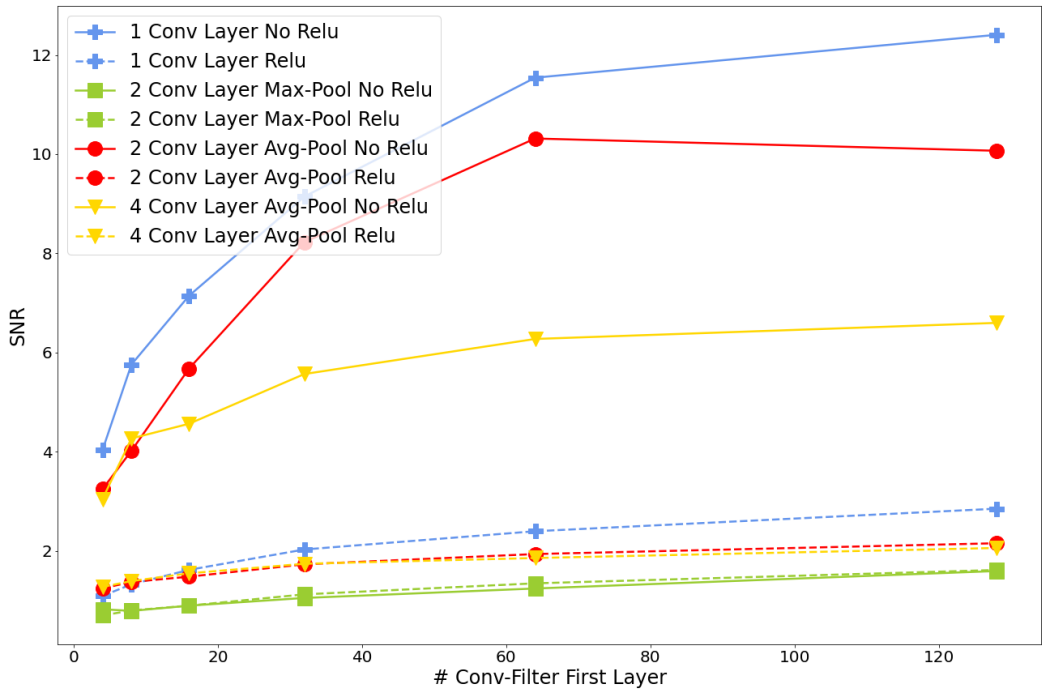


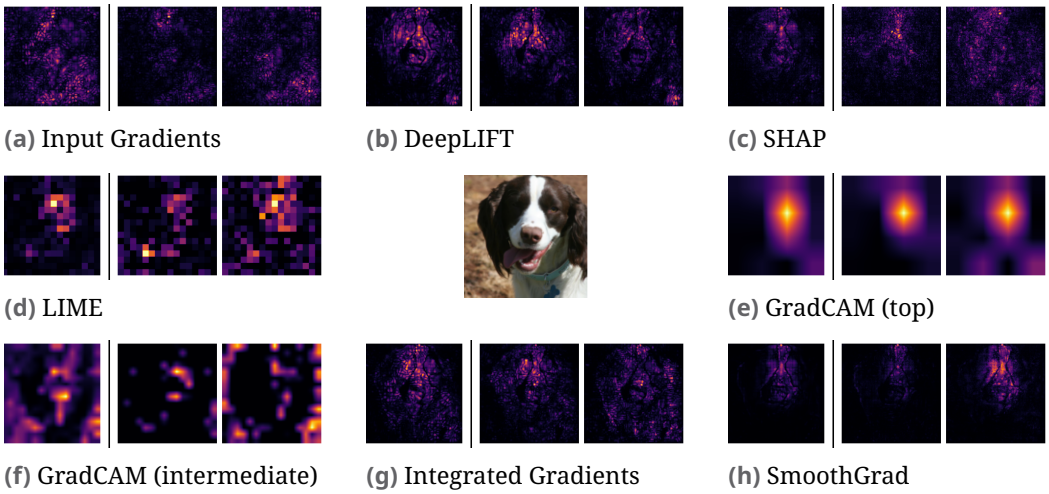
Fig. 5.4.: SNR for different architectures. Increasing the width tends to increase the SNR, while adding depth and sources of non-linearity tend to decrease SNR.

training images and 12.000 validation images divided into 10 different classes of cloth.

We vary width and depth, and toggle sources of nonlinearity: Starting from a baseline model of a single 3×3 convolution layer followed by a linear classifier, we iteratively add more layers and/or increase the number of feature channels in the convolutional filters. We also compare networks with and without max-pooling, and with and without ReLU layers.

Width is varied from 4 to 128 filters, while depth varies from a single to 4 convolutional layers. For every combination of these parameters, we train 20 networks with random initialization and calculate the mean SNR over the complete validation set.

Results are shown in Fig. 5.4. The number of channels (width) used in the first layer is plotted on the x-axis, while the y-axis indicates the mean SNR over all validation samples. The solid lines refer to linear networks without ReLU activation, while the dashed lines correspond to networks with ReLU activation.



(a) Input Gradients (b) DeepLIFT (c) SHAP

(d) LIME (e) GradCAM (top)

(f) GradCAM (intermediate) (g) Integrated Gradients (h) SmoothGrad

Fig. 5.5.: Comparison of saliency maps for multiple saliency methods on differently initialized models. All methods show variability in the produced results. First column: mean over 30 models, column 2-3: single model with random initialization.

Width and depth: In each architecture, increasing the width w improves SNR, visually consistent with a $\sqrt{w}$-increase due to the averaging of multiple computation paths. Depth decreases SNR consistently, as already expected from the previous experiments.

Linear vs. nonlinear: The most prominent result is the significant drop in SNR once nonlinearities are introduced (with max-pooling and ReLUs both showing similar effects). The convex 1-layer and 2-layer with averaged pooling without non-linearity show the best SNR by a large margin. Interestingly, improvements are also visible for the convex networks, indicating that some initialization noise due to imperfect optimization might still be present, although at a much lower level than in nonlinear architectures. Our observations show that both nonconvexity and nonlinearity increase initialization noise (including non-convex [28] linear multi-layer networks, solid-yellow-triangle curve). The results are in principle consistent with both an amplification of numerical noise as well as convergence to different local minima as the main source of noise.

Activation functions: One might conjecture that non-smooth activations might be a source of noise. For this reason, we employ softmax in the VGG19-results on Imagenette in Table 5.1, with results not differing qualitatively from ReLU, indicating that results hold for alternative, sufficiently nonlinear activation functions.

	VGG19	ResNet18	ResNet101
Input Gradient	0.436	0.555	0.435
Integrated Gradient*	0.612	0.742	0.744
Smoothgrad**	0.627	0.789	0.878
GradCAM (top)	0.816	1.593	1.373
GradCAM (interm.)	0.612	1.112	0.680
SHAP	0.252	0.365	0.308
DeepLIFT	0.453	0.628	0.554
LIME**	0.609	0.772	0.826

(a) CIFAR10

	VGG19	ResNet18	ResNet101
Input Gradient	0.585	0.589	0.444
Integrated Gradient*	0.725	0.695	0.577
Smoothgrad**	0.724	0.681	0.470
GradCAM (top)	1.370	2.363	3.171
GradCAM (interm.)	0.748	1.171	1.253
SHAP	0.418	0.386	0.327
DeepLIFT	0.978	0.650	0.587
LIME**	0.919	0.997	0.982

(b) Imagenette

Tab. 5.1.: Mean SNR over validation set for different saliency methods and model architectures. Due to increased computational costs we use only a subset for methods marked with "*" (10% of the validation set) and "**" (50 random samples of validation set).

Normalization methods: Normalization layers are crucial for stabilizing generalization performance in deep learning. To assess their impact on noise levels, we conducted experiments using two ensembles of ResNet18 models: one with batch normalization (BN) and the other with instance normalization (IN). The results showed that the alternative normalization method had only minimal differences in performance (average SNR of BN: 0.694 vs. IN: 0.702).

5.3 Further Saliency Methods

So far, we have only studied plain image gradients of logits. While conceptually important, gradients are not very suitable for attributing network decisions. We thus extend our experiments to several popular saliency approaches that operate by perturbations that resemble augmented gradient or finite difference computations. Specifically, we run *DeepLIFT*, *SHAP*, *LIME*, *Integrated Gradients*, *GradCAM* and *SmoothGrad* on top of *Integrated Gra-*

dients. For this experiment, we use ensembles of 30 models for estimating mean and variance.

Signal to noise ratios: Table 5.1 shows the resulting SNR values on *Imagenette* and *CIFAR10* for different architectures. Strikingly, only GradCAM was able to achieve an SNR above 1. Lower layer GradCAMs (tested on 2nd lowest ResBlock-layer of four) fare, expectedly, worse than top-level visualizations.

On CIFAR10, Residual Networks seem to consistently generate better SNR values than VGG19, but this pattern does not hold on Imagenette. Notably SHAP, which has a strong theoretical justification, performs the worst. This might not be surprising, as our results do not imply that any of these methods is a bad saliency method, but that saliency methods in general tend to capture information about the model, and not the phenomenon modelled. A high specificity to the most relevant features exploited by the model might therefore plausibly increase the visibility of the influence of initialization.

Qualitative comparison: Again, in order to understand the practical implications, we search through example images manually for particularly large variations (as these might lead to misinterpretations). Fig. 5.5 shows examples of varying saliency maps for different methods and initialization, applied to the dog ("English Springer") image in the center. The first column of every subfigure shows the mean saliency map over 30 models, while the second and third columns display the saliency map of two manually chosen networks (an exhaustive account is provided in the supplementary material). The reduced variability of top-level GradCAM is clearly visible, but even here, different initializations might lead to attributions distinct enough that they might conduct to inconsistent interpretations (in particular, taking into account the low resolution of the output). Intermediate-level GradCAM and the popular SHAP and LIME show strong structural differences.

6 Implications for Saliency-Based Interpretability

Assessing our overall findings, it is very important to state that this work at its core shows a negative result: Saliency maps based on gradients as well as popular more sophisticated attribution methods vary substantially with network initialization. Consequently, it cautions the reader to draw conclusions about the nature of the phenomenon observed from single model instances only, as some of the structures obtained originate in initialization randomness rather than uniquely reflecting training data properties. Also, the observed variability does not imply that attribution results are wrong, but we can be sure that they must be at least incomplete at times.

Importantly, one should also not make the converse conclusion: While marginalization is able to reliably dampen initialization noise to obtain clearer signals[1] this eliminates only one but not all potential extrinsic factors (such as architectural limitations and hyperparameters), and cannot address conceptual limitations such as the limited sensitivity of gradient-based saliency maps.

If marginalization is not possible (for example due to the increased training costs), the (popular) top-level GradCAM results appear to be least affected by initialization noise. Nonetheless, the lower-level variability we have observed might still be problematic in certain critical applications.

Limitations & Future work: Our work studies only a limited range of data sets and architectures, and counter-examples in Fig. 1 and 6 are manually curated. As these serve primarily as counter-examples, we consider this nonetheless sufficient to show the presence of a potential problem. However, a study of a broader class of attribution methods (such as masking-based

[1] Averaging 50 networks yields a 7-fold and 30 networks a 5.5-fold increase in signal to noise ratio (SNR) by the $\sqrt{n}^{-1}$-law, pushing all SNR values well above 1.0.

and attention-based methods) as well as strongly divergent architectures (such as vision transformers [12]) is still subject to future work.

The employed marginalization technique is rather inefficient; more sophisticated schemes for sampling network ensembles exist already in the literature [61]. Our approach has been motivated by simplicity and elimination of hidden dependency, not practical efficiency.

A broader study of external randomness due to choices of architectures and hyper-parameters would also be an interesting direction for future research, aiming at drawing attribution conclusions in a more automated fashion.

Part II

Information Extraction in the Context of Cloud Structures

7 Overview

Chapter 7 to 10 are based on the following paper:

“What Atmospheric Data Knows About Clouds: An Information-Theoretic Approach”
A.-C. Woerl, M. Wand, and P. Spichtinger
In preparation

Several text passages and images were taken directly from it. The experimental setting was developed in collaboration with Michael Wand and Peter Spichtinger. I implemented the whole software pipeline, performed and evaluated the experiments, and wrote the main parts of the manuscript. Due to duplications in related work, the foundations are transferred to Chapter 2 and can be reviewed there.

This part moves beyond viewing neural networks as purely predictive models and instead treats them as diagnostic tools for extracting structured information from high-dimensional atmospheric data.

We argue that generalization across spatial and temporal shifts as well as across different physical entities can serve as an empirical filter for identifying transferable information content, thereby distinguishing physically meaningful dependencies from memorization effects. Within this framework, gradient-based saliency methods are applied to the reconstruction of cloud structures in order to localize and quantify the atmospheric variables most relevant for cloud formation.

7.1 Abstract

Understanding and accurately modeling the formation of clouds in Earth’s atmosphere remains a challenging problem in atmospheric physics. While the fundamental physical components are known, our understanding of the overall picture remains incomplete due to its inherent complexity, which

arises in particular interactions of many different processes at vastly different scales.

In this work, we address this problem area by asking a targeted question: How much does an atmospheric model, represented in a simulation or reanalysis, actually "knows" about the clouds that will form? And, possibly, which aspects of the atmospheric state are most critical for determining the outcome. We view the question from an information-theoretic rather than physical point of view: Given data about certain physical conditions in the atmosphere, how much can this tell us about clouds forming, and which parameters are particularly sensitive, i.e., would affect the outcome the most if altered? Answering such questions purely based on statistical dependencies has become possible with modern representation learning methods based on deep networks. Deep learning nowadays permits building remarkably predictive models from example data for many different natural phenomena. As such a data-driven approach is able to learn unmodeled structures and dependencies, it has, in principle, the potential to even outperform a carefully designed physical model.

Our statistically learned model reconstructs and predicts cloud structures observed by satellites, taking atmospheric reanalysis data (and, optionally, predictive simulation data) as input. The model accurately predicts 2D cloud images, significantly outperforming the optical thickness estimates derived from the actual physical model. Curiously, withholding several of the physical input fields from the models during statistical learning often shows surprisingly little impact on prediction accuracy, suggesting that information required for predicting clouds is imprinted in a distributed fashion in the data when in an information-theoretic perspective when taking statistical knowledge about the distribution of real-world atmospheric conditions into account. The study also explores prediction of future cloud formations, demonstrating the model's capacity to forecast cloud structures, with varying degrees of temporal degradation in accuracy across different climate regimes. Finally, in order to study the sensitivity of the learned regressor to parameter changes, gradient-based saliency methods are applied, which yield attributions that seem to align with basic physical principles.

7.2 Introduction

Clouds play a critical role in shaping the weather and climate systems on Earth. Precise modeling of cloud microphysics is essential for reliable weather forecasting and climate modeling. However, the physical models that underlie weather and climate predictions often struggle with this task due to the complexity of cloud processes and their interaction across different timescales and spatial scales. A major goal for the field of atmospheric physics is to achieve a deeper understanding of the underlying physical processes.

Traditionally, cloud processes are most often modeled using physically motivated parameterizations. This means a coarser-scale simulation is coupled to an effective model that tries to describe the smaller-scale effects of clouds at this coarse scale. Despite drawing from decades of expertise and experience, these types of approaches still struggle to make accurate predictions in some regimes, and yet more research is required to understand the limitations better. Our work tries to provide an unconventional perspective on this question, using the newly available tools from statistical learning with deep networks.

We try to understand how much information on cloud structures can actually be found in such a model – in our case, we use the Integrated Forecast System (IFS) of the European Center for Medium Term Weather Forecasting (ECWMF). Thereby, information is understood operationally, not in a Shannon sense. We employ a statistical, purely data-driven approach: We use machine learning with deep networks (a simple/classical U-Net variant) and apply it to reanalysis (ECWMF ERA5, training, and inference) and simulation (IFS forecasts, inference only) data as input. Satellite observations (NASA Aqua/Terra 2D images in the visible spectrum, taken over open water) provide the “ground-truth” outputs, for training and evaluation. The question then becomes whether we would be able to extract more accurate information about cloud structures from the inputted atmospheric data using regression than what the underlying physical model itself suggests. Specifically, we study
(1) How well can we reconstruct the appearance of clouds from coarse-scale physical simulation data?
(2) How well does the model generalize across (a) space, time, physical variables, and data sources, and (b) how much does accuracy degrade with lead times in case of predicting future cloud formation?

(3) What insights into the underlying physical processes can we gain by studying how the statistical model operates?

We find the answer to *question (1)* to be quite surprising: The statistical model almost universally outperforms the physical model by a significant margin (roughly an average factor of 1.4 in normalized cross-correlation (NCC)). While it might not be surprising that a learned model, which can learn unmodeled effects in addition to the physical rules included in the model, can perform better, the magnitude of the improvement seems quite substantial.

Question (2) examines how generalizable the statistical models are: We (a) investigate geographical generalization by training models in restricted regions and testing them elsewhere, revealing high contrasts between regions. While models trained in temperate zones often generalize well across different climates, those trained in the tropics tend to specialize in these conditions. Next, we assess physical quantity generalization by limiting the model's input to subsets of physical variables, which again, yields surprising results: Despite this reduction, performance drops only modestly in most cases, suggesting that relevant physical information is redundantly encoded across multiple variables and distributed in the data.

When asking (b) how far into the future current atmospheric states retain predictive power, we obtain results that strongly vary by region, with temperate zones showing a more rapid decay in predictability than the tropics. Nonetheless, the statistical model still outperforms the physical optical thickness with zero lead time when itself subjected to lead times of up to $\sim$20h in the North Atlantic and for several days in the tropics.

Further, as an additional sanity check concerning generalization, we apply the model originally trained on reanalysis data to numerical forecasts outside the scope of the training data. While this question is interesting on its own, as it resembles an actual operational application scenario, we also run this experiment to exclude potential data leakage from data assimilation as a potential hidden causal: ERA5 includes the same satellite data (Aqua, Terra) our model tries to predict during its data-fitting procedure in the reanalysis, thus posing a risk of data leakage. However, we find that our statistical model is still able to consistently outperform physical baseline forecasts by a substantial margin in the predictive setting, indicating that it captures generalizable, physically meaningful patterns rather than simply learning to "decode" information imprinted during reanalysis.

Finally, with respect to *question (3)*, investigating the internal logic of the model, we use gradient-based attribution methods to estimate the importance of each variable and spatial region for the prediction. While such methods are known to be problematic in understanding supervisedly-trained discriminative models, the regression problem in our application domain is more suitable for gradient-based sensitivity estimates. The resulting saliency patterns are consistent with established physical understanding – highlighting, for example, the role of humidity at medium model levels and surface temperature. For moderate lead times, these attributions shift with wind patterns, potentially indicating a sensitivity to humidity transport. While no causal proof, these findings suggest that the model captures physically meaningful dependencies in a fully data-driven manner. However, we do not find any apparent novel or particularly surprising patterns in this sensitivity analysis.

In summary, with an objective to understand *physical* models better, this work frames cloud modeling as an information-theoretic problem rather than as systems of differential equations grounded in physics and shows that data-driven models can exceed physical parameterization in accuracy, reveal interpretable links to known physical processes, and demonstrate that cloud information is encoded redundantly in physical fields. The conclusion, and answer to our initial question, is that the simulation actually "knows" more about clouds than the physical model it is built from when taking the distribution of real-world atmospheric conditions into account.

8 Data and Methods

This chapter introduces the data, models, and methodological framework used to analyze cloud reconstruction from atmospheric state variables. In addition to the technical setup, it defines the information-theoretic perspective adopted in this part and describes how interpretability and visualization are used to study cloud-relevant information in the data.

8.1 Data

ECMWF Reanalysis v5 (ERA5): In our work, we try to predict cloud structures based on physical conditions. Therefore, we use the ERA5 dataset [@1] of global atmospheric reanalyses (i.e., a best fit of an atmospheric simulation against most available observational data). It offers high-resolution data with a spatial resolution about 31 kilometers (0.25 degrees in longitude and latitude) and hourly temporal resolution. For vertical resolution, we choose model levels, which resolve the atmosphere from the surface up to a height of approximately 24 kilometers. The ERA5 dataset covers a variety of physical quantities. We want to focus on the most general relations between entities and minimize the influence of predefined knowledge, so we restrict ourselves to the most basic ones: temperature t, specific humidity q, cloud liquid and ice water content $clwc$ respectively $ciwc$, horizontal wind components u, v, and vertical velocity w. ERA5 provides a scalar cloud cover field, representing cloud density at various model levels. A simple estimate of a 2D cloud image can be obtained by integrating in height; however, we use the recommended optical thickness estimate to compute the model's prediction of a 2D top view of clouds.

Optical Thickness: As a baseline, we use the optical thickness derived from ERA5 data. Optical thickness describes the reduction in radiation intensity due to scattering and absorption by cloud particles and is calculated by integrating the extinction coefficients vertically. In our formulation, contributions from liquid and ice phases of clouds are treated separately and then

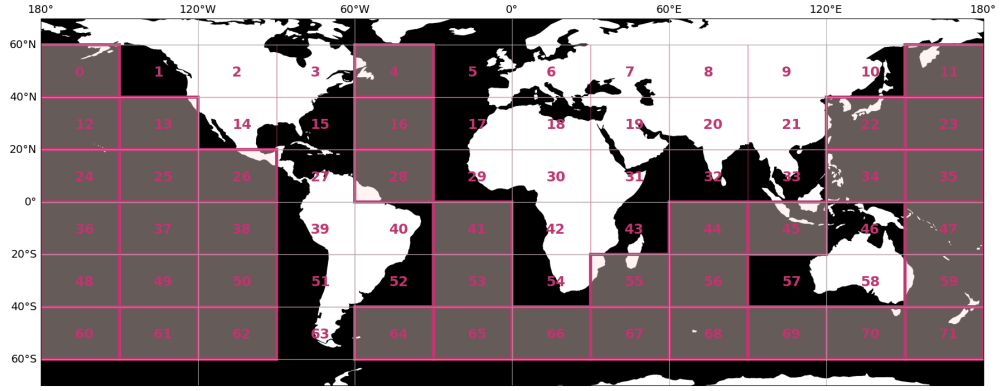


Fig. 8.1.: Geographical segmentation of the globe into 39 equally-sized regions mostly over water, highlighted in pink. These regions are used to train and evaluate a neural model to predict cloud structures.

combined along the vertical dimension. Further details can be found in Appendix A.

Satellite Images: To obtain the visual appearance of clouds, we use satellite images obtained from polar orbiting satellites Terra and Aqua, hosted by the NASA. These satellites capture multi-spectral imagery in various spectral bands, providing comprehensive coverage of Earth's surface and atmosphere. In this work, we restrict ourselves to two-dimensional, true color-corrected reflectance images provided by NASA's Global Imagery Browse Services (GIBS) [@3] which correspond to natural-looking images by combining visible bands from the satellite. The satellites cross each location on Earth two times a day. Since we are interested in the visible spectrum, we only use the daytime passing.

Integrated Forecasting System (IFS): In addition to ERA5, we also use a second source of physical approximation data – the forecasts produced by the ECMWF based on its Integrated Forecast Model, IFS. We should first note that ERA5 itself uses IFS as its underlying dynamical model, i.e., the physical representation matches directly. When used in forecast applications, the forecasts combine assimilated initial conditions to simulate the likely evolution of the atmosphere (we do not consider ensembles in this paper, only the main run). To evaluate our model's ability to generalize beyond reanalysis data and mitigate the influence of data assimilation (particularly from satellite imagery), we use IFS forecast outputs as an alternative input source. Forecasts are initialized at 00:00 and 12:00 UTC, and we extract hourly lead times from 7 to 48 hours, covering six representative days throughout 2014. To ensure comparability, we use the same spatial resolution, physical vari-

ables, and preprocessing steps as for ERA5. All evaluations are restricted to oceanic regions to maintain consistency with our main training domain.

Data Preparation: Cloud structures are medium-scale phenomena that do not extend over the entire globe. For this reason, we divide the world into smaller regions of $30°$ longitude and $20°$ latitude and run our algorithm on these subpatches. To avoid the distortion at the polar caps, we limit ourselves to the range of $60°$ South and $60°$ North. To clearly separate the clouds from the background, we only consider regions that are mostly over water. Fig. 8.1 shows the resulting tiling with used regions highlighted in pink. Variations in brightness and contrast from different incident solar illumination remain in the images; we thus employ a scale-normalized matching function (normalized cross-correlation (NCC), see sect. 8.3) to facilitate invariance. We use 7 years of data (2014–2020), split into training and validation sets, and map satellite and ERA5 data accordingly, using the spatial resolution from ERA5 for our satellite images as well. To avoid memorizing effects, we use one whole year (2014) as a validation set that the model hasn't seen in the training stage. To reduce the amount of necessary memory and computation time, we only use every second model level, resulting in 49 levels per physical quantity. We justify our selection by the fact that the vector fields of individual height layers exhibit minimal variation, making every second level sufficient to capture the atmospheric conditions accurately.

8.2 Model Architecture

Cloud Structure Prediction: If cloud fields solely rely on physical conditions, then reconstructing them from reanalysis data becomes a question of extracting the relevant information embedded within these quantities. Conventional approaches, such as optical thickness calculations, use a suitable optical model to integrate over liquid and ice water to generate cloud fields. In contrast, our statistical model uses a purely data-driven approach and trains a regression model to learn these mappings directly. This allows us to test not only the feasibility of reconstruction but also the extent to which physically meaningful cloud structures are inferable from coarse atmospheric data.

As a regression model, we use a U-Net architecture ([40], see fig. 8.2 for details), which is a convolutional neural network commonly used for segmentation tasks. The architecture consists of a contracting path that captures

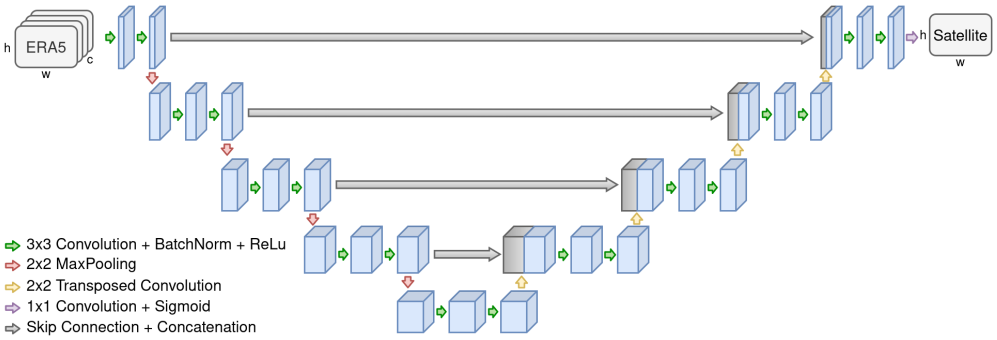


Fig. 8.2.: Schematic of our model using atmospheric data to predict satellite cloud images. The model uses an U-Net architecture, down- and upscaling the input four times.

context and an expansive path that enables precise localization. The architecture's skip connections allow for the combination of high-level features with low-level spatial information, making it well-suited for reconstructing complex structures like clouds from atmospheric data. Our model is trained to predict the appearance of cloud fields as seen by satellites, using a combined loss of mean squared error and structural similarity. This combination ensures both capturing the large, underlying structure of clouds while encouraging the model to consider smaller details. We use ERA5 reanalysis data as input, treating it as the best available approximation of the atmospheric state. Thereby, the model operates on high-dimensional subpatches of shape $c \times w \times h = 343 \times 120 \times 80$, where c represents seven physical variables across 49 vertical levels stacked together. Grayscale satellite images of shape 120×80 serve as the target. We train our network for 35 epochs, employing an ADAM optimizer [25] and integrating a one-cycle learning rate scheduler [52] since latest research has shown this strategy to improve learning speed.

Horizon of Predictability: Beyond showing that current atmospheric data contains enough meaningful information to predict cloud formations, we aim to quantify the temporal persistence of such information – i.e. how far into the future relevant atmospheric signals remain decodable for cloud formations. To explore this, we select satellite observations at a target time t_0 and combine them with physical data from earlier time points $t = t_0 - t_1$, where t_1 ranges from one to 96 hours. This setup enables us to evaluate how predictive information decays over time, providing an accurate representation of the time horizon within atmospheric data remains useful for forecasting cloud structures. We would like to emphasize that we are not using

a forecasting model, but rather time-fixed reanalysis data with a time lag to predict cloud images.

8.3 Evaluation Metrics & Algorithms

Normalized Cross-Correlation: To evaluate the quality of our cloud structure prediction, we use the NCC [19] to measure the similarity between our prediction and the real satellite image. The NCC is a commonly used measure capturing linear correlation between two images, yielding outputs between -1 and 1. It is invariant to shifts and scaling of either of the signals compared (i.e., invariant under changes in brightness and contrast of the images considered). This is crucial for our application as we do not normalize data for incident solar illumination.

Multi-Scale Structural Similarity Index Measure: Another possibility of quantifying the quality of our predictions is the structural similarity index measure (SSIM) [59] that evaluates the structural similarity between images. While the SSIM is based on a single scale, an extended version, the multi-scale structural similarity index measure (MS-SSIM) [60], compares images across multiple scales, providing a more comprehensive and nuanced assessment. Unlike the original approach, where the results from different scales are summed to produce a single value, we also treat each scale independently, considering them separately in our analysis to gain more detailed information about the accuracy on different scales. In contrast to NCC, SSIM is not fully invariant to global brightness and contrast shifts, but is robust against moderate changes in illumination, making it suitable for our application.

Multidimensional Scaling: "Classical" multidimensional scaling (MDS) is a dimensionality reduction technique that embeds a set of data points into a lower-dimensional space while preserving the pairwise dissimilarities as closely as possible [57]. By providing a low-dimensional visualization, it simplifies the analysis and interpretation of complex, high-dimensional data. We apply metric MDS to a symmetric distance matrix derived from model performance differences, allowing us to visualize generalization relationships between geographically trained models.

8.4 Information-theoretic Background

A central concept of this work is the measurement of information about cloud formation contained in atmospheric state variables. Importantly, we do not attempt to quantify abstract information-theoretic quantities such as Shannon entropy or mutual information. Instead, we adopt an operational definition of information based in decodability: An atmospheric variable is said to contain information about cloud structure if a statistical model trained on atmospheric data can reliably reconstruct cloud patterns from it. Information is therefore the part of the input signal that can be systematically transformed into accurate predictions of cloud appearance.

This definition has several implications. First, information is not treated as an intrinsic property in isolation, but as a relational quantity between inputs and a specific prediction target. Second, the measured information content depends on the expressiveness of the model class used for decoding. Consequently, the estimated information should be interpreted as model-dependent and task-specific.

All models employed in this work are regression-based predictive models rather than generative models. This restriction ensures that reconstructed cloud structures must be inferred from statistical dependencies present in the input variables. The models cannot freely generate cloud patterns independent of the provided atmospheric fields. Under this predictive setup, reconstruction accuracy provides an empirical lower bound on the decodable information about clouds contained in the atmospheric data. If a sufficiently expressive regression model fails to reconstruct certain structures, this indicates that the corresponding information is either weak, non-systematic, or not accessible within the chosen representation framework.

8.5 Saliency Analysis

Interpreting the behavior of neural networks remains a fundamental challenge, since these models encode complex, non-linear functions in a high-dimensional parameter space. From a information-theoretic perspective, we try to answer the question which parts of the input contribute most significantly to the predictive output of the model. To address this, we employ saliency-based attribution methods, which use gradients of the model's out-

put with respect to the input feature, thereby offering a proxy for information flow through the network.

Classically, such methods trace how perturbations in input space would alter the output, identifying regions that the model thinks most informative. Lately, saliency maps have been critiqued for their reliability to reveal causal connections between input and output in classification tasks ([1], [62]). Other work indicates that this phenomenon can be attributed to the process of learning incomplete features. Subsequent work has shown that interpretability improves when models are trained under generative or adversarially robust regimes ([4], [58]), since this guides the network to learn a more comprehensive set of features. In our setting, where the network learns to reconstruct cloud structures rather than assign discrete class labels, the regression-based formulation provides a more natural context for meaningful attribution.

Our analysis is based on a U-Net architecture. Due to its hierarchical encoding and decoding paths, the U-Net architecture possesses a large receptive field and is able to capture multi-scale dependencies. The choice of a U-Net reflects the hypothesis that cloud-relevant information is spatially distributed and multi-scale rather than local. To trace these dependencies, we compute input gradients for each output pixel, yielding saliency maps of shape $343 \times 120 \times 80$, representing the model's sensitivity to each physical variable across space and altitude.

However, the high dimensionality of these maps introduces significant computational and memory overhead. To mitigate this, we restrict the attribution analysis to a spatial neighborhood of 30×30 pixels around each output pixel (approximately 900 kilometers in Earth distance), under the assumption that the influence of distant regions is marginal. We then normalize saliency maps across channels to emphasize relative importance among physical fields and simplify comparisons across regions. This process is visualized in fig. 8.3 and forms the foundation for extracting dominant information flows in cloud structure prediction.

8.6 Visualization

As a useful side product of the analysis pipeline, we developed an interactive visualization tool to facilitate the exploration and interpretation of the high-dimensional attribution data. The tool combines direct visualization

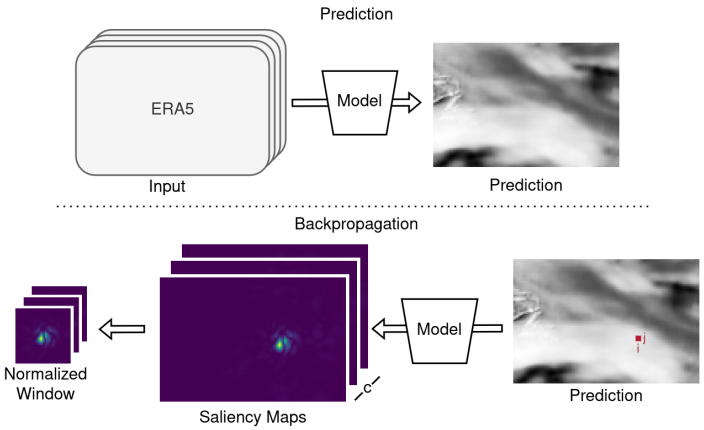


Fig. 8.3.: Saliency maps are generated using model prediction and pixelwise backpropagation. The resulting saliency maps are cropped and normalized afterwards.

of individual attribution maps with dimensionality reduction and clustering techniques, thereby supporting both detailed case-based inspection and the identification of broader structural patterns.

A central challenge is the sheer number of saliency maps produced by the model. Since every output pixel is associated with its own attribution map, even a single 120×80 image together with a 30×30 attribution window yields thousands of high-dimensional explanations. A direct manual inspection of all these maps is therefore impractical. The visualization tool was designed to address this problem by linking local attribution information for individual pixels to more compact summaries across the full dataset.

The tool supports both volumetric (3D) visualization of attribution maps and reduced 2D summaries. In the 3D view, attribution values are displayed in their spatial context, allowing the user to inspect how contributions are distributed across the attribution window and across altitude levels. To suppress visually distracting low-magnitude responses and emphasize the dominant structures, a threshold can be applied to the 3D representation, which hides small attribution values. For the 2D summary, attribution values are aggregated along the horizontal dimensions and considered separately for each physical quantity. The resulting altitude-dependent relevance profiles are normalized and displayed as vertical curves with highlighted relevance ranges, providing a compact view of which variables and heights are most informative for the selected prediction. While this projection substantially simplifies interpretation, it necessarily discards information about the horizontal spatial distribution of attribution.

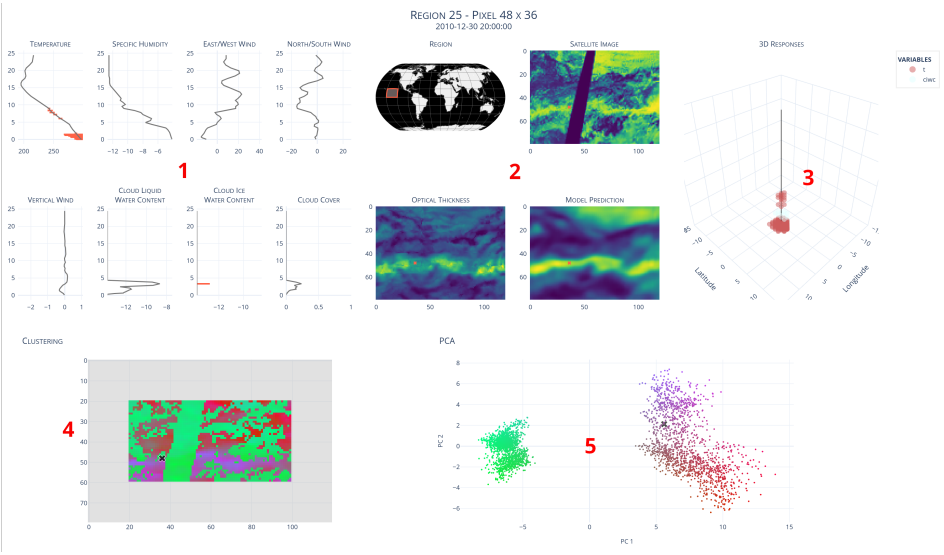


Fig. 8.4.: Screenshot of the visualization tool. The panels show (1) altitude-resolved atmospheric profiles with attribution summaries, (2) the geographic and observational context including satellite image, estimated optical thickness, and model output, (3) a three-dimensional attribution map, and (4,5) clustering-based analysis of attribution maps in reduced-dimensional space.

To move beyond individual attribution maps, the tool also includes a global analysis of attribution patterns across all output pixels. For this purpose, Principal Component Analysis (PCA) is applied to the attribution maps in order to extract dominant modes of variation. The resulting low-dimensional representation reveals common structural patterns and enables clustering of similar attribution configurations. In this way, the tool connects single-pixel explanations to broader families of attribution behavior.

Fig. 8.4 provides an overview of the visualization tool, which consists of five linked visual elements designed to support interactive interpretation of the model's internal information usage:

1. **Vertical Profile Viewer (top-left):** Displays altitude-resolved profiles for key meteorological variables, including temperature, specific humidity, horizontal and vertical wind components, cloud liquid water content, cloud ice water content, and cloud cover. These profiles provide insights into the vertical structure of the atmosphere and cloud formation processes at the selected location. Orange bars highlight the variables and altitudes that the model thinks are most informative based on saliency.

2. **Geographic Context Panel (center):** Provides the spatial context of the selected sample by comparing the input satellite image, the estimated optical thickness, and the model reconstruction. The orange cross marks the output location for which the attribution analysis is currently shown.
3. **3D Attribution Map (right):** Visualizes the attribution values in three dimensions, making it possible to inspect how relevant input information is distributed spatially and vertically around the selected output pixel. This view forms the basis for the condensed vertical summaries shown in panel 1.
4. **Clustering Panel (bottom-left):** Shows representative clustering results for attribution maps after dimensionality reduction. This makes it possible to identify recurring attribution structures and compare how different samples are grouped.
5. **PCA Embedding (bottom-right):** Displays a two-dimensional PCA projection of the attribution maps. Colors indicate cluster membership and correspond to panel 4, providing an overview of how attribution patterns are distributed across the dataset.

Overall, the visualization tool serves both as a diagnostic interface and as a hypothesis-generating instrument. It enables domain experts to inspect individual attribution maps, compare them across locations and samples, and relate recurring attribution structures to known physical processes.

9 Results

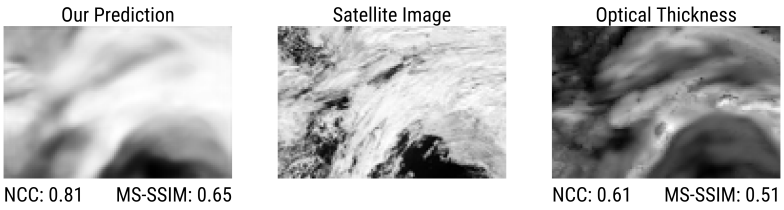
This chapter presents the main empirical results on cloud reconstruction and information extraction from atmospheric data. It first evaluates the predictive quality of the neural network models and then studies how cloud-relevant information generalizes across regions, variables, time, and data sources.

9.1 Cloud Structure Prediction

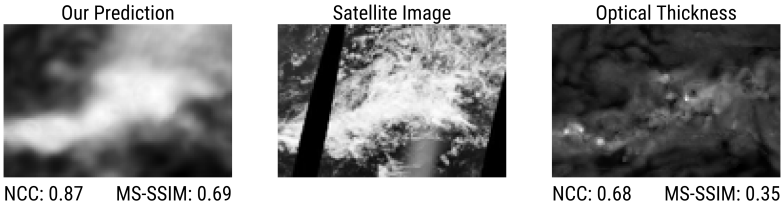
Accurate weather forecasting depends on good representations of cloud microphysics, which is by now only understood incompletely. We approach this challenge through a data-driven lens, aiming to recover cloud structures by learning directly from observations and reanalysis data. Rather than relying on explicit physical equations, we train a regression model on ERA5 reanalysis data to extract the predictive information implicitly encoded in the atmospheric state. As a reference, we compare our model's output against conventional optical thickness calculations derived from physical approximations.

Fig. 9.1 presents qualitative and quantitative comparisons across four climatologically different regions. We choose these regions to capture a range of atmospheric dynamics, from the highly variable North Atlantic to the more stable Equatorial zone, the South Pacific, and the Indian Ocean. Each triplet includes true satellite imagery (center), our model's prediction (left), and the optical thickness estimate (right). To objectively evaluate quality, we calculate normalized cross-correlation (NCC) and multi-scale structural similarity index measure (MS-SSIM) scores for all samples for both our prediction and the optical thickness. They are shown beneath each image triplet.

Both – the visual inspection and the calculated metrics – show that our model provides a better representation of the overall cloud structure even if finer details are not perfectly captured. This suggests that traditional physical-based equations underutilize available information, while data-driven



(a) North Atlantic



(b) Equator

(c) South Pacific

(d) Indian Ocean

Fig. 9.1.: Comparison of cloud structures generated from our model against the satellite image and the optical thickness for different regions on Earth. Alongside a visual comparison, we calculate the NCC and the MS-SSIM. Left: Our model prediction, Middle: Satellite image, Right: Optical thickness

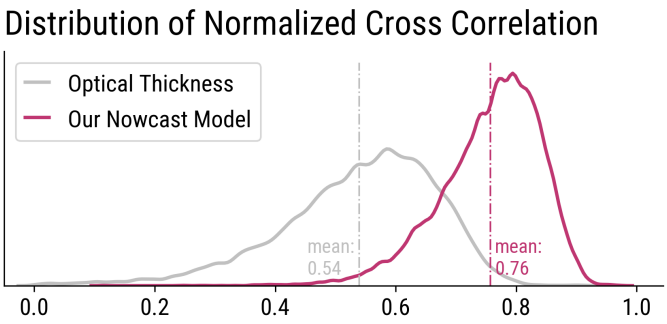


Fig. 9.2.: Distribution of NCC scores on satellite validation data comparing the optical thickness baseline derived from ERA5 and the proposed nowcast model. The model consistently achieves higher correlation values, indicating improved cloud reconstruction performance.

approaches can uncover richer structure embedded in the atmospheric fields.

We further validate these findings by evaluating the NCC across the entire validation dataset, shown in fig. 9.2. Our model (magenta) achieves a mean NCC of 0.76, significantly outperforming optical thickness baseline (gray) at 0.54. Looking at the distribution of NCC values reveals a narrower spread for our model, indicating not only a better average performance but also greater robustness across diverse atmospheric conditions.

Although the performance gains are significant, it is important to acknowledge the limitations of our model. The model may fail to localize cloud features, particularly small-scale details. As expected, it also cannot overcome the band limitation of the input data, thus cannot reconstruct sub-grid-scale signals that would still be visible in the (higher resolution) satellite images. It seems plausible that this is impossible due to the insufficient resolution of ERA5. Our model is not generative (i.e., it is intentionally just a regressor and does not learn a probability distribution as its output). Thus, it only predicts averages and does not attempt to fill in plausibly structured (in particular small-scale) information that cannot be reconstructed deterministically from data.

9.2 Generalization

To assess whether our learned models truly extract transferable information rather than merely memorizing, we design four complementary generalization tests:

First, we investigate *geographic generalization* by training and evaluating the models across distinct climate regimes. This analysis examines whether predictive relationships learned in one region remain informative in others, thereby testing whether the extracted dependencies reflect local artifacts or broader physical structures.

Second, we analyze *physical quantity generalization* by systematically varying the set of atmospheric input variables. This allows us to evaluate how predictive information is distributed across different physical variables and to identify which quantities contribute robustly to accurate cloud reconstruction. In this context, informativeness is interpreted as the contribution of a variable to transferable predictive performance.

Third, we address *temporal generalization* by introducing time lags between atmospheric inputs and cloud targets. This experiment probes how long predictive information persists in the atmospheric state and whether the model captures temporally stable dependencies rather than instantaneous correlations.

Finally, we test *data-source generalization* by evaluating the trained models on an independent input source, i.e. numerical weather forecasts produced by the Integrated Forecasting System (IFS). Since reanalysis and forecast data differ in their construction and error characteristics, consistent performance across both datasets provides evidence that the model relies on physically meaningful relationships rather than dataset-specific artifacts.

Taken together, generalization across regions, time scales, physical inputs, and data sources acts as an empirical information filter. Only those input-output dependencies that are stable under these distribution shifts can be considered transferrable. In this operational sense, successful generalization indicates that the model captures structured information embedded in the atmospheric state, rather than memorizing patterns tied to a specific dataset or sampling configuration.

9.2.1 Geographical Transferability

To derive physical meaningful conclusions from machine learning models, it is essential to demonstrate that predictive performance does not arise from memorization effects. Beyond enforcing a strict temporal separation

between training and validation data, we therefore asses geographical generalization.

We train 39 distinct models M_i, each restricted to a single geographic region (see fig. 8.1 for selected regions). Each model is subsequently evaluated on all other regions V_j. Cross-regional performance is quantified using both NCC and MS-SSIM, enabling a systematic assessment of transferability across climate regimes.

To illustrate the diversity in generalization behavior, we focus on two contrasting cases: model M_4, trained over the North Atlantic mid-latitudes, and model M_{25}, trained in an equatorial region. Fig. 9.3 shows their relative average performance compared to the optical thickness baseline, where red indicates improvement and blue indicates deterioration.

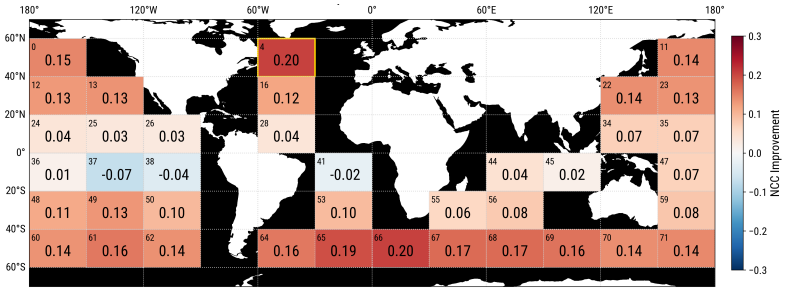
Model M_4, trained in the mid-latitudes, generalizes robustly across most regions, particularly within temperate climate zones, outperforming the baseline by up to 0.2 in NCC (fig. 9.3a). However, its performance decreases toward tropical regions and parts of South America.

In contrast, model M_{25}, trained near the Equator, achieves strong performance within tropical regions but deteriorates noticeably toward the poles (fig. 9.3b), in several cases even underperforming the physical baseline.

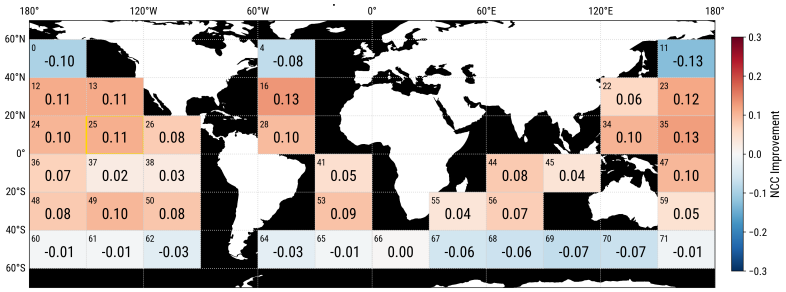
These contrasting behaviors are consistent with known differences in atmospheric dynamics across climate regimes. Mid-latitudes are characterized by a high variability (e.g. frontal zones, continental influence, meridional temperature gradients) of atmospheric conditions. Training in such dynamically diverse environments may expose the model to a broader spectrum of cloud-forming mechanisms, thereby encouraging the learning of more transferable representations.

In contrast, tropical regions are often dominated by persistent, large-scale convective activities with comparatively homogeneous vertical structures. While this can yield high predictive accuracy within similar dynamic environments, models trained exclusively under such conditions may learn representations that are less adaptable to regimes affected by different physical processes.

It is important to note that the optical thickness baseline itself exhibits regional performance variations. In tropical regions, where large, homogeneous convective cloud systems dominate, the baseline performs relatively well. In temperate zones, however, cloud structures are typically flatter and



(a) North Atlantic - Model M_4



(b) Equator - Model M_{25}

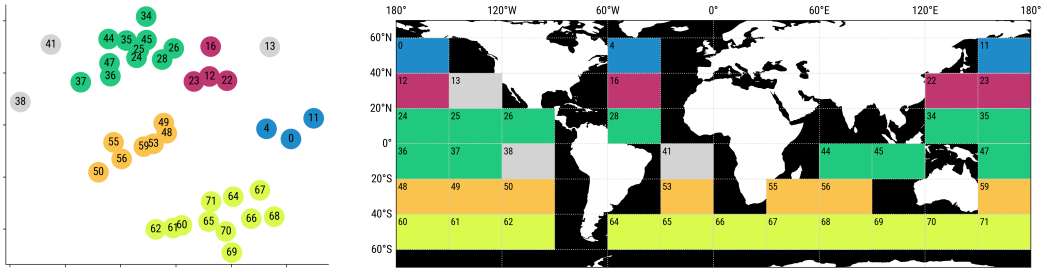
Fig. 9.3.: Comparison of two regional models trained on: (a) the North Atlantic (model M_4) and (b) the Equator (model M_{25}). Each model is evaluated in all other regions. Colors represent the relative improvement in NCC over the optical thickness baseline.

more heterogeneous, reducing baseline accuracy. Consequently, smaller absolute improvements relative to the baseline in tropical regions do not necessarily indicate weaker model performance; rather, they reflect a strong baseline in those regimes.

Beyond analyzing individual cross-regional performances, we aim to identify broader structural patterns in geographic generalization. Specifically, we investigate whether regions exhibit similar global transfer behavior when their locally trained models are applied elsewhere.

To this end, we represent each model M_i by its full cross-regional performance vector, defined as the set of absolute NCC scores obtained when evaluating M_i on all regions V_j. This vector can be interpreted as the model's generalization profile, capturing how its predictive capability transfers across different climate regimes. Pairwise distances between these profiles are computed to quantify similarities in transfer behavior.

The resulting distance matrix is symmetrized and normalized, and we apply multidimensional scaling (MDS) to obtain a two-dimensional embedding that



(a) A 2D embedding using MDS to group regions by generalization similarity. Points close together indicate regions whose models generalize similarly (grey=unassigned).

(b) Spatial map where regions are color coded according to their MDS embedding, revealing a latitudinal strip pattern in generalization behavior.

Fig. 9.4.: Similarity analysis of geographical generalization behavior using NCC and MS-SSIM metrics as distance measurements. MDS is used to provide a simplified 2D representation, revealing similarity clusters.

approximately preserves the original similarity structure (fig. 9.4a). In this embedding space, models with similar generalization behavior appear close to each other. For improved interpretability, we cluster the embedded points and assign cluster labels that are subsequently projected back onto their geographic locations (fig. 9.4b). Models that do not clearly belong to any cluster are marked in gray.

The spatial projection reveals a distinctive latitudinal banding pattern: Models trained at similar latitudes tend to exhibit similar generalization behavior. This indicates that large-scale atmospheric variability, which is strongly structured along zonal bands, plays a dominant role in shaping the transferability of learned representations.

While clustering results depend on the chosen algorithm and hyperparameters, the latitudinal organization of the embedding proves robust across different methods. In our analysis, we use DBSCAN from Scikit-learn [13, 43] with `min_samples` of 3 and `eps` of 0.12, but similar banding patterns emerge under alternative clustering choices.

Interestingly, models trained in the southern mid-latitudes, particularly between $20°$ and $40°$S, achieve the highest overall generalization performance, while models trained in equatorial regions generally exhibit weaker generalization. This observation reinforces the findings from the representative models M_4 and M_{25}, and is consistent with the hypothesis that highly dynamic atmospheric regimes promote the learning of more transferable rep-

resentations. Regions characterized by comparatively homogeneous large-scale dynamics, on the other hand, appear to produce models that are locally specialized but less globally adaptable.

These findings suggest that the geographical composition of the training data plays a crucial role in shaping generalization properties. Rather than training on isolated regions, selecting training domains that span dynamically diverse latitudinal bands may promote broader transferability across climatic regimes.

9.2.2 Variable-wise Information Redundancy

Next, we address a fundamental question from an operational information-theoretic perspective: To what extent is predictive information about cloud structures redundantly distributed across atmospheric variables, and how much can be reconstructed from isolated signals? To explore this, we construct a series of *minimal models*, each trained using only a single atmospheric variable from the original multivariate input set; i.e. temperature, specific humidity, cloud liquid and ice water content, horizontal wind components and vertical velocity. If multiple physical variables independently allow reconstruction of similar cloud structures, this indicates that cloud-relevant information is distributed and redundantly encoded within the atmospheric state.

For each regional model M_i, we therefore train seven corresponding minimal models $MM_{(i,var)}$, each using only a single atmospheric variable $var \in \{t, q, clwc, ciwc, u, v, w\}$, corresponding to temperature, specific humidity, cloud liquid water content, cloud ice water content, zonal and meridional wind components, and vertical velocity, respectively. This setup allows us not only to quantify overall redundancy, but also analyze how redundancy patterns vary geographically. The quality of each minimal model $MM_{(i,var)}$ is assessed via its NCC score on the full validation dataset and compared to the corresponding multivariate model M_i. Fig. 9.5 shows the deterioration in NCC performance relative to the full multivariate model, averaged over samples for different regions.

As expected, restricting the input to a single variable reduces performance. However, several minimal models retain substantial predictive capability,

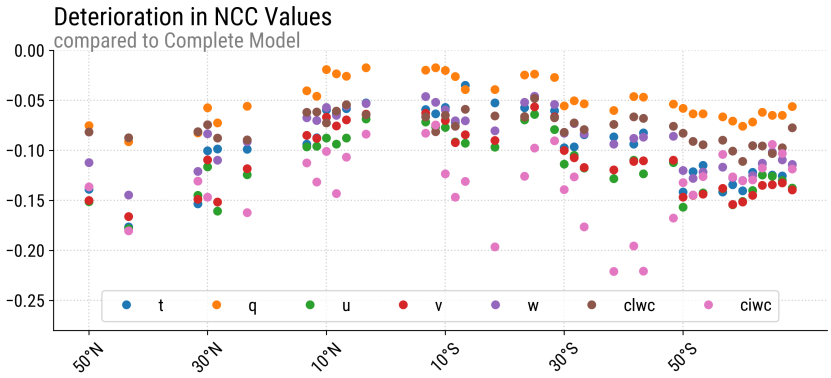


Fig. 9.5.: Average NCC drop for minimal models trained on a single physical quantity compared to the full-input baseline. Specific humidity and ice water content contributes the most to reconstruction accuracy, especially in the tropics.

particularly in tropical regimes. This indicates that, in these regions, cloud-relevant information is strongly concentrated in individual physical quantities and can be decoded even in isolation. In contrast, the largest performance deterioration is observed in the mid-latitudes. This suggests that cloud formation in these regions depends more strongly on the joint configuration of multiple atmospheric variables, rather than being redundantly encoded in single fields.

Among all variables, specific humidity contributes the most to cloud reconstruction quality across regions. This aligns with the physical expectation, since cloud formation fundamentally depends on moisture availability and condensation processes. Liquid water content outperforms ice water content, consistent with the global predominance of liquid clouds, which cover approximately $60 - 70\%$ of the Earth's surface, compared to roughly $20 - 30\%$ for ice clouds [37]. Temperature shows decent predictive capacity in tropical regions but contributes less in mid-latitudes, suggesting that temperature alone is insufficient to capture the dynamical complexity of extratropical cloud systems. Among wind variables, vertical velocity retains a measurable predictive signal, reflecting its role in convective lifting processes, whereas horizontal wind components exhibit the weakest isolated performance. This indicates that advective structure alone is less informative without accompanying thermodynamic context.

Fig. 9.6 illustrates a representative example comparing all minimal models for a scene near the Equator. The first column shows the observed satellite image together with the corresponding model prediction in the middle, given

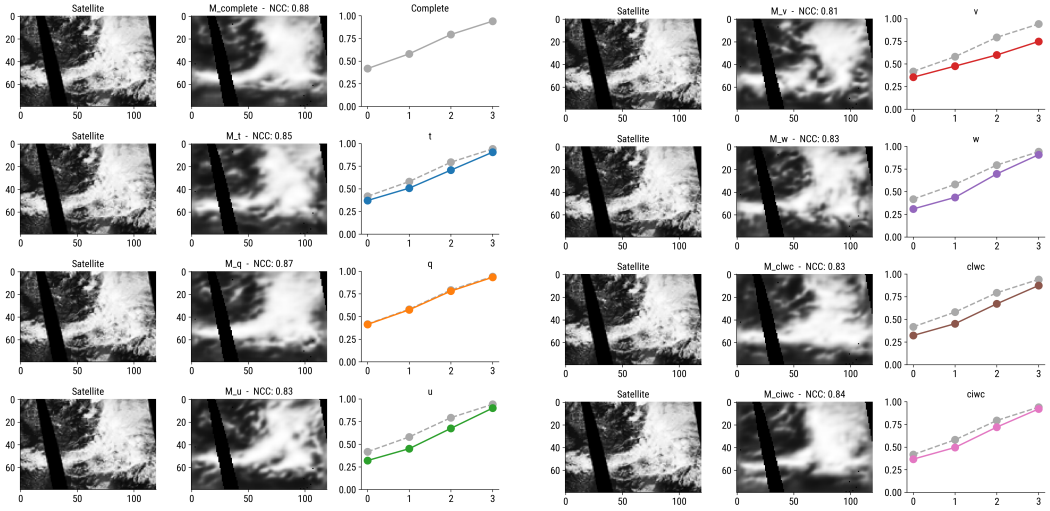


Fig. 9.6.: Qualitative comparison of minimal models across scales. Example from the Equator shows triplets of original satellite (left), the model's prediction (middle) and MS-SSIM scores for four different scales (right). All minimal models capture the coarse structure but differ in the fine-grained details.

the calculated NCC above each image. To better capture subtle differences beyond pixel-wise correlation, we additionally compute the MS-SSIM across four resolutions. Starting from the native ERA5 grid scale (31 km), the image is successively downsampled by factors of two, yielding three increasingly coarse representations. The resulting multi-scale similarity curves are shown in the right panel.

The multi-variate model (gray curve) achieves the highest similarity across all spatial scales, confirming that combining variables improves reconstruction consistently from coarse to fine structure. Among the minimal models, the specific humidity model MM_q comes closest to the full model at all resolutions. This indicates that moisture alone already encodes a substantial fraction of both large-scale organization and fine-grained cloud texture in tropical regimes.

Other models exhibit scale-dependent behavior. For instance, the vertical velocity model MM_w captures much of the large-scale organization, as reflected by competitive MS-SSIM values at coarse resolution, but deteriorates at finer scales. This suggests that dynamical lifting contains information about mesoscale structure but lacks sufficient thermodynamic detail to reproduce fine-scale cloud variability. Similar scale-specific limitations are visible for other isolated variables.

Taken together, these results support the hypothesis that atmospheric variables encode cloud-relevant information in a partially redundant yet complementary manner. While specific humidity dominates in overall decodability, other variables contribute structurally distinct components that become particularly relevant at certain spatial scales. The atmospheric state therefore appears to store information about clouds in a distributed encoding, where different physical fields emphasize different aspects of the same cloud system.

We emphasize, however, that assessing similarity in high-dimensional cloud images remains inherently challenging. Metrics such as NCC and MS-SSIM provide quantitative guidance but should be interpreted alongside visual inspection. Future work should focus on disentangle cross-variable dependencies more explicitly, for example by analyzing conditional contributions or synergistic effects. Such investigations may help clarify how minimal predictors can inform both data-driven modeling strategies and theoretical understanding of cloud formation processes.

9.2.3 Temporal Persistence of Predictive Information

While our initial analysis focuses on cloud field reconstruction based on atmospheric situation at time t_0, we now investigate how far into the future this information remains predictive. This allows us to operationally define a horizon of predictability: the temporal boundary beyond which prior atmospheric conditions no longer contain sufficient information to reliably reconstruct future clouds.

We evaluate this across the same two meteorologically distinct regions as in the previous section – the highly dynamic North Atlantic and the relatively stable Equatorial zone. For both regions, we generate cloud field predictions at time t_0 using physical input data from earlier times $t = t_0 - t_1$, where t_1 spans from 0 to 96 hours. A separate model is trained for each time lag. Prediction quality is quantified using NCC with observed satellite images at t_0.

Importantly, this setup does not try to measure forecast errors, but the amount of predictive information about clouds in the future. Therefore, we use reanalysis data as input – meaning the physical fields represent the true atmospheric state at time t – and combine them with actual satellite images

at time t_0. This isolates the temporal decay of predictive information in the atmosphere from errors introduced by numerical prediction systems.

As shown in Fig. 9.7, the two regions exhibit markedly different temporal decay patterns. Over the Equator, predictive accuracy declines gradually but consistently remains above the optical thickness (at lead time zero) baseline, even at extended time lags. This slow decay reflects the relatively persistent large-scale organization of tropical convection and moisture fields, where dynamical variability is weaker and thermodynamic structure evolves more gradually. In contrast, the North Atlantic shows a rapid deterioration in predictive skill. After a 24-hour offset, the NCC drops below the baseline of the optical thickness at lead time zero and continues to decline further. Beyond roughly 48 hours, the atmospheric state at time t becomes effectively uninformative for reconstructing cloud appearance at time t_0. This sharp decay is consistent with the strong dynamics in mid-latitudes, where cloud systems reorganize on short timescales.

These results emphasize how regional atmospheric dynamics modulate the persistence of predictive signals. Highly variable regions like the North Atlantic limit the temporal reach of data-driven models, while more stable regions such as the tropics retain predictive information over longer horizons. Understanding this temporal structure of decodability has practical implications for data selection, model design, and the interpretation of learned representations.

9.2.4 Cross-Dataset Generalization

Data assimilation combines numerical weather prediction models with observational data to produce the best available estimate of the atmospheric state. Over oceanic regions – our area of focus – satellite observations dominate the assimilation streams. As a result, ERA5 reanalysis data incorporates information from the same satellite instruments (e.g. Terra and Aqua) that provide the cloud imagery used as ground truth in this study. This raises a critical concern: a model trained on ERA5 might partially retrieve satellite-imprinted information rather than learning physically meaningful relationships between atmospheric variables and cloud structures. To address this potential leakage, we evaluate our model on independent input data from the IFS. Unlike reanalysis data, IFS forecasts progressively lose direct obser-

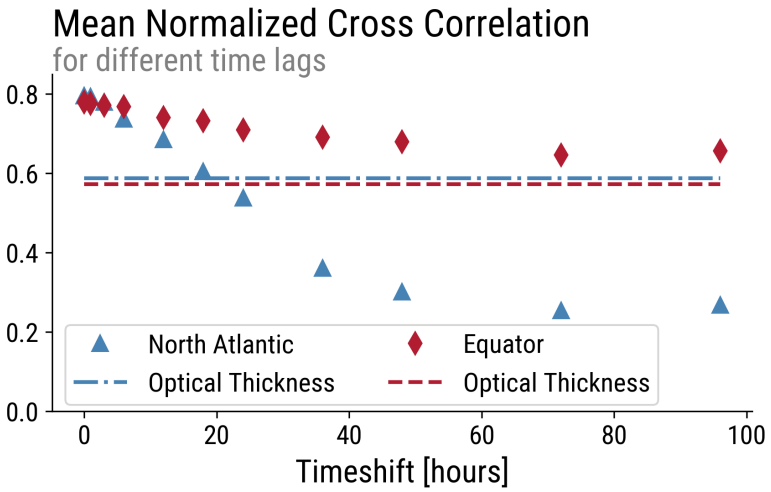


Fig. 9.7.: Prediction quality measured in average NCC over increasing time lags between atmospheric input and observed cloud fields, comparing two climatic zones. The North Atlantic shows a steep decline in accuracy after 24 hours, while the Equator retains meaningful predictive power up to 96 hours. Dashed lines show the optical thickness baseline for both regions.

vational constraints as lead time increases, making them less influenced by the original assimilated satellite imagery.

Specifically, we evaluate the non-regional baseline model – trained on a broad longitudinal stripe of ERA5 data – on previously unseen IFS forecasts for six representative days in 2014. These dates were selected randomly a priori, not based on model performance. We use forecasts initialized at $00:00$ and $12:00$ UTC and focus on lead times between 7 and 48 hours. This window minimizes the immediate influence of data assimilation while avoiding excessive forecast degradation.

We first asses prediction performance via NCC between the model's output and satellite imagery, comparing IFS and ERA5 data as input. It is important to distinguish the temporal meaning of both inputs in this comparison. While ERA5 represents the best possible estimate of the atmospheric state at the verification time, IFS inputs corresponds to true forecasts: the atmospheric state evolved forward in time for 7-48 hours without further assimilation. Lead time therefore applies only to IFS, not to ERA5.

Fig. 9.8 shows the mean NCC values averaged across IFS lead times between 7 and 48 hours, with the horizontal lines representing the overall mean. Within the first hours, the model's prediction quality remains relatively stable. As expected, performance degrades with increasing IFS forecast lead time. Nev-

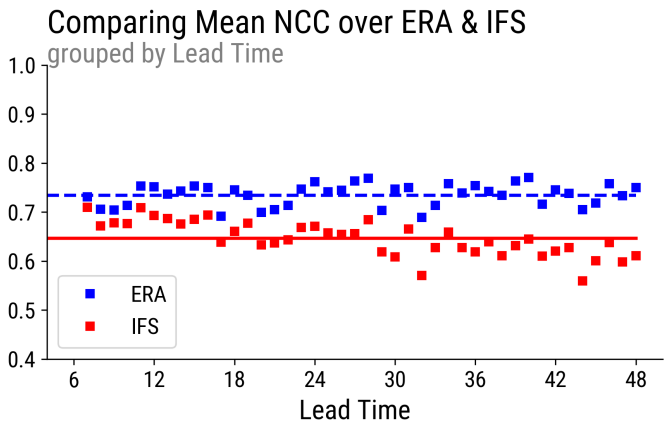


Fig. 9.8.: Average NCC across varying lead times (7-48 hours) for ERA5 and IFS input data. The model was trained on ERA5 and generalizes to IFS inputs with only moderate deterioration, demonstrating its ability to operate on forecast data not seen during training.

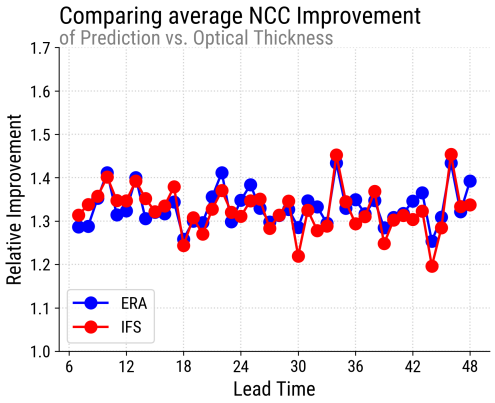
ertheless, reconstruction quality remains robust. The model achieves a mean NCC of 0.73 on ERA5 and 0.65 of IFS.

From an information-theoretic perspective, this experiment functions as a domain shift test. If the model had merely learned to decode assimilation artifacts specific to ERA5, performance would collapse when applied to IFS input. The sustained predictive skill instead suggests that the learned mapping captures physically consistent patterns embedded in the atmospheric state, rather than dataset-specific shortcuts.

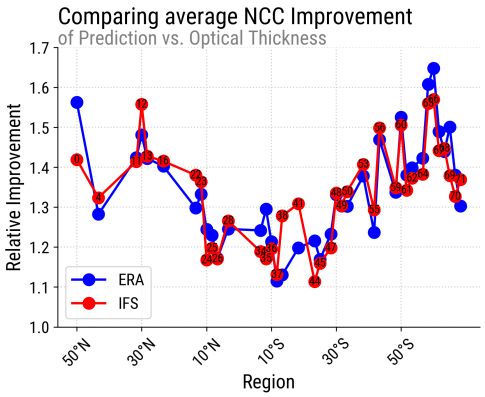
To disentangle differences arising from forecast lead times and focus on the impact of the input data source, we evaluate the model's relative improvement over physical cloud estimates represented by optical thickness. For each input dataset, we compute the NCC between optical thickness and satellite imagery and compare it to the NCC achieved by the neural network output. The ratio between these two scores quantifies the additional information extracted by the learned model beyond what is directly encoded in the physical proxy. This relative improvement is calculated separately for ERA5 and IFS inputs.

Fig. 9.9 summarizes the results. The model consistently outperforms the optical thickness baseline. On average, using ERA5 input yields an improvement of approximately 43%, while IFS input still results in a gain of roughly 38%.

To gain deeper insights, we analyze the improvements both as a function of IFS lead time (fig. 9.9a) and as a function of geographic region (fig. 9.9b). The



(a) Relative NCC improvement averaged over lead times for ERA5 and IFS input showing comparable results for both datasets.



(b) Relative NCC improvement averaged over regions for ERA5 and IFS input showing comparable results for both datasets. Large regional differences are shown between tropical and temperate zones.

Fig. 9.9.: Relative prediction improvement over optical thickness baseline for ERA5 and IFS. Please note that gains remain consistently positive; the statistical model outperforms optical thickness significantly also in forecast setting.

key observation is the structural similarity between ERA5- and IFS-driven improvements. Across both lead times and regions, the neural network consistently outperforms the optical thickness baseline for both input types. This consistency indicates that the model's predictive capability does not rely solely on assimilation-imprinted satellite signals specific to ERA5, but instead reflects learned features of physically meaningful atmospheric structure.

While the absolute NCC values show a slight decline towards longer lead times, the relative improvement over the optical thickness remains stable for both ERA5 and IFS input data. Across lead times, the improvements typically range between 20 and 50%, without a clear systematic dependence on forecast range.

In contrast, fig. 9.9b reveals a pronounced latitudinal dependence. Near the Equator, both ERA5 and IFS inputs show only modest improvements (around 10% in NCC). The relative gain increase significantly toward the poles,reaching values of roughly 40%. This pattern suggests that in tropical regimes, optical thickness already captures a large fraction of the cloud-relevant information contained in the atmospheric state. The neural network can refine this estimate but adds comparatively limited additional structure. In dynamically active mid- and high-latitude regions, cloud forma-

tion depends on more complex interactions among atmospheric variables. These nonlinear relationships are not fully represented by optical thickness alone, allowing the neural network to provide substantially larger improvements.

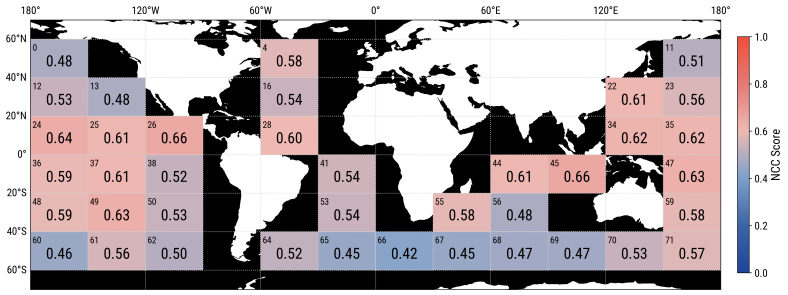
Since the high variability of optical thickness in terms to varying regions is well known, we consider not only the relative change in prediction quality, but also the absolute values. Fig. 9.10 displays the spatial distribution of the average NCC values for both the optical thickness and the model's predictions using ERA5 and IFS input data. The maps indicates that the physical estimation of optical thickness performs relatively well in tropical regions but declines in the temperate zones, particularly in the Southern Hemisphere (figures 9.10a and 9.10c). This pattern is less pronounced for IFS input data but still visible. Conversely, the model achieves higher prediction accuracy in these temperate regions than in the tropics, especially when applied to IFS inputs. As a result, the overall improvement in cloud reconstruction is more limited in tropical areas, where the baseline already performs well, and more pronounced in regions where the physical model struggles. This explains the drop in improvement for tropical regions.

These findings support the interpretation of neural networks as information extraction tools that uncover physically meaningful structure in atmospheric data rather than merely reproducing assimilated observational signals.

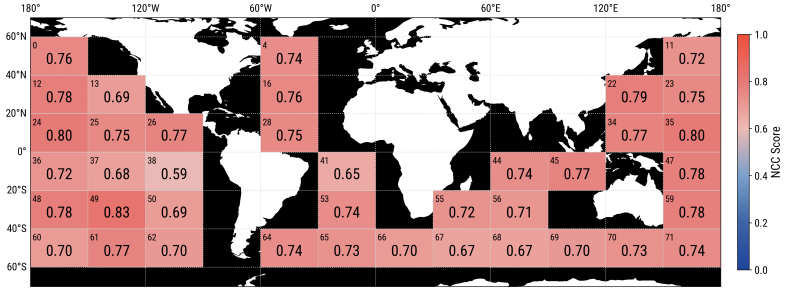
9.3 Attribution-Based Identification of Predictive Variables

We have shown that reconstructing cloud fields from the current atmospheric state is not only feasible but also surpasses the performance of conventional physical-based methods. In the framework of this thesis, such predictive capability is interpreted as evidence that the atmospheric state contains decodable information about cloud structures. However, despite their predictive success, neural networks represent complex nonlinear mappings implicitly, which poses a significant challenge for interpreting the underlying physical relationships. Moreover, the large volume of attribution data generated by pixel-wise explanations complicates systematic analysis.

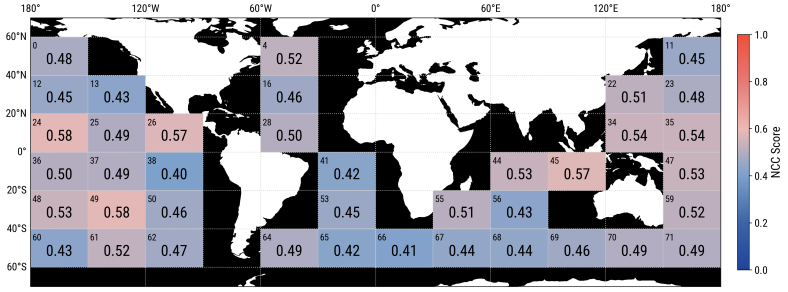
To address these challenges, we apply gradient-based saliency methods to analyze how the trained models utilize information contained in the atmosphe-



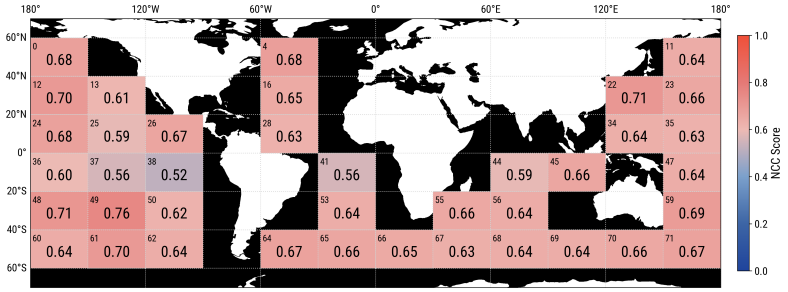
(a) Regional averaged NCC of optical thickness for ERA5 showing high performance in tropical regimes while being slightly worse especially in the South.



(b) Regional averaged NCC of model's output for ERA5 showing a strong performance in all areas.

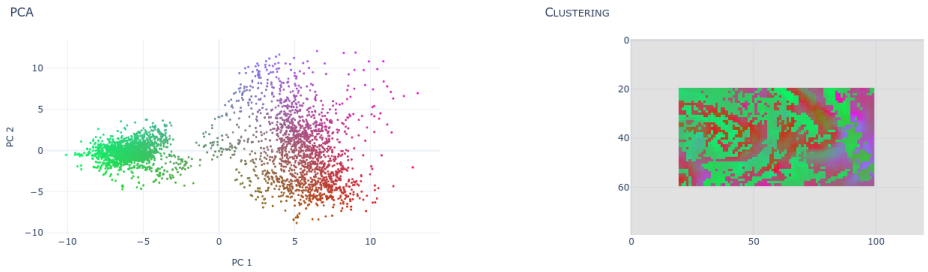


(c) Regional averaged NCC of optical thickness for IFS showing weaker accuracy.



(d) Regional averaged NCC of model's output for IFS showing good performance with some weaknesses in the tropics.

Fig. 9.10.: Comparison of average NCC values per region calculated for different input types (ERA5 vs. IFS) and methods (optical thickness vs. our model's prediction).



(a) 2D embedding of the attributions. The colors match the x- and y-coordinates.

(b) Model prediction of the same situation. Colors are chosen accordingly to the 2D embedding of the attributions.

Fig. 9.11.: Analyzing the attribution of our model for a specific situation in the south pacific, regarding the PCA embedding.

ric input variables. Our custom visualization tool enables interactive exploration of these saliency maps in both two- and three-dimensional representations, providing a first step toward understanding the informational pathways used by the model. Although the analysis is exploratory at this stage, the results reveal consistent patterns that provide insight into the distribution of predictive signals across physical inputs. It should be emphasized that saliency maps cannot identify physical causalities, but characterize how the learned decoder utilized available information in the input space.

When generating saliency maps, a three-dimensional attribution map is produced for each input variable and output pixel within a defined spatial window. For our setup – with regions of 80×120 pixels and a 20-pixel offset – this results in roughly $3,200$ saliency maps per example. To move beyond local pixel-level explanations and identify broader structural patterns in these high-dimensional attribution fields, we apply a two-dimensional Principal Component Analysis (PCA) embedding to the aggregated attribution data. This dimensionality reduction highlights the dominant modes of variation in attribution space and enables a systematic comparison across spatial locations.

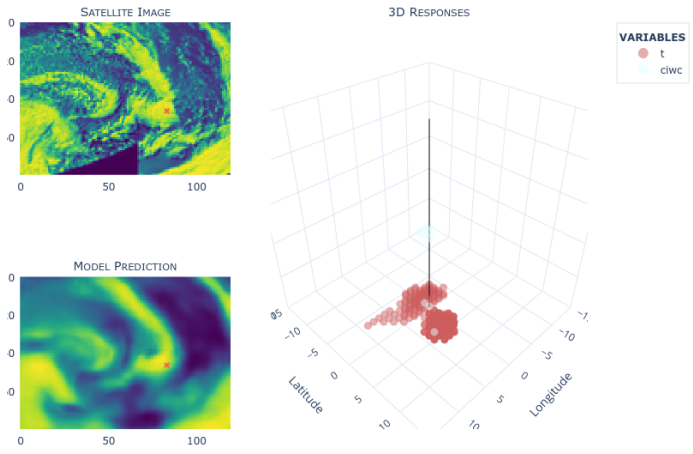
The resulting PCA embedding consistently reveals two well-separated clusters (see fig. 9.11a), suggesting that the model distinguishes between at least two distinct predictive regimes. This clustering pattern is robust across examples, although the spatial extent of each cluster varies. Some clusters remain tightly localized, while others span larger regions, reflecting the varying spatial footprints of underlying physical processes. More cluster examples are shown in the appendix B.1.

A closer inspection reveals that these dominant attribution modes correspond to two key physical drivers: near-surface temperature, which modulates convection processes, and mid-tropospheric specific humidity (up to approximately 12 km altitude), which governs moisture availability for condensation and cloud formation. To further interpret these clusters, we color-code the PCA points according to their spatial position – using a gradient from green to red along the x-axis and increasing blue intensity along the y-axis – and project these colors back onto the predicted cloud fields (see fig. 9.11b). This visualization reveals a clear alignment between attribution clusters and the presence or absence of cloud structures, confirming a strong link between the model's internal attribution and physical cloud regimes. Some examples of attribution maps for different regions are shown in the appendix B.1.

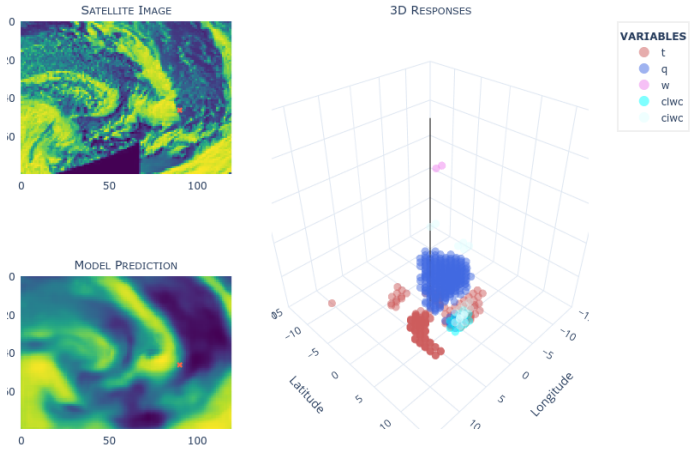
Examining specific regions provides additional qualitative insights. For example, in the South Pacific we observe distinct attribution patterns depending on the spatial relationship to cloud structures. Within clouds, the model primarily attributes its predictions to near-surface temperature with a relatively homogeneous spatial distribution (see fig. 9.12a). Near cloud edges, attribution shifts toward specific humidity in the mid-troposphere (between 3-10 km altitude), while temperature contributions begin to reflect the spatial outline of the cloud (see fig. 9.12b). In cloud-free regions containing scattered "popcorn"-like structures, the attribution maps exhibit similarly fragmented spatial patterns (see fig. 9.12c).

Given the known limitations of saliency methods, we also assess the robustness of these findings by training 25 independent model instances and comparing their attribution patterns. Encouragingly, the mean attribution signal remains largely stable across models, indicating that the dominant attribution structures are not driven by random initialization effects. In particular, attribution signals at altitudes above 15 km appear only sporadically across individual models and tend to vanish when averaged across the ensemble. This behavior suggests that such high-altitude signals are likely artifacts of training noise rather than physically meaningful relationships.

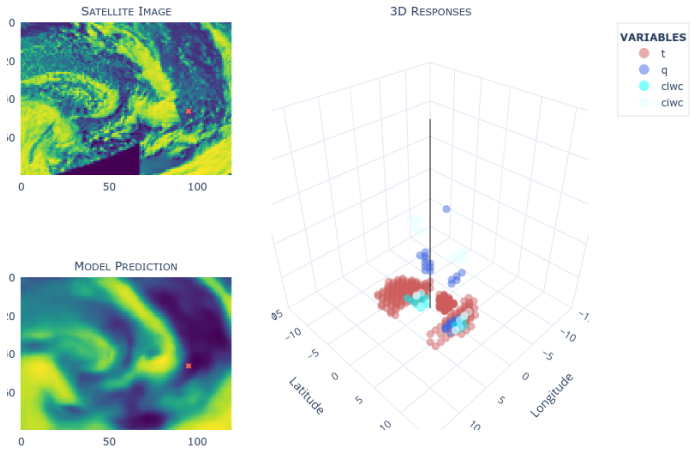
So far, our analysis has focused on explaining the factors that influence reconstruction of the current atmospheric state. However, a more consequential and still open question concerns which physical processes are most relevant for predicting the future evolution of cloud fields. Identifying the temporal pathways through which predictive information propagates in the atmosphere therefore represents an important direction for future research.



(a) Attribution for cloud prediction inside a cloud.



(b) Attribution for cloud prediction at the edge of a cloud.



(c) Attribution for cloud prediction without a cloud.

Fig. 9.12.: Analyzing the attribution of our model for a specific situation in the south pacific, regarding the attributions inside and outside a cloud.

10 Insights into Cloud-Relevant Information

Modeling the microphysics of cloud structures remains quite challenging, yet it is crucial for improving weather forecasts and climate modeling. In this work, we approached the problem from an information-theoretic perspective, analyzing the extent to which cloud structures can be reconstructed and predicted using data-driven methods. We addressed the following three research questions:

(1) How well can we reconstruct the appearance of clouds from coarse-scale physical simulation data? We show that a convolutional neural network trained on reanalysis data can reconstruct cloud formations with higher accuracy than traditional physical estimates, such as optical thickness. This suggests that the information needed for accurate cloud prediction is embedded in the atmospheric state but not fully exploited by physical parameterizations. While fine-scale cloud features remain challenging to resolve, this gap may narrow with higher-resolution atmospheric inputs.

(2a) How well does the model generalize across space, time, physical variables and data sources? We systematically assessed four generalization dimensions to probe the structure and distribution of predictive information in the data. Analyzing a variety of different forms of generalization offers new insights into the information contained in the data, potentially informing physical research questions.

Testing our model on geographical regions it hasn't seen before reveals different strong general patterns contained in atmospheric data. Our analysis shows that generalization ability is highly related to latitudinal location. While tropical regions tend to produce models that are highly specialized for this climatic situation, temperate zones contain a more comprehensive representation of cloud generating processes.

Through the use of minimal models – each trained on a single physical quantity – we show that even isolated inputs such as specific humidity or temper-

ature carry substantial predictive power. Among all variables, specific humidity generally yields the highest reconstruction quality, though variability across cases indicates that cloud predictability is governed by complex, content-dependent interactions between variables. These findings highlight the potential of machine learning to reveal latent structures not captured by existing physical models as it is able to exploit hidden dependencies and the constraints and structures of real-world data distributions that do not cover the whole range of possible physical processes described by a handcrafted simulation model. It also shows that, under real-world data conditions, the model actually "knows" more about the clouds in a statistical (i.e., information-theoretic) sense than a general physical model would suggest.

(2b) How much does accuracy degrade with lead times in case of predicting future cloud formation? We extend our analysis to the temporal stability of predictive signals. Using atmospheric data from previous states, we evaluated how well future cloud fields can be reconstructed. Results show that predictive information degrades more rapidly in highly dynamic regions (e.g., the North Atlantic), with the model's skill falling below baseline levels after about 20 hours. In contrast, in more stable regions such as the Equator, meaningful predictions remain feasible up to 96 hours. This is in line with the fact that the horizon of predictability is highly dependent on regional meteorological variability.

Observational data is embedded into reanalysis data by data assimilation. To exclude that our model only extracts this information previously embedded into the data, we apply our model to a dataset it hasn't seen during training, arising from Integrated Forecasting System (IFS). Comparisons against the physical baselines confirm our model's ability to improve cloud structure prediction. In addition to excluding data leakage, this might also point towards a potential, promising direct application of such a model.

(3) What insights into the underlying physical processes can we gain? While neural networks do not yield explicit, human-interpretable equations, we applied gradient-based saliency methods to analyze the internal logic of our model. These methods highlight the most influential regions and variables for cloud reconstruction. Our analysis consistently identifies near-surface temperature and mid-level specific humidity as dominant contributors – both known drivers of cloud formations. Importantly, these findings are derived in a purely data-driven manner, illustrating how computational tools can complement physical reasoning and help refine our understanding of atmospheric processes.

However, the results must be interpreted with caution. Gradient-based saliency methods have inherent limitations: they indicate the direction of change needed in the input to affect the output, but not the increase or decrease of that change. Additionally, they emphasize regions with high sensitivity, which does not necessarily equate to causal importance. As demonstrated (quite vividly) by our minimal model experiments, other physical effects – though less prominent in attribution maps – can still play a significant predictive role. Thus, while saliency analysis provides valuable clues, it should be seen as one layer of interpretation rather than a definitive explanation.

Limitations and Future Work Currently, our model is restricted to oceanic regions, where the distinction between clouds and surface is rather simple. Extending the method to land regions will introduce additional challenges and new techniques to include orography. While our approach does not produce explicit physical equations, it provides valuable statistical evidence about the encoding and flow of predictive information in the atmosphere. It is rather general and can be applied in the future to other observational sources – such as radar data, aircraft measurements or high-resolution simulations – to further enhance physical interpretability.

Part III

Reliability of Saliency Methods for Scientific Interpretation

11 Conceptual Integration

In this part, we apply the fundamental robustness tests developed earlier to a real-world application in order to evaluate the reliability of saliency methods as diagnostic tools in practice.

In Part I, we established basic robustness criteria that saliency methods must satisfy if neural networks are to be used as diagnostic instruments. Our analysis focused particularly on the influence of random model initialization in image classification tasks. We found that saliency maps produced by both gradient-based and more sophisticated attribution methods can exhibit substantial variability across independently initialized models. This variability does not necessarily imply that attribution results are incorrect, but it indicates that individual explanations may highlight incomplete or unstable features. These findings emphasize the need for caution when interpreting saliency maps and using them to extract scientific insights.

In contrast, Part II investigated neural networks as tools for extracting information from high-dimensional real-world data. Using models trained on large-scale atmospheric datasets, we examined which aspects of cloud formation are already encoded in the atmospheric state variables. By shifting from traditional physics-based equations towards a data-driven approach, we demonstrated that even isolated physical quantities carry substantial predictive information. These results highlight the potential of machine learning to uncover latent relationships that are not explicitly represented in existing physical models.

Taken together, these findings expose an underlying tension. On the one hand, Part I shows that saliency maps can be highly unstable. On the other hand, Part II suggests that these maps have the potential to reveal physically meaningful structures in complex atmospheric data. This naturally raises a critical question: if saliency methods appear to reveal interpretable physical structures, to what extent can these interpretations be considered robust?

Addressing this question is the objective of the present part. The synthesis serves two purposes: First, it tests the generality of the proposed robustness

framework by applying it to a different scientific domain and to a regression-based prediction task rather than image classification. Second, it allows us to evaluate whether the physically interpretable attribution patterns identified in the cloud reconstruction experiments represent stable properties of the learned models or whether they may arise from random initialization effects. This analysis therefore represents a critical step toward establishing neural networks as reliable scientific instruments for investigating complex physical systems.

In the following chapter, we implemented this synthesis in practice. We apply the robustness framework to an ensemble of cloud prediction models and quantitatively evaluate the stability of their saliency maps. This analysis bridges the methodological insights on initialization robustness with the physical interpretation of atmospheric patterns, establishing the foundation for the broader discussion presented in Chapter 13. Finally, Chapter 14 summarizes the main findings of this work.

12 Robustness of Saliency Methods in Physical Regression Tasks

The previous chapter identified a conceptual tension between data-driven physical interpretability and the initialization sensitivity of saliency methods. In this chapter, we examine the reliability of saliency maps in regression-based scientific tasks. Is physical interpretability stable, or is it merely an artefact of initialization noise?

To address this question, we apply the robustness framework from Part I to the cloud prediction models introduced in Part II, thereby providing an empirical validation of the earlier results. We restrict our analysis to initialization noise and exclude other potential sources of variability, i.e. differences in model architecture or training schemes.

According to prior studies cited in Chapter 2, learned features and their corresponding attribution patterns tend to be more stable in regression settings than in classification tasks. Therefore, we hypothesize that initialization noise has a smaller impact on cloud prediction models, resulting in more robust and interpretable saliency maps that can be used to draw conclusions about physically atmospheric processes.

12.1 Experimental Setup

Part II analyzed the information content of atmospheric states using models trained on spatially localized regions. However, Section 9.2.1 showed that generalization behavior is primarily structured by latitude rather than longitude. To obtain a representative yet computationally efficient setup, we therefore decided to move away from single-region models and instead train a model on a full longitudinal stripe. Specifically, we select the region be-

tween $180°$ and $150°$ West, covering predominantly oceanic areas and six latitudinal subregions (see fig. 8.1 for details).

To assess attribution robustness, we follow the approach from Part I and train an ensemble of 40 independently initialized U-Net models. All models share identical architecture, training data, and hyperparameters, and differ only in their random initialization, isolating initialization noise as the sole source of variability.

Models are trained using ADAM optimizer with a One-Cycle learning rate schedule (max LR 8×10^{-3}, batch size 16) and He initialization. As the training objective, we use a weighted combination of structural similarity index measure (SSIM) and L1 loss with weights 0.8 and 0.2, respectively. The SSIM component emphasizes structural similarity between predicted and observed cloud fields, which is particularly important for capturing coherent cloud patterns, while the L1 loss stabilizes training by penalizing pixel-wise deviations.

Our training dataset consists of data from the years 2010 to 2013, while the validation set is drawn from the year 2016. This temporal separation ensures that the model cannot memorize specific atmospheric situations but must learn generalizable relationships between atmospheric variables and cloud structures.

For each model and input sample, we compute gradient-based saliency maps that measure the sensitivity of the predicted cloud field with respect to the input variables. Concretely, for each output pixel, we evaluate gradients within a local spatial window, resulting in high-dimensional attribution tensors over spatial offsets, variables, and vertical levels.

We consider only the magnitude of the gradients, discarding their sign. This focuses the analysis on the strength of influence rather than its direction and avoids cancellation effects when aggregating across models.

To ensure comparability across variables and samples, attribution values are normalized per sample by their maximum absolute value. Furthermore, to reduce computational cost, attributions are evaluated on a spatial grid with step size 4. Since all subsequent analysis aggregates over spatial dimensions, this subsampling does not affect the resulting statistics.

To quantify robustness with respect to initialization, we aggregate attribution tensors across the ensemble of 40 independently trained models. For

each sample, we compute the mean μ and standard deviation σ of the attribution values across models.

We then reduce the attribution tensors by aggregating over spatial dimensions, resulting in a compact representation of attribution strength for each physical variable and vertical level. Formally, we compute

$$\mu(v, lv) = \sqrt{\sum_{w,h,win_x,win_y} m^2}, \quad \sigma(v, lv) = \sqrt{\sum_{w,h,win_x,win_y} s^2},$$

where m and s denote the mean and standard deviation across models before spatial aggregation.

This aggregation yields a representation that captures how strongly and consistently each variable contributes to the prediction across model initializations. It sacrifices spatial precision in favor of global interpretability, enabling a variable-level analysis of attribution stability.

Based on these quantities, we define the attribution-wise signal to noise ratio (SNR) as

$$\mathrm{SNR}_{\mathrm{attr}}(v, lv) = \frac{\mu(v, lv)}{\sigma(v, lv)}. \tag{12.1}$$

Entries with negligible mean attribution are excluded to avoid misinterpreting inactive features as noise-dominated. The final results are obtained by averaging across samples. This metric directly quantifies the stability of attribution patterns across model initializations and allows for a variable- and altitude-resolved analysis of interpretability.

12.2 Results

We evaluate attribution robustness across an ensemble of 40 independently initialized models. All models achieve comparable predictive performance (normalized cross-correlation (NCC) ≈ 0.74), ensuring that observed differences in attribution are driven by initialization rather than model quality. Due to computational constraints, the analysis is performed on 60 randomly selected validation samples.

The attribution-wise SNR aggregated across samples is shown in Fig. 12.1. High SNR values indicate that attribution signals are consistent across model

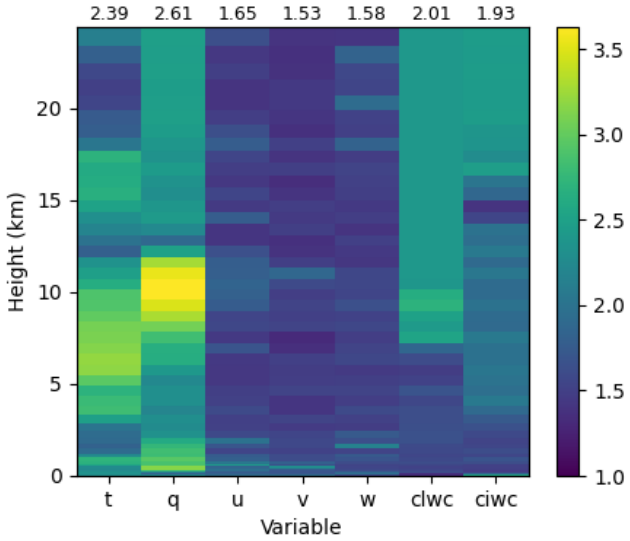


Fig. 12.1.: Attribution-wise SNR across variables and vertical levels, averaged over all samples and spatial locations. High values highlight variables with stable attribution patterns across model initializations.

initializations, while low values correspond either to noisy or inactive features. The averaged SNR across all heights for each variable is shown at the top of each column. For further inspection, attribution-wise SNR maps for several individual samples are presented in Appendix B.2. These examples confirm that the overall attribution patterns remain largely consistent across samples.

A clear structure emerges: specific humidity q, temperature t, and cloud water content $clwc$ exhibit consistently high SNR values, often exceeding 2. In particular, mid-tropospheric humidity and near-surface temperature show the highest stability. These findings align closely with the results from Part II, where these variables were identified as the most informative predictors for cloud reconstruction. The consistency across independently initialized models indicates that these attribution patterns represent stable structures learned from the data rather than artifacts of random initialization. In contrast, wind-related variables exhibit lower SNR values across most altitudes, suggesting a weaker and less consistent contribution to the prediction.

To distinguish between instability and irrelevance, we analyze the absolute mean and standard deviation of attribution values (Fig. 12.2). Regions with low SNR are characterized by both low mean and low variability. This indicates that these variables contribute only marginally to the prediction, rather

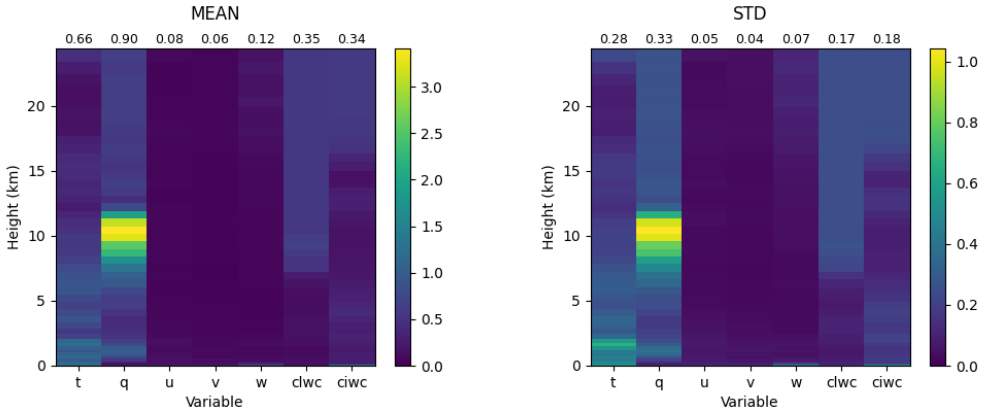


Fig. 12.2.: Mean (left) and standard deviation (right) of attribution values across variables and vertical levels. Regions with low SNR are characterized by both low mean and low variability, indicating weak relevance rather than instability.

than being dominated by noise. In other words, low SNR seems to reflect weak relevance rather than instability in this specific context.

Overall, the results show that attribution patterns for the most physically relevant variables are stable across model initializations. Compared to the strong variability observed in classification settings in Part I, attribution maps in this regression task exhibit substantially higher robustness. These findings are consistent with the hypothesis that regression-based models may yield more stable attribution patterns than classification settings, enabling more reliable interpretation of learned relationships.

13 Discussion

The previous chapters addressed complementary aspects of a common question: under which conditions, and to what extent, can neural networks serve as tools for extracting scientifically useful information from complex data? The results suggest that this is indeed possible, but only if information extraction is approached from more than one perspective. In particular, this work has shown that two complementary questions must be considered: first, whether model interpretations are robust enough to support scientific analysis, and second, whether the predictive information contained in the data can be localized, restricted, and characterized through systematic manipulation of the input space. This chapter synthesizes the results of the present work, places them in a broader methodological context, and discusses the limitations of the proposed approach.

13.1 Information Extraction with Neural Networks

Taken together, the three parts of this thesis support a view of neural networks that goes beyond their role as powerful prediction models. In the present context, their scientific usefulness lies in their capacity to probe the information content of atmospheric data. More specifically, they can be used to investigate whether coarse-scale atmospheric fields contain sufficient information for cloud-related prediction, how this information is distributed across variables, regions, and temporal contexts, and whether the patterns highlighted by the models are stable enough to support interpretation. From this perspective, prediction is not the endpoint of the analysis, but the starting point for asking what the model must have extracted from the data in order to perform well. In other words, high predictive skill demonstrates that relevant information is present in the data, but it does not by itself reveal how this information is distributed, which parts of the input are sufficient for prediction, or whether the learned relevance patterns are stable enough to support scientific interpretation.

A first important contribution of this thesis is to show that such information should be treated with caution. Part I demonstrated that attribution-based saliency maps are sensitive to random model initialization. Even when independently trained networks achieve similar predictive performance, the resulting attribution patterns can differ substantially. This result has an important conceptual consequence: if saliency maps are used as a means of extracting information about the decision process from trained models, then the extracted information is not automatically unique or stable. Instead, interpretability itself becomes an objective of uncertainty analysis. This should not be taken to deny the value of saliency maps. However, they cannot be treated as transparent windows into model decision-making. Their value depends on whether consistent patterns can be distinguished from variability introduced by the training process.

Interpretability is only one possible route for information extraction. Part II shows that neural networks can also be used in a broader and, in some sense, more direct way to study the information content of the data itself. In the cloud prediction setting considered here, the strong predictive performance of the models indicates that coarse-scale atmospheric fields contain substantial cloud-relevant information, more than is currently exploited by traditional parameterizations. By systematically varying the input space across variables, regions, and temporal settings, the thesis examines how much information is actually required to maintain useful prediction. These reduced-input experiments reveal that predictive skill remains partly preserved even under substantial restrictions. This suggests that cloud-relevant information is distributed across the atmospheric state and partly redundant.

This restriction-based approach complements the attribution-based analysis. Saliency methods primarily address local relevance within trained models: they indicate which parts of the input appear to have the strongest influence on a given prediction. Input restrictions, by contrast, probe the sufficiency and distribution of predictive information more globally: they estimate which subsets of the input still allow the model to perform well, and therefore what this implies about how information is organized in the data. The two perspectives therefore address different but related questions. One concerns how the trained model allocates relevance, the other concerns how much useful information is available in different parts of the input space. Considering both together provides a richer picture of information extraction than either approach could offer alone.

The concerns raised in Part I make clear that attribution patterns should be verified in terms of reliability. Part III addresses this issue by asking whether the attribution-based information extracted from cloud prediction models is sufficiently stable to support scientific interpretation in this regression setting. The signal to noise based analysis show that, although uncertainty remains, attribution patterns are considerably more stable here than in previously studied classification settings. Although this is not proof, it suggests that saliency based interpretation might be strongly task-dependent and more reliable in regression settings. This could be explained by smoother target functions and less ambiguous decision boundaries in regression setting, leading to more stable features and therefore more stable attribution patterns.

Across all three parts, a coherent picture emerges. Part I establishes that interpretability-based information extraction is limited by model-dependent variability and therefore requires explicit robustness assessment. Part II shows that neural networks can nevertheless be scientifically informative because they reveal that cloud-relevant information is present, distributed, and partly redundant in coarse-scale atmospheric data. Part III then demonstrates that, in this setting, interpretability is not only possible but can attain a meaningful degree of reliability. The contribution of the thesis is therefore not to claim that neural networks provide direct explanations of atmospheric physics, but to show that they can function as diagnostic tools for identifying robust, prediction-relevant structure in complex data.

It is important to emphasize once again that neural networks do not reveal causal relationships. The patterns identified through saliency analysis and input restriction reflect statistical structure learned from the data, not physically guaranteed laws. Nevertheless, such structure can be scientifically valuable. It can indicate where relevant information resides, how strongly it is distributed across the atmospheric state, and which subsets of the data are sufficient for useful prediction. In this sense, neural networks can support scientific hypothesis generation and exploratory analysis, provided that their outputs are interpreted as evidence about data structure rather than as self-sufficient causal explanations.

In summary, neural networks should neither be understood merely as black-box predictors nor be granted an explanatory status they cannot justify. Instead, they could be used as tools for probing information content of complex datasets. To use them responsibly in this role requires both methodological caution and complementary analysis strategies. Attribution methods can

highlight potentially relevant structures, but their uncertainty must be assessed explicitly. Input restriction can reveal how predictive information is distributed and how much of it is retained under systematic constraints. Considered together, these approaches make it possible to use neural networks not just to predict atmospheric phenomena, but to study what the data themselves contain and how this information can be extracted in a robust and scientifically meaningful way.

13.2 Limitations & Future Work

Despite these promising results, several limitations must be acknowledged.

First, our analysis concentrated primarily on gradient-based saliency methods and a specific convolutional neural network architecture. While these methods are widely used and well suited to the problem, they represent only a subset of available interpretability approaches. It therefore remains an open question to what extent the observed robustness properties generalize to other model classes, such as transformer-based architectures, hybrid physics-informed models or alternative attribution methods that do not rely on gradients. Such comparisons would help determine which aspects of the observed behavior are method-specific and which reflect broader properties of information extraction in machine learning models.

Second, while machine learning models can serve as powerful tools for information extraction, they do not inherently encode physical laws or constraints. As a result, the patterns identified through saliency-based analysis reflect statistical dependencies learned from the data rather than physically enforced relationships. Therefore, physically plausible attributions cannot immediately be interpreted as guarantees of causal relationships. Future work could therefore combine the present approach with more strongly physics-informed modeling, causal inference strategies, or targeted process-based experiments in order to better connect extracted model behavior with underlying physical mechanisms.

Third, the empirical analysis in this work is limited to reanalysis data, which is characterized by relatively coarse spatial resolution and model-dependent uncertainties. These factors may restrict the achievable accuracy of cloud reconstruction and the granularity of interpretability analyses. In addition, the study focuses exclusively on oceanic regions, where surface properties

are comparatively homogeneous. Extending the framework to land surfaces could provide further insights and test the robustness of the conclusions.

Finally, while this thesis concentrates on cloud structure prediction, the proposed methodology is not restricted to this application. Future research could apply the same framework to other atmospheric processes such as convection or precipitation formation, or to different observational modalities, including radar or multi-spectral satellite data. Such extensions would help to assess the general applicability of the proposed interpretability and robustness analysis across a broader range of geophysical phenomena.

14 Conclusions on Neural Networks as Diagnostic Tools for Information Extraction

This thesis investigated under which conditions neural networks can be used as diagnostic tools for extracting scientifically useful information from complex atmospheric data. The main conclusion is that such use is possible, but only if information extraction is not inferred from predictive performance alone. Instead, it must be examined from two complementary perspectives: first, through interpretability methods whose reliability has to be assessed explicitly, and second, through systematic restriction of the input space in order to probe how predictive information is distributed in the data.

A central contribution of this work is the demonstration that saliency-based interpretation should itself be treated as a source of uncertainty. In particular, random initialization can induce substantial variability in attribution patterns even when predictive performance remains similar. This means that attribution maps should not be interpreted as uniquely reliable descriptions of model behavior without additional validation. At the same time, the results also show that such instability is not uniform across settings: in the atmospheric regression problem studied here, attribution patterns were considerably more stable than in classification tasks. This suggests that the reliability of interpretability may depend strongly on the structure of the prediction problem.

Additionally, this thesis showed that neural networks can be used to study the information content of atmospheric data more directly. Reduced-input analyses revealed that important cloud-relevant predictive information is distributed across the atmospheric state and remains partly preserved even under substantial input restriction. These findings indicate that the models do not rely on a single subset of variables, but capture broader and partly redundant statistical structure in the data. In this sense, neural networks

can contribute not only to prediction, but also to the analysis of where useful information is embedded, even though they do not provide direct causal explanations.

Overall, this thesis concludes that neural networks should be understood neither as mere black-box predictors nor as direct explanatory models of physical processes. Rather, they can function as scientifically useful diagnostic tools for identifying robust, prediction-relevant structure in complex datasets. Their value lies not in providing causal explanations, but in supporting the analysis of information content, guiding hypothesis generation, and revealing how predictive information is distributed in scientific data when combined with robustness-aware interpretation.

Bibliography

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. “Sanity checks for saliency maps”. In: *Advances in neural information processing systems* 31 (2018). DOI: 10.48550/arXiv.1810.03292 (cit. on pp. 9, 19, 49).

[2] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. “Gradient-Based Attribution Methods”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. Cham: Springer International Publishing, 2019, pp. 169–191. DOI: 10.1007/978-3-030-28954-6_9 (cit. on pp. 6, 24, 26).

[3] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. *Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast*. 2022. DOI: 10.48550/arXiv.2211.02556 (cit. on p. 11).

[4] M. Böhle, M. Fritz, and B. Schiele. *B-cos Networks: Alignment is All We Need for Interpretability*. June 2022. DOI: 10.1109/CVPR52688.2022.01008 (cit. on pp. 10, 49).

[5] N. D. Brenowitz and C. S. Bretherton. “Spatially extended tests of a neural network parametrization trained by coarse-graining”. In: *Journal of Advances in Modeling Earth Systems* 11.8 (2019), pp. 2728–2744. DOI: 10.1029/2019MS001711 (cit. on p. 11).

[6] R. J. Chase, K. Haynes, L. V. Hoef, and I. Ebert-Uphoff. *Score-based diffusion nowcasting of GOES imagery*. 2025. DOI: 10.48550/arXiv.2505.10432 (cit. on p. 10).

[7] H. Chen, X. Zhong, Q. Zhai, X. Li, Y. W. Chan, P. W. Chan, Y. Huang, H. Li, and X. Shi. *Skillful Nowcasting of Convective Clouds With a Cascade Diffusion Model*. 2025. DOI: 10.48550/arXiv.2502.10957 (cit. on pp. 10, 11).

[8] J. Choi, D. Seo, J. Jung, Y. Han, J. Oh, and C. Lee. “Cloud Detection Using a UNet3+ Model with a Hybrid Swin Transformer and EfficientNet (UNet3+ STE) for Very-High-Resolution Satellite Imagery”. In: *Remote Sensing* 16.20 (2024), p. 3880. DOI: 10.3390/rs16203880 (cit. on p. 10).

[9] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold. “A historical perspective of explainable Artificial Intelligence”. In: *WIREs Data Mining and Knowledge Discovery* 11.1 (2021), e1391. DOI: 10.1002/widm.1391 (cit. on p. 16).

[10] A.-K. Dombrowski, M. Alber, C. J. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. *Explanations can be manipulated and geometry is to blame*. 2019. DOI: 10.48550/arXiv.1906.07983 (cit. on p. 9).

[11] F. Doshi-Velez and B. Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. DOI: 10.48550/arXiv.1702.08608 (cit. on pp. 5, 9).

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. DOI: 10.48550/arXiv.2010.11929 (cit. on p. 34).

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *kdd*. Vol. 96. 34. 1996, pp. 226–231 (cit. on p. 59).

[14] C. Fang, Y. Gu, W. Zhang, and T. Zhang. "Convex Formulation of Overparameterized Deep Neural Networks". In: *IEEE Transactions on Information Theory* 68.8 (2022), pp. 5340–5352. DOI: 10.1109/TIT.2022.3163341 (cit. on p. 18).

[15] S. Foersch, C. Glasner, A.-C. Woerl, M. Eckstein, D.-C. Wagner, S. Schulz, F. Kellers, A. Fernandez, K. Tserea, M. Kloth, et al. "Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer". In: *Nature medicine* 29.2 (2023), pp. 430–439. DOI: 10.1038/s41591-022-02134-1 (cit. on p. 4).

[16] R. C. Fong and A. Vedaldi. "Interpretable Explanations of Black Boxes by Meaningful Perturbation". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 3449–3457. DOI: 10.1109/ICCV.2017.371 (cit. on p. 8).

[17] T. Freiesleben, G. König, C. Molnar, and Á. Tejero-Cantero. "Scientific Inference with Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena". In: *Minds and Machines* 34.3 (2024). DOI: 10.1007/s11023-024-09691-z (cit. on pp. 16, 21).

[18] A. Ghorbani, A. Abid, and J. Zou. "Interpretation of Neural Networks Is Fragile". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 3681–3688. DOI: 10.1609/aaai.v33i01.33013681 (cit. on p. 9).

[19] A. A. Goshtasby. "Similarity and dissimilarity measures". In: *Image registration: Principles, tools and methods* (2012), pp. 7–66. DOI: 10.1007/978-1-4471-2458-0_2 (cit. on p. 47).

[20] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/arXiv.1512.03385 (cit. on p. 23).

[21] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek. "Explainable AI Methods - A Brief Overview". In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Ed. by A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek. Cham: Springer International Publishing, 2022, pp. 13–38. DOI: 10.1007/978-3-031-04083-2_2 (cit. on p. 16).

[22] S. Khorram, T. Lawson, and L. Fuxin. “IGOS++: Integrated Gradient Optimized Saliency by Bilateral Perturbations”. In: *Proceedings of the Conference on Health, Inference, and Learning*. CHIL ’21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 174–182. DOI: 10.1145/3450439.3451865 (cit. on p. 8).

[23] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon. “Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps”. In: (2019), pp. 4149–4157. DOI: 10.1109/ICCVW.2019.00510 (cit. on p. 7).

[24] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. *The (Un)reliability of saliency methods*. 2019. DOI: 10.1007/978-3-030-28954-6_14 (cit. on p. 9).

[25] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/arXiv.1412.6980 (cit. on pp. 25, 46).

[26] A. Krizhevsky et al. “Learning multiple layers of features from tiny images”. In: (2009) (cit. on p. 23).

[27] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, et al. “GraphCast: Learning skillful medium-range global weather forecasting”. In: *Science* 382.6677 (2023), pp. 1416–1421. DOI: 10.1126/science.adi2336 (cit. on p. 11).

[28] T. Laurent and J. von Brecht. “Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 2902–2907 (cit. on p. 29).

[29] Z. C. Lipton. “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery.” In: *Queue* 16.3 (June 2018), pp. 31–57. DOI: 10.1145/3236386.3241340 (cit. on pp. 5, 9).

[30] C. Liu, L. Zhu, and M. Belkin. *On the linearity of large non-linear models: when and why the tangent kernel is constant*. 2020. DOI: 10.48550/arXiv.2010.01092 (cit. on p. 18).

[31] S. Lundberg and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. DOI: 10.48550/arXiv.1705.07874 (cit. on pp. 8, 18).

[32] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3rd ed. 2025 (cit. on pp. 5, 6).

[33] C. Molnar, T. Freiesleben, G. König, G. Casalicchio, M. N. Wright, and B. Bischl. *Relating the partial dependence plot and permutation feature importance to the data generating process*. 2021. DOI: 10.48550/arXiv.2109.01433 (cit. on p. 9).

[34] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl. “General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models”. In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Ed. by A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek. Cham: Springer International Publishing, 2022, pp. 39–68. DOI: 10.1007/978-3-031-04083-2_4 (cit. on p. 9).

[35] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam. “Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models”. In: (2019). DOI: 10.48550/arXiv.1908.01224 (cit. on p. 6).

[36] M. Partio, L. Hieta, and A. Kokkonen. *CloudCast – Total Cloud Cover Nowcasting with Machine Learning*. 2025. DOI: 10.48550/arXiv.2410.21329 (cit. on p. 10).

[37] M. Quante. “Anmerkungen über die Eiswolken und ihre Bedeutung”. In: *Warnsignal Klima: Das Eis der Erde* (2015), pp. 99–103. DOI: 10.2312/warnsignal.klima.eis-der-erde.15 (cit. on p. 61).

[38] S. Rasp, M. S. Pritchard, and P. Gentine. “Deep learning to represent subgrid processes in climate models”. In: *Proceedings of the national academy of sciences* 115.39 (2018), pp. 9684–9689. DOI: 10.1073/pnas.1810286115 (cit. on p. 11).

[39] M. T. Ribeiro, S. Singh, and C. Guestrin. “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778 (cit. on pp. 8, 16, 18).

[40] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. 2015. DOI: 10.1007/978-3-319-24574-4_28 (cit. on p. 45).

[41] J. Rosemeier, M. Baumgartner, and P. Spichtinger. “Intercomparison of Warm-Rain Bulk Microphysics Schemes using Asymptotics”. In: *Mathematics of Climate and Weather Forecasting* 4.1 (2018), pp. 104–124. DOI: 10.1515/mcwf-2018-0005 (cit. on p. 12).

[42] C. Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x (cit. on p. 9).

[43] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. “DBSCAN revisited, revisited: why and how you should (still) use DBSCAN”. In: *ACM Transactions on Database Systems (TODS)* 42.3 (2017), pp. 1–21. DOI: 10.1145/3068335 (cit. on p. 59).

[44] S. Schulz, A.-C. Woerl, F. Jungmann, C. Glasner, P. Stenzel, S. Strobl, A. Fernandez, D.-C. Wagner, A. Haferkamp, P. Mildenberger, et al. “Multimodal deep learning for prognosis prediction in renal cancer”. In: *Frontiers in oncology* 11 (2021), p. 788740. DOI: 10.3389/fonc.2021.788740 (cit. on p. 4).

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74 (cit. on pp. 7, 16, 18).

[46] A. Shaban-Nejad, M. Michalowski, J. S. Brownstein, and D. L. Buckeridge. “Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare”. In: *IEEE Journal of Biomedical and Health Informatics* 25.7 (2021), pp. 2374–2375. DOI: 10.1109/JBHI.2021.3088832 (cit. on p. 16).

[47] R. Sharma, M. Gupta, and G. Kapoor. “Some better bounds on the variance with applications”. In: *Journal of Mathematical Inequalities* 3 (2010), pp. 355–363. DOI: 10.7153/jmi-04-32 (cit. on p. 21).

[48] A. Shrikumar, P. Greenside, and A. Kundaje. *Learning Important Features Through Propagating Activation Differences*. 2019. DOI: 10.48550/arXiv.1704.02685 (cit. on pp. 7, 8, 16, 18).

[49] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*. 2014. DOI: 10.48550/arXiv.1312.6034 (cit. on pp. 6, 16).

[50] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. DOI: 10.48550/arXiv.1409.1556 (cit. on p. 23).

[51] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. “SmoothGrad: removing noise by adding noise”. In: *CoRR* abs/1706.03825 (2017). DOI: 10.48550/arXiv.1706.03825 (cit. on pp. 6, 16).

[52] L. N. Smith and N. Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. 2017. DOI: 10.48550/arXiv.1708.07120 (cit. on pp. 25, 46).

[53] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. “Striving for Simplicity: The All Convolutional Net”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015. DOI: 10.48550/arXiv.1412.6806 (cit. on p. 6).

[54] B. Stevens, S. Bony, H. Brogniez, L. Hentgen, C. Hohenegger, C. Kiemle, T. S. L’Ecuyer, A. K. Naumann, et al. “Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds”. In: *Quarterly Journal of the Royal Meteorological Society* 146.726 (2020), pp. 141–152. DOI: 10.1002/qj.3662 (cit. on p. 12).

[55] X. Sun, X. Zhong, X. Xu, Y. Huang, H. Li, J. D. Neelin, D. Chen, J. Feng, W. Han, L. Wu, and Y. Qi. *FuXi Weather: A data-to-forecast machine learning system for global weather*. 2024. DOI: 10.48550/arXiv.2408.05472 (cit. on p. 11).

[56] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic Attribution for Deep Networks*. 2017. DOI: 10.48550/arXiv.1703.01365 (cit. on pp. 7, 16, 18, 26).

[57] W. S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (1952), pp. 401–419. DOI: 10.1007/BF02288916 (cit. on p. 47).

[58] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. *Robustness May Be at Odds with Accuracy*. 2019. DOI: 10.48550/arXiv.1805.12152 (cit. on pp. 10, 49).

[59] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861 (cit. on p. 47).

[60] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. 2003, pp. 1398–1402. DOI: 10.1109/ACSSC.2003.1292216 (cit. on p. 47).

[61] A. G. Wilson and P. Izmailov. "Bayesian Deep Learning and a Probabilistic Perspective of Generalization". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 4697–4708 (cit. on pp. 18, 21, 34).

[62] A.-C. Woerl, J. Disselhoff, and M. Wand. "Initialization Noise in Image Gradients and Saliency Maps". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 1766–1775 (cit. on pp. 3, 15, 49).

[63] A.-C. Woerl, M. Eckstein, J. Geiger, D.-C. Wagner, T. Daher, P. Stenzel, A. Fernandez, A. Hartmann, M. Wand, W. Roth, and S. Foersch. "Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides". In: *European Urology* 78 (Apr. 2020). DOI: 10.1016/j.eururo.2020.04.023 (cit. on pp. 4, 16).

[64] A.-C. Woerl, M. Wand, and P. Spichtinger. "Towards a Data-Driven Understanding Of Cloud Structure Formation". In: *Workshop on Tackling Climate Change with Machine Learning at ICLR 2024*. 2024 (cit. on p. 4).

[65] A.-C. Woerl, M. Wand, and P. Spichtinger. "What Atmospheric Data Knows About Clouds: An Information-Theoretic Approach". In preparation. 2026 (cit. on pp. 4, 37).

[66] H. Xiao, K. Rasul, and R. Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. DOI: 10.48550/arXiv.1708.07747 (cit. on p. 23).

[67] V. Zantedeschi, F. Falasca, A. Douglas, R. Strange, M. J. Kusner, and D. Watson-Parris. *Cumulo: A Dataset for Learning Cloud Classes*. 2022. DOI: 10.48550/arXiv.1911.04227 (cit. on p. 10).

[68] M. D. Zeiler and R. Fergus. "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, 2014, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53 (cit. on pp. 6, 8).

[69] G. Zhai and X. Min. "Perceptual image quality assessment: a survey". In: *Science China Information Sciences* 63.11 (2020), pp. 1–52. DOI: 10.1007/s11432-019-2757-1 (cit. on p. 22).

[70] Y. Zhang, P. Tino, A. Leonardis, and K. Tang. "A Survey on Neural Network Interpretability". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (Oct. 2021), pp. 726–742. DOI: 10.1109/tetci.2021.3100641 (cit. on pp. 5–7).

Webpages

[@1] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, et al. *ERA5 hourly data on single levels from 1940 to present*. 2023. DOI: 10.24381/cds.adbb2d47. (Visited on June 12, 2024) (cit. on p. 43).

[@2] J. Howard. *Imagenette*, https://github.com/fastai/imagenette/. https://github.com/fastai/imagenette/ (visited on June 7, 2021) (cit. on pp. 17, 23).

[@3] NASA. *True Color - Corrected Reflectance (MODIS / Aqua)*. 2023. https://gis.earthdata.nasa.gov/portal/home/item.html?id=7e2b03d9331b4df785f7c64959df8fd8 (visited on Aug. 7, 2025) (cit. on p. 44).

List of Figures

List of Tables

List of Abbreviations

IFS	Integrated Forecasting System
MDS	multidimensional scaling
ML	machine learning
MS-SSIM	multi-scale structural similarity index measure
NCC	normalized cross-correlation
NWP	Numerical Weather Prediction
PCA	Principal Component Analysis
SNR	signal to noise ratio
SSIM	structural similarity index measure

Appendix

A Optical Thickness Calculation

In this work, the optical thickness τ is computed from liquid and ice water content obtained from ERA5 following a parameterized radiative approximation. The formulation is based on effective particle sizes and vertical integration of extinction coefficients, consisting of liquid and ice water optical thickness. The total optical thickness is defined as the (discrete) vertical integral over liquid and ice contributions:

$$\tau = \sum_{z} \left(\tau_l(z) + \tau_i(z)\right), \tag{A.1}$$

with τ_l and τ_i being dependent of the model levels z. All quantities are expressed in SI units unless stated otherwise.

A.1 Optical Thickness of Liquid Water Phase Clouds

The liquid water content (LWC) is given by

$$\mathrm{LWC} = \max(q_l \cdot \rho, 0), \tag{A.2}$$

where q_l denotes the specific cloud liquid water content $clwc$ extracted from ERA5 and ρ the air density. The air density is computed using the ideal gas law:

$$\rho = \frac{p}{R \cdot T}, \tag{A.3}$$

where p denotes pressure, T temperature, and R the specific gas constant for dry air.

The optical thickness of liquid water clouds is determined by the effective droplet radius, which is computed as:

$$r_{\mathrm{eff}} = \left(\frac{3\,\mathrm{LWC}}{4\pi\rho_l k N_d}\right)^{1/3}, \tag{A.4}$$

with droplet number concentration

$$N_d = (-1.15 \cdot 10^{-3} \cdot N_a^2 + 0.963 \cdot N_a + 5.3) \cdot 10^6, \tag{A.5}$$

and the skew factor of the droplet distribution

$$k = \frac{(1+D^2)^3}{(1+3D^2)^2}. \tag{A.6}$$

We restrict the effective droplet radius to be between $r_{\min}$ and $r_{\max}$.

The liquid optical thickness is then given by

$$\tau_l = \mathrm{LWC} \cdot 10^3 \cdot \Delta z \cdot \left(a_l + \frac{b_l}{r_{\mathrm{eff}} \cdot 10^6}\right), \tag{A.7}$$

where Δz denotes the vertical thickness of each atmospheric model layer.

A.2 Optical Thickness of Ice Water Phase Clouds

The ice water content (IWC) is computed as

$$\mathrm{IWC} = \max(q_i \cdot \rho \cdot 10^{-3}, 0), \tag{A.8}$$

where q_i is the specific cloud ice water content $ciwc$ extracted from ERA5.

The effective ice particle diameter is defined as

$$D_{\mathrm{eff}} = (c_0 + c_1(T - T_0))\,(a(\mathrm{IWC}) + b(\mathrm{IWC})(T - T_1))\,, \tag{A.9}$$

with

$$a(x) = a_0 x^{a_1}, \quad b(x) = b_0 x^{b_1}. \tag{A.10}$$

We restrict the effective ice particle diameter to be between $D_{\min}$ and $D_{\max}$.

The ice optical thickness is then computed as

$$\tau_i = \mathrm{IWC} \cdot \Delta z \cdot \left(a_i + \frac{b_i}{D_{\mathrm{eff}}}\right). \tag{A.11}$$

A.3 Constants

In our calculations, we use the following constants:

$$
\begin{aligned}
&\textbf{General:} && R = 287.05 \\
&\textbf{Liquid phase:} && a_l = 2.817 \cdot 10^{-2},\ b_l = 1.305,\ \rho_l = 1000,\ N_a = 50, \\
& && D = 0.33,\ r_{\min} = 4 \cdot 10^{-6},\ r_{\max} = 30 \cdot 10^{-6}, \\
&\textbf{Ice phase:} && a_i = -1.30817 \cdot 10^{-4},\ b_i = 2.52883, \\
& && a_0 = 45.8966,\ a_1 = 0.2214,\ b_0 = 0.7957,\ b_1 = 0.2535, \\
& && c_0 = 1.2351,\ c_1 = 0.0105,\ T_0 = 273.15,\ T_1 = 83.15, \\
& && D_{\min} = 20 + 40\cos(\phi),\ D_{\max} = 155
\end{aligned}
\tag{A.12}
$$

B Attribution-Based Predictive Input Identification

B.1 Additional Attribution Samples

The following section presents additional examples of clustering results for randomly selected samples from different regions (see fig. B.1). Although the cluster shapes vary across samples, a clear separation into two distinct regions can be observed throughout. These examples are intended to illustrate that the overall clustering behavior remains qualitatively consistent despite regional and sample-specific differences.

Furthermore, screenshots of the implemented visualization tool are provided (see fig. B.2 to B.6). Each figure shows 2D and 3D attribution maps for randomly selected pixels. In order to illustrate the variability of the results, examples from five different regions are presented. The visualizations are meant to provide a qualitative impression of the spatial attribution patterns and of how these patterns can differ across regions and target pixels. The dominance of mid-level humidity and near-surface temperature can be seen across many samples.

B.2 Additional SNR Samples

The following section provides additional examples of attribution-wise signal to noise ratio (SNR) for randomly selected samples from different regions (see fig. B.7 to B.12). These examples are intended to complement the aggregated results shown in the main text and to illustrate the extent of sample-to-sample variability. Although minor local differences are present, the overall attribution patterns remain largely consistent across samples and regions. This is consistent with the main finding that the attribution-wise SNR does not vary strongly between individual samples.

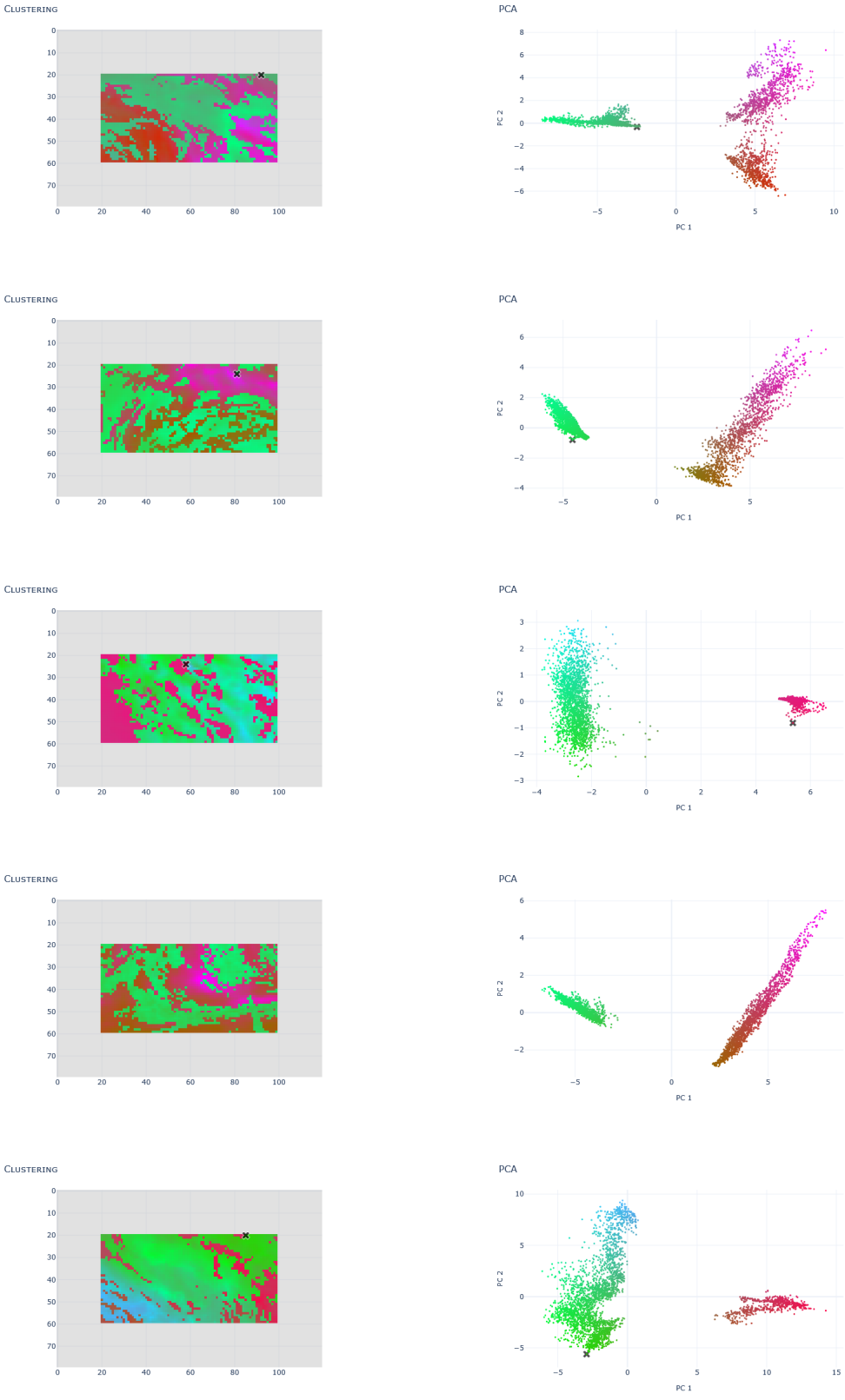


Fig. B.1.: PCA clustering of attribution maps from five different samples around the globe and their corresponding cloud image. Every sample is showing a clear separation.

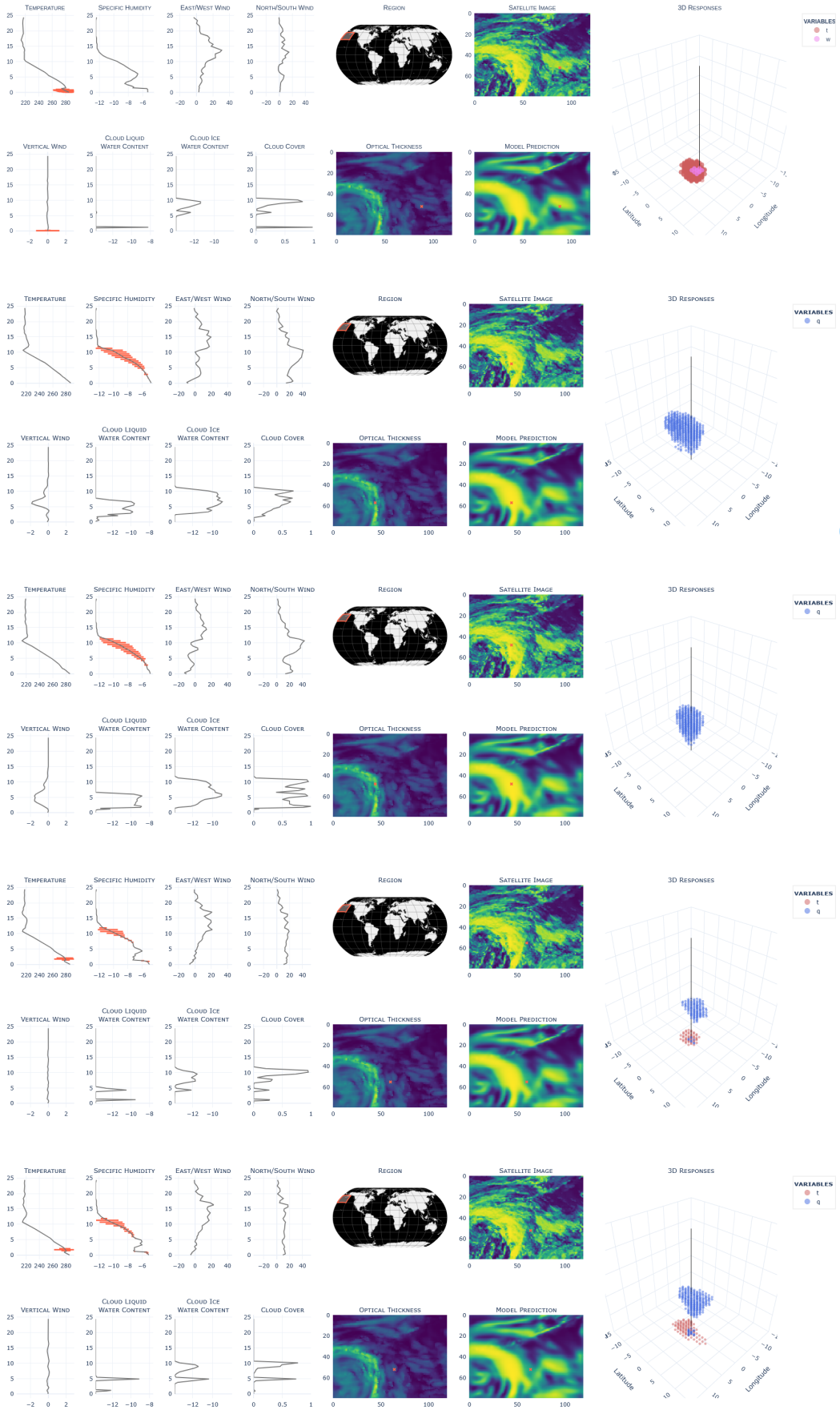


Fig. B.2.: 5 different attribution maps taken from a sample in the North Pacific. Each map represents one specific location on the same image.

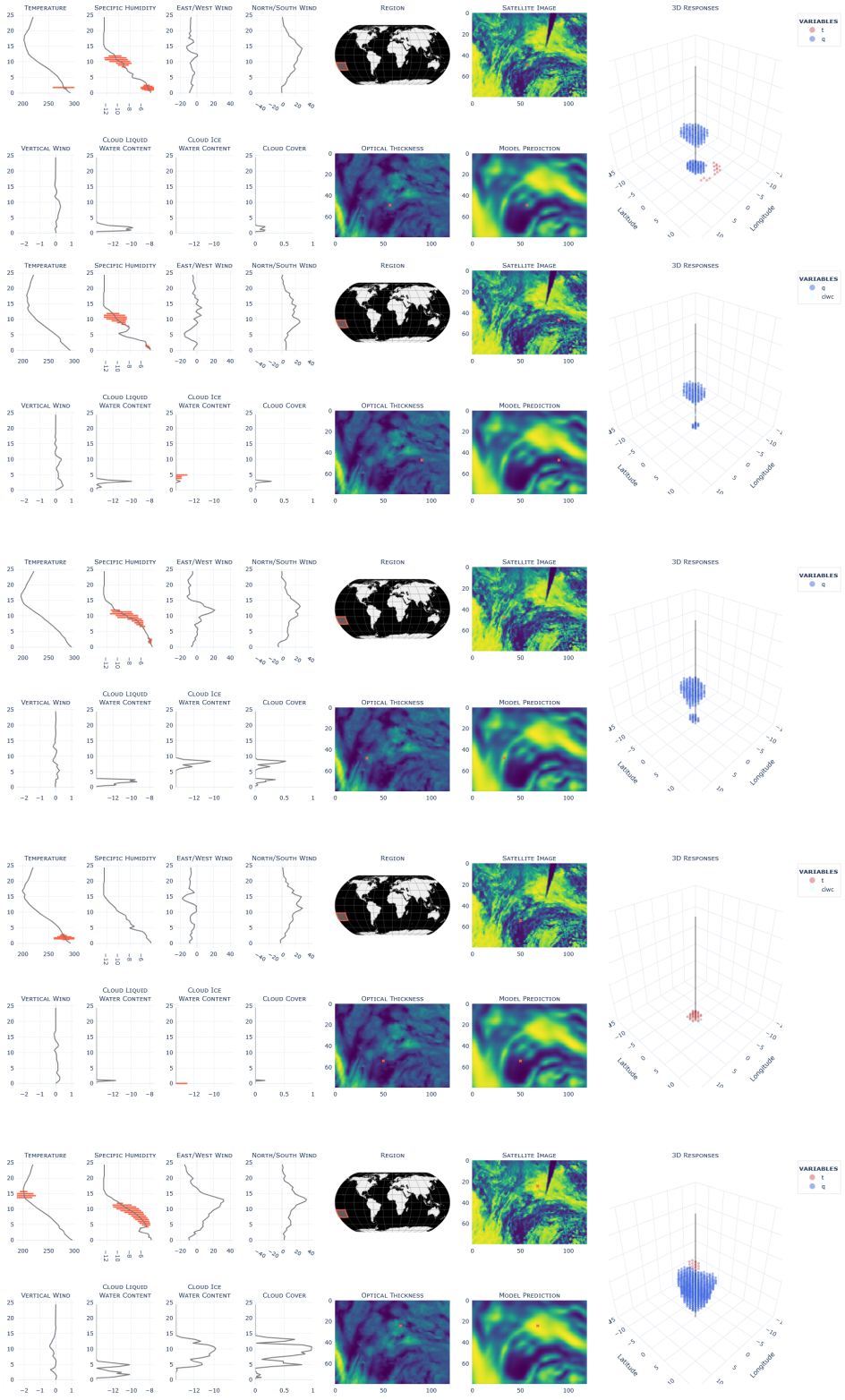


Fig. B.3.: 5 different attribution maps taken from a sample in the South Pacific. Each map represents one specific location on the same image.

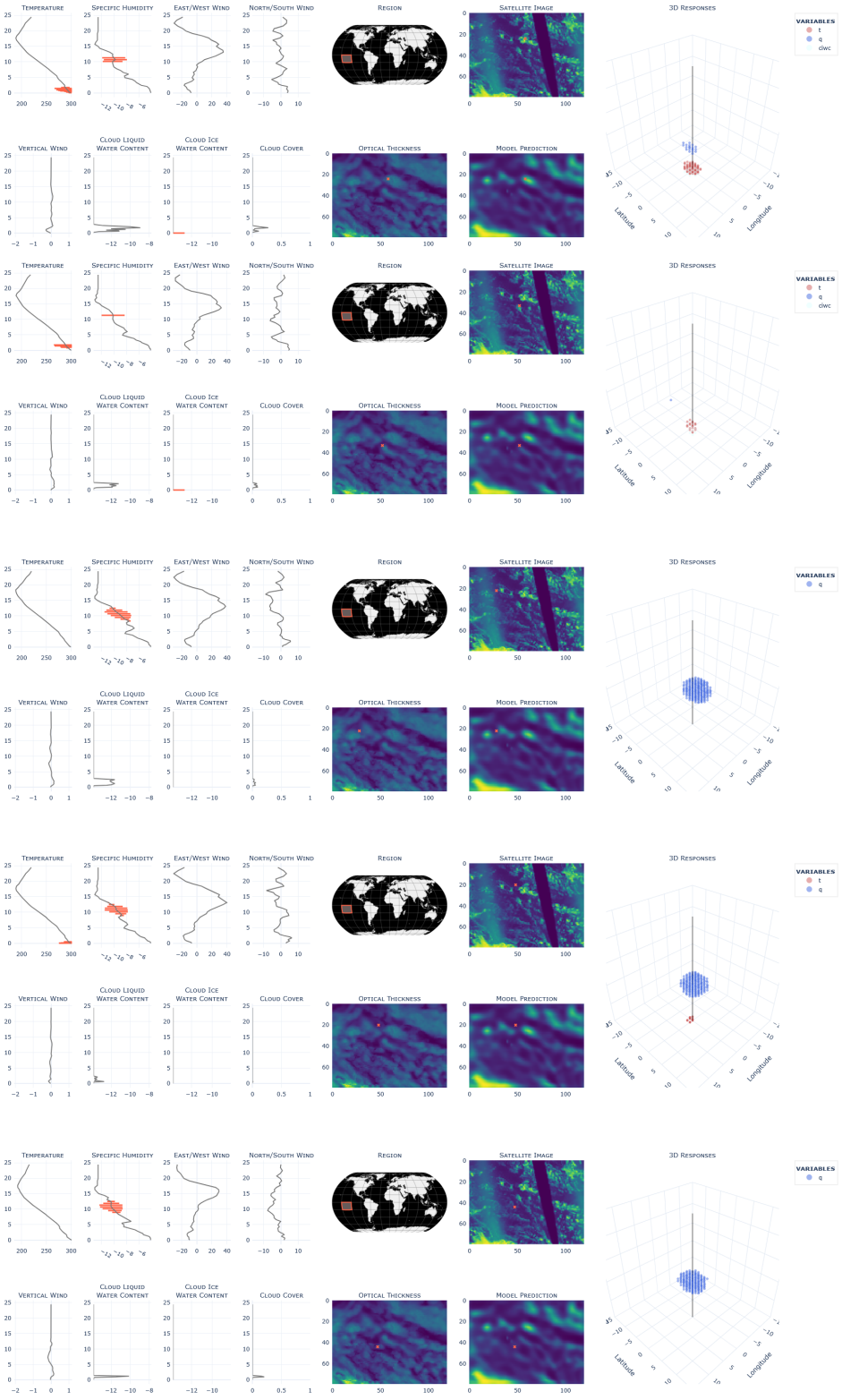


Fig. B.4.: 5 different attribution maps taken from a sample at the Equator. Each map represents one specific location on the same image.

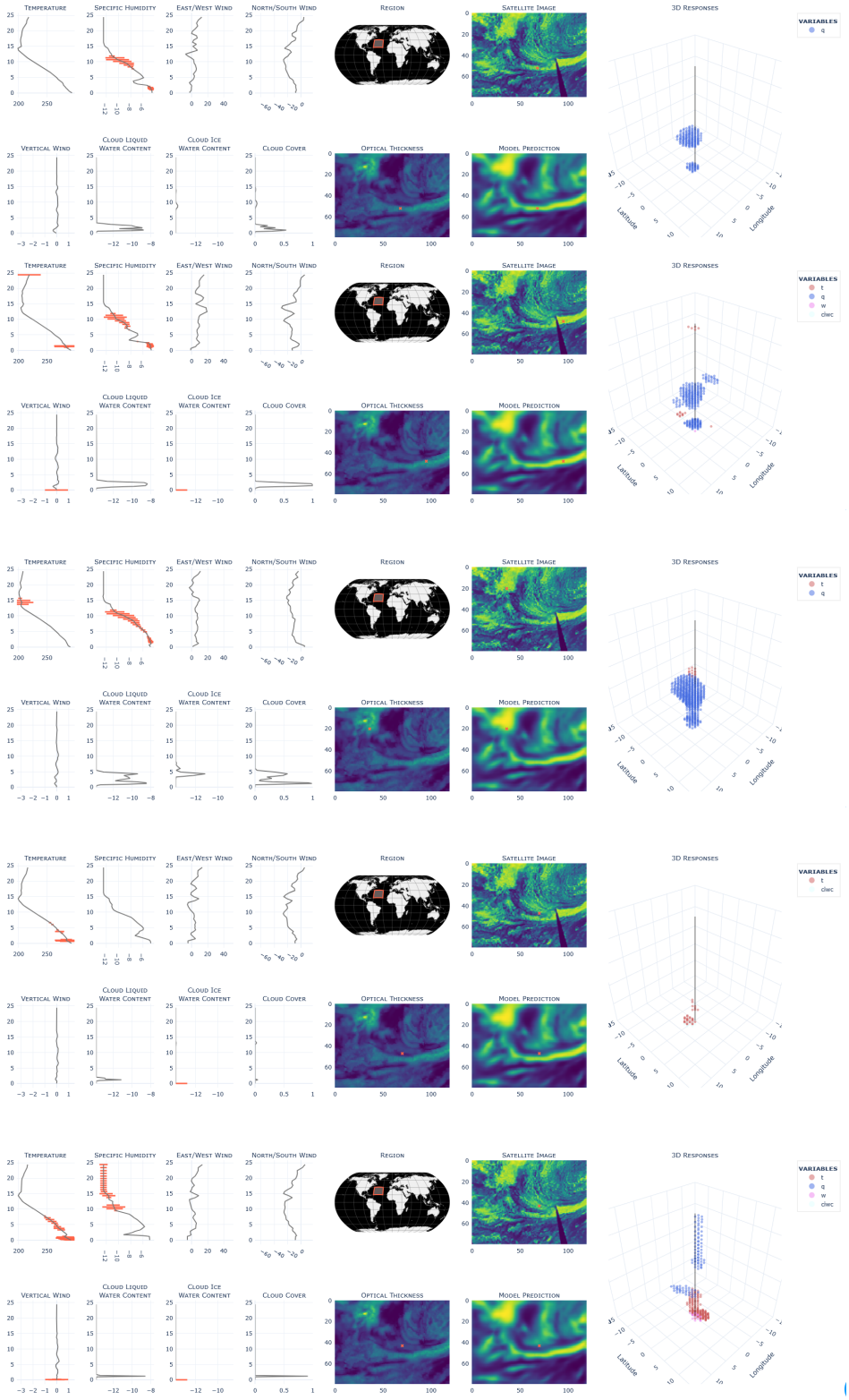


Fig. B.5.: 5 different attribution maps taken from a sample in the North Atlantic. Each map represents one specific location on the same image.

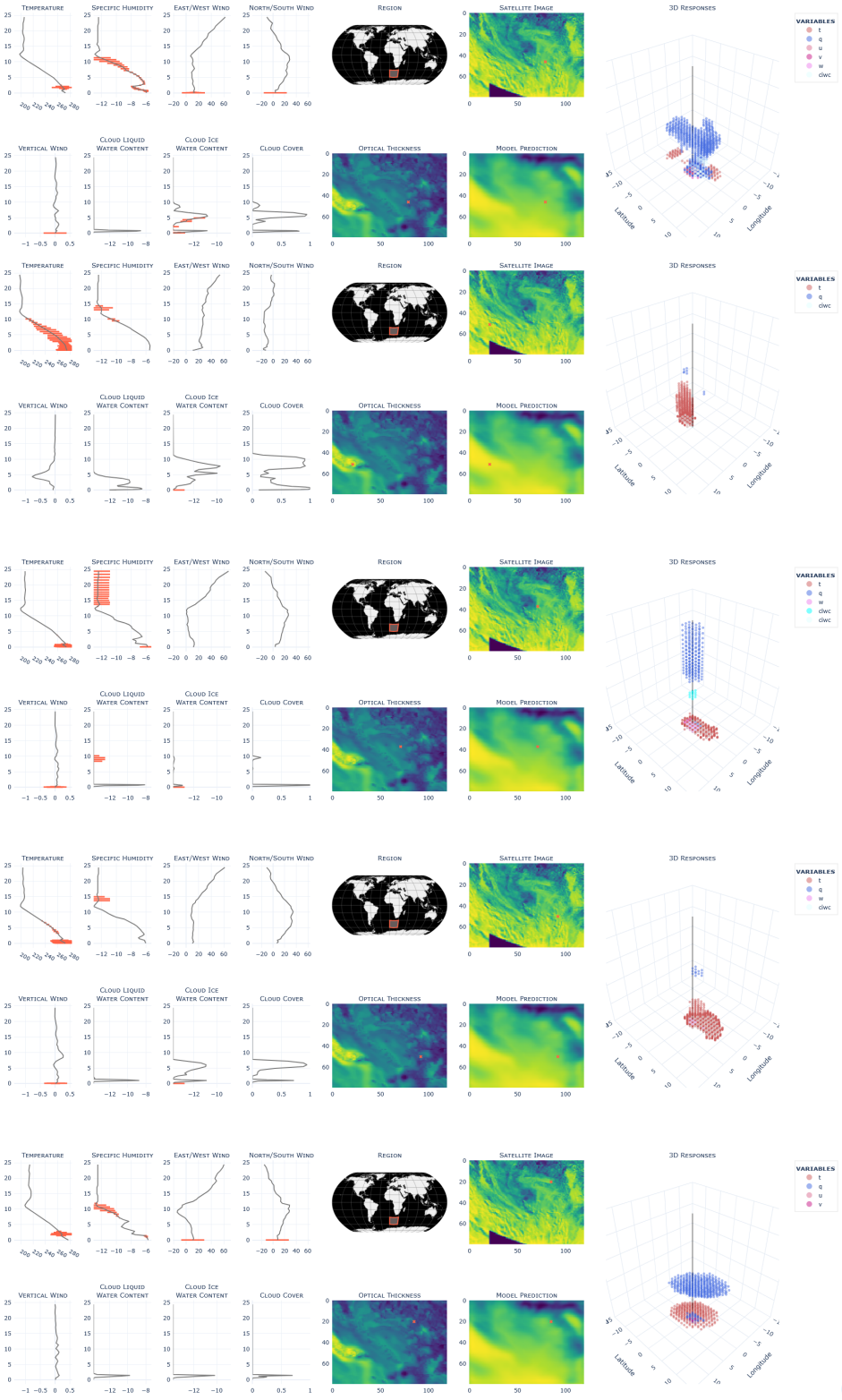


Fig. B.6.: 5 different attribution maps taken from a sample in the South Atlantic. Each map represents one specific location on the same image.

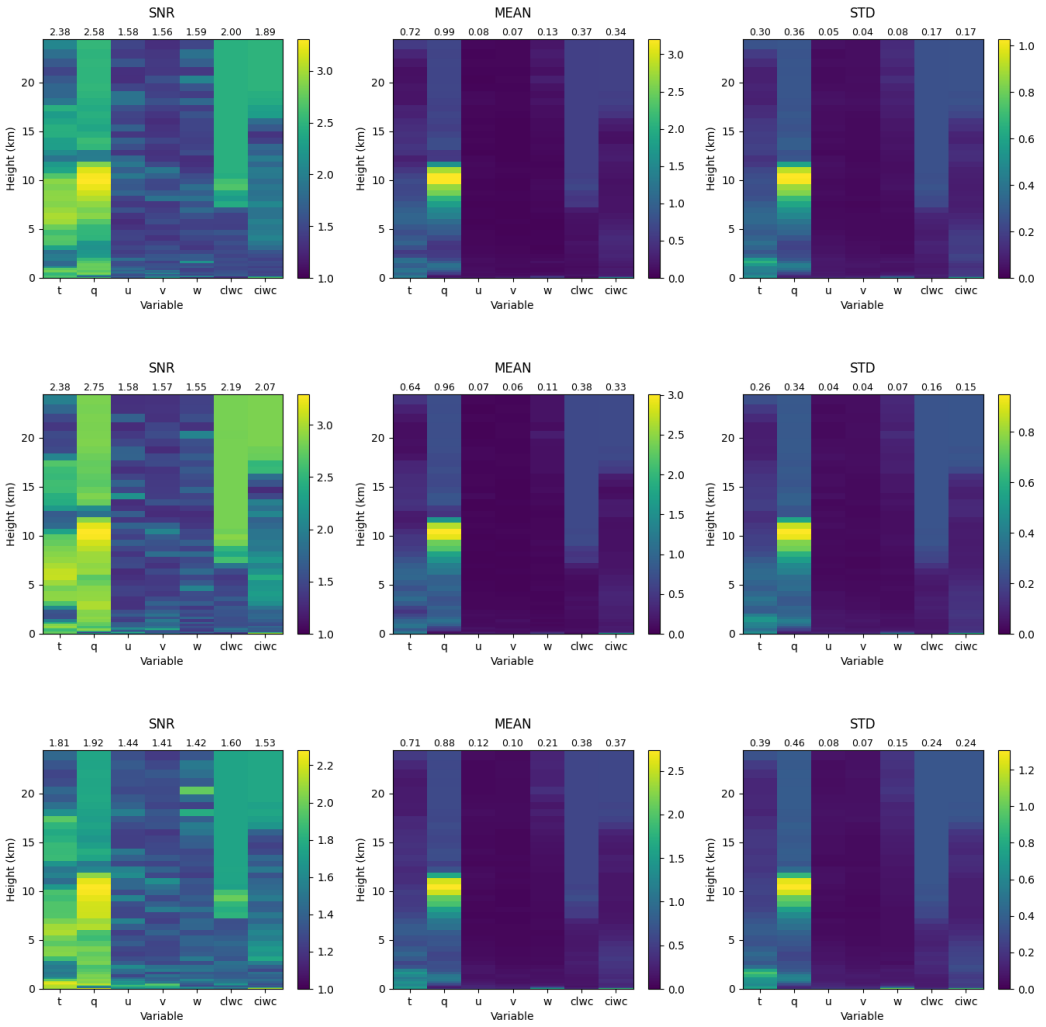


Fig. B.7.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $40°$ and $60°$ North.

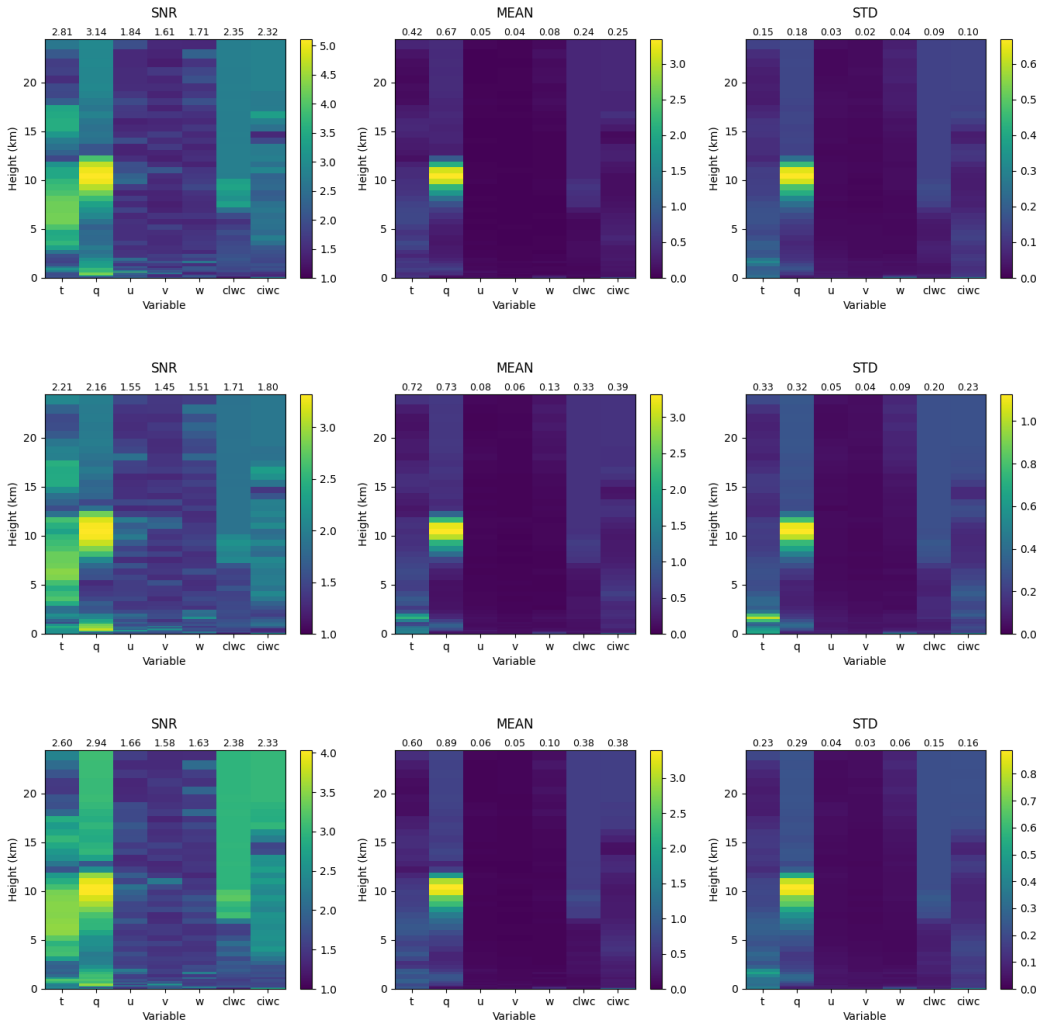


Fig. B.8.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $20°$ and $40°$ North.

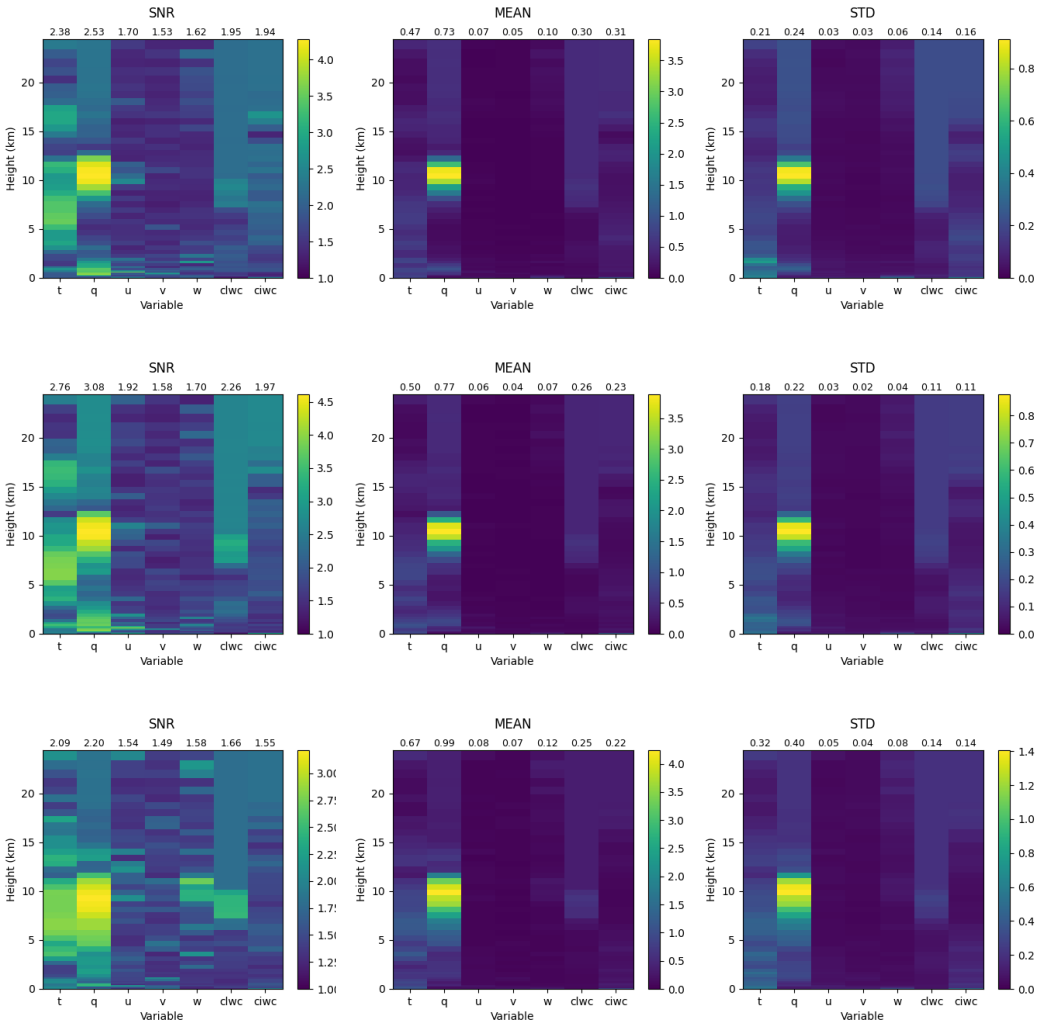


Fig. B.9.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $0°$ and $20°$ North.

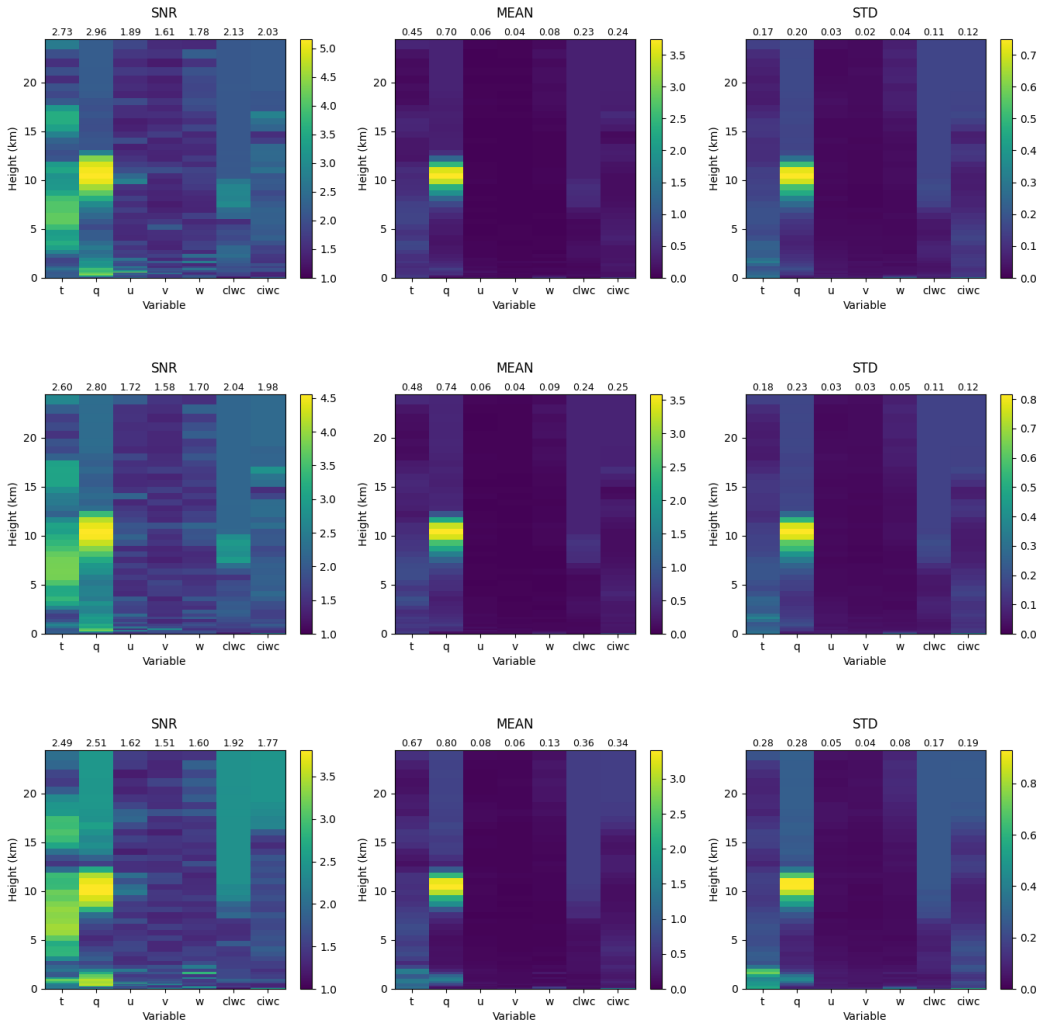


Fig. B.10.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $0°$ and $20°$ South.

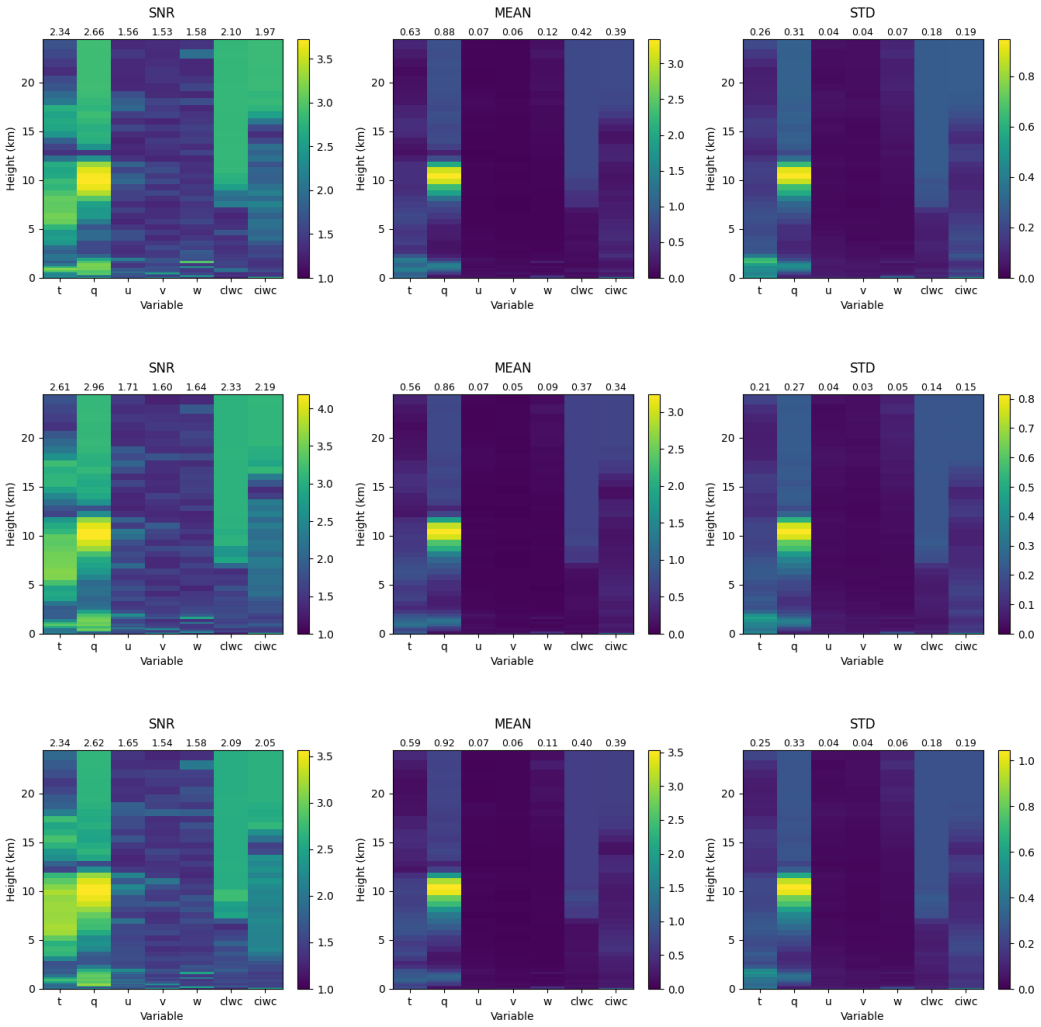


Fig. B.11.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $20°$ and $40°$ South.

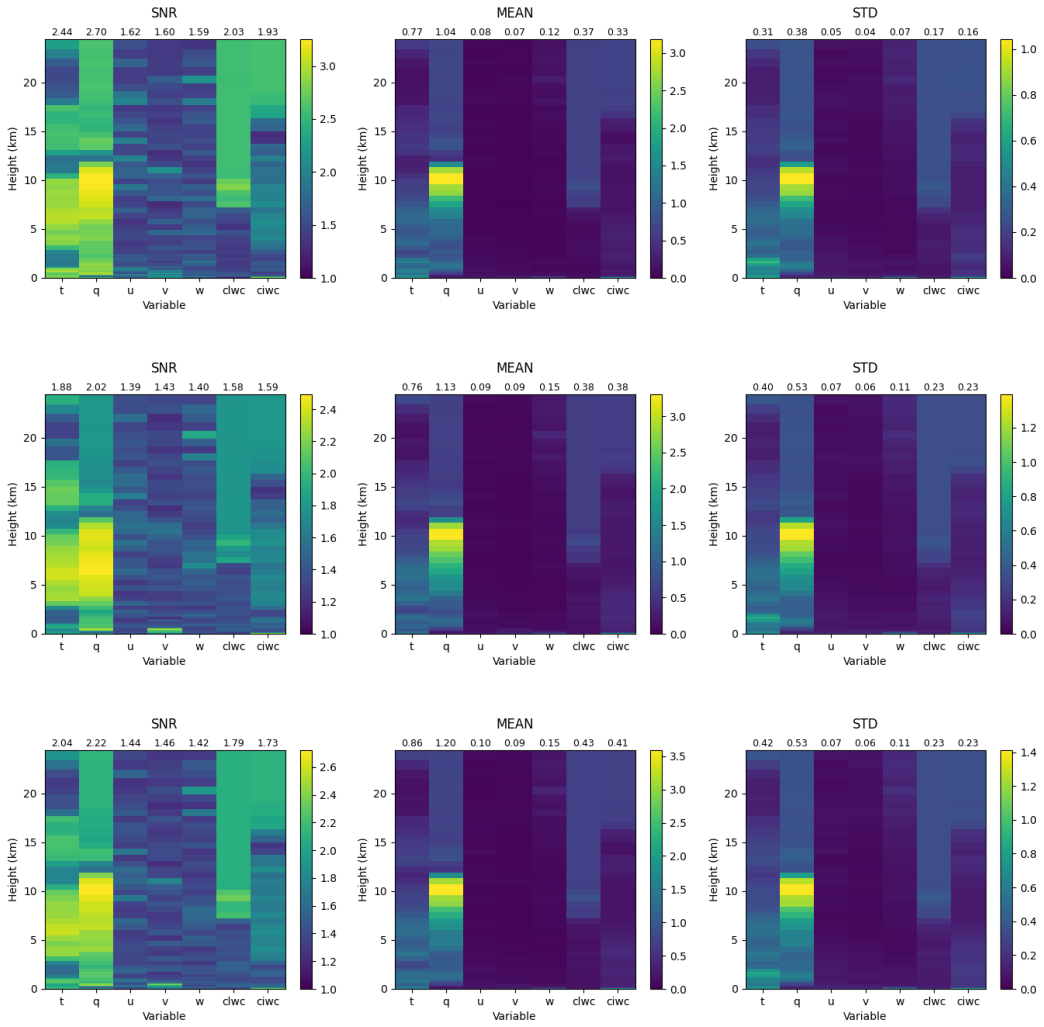


Fig. B.12.: SNR (left), mean (middle) and standard deviation (right) of attribution values across variables and vertical levels for one sample per row. Regions are based between $40°$ and $60°$ South.

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Adaptations to the style of the Institute of Computer Science can be found at `https://gitlab.rlp.net/institut-fur-informatik/cleanthesis-jgu`.

Declaration of Originality

I hereby declare

that I have written the submitted thesis independently using only the resources and sources listed (this includes AI-based applications or tools[1]). All direct or indirect quotations and paraphrased texts have been marked and cited appropriately. I certify that I did not use any tools whose use was explicitly prohibited by the examiner.

I documented the AI tools I used in the appendix "'Use of AI Tools.'"

By submitting this work, I take responsibility for the entirety of its content. I thereby also take responsibility for any and all AI-generated content included in my work. I have checked the veracity of the included (AI-generated) content to the best of my knowledge.

I have not submitted this work in an identical or similar form to earn credit in any other context.

I am aware that a violation of the above specifications has consequences in terms of examination law and in particular can lead to the coursework or examination being marked as "failed." Enrollment can be revoked for up to two years if a student has cheated in coursework or examinations two or more times (section 69 subsections 4 and 5 University Act, HochSchG).

Mainz, 28.04.2026

Ann-Christin Wörl

[1] Further information on AI-based applications or tools is available here (in German): https://digitale-lehre.uni-mainz.de/lehren-pruefen/ki-in-der-hochschulbildung/

Use of AI tools.

AI Tool	Used for	Reason	When
DeepL Write	reformulation of my drafts	to improve readability and correct grammar	throughout the entire work
ChatGPT	reformulation of my drafts	to improve readability and correct grammar	throughout the entire work
GitHub CoPilot	code generation	to generate basic code for visualization purposes	analysis of experiments