SnowProfGPT: A Transfer Learning Approach to Reconstruct Ground-Cluttered Near-Surface Snowfall Profiles Globally from Spaceborne Radar

by

Ding Li

A research paper
presented to the University of Waterloo
in fulfillment of the requirements for the degree of
Master of Mathematics in Computational Mathematics

Waterloo, Ontario, Canada
2025

Author's Declaration

I hereby declare that I am the sole author of this research paper. This is a true copy of the paper, including any required final revisions, as accepted by my supervisor and reader.

I understand that my research paper may be made electronically available to the public on the Computational Mathematics Past Research Projects webpage.

Statement of Contributions

This research paper was completed under the supervision of Professor Chris Fletcher (Faculty of Environment) and Professor Marek Stastna (Faculty of Mathematics). All core ideas, data processing, model development, analysis, writing, and visualization were conducted by the author. Professor Chris Fletcher provided conceptual guidance, technical feedback, and oversight throughout the project.

Abstract

Accurately reconstructing near-surface snowfall profiles from spaceborne radar remains a longstanding challenge due to ground clutter contamination in the lowest ∼1 km of the radar column. This "clutter zone" (CZ) is systematically excluded from current satellite retrievals, introducing critical biases in snowfall estimation and downstream hydrological modeling. We propose **SnowProfGPT**, a self-attention–based autoregressive model trained under a generative pretraining and transfer learning framework to globally reconstruct snowfall-rate profiles within the CZ using CloudSat observations.

The model is pretrained on over **50 million** clutter-free CloudSat scenes and fine-tuned on a synthetic ground-based (GB) subset designed to simulate real-world ground-truth CZ data availability. This training strategy mitigates domain shift along both vertical and horizontal dimensions by leveraging high-quality observations from the well-observed upper profile. To evaluate performance, we introduce a novel *Synthetic Clutter Zone* (SCZ) as a proxy for the true CZ and benchmark accuracy across geolocations, surface types, and meteorological conditions. SnowProfGPT improves global R^2 in the lowest SCZ bin from **2% to 28%**, capturing **26 percentage points** more snowfall-rate variability than the official replication baseline. It also reduces near-surface mean absolute error by **∼0.015 mm/hr**, with the largest gains observed over open ocean, where R^2 improves from $< 0\%$ to over **20%**. A model-aware generalization analysis using SHAP reveals that SnowProfGPT is highly stable and generalizable in predicting the lowest bin in the SCZ across most footprint attributes with modest overestimation. We exceptionally find negative and positive dependency only in regions of extreme latitude and surface elevation respectively. These findings demonstrate that autoregressive modeling combined with large-scale pretraining enables improved and acceptably generalizable snowfall reconstruction compared with the baseline replication method, even in domains with limited ground truth.

This work establishes the first global-scale CZ reconstruction from satellite radar and introduces a transferable training paradigm for geophysical inpainting under extreme data sparsity. SnowProfGPT sets a new benchmark for near-surface snowfall estimation and offers a foundation for future retrieval models across cluttered or partially observed atmospheric domains.

Acknowledgements

I would like to express my sincere gratitude to my research supervisor, Professor Chris Fletcher, for his invaluable guidance and support throughout this project and for helping shape my research skills and mindset. His interdisciplinary insight, thoughtful mentorship, and consistent feedback were vital to this work. He has been a role model for me in learning how to conduct research effectively and manage my work with purpose and efficiency. I am also deeply grateful for his financial support and for crafting a research project that was exceptionally well-suited to my background and interests, both of which were essential to the successful completion of this degree and my potential career in academia. His influence on me has been deeply positive and will remain lifelong.

I would also like to thank Professor Marek Stastna, who served as my co-supervisor during the program, for his availability and support throughout my general studies.

I am thankful to Dr. Chuyin Tian and other members of Professor Fletcher's research group for their valuable discussions and constructive feedback, which helped me improve both my research and presentation skills.

Finally, I would like to thank my family and close friends for their consistent support, emotionally, financially, and in life decisions, throughout my academic journey. In particular, I wish to acknowledge my parents, grandparents, cousin, and Xinyu Cai, as well as my close friends Feiyang Chen, Yuxingchen Pei, and Zixu Fan, for their unwavering encouragement. I would also like to thank my high school mathematics teacher for inspiring my early interest and confidence in the subject. In addition, I am grateful to the Wu family for their warm, family-like support during my time in Canada. I would also like to thank the University of Waterloo for providing a life-changing undergraduate and graduate education.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

List of Figures ix

List of Tables xii

1 Introduction 1

1.1 Operational Treatment of Clutter Zones in Satellite Snowfall Retrievals . . 3

1.2 Domain Shift: A Key Limitation in Generalizing Profile Models 4

1.3 Limitations of Prior Methods Under Domain Shift 6

1.3.1 Unsupervised self-contained (Baseline) 7

1.3.2 Assimilation models 8

1.3.3 Self-contained models 9

1.3.4 Summary 10

1.4 Research Gaps and Objectives 11

2 Data **15**

2.1 Snowfall Profiles and Auxiliary Data from CloudSat 16

2.2 Auxiliary Meteorological Profiles from ECMWF 17

3 Methodology **20**

3.1 Inductive Bias: An Identified Underlying Relationship 20

3.2 SnowProfGPT: A Self-Attention-Based Autoregressive Model for Snowfall-Rate Inpainting 25

3.2.1 Forward model definition: Inputs and outputs 25

3.2.2 Tokenization and Embedding: Unifying heterogeneous meteorological inputs into a shared contextual feature space 28

3.2.3 Self-Attention Encoder: Contextual Representation of the Meteorological Scene 29

3.2.4 Self- and Cross-Attention Decoder: Autoregressive Meteorological Scene Inpainter 31

3.2.5 Overview 33

3.3 Generative Pre-training: A Transfer Learning Approach 35

3.3.1 Self-Supervised Pretraining on Unlabeled CloudSat Observations . . 36

3.3.2 Multi-Stage Pretraining and Generalization Regularizations 38

3.3.3 Transfer Learning: From CS to GB 40

3.3.4 Summary 41

3.4 Synthetic Clutter Zone: A Proxy for Model Assessment 42

3.4.1 Generalizability Evaluation: Through Residual Modeling 43

4 Results **49**

4.1 Model global performance 49

4.2 Generalizability analysis 50

4.2.1 Spatial heterogeneity of SnowProfGPT predictions 50

4.2.2 SnowProfGPT Performance Across Surface Types 52

4.2.3 Interpretable Generalization Insights via SHAP and Residual Modeling 52

5 Conclusions and Discussion **59**

5.1 Key Contributions and Findings 60

5.2 Limitations and Future Directions 60

5.3 Closing Remarks 61

References **63**

List of Figures

1.1 Illustration of the CloudSat mission. The onboard Cloud Profiling Radar (CPR) provides vertically resolved profiles of cloud and precipitation structure with a native vertical resolution of 240 m per range bin. Source: NASA/JPL-Caltech/CloudSat Data Processing Center [58]. 2

1.2 Surface responses to CloudSat's radar observations (top) and the oversampling process near the surface (bottom). Source: Adapted from Marchand et al. (2008) [45]. 13

1.3 Conceptual illustration of CloudSat observational domains and domain shift. **Top left:** Horizontal view showing the global extent of CloudSat coverage, with three representative ground-based snowfall observation sites (e.g., ARM KaZR) marked as dots (GB). **Bottom left:** Vertical cross-sectional view along latitude and height, representing the CloudSat profile geometry. The lowest pink layer indicates the clutter zone (CZ); the yellow region above represents the well-observed CloudSat domain (CS). **Right:** Combined 3D view integrating both horizontal and vertical perspectives. The blue surface layer represents land and ocean boundaries. 14

2.1 Spatio-temporal overview of the sampled test dataset. Local times are converted from the original TAI/UTC timestamps using the time zones inferred from geolocation. 19

3.1 Empirical distributions sampled from a random granule of CloudSat snowfall profiles (snowfall events only). **Left:** Residual distribution of snowfall-rate differences between adjacent vertical bins, $S(R+dR)-S(R)$, approximating the first-order local gradient. **Right:** Marginal distribution of snowfall rates $S(R)$ across all valid bins. The left panel supports a Markov assumption by showing that local vertical changes in snowfall rate are approximately zero-mean and narrowly distributed. 46

3.2 Overview of the SnowProfGPT model architecture. The diagram shows the core design components. For each scene, a 5-by-5 window of adjacent footprints is sampled from a CloudSat granule. The scene is tokenized using a 3D tokenizer along the footprint (width), variable (channel), and vertical bin dimensions. Positional encodings for width, height, and channel are added in the same order. Each token is linearly embedded into a vector space and transformed into Query, Key, and Value representations through multi-head projection. The encoder applies multi-head self-attention to compute contextualized token embeddings. The decoder uses masked self-attention to process previously predicted snowfall tokens and applies cross-attention to incorporate encoder outputs. Tokens are inpainted autoregressively from top to bottom and left to right. Each predicted token is projected back into snowfall-rate space and placed into its original location via the inverse tokenizer. The spatial window is fixed at 5-by-5. Encoder memory and decoder inputs are dynamically updated throughout decoding. 47

3.3 Overview of the SnowProfGPT training paradigm. Yellow regions represent well-observed vertical bins in the CloudSat-observed domain (CS), located immediately above the clutter zone (CZ). Red regions denote masked snowfall-rate bins used as autoregressive training targets. Pink regions represent the clutter zone, which is the focus of reconstruction. The top panel shows the horizontal footprint sampling scheme, where adjacent CloudSat footprints are grouped to form each scene. The middle panel shows the vertical bin sampling strategy used to define the scene depth and mask placement. The bottom panel illustrates how the scene is reassembled in three dimensions. The blue pipeline indicates the sequential pretraining and fine-tuning process used in the transfer learning scheme. 48

4.1 Global performance comparison between the baseline (replication) model and the proposed inpainting model (SnowProfGPT), evaluated over snowfall footprints only. **Left:** Global coefficient of determination (R^2) for each height bin in the inpainted clutter zone (SCZ). **Middle:** Global mean absolute error (MAE) in snowfall-rate prediction. **Right:** Global bias (mean signed error) relative to reference snowfall rates. "Bin 1" through "Bin 4" correspond to the four lowest bins in the SCZ, listed in descending order from the surface. "Combined" denotes aggregate metrics across all SCZ bins. 51

4.2 Spatial distribution of mean absolute error (MAE) for the inpainting model (SnowProfGPT) and the baseline (replication) model, evaluated over snowfall footprints in the lowest bin of the SCZ. Each point reflects the MAE within a 2-km spatial window aggregated across time. Circled anomalies ("spark patches") indicate regions where performance degrades significantly. Results shown for polar regions only. 55

4.3 Spatial distribution of Q-sigma for the inpainting and baseline models in the lowest SCZ bin. Q-sigma normalizes residuals by local snowfall variability, highlighting error stability across geolocations. Outliers below the 1st percentile and above the 99th percentile are excluded for clarity. 56

4.4 Comparison of model performance across surface types. Bars indicate R^2 for the inpainting and baseline models in the lowest SCZ bin, stratified by surface type: open ocean, ice, and land. Surface-type effects reflect both model sensitivity and underlying retrieval challenges. 57

4.5 Residual sensitivity analysis using the Residual Regression Model (RRM). **Left:** SHAP-based feature importance ranking for footprint-level conditions impacting model performance, based on a residual model trained to predict inpainting errors with 59% R^2 on a held-out test set. **Right:** Adjusted SHAP violin plot showing how each attribute influences residual directionality. SHAP values are re-centered such that zero corresponds to no error, with positive (negative) values indicating a tendency toward overestimation (underestimation). 58

List of Tables

1.1 Summary of representative snowfall reconstruction methods under a statistical learning framework. Here, CS, GB, and CZ denote the CloudSat-observed domain, the ground-based observational domain, and the clutter zone to be reconstructed, respectively. 7

3.1 Core architectural components and hyperparameter configurations of the SnowProfGPT model. Embeddings are applied separately for each channel, footprint, and auxiliary metadata feature. The model uses a multi-head transformer encoder–decoder architecture, with autoregressive decoding over four vertical bins. All embeddings operate in a shared latent space of 512 dimensions. 26

3.2 Training strategy and hyperparameter configurations used for pretraining and fine-tuning SnowProfGPT. The model is pretrained on CloudSat observations above the SCZ using a self-supervised autoregressive task, then fine-tuned on a small synthetic ground-based subset. Pretraining is performed in three stages with gradually increasing task complexity. Fine-tuning disables all regularization and uses fixed weights and learning rate. 38

Chapter 1

Introduction

Atmospheric hydrometeors play a vital role in Earth's water cycle and energy budget. In particular, hydrometeors near the surface interact strongly with orographic components of the hydrosphere, such as sea ice and surface temperature, and thus influence key boundary-layer climatic processes. Inadequate observation of these near-surface hydrometeors limits our ability to understand and project hydroclimatological dynamics, which depend critically on accurate observations for both validation and model constraints. However, ground-based radar networks provide only sparse coverage, particularly in remote Arctic regions. To overcome these spatial limitations, spaceborne cloud profiling radars (CPRs) offer global to quasi-global observational capabilities [47, 32, 17].

Among spaceborne missions, NASA's *CloudSat* carries a Cloud Profiling Radar (CPR) with sufficient sensitivity to detect both clouds and snowfall using a unique 94 GHz radar system. This active remote sensing instrument operates in the W-band microwave spectrum, allowing it to backscatter signals from light hydrometeors without being blocked as visible wavelengths. As a result, it provides vertically resolved profiles of cloud and precipitation structure with a native vertical resolution of 240 m per range bin [56] (as illustrated in Fig. 1.1). These characteristics make CloudSat particularly well suited for studying the vertical development of aloft snowfall and its relationship with the precipitating cloud. Over the past decade, a suite of snowfall retrieval algorithms and data products has been developed based on CloudSat's CPR, often in collaboration with other instruments in NASA's A-Train and C-Train satellite constellations [15, 14, 13, 23, 46, 56, 65, 66]. These algorithms integrate CPR radar reflectivity profiles with auxiliary observations to detect precipitation occurrence, classify phase (e.g., rain or snow), and estimate quantitative precipitation rates, such as **snowfall intensity**, the primary focus of this study.

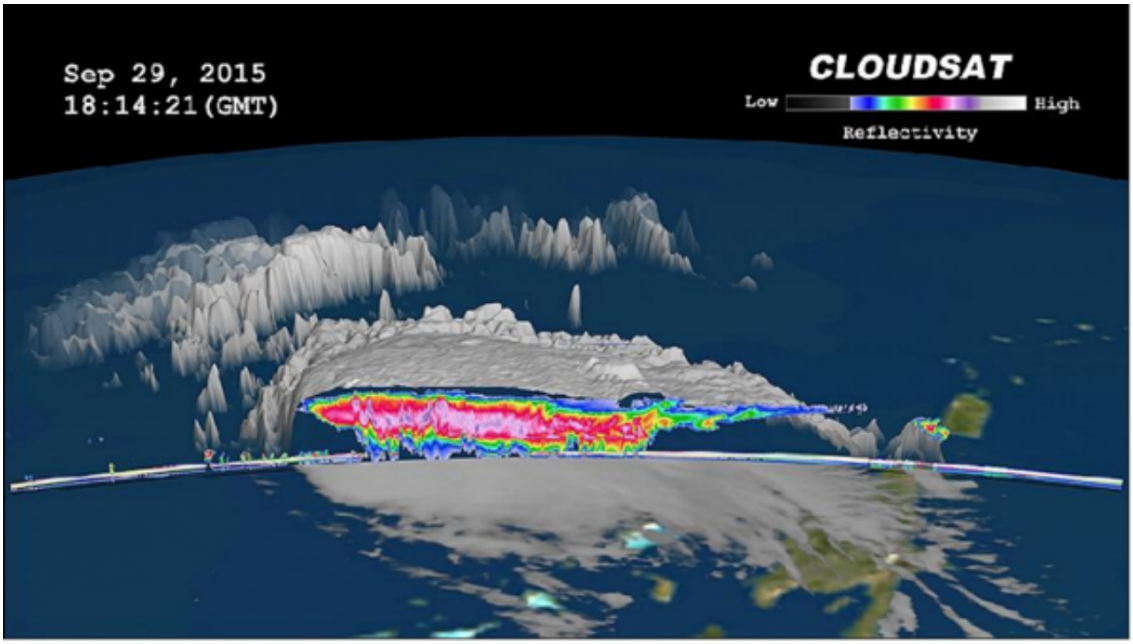

Figure 1.1: Illustration of the CloudSat mission. The onboard Cloud Profiling Radar (CPR) provides vertically resolved profiles of cloud and precipitation structure with a native vertical resolution of 240 m per range bin. Source: NASA/JPL-Caltech/CloudSat Data Processing Center [58].

Despite its high sensitivity, the parametric design of CloudSat's CPR results in significant vulnerability to ground clutter, a limitation shared by other spaceborne radars that detect weak scatterers such as hydrometeors [56, 37]. Operating at a frequency of 94 GHz, the radar receives surface returns that are typically 2–5 orders of magnitude stronger than those from snowfall or shallow clouds. As a result, strong surface echoes contaminate the lower atmospheric bins, obscuring the hydrometeor signals of interest near the surface boundary layer [46]. As shown in the upper panel of Fig. 1.2), surface responses can reach approximately 40 dBZ at the lowest range bin, whereas snowfall reflectivity is typically thresholded at -15 dBZ for detection [36, 23]. Lamer et al. [37] report a similar median surface echo strength over land, with variability constrained to less than 2 dBZ due to factors such as surface heterogeneity and satellite altitude. Collectively, these statistics underscore the practical infeasibility of detecting snowfall signals within the surface bin.

Furthermore, surface clutter can propagate vertically into the upper bins due to CloudSat's vertical oversampling scheme (illustrated in the lower panel of Fig. 1.2), leading to

clutter signatures superimposed on the hydrometeor profiles within the lowest four bins, and occasionally beyond [46]. According to the upper panel of Fig. 1.2, profiles become largely free of clutter at approximately the fifth bin above the surface. Lamer et al. [37] report a median surface echo strength of -27 dBZ at 0.75 km altitude, approaching CloudSat's minimum detectable signal (MDS) threshold of -30 dBZ [46]. This suggests that more than half of the profiles at the third bin above the surface are likely to be minimally affected by clutter. However, when examining individual profile columns, granule-level statistics become insufficient, as surface reflectivity varies considerably across footprints. The clear-sky surface return observed at one footprint does not reliably predict the clutter contamination at another. To resolve snowfall near the surface, it is therefore essential to determine, on a per-bin basis, whether clutter is present and to what extent it contaminates the hydrometeor signal.

Surface clutter contamination in CloudSat's CPR reflectivity profiles propagates into all downstream data products. In this study, however, we focus specifically on **snowfall rate** as a case study, due to its interpretability and suitability for proof-of-concept demonstration.

1.1 Operational Treatment of Clutter Zones in Satellite Snowfall Retrievals

Although the official Data Processing Center (DPC) provides both a surface clutter identification (SCI) algorithm and a surface clutter estimation (SCE) algorithm for the 2B-GEOPROF radar profile product [56, 57, 46], the snowfall retrieval community has traditionally excluded the lowest 1 km of radar profiles when retrieving snowfall occurrence and intensity. This exclusion stems primarily from two limitations. First, the SCI algorithm is highly conservative, designed to minimize false positives in hydrometeor detection at lower bins [46]. While this ensures that most identified hydrometeors are genuine, it also results in the misclassification of many bins that may contain significant hydrometeor signals as clutter. Second, the SCE algorithm relies on simplified assumptions: flat terrain, weak attenuation, and the absence of multiple scattering [57, 56], which often do not hold in practice. Violations of these assumptions, particularly over land, lead to unreliable clutter estimates [43]. Our analysis of a representative CloudSat granule shows that only approximately 85% of footprints have tolerable clutter estimation confidence, markedly lower than the over 90% confidence typically observed over water surfaces (excluding sea ice).

To address ground clutter, the snowfall retrieval community has long adopted a practical

convention: discarding the lowest portion of the radar profile most likely to be contaminated [43]. In 2008, Liu proposed defining the fifth range bin above the surface over land (and the sixth over ocean) as the "near-surface bin" (NSB), using its retrieved snowfall rate as a proxy for surface snowfall and excluding all bins below [41]. This convention was subsequently adopted by Kulie and Bennartz [36] and has since been implemented in nearly all CloudSat snowfall retrieval products, including the most recent version of 2C-SNOW released by the DPC [65, 66]. While these lower bins are not explicitly flagged as cluttered, they are implicitly treated as such by being entirely ignored, thus defining a nominal "clutter zone" (CZ, or blind zone in some other contexts). The NSB snowfall is then simply replicated downward to approximate surface values, under the assumption that it represents the last well-observed bin. However, this approach has well-recognized limitations: it systematically omits shallow snowfall events and risks overestimation due to virga, and its accuracy deteriorates significantly in meteorologically heterogeneous boundary layers [43, 32, 33, 61].

Accordingly, this study follows the established convention of removing the lowest portion of the radar profile identified as the surface clutter zone. We then formulate the reconstruction problem as a profile inpainting task following the work done by King et al. in 2024 [32], using the clutter-free portion of the radar observations along with auxiliary meteorological variables to infer the missing values. Unlike conventional inpainting problems where missing values can occur arbitrarily throughout the domain [18], our task involves a structurally constrained gap: only the lowest part of each profile is missing. This constraint prohibits interpolation-based approaches that rely on values below the missing bins. Instead, the reconstruction must extrapolate downward from the well-observed upper portion of the profile, making it both an inpainting and a sequence-prediction problem [27]. Since the goal is to estimate continuous snowfall rates, the task is framed as a real-valued regression problem. Several unique challenges arise in this context, which are addressed in the following subsections.

1.2 Domain Shift: A Key Limitation in Generalizing Profile Models

As noted in King et al. [32], who reconstructed near-surface radar reflectivity profiles at a single ground-based site, transfer learning should be considered to address the degradation in model performance when applied to locations different from where it was trained (i.e., the leave-site setting). Fundamentally, predictions from a regression model aim to estimate the

conditional expectation of a target variable given some known inputs—that is, the integral of all possible labels under the conditional distribution of the data [22, 26].

Domain shift arises when the distribution of data used for training differs from the distribution where the model is later applied [35, 54, 22, 26]. This gap, encompassing differences in covariates, response variables, or the underlying functional relationship, introduces bias and limits generalization. In our context, the task of reconstructing near-surface snowfall (i.e., inpainting the clutter zone) can be expressed

$$\hat{S} = f(v) = \mathbb{E}(S \mid v)$$

Here, S denotes the random vector of snowfall rates in the lowest four bins of the clutter zone (the "prediction chunk"), and v is a feature vector comprising well-observed information used to infer S. The features v may include retrieved snowfall profiles above the clutter zone and collocated auxiliary data such as ECMWF model fields [51, 1], ground-based observations, or products from other A-Train/C-Train-constellation sensors [6]. The model f is the regression function determined by the data-generating distribution for (V, S), so different distributions yield different underlying relationships. Under standard regression assumptions, the optimal regression function under the training distribution is simply the conditional mean:

$$f_{\mathcal{D}_{\text{train}}}(v) = \mathbb{E}_{\mathcal{D}_{\text{train}}}[S \mid v].$$

[22, 26].

However, when the application domain $\mathcal{D}_{\text{test}}$ (e.g., the global Earth system) differs from the training domain $\mathcal{D}_{\text{train}}$ (e.g., a specific ground-based site), prediction errors arise due to domain shift. The generalization error at a test covariate $v' \sim \mathcal{D}_{\text{test}}$ is

$$\varepsilon_s = \left\| f_{\mathcal{D}_{\text{train}}}(v') - \mathbb{E}_{\mathcal{D}_{\text{test}}}[S \mid v'] \right\| = \left\| \mathbb{E}_{\mathcal{D}_{\text{train}}}[S \mid v'] - \mathbb{E}_{\mathcal{D}_{\text{test}}}[S \mid v'] \right\|, \tag{1.1}$$

where $\|\cdot\|$ denotes an appropriate norm (e.g., absolute value for a scalar target).

This error term ε_s quantifies the distance between the model's learned expectation and the true conditional expectation under the test distribution, and it is expected to be significantly nonzero when domain shift is severe. In our setting, the domain shift problem is particularly pronounced because ground-based snowfall observations are geographically sparse and concentrated in specific regions [32, 47, 17], whereas our goal is to apply the model globally, including remote, poorly sampled regions such as the open ocean. Overcoming this distributional mismatch is therefore essential for achieving robust global inpainting of snowfall profiles in the clutter zone.

Particularly, in this task, domain shifts arise along three key dimensions: horizontal, vertical, and temporal. As illustrated in the right panel of Fig. 1.3, the yellow region denotes the domain where CloudSat and other spaceborne CPRs provide reliable observations; we refer to this region as CS (CloudSat-observed space). In contrast, the pink layer represents the surface clutter zone, where CloudSat measurements are unreliable and must be reconstructed; this region is denoted as CZ (clutter zone). By construction, CZ and CS are mutually exclusive. The thin vertical cylinders in the figure represent the sparse ground-based radar sites, denoted as GB, which provide localized but vertically complete observations. At present, full observational access is only available in CS and GB, while the vast majority of CZ remains unobserved. As will be discussed in the next subsection, most existing methods are trained exclusively on GB, which is geographically sparse and often unavailable in remote regions such as the open ocean [32, 47, 17]. Although GB provides complete vertical profiles and thus minimizes vertical domain shift, its limited horizontal coverage raises concerns about spatial generalization. For instance, a model trained on Arctic ground-based data is unlikely to perform well when applied to tropical or oceanic regimes, due to the identified domain-shift effect in Equation 1.1.

To the best of our knowledge, nearly all existing methods for near-surface snowfall retrieval have been trained exclusively on GB data. While these models benefit from vertically complete profiles, their limited spatial coverage raises serious concerns about generalizability. In particular, as we will show in the next subsection, they often fail when applied beyond their original training domains, where observational conditions differ substantially. This reveals a critical gap in the current literature: existing models are not designed to transfer across spatial or observational domains, limiting their utility in global applications. These limitations motivate our first research question and form the basis for the transfer learning strategy proposed in this study.

1.3 Limitations of Prior Methods Under Domain Shift

This section reviews existing approaches to snowfall profile reconstruction through the lens of statistical learning, as motivated in Section 1.2. We compare several representative models, highlight their input requirements and assumptions, and assess their vulnerability to domain shift. Table 1.3.1 summarizes the general formulation and core limitations of each method, including both historical approaches and the strategy proposed in this study. Particular attention is paid to the underlying learning paradigms (e.g., supervised, unsupervised), as well as to the horizontal and vertical spatial constraints introduced by different data sources for both training and inferences.

Table 1.1: Summary of representative snowfall reconstruction methods under a statistical learning framework. Here, CS, GB, and CZ denote the CloudSat-observed domain, the ground-based observational domain, and the clutter zone to be reconstructed, respectively.

Method Class	**Model Input**	**Training Data**	**General Formulation**	**Key Limitations**
Unsupervised Self-contained (Baseline)	CS	—	$\hat{s} = \arg\min_{s_i} \|p(s_i) - (x, y, h, t)\|$	Assumes spatio-temporal homogeneity; lacks model flexibility
Supervised Assimilation	CS, GB	$D'_{\mathrm{CS}}, D_{\mathrm{GB}}$	$\hat{s} = \mathbb{E}_{\mathcal{D}_{\mathrm{CS}\cap\mathrm{GB}}}[S \mid v_{\mathrm{CS}}, v_{\mathrm{GB}}]$	Domain shift; requires collocated auxiliary data
Supervised Self-contained	CS	$D'_{\mathrm{CS}}, D_{\mathrm{GB}}$	$\hat{s} = \mathbb{E}_{\mathcal{D}_{\mathrm{CS}\cap\mathrm{GB}}}[S \mid v_{\mathrm{CS}}]$	Domain shift; error-prone due to collocation mismatch
CS Pretraining + GB Finetuning	CS	$D_{\mathrm{CS}}, D_{\mathrm{GB}}$	$\hat{s} = \mathbb{E}_{\mathcal{D}_{\mathrm{CS}\times\mathrm{GB}}}[S \mid v_{\mathrm{CS}}]$	Underexplored transferability; effectiveness varies with domain gap

1.3.1 Unsupervised self-contained (Baseline)

We begin by reviewing the algorithm used in the official snowfall profile data product released by the CloudSat Data Processing Center (DPC), which remains one of the few approaches not subject to domain shift. In this default algorithm that is also adopted by most other CloudSat-based snowfall retrievals, surface snowfall and snowfall within the boundary layer are estimated by replicating the snowfall rate observed in the near-surface bin (NSB) [31, 65, 66, 23]. The NSB is defined as the lowest radar bin expected to be free of ground clutter, and thus assumed to be the last reliably observed bin above the clutter zone.

This approach can be interpreted as a 1-nearest-neighbor model that selects the "closest" snowfall rate from the spatio-temporal domain. The prediction function is given by:

$$\hat{s} = f(x, y, h, t) = \arg\min_{s_i} \|p(s_i) - (x, y, h, t)\| \tag{1.2}$$

where (x, y, h, t) denote the latitude, longitude, height, and timestamp of the target radar bin, $\|\cdot\|$ is a dimensionless distance metric, and $p(s_i) = (x_i, y_i, h_i, t_i)$ is the location of a candidate snowfall value s_i. This method is physically motivated: snowfall tends to vary smoothly in space and time due to the continuity of meteorological fields, and spatio-temporal autocorrelation is typically strong. The method is further justified by the quasi-Markov property of snowfall rate fields (as discussed in Section 2.1), which implies that the NSB snowfall is often the most informative predictor for bins within the clutter zone.

As an unsupervised, training-free method, this approach should offer high generalizability and minimal prediction variance based on statistical-learning theory [22]. However, its lack of flexibility becomes a major limitation in meteorologically heterogeneous regions, where the assumption of vertical homogeneity fails. In such cases, the replication model can produce large biases, especially when complex vertical gradients or shallow snow layers are present [43, 32, 33]. More flexible models are therefore needed to capture higher-order dependencies and richer meteorological structure from broader observational contexts.

Beyond replication-based approaches, two major model families are commonly used in the CloudSat snowfall community to address the discrepancy between NSB observations and surface snowfall: (1) assimilation models and (2) self-contained models. The choice between them typically depends on the availability of auxiliary data and the desired application scope.

1.3.2 Assimilation models

Assimilation models incorporate collocated external data, such as surface elevation, temperature, or reflectivity from other instruments, as auxiliary inputs to improve snowfall prediction in the clutter zone. These models are typically used in localized case studies where high-quality, vertically resolved ground-based observations are available. For example, Bennartz et al. corrected DEM-related errors in CloudSat snowfall retrievals over Greenland by using high-resolution regional airborne elevation data to adjust the surface bin classification locally [5]. Similar corrections were later integrated into CloudSat's 2B-GEOPROF and related products [44]. The general formulation of an assimilation model is:

$$\hat{s} = \mathbb{E}_{\mathcal{D}_{\mathrm{CS}\cap\mathrm{GB}}}\big[S \mid v_{\mathrm{CS}},\, v_{\mathrm{GB}}\big]$$

where CS $\cap$ GB denotes only the subset of CloudSat scenes that are spatially and temporally collocatable with ground-based measurements, and v_{CS}, v_{GB} are the input features from the overlapping domains. Although the model nominally draws from CloudSat observations, it can only be applied in regions where corresponding ground-based features v_{GB} are available. This severely limits the application range of assimilation models. Even in areas with abundant CloudSat coverage, the model is effectively inoperable without concurrent ground-based data. In addition, collocation errors pose a substantial challenge. These errors arise from mismatches in space and time when aligning auxiliary ground-based observations with CloudSat footprints, often formalized by a nearest-neighbor search:

$$v_{\mathrm{GB}}(x,y,h,t) = c(v_{\mathrm{CS}}(x,y,h,t)) = \arg\min_{v\in D_{\mathrm{GB}}} \|p(v) - (x,y,h,t)\|$$

Here, $p(v)$ denotes the spatio-temporal coordinates of the auxiliary input, and $c(\cdot)$ defines the collocation function. When the match is imperfect, which is often the case, the resulting residual error affects both model training and inference. Kodamana and Fletcher, for example, found that such collocation mismatches significantly underestimated the frequency of extreme virga events and proposed a distance-weighted correction to mitigate the effect [33].

In summary, assimilation models face three core limitations: (1) a highly constrained application range due to dependence on ground-based inputs, (2) sensitivity to collocation mismatch, and (3) potential domain shift even within the collocated dataset.

1.3.3 Self-contained models

Self-contained models represent a more scalable alternative to assimilation approaches, as they rely primarily on CloudSat data or other satellite-based inputs that are easier to collocate, such as products from the A-Train or C-Train constellations. These models predict snowfall in the clutter zone using reflectivity profiles observed above the CZ. The default SnowProf algorithm released by the DPC is an unsupervised self-contained method, but more flexible supervised variants have also been developed.

For example, Bennartz et al. fit a regression model using ground-based MMCR observations to correct for height bias in the Z–S relationship between reflectivity near the surface (135 m) and in the NSB (1000–1500 m), enabling estimation of surface snowfall from NSB values [5]. Milani et al. similarly replaced anomalous snowfall values in 2C-SNOW-PROFILE with those from the nearest valid bin above the clutter-affected region [48]. Such models generally take the form:

$$\hat{s} = \mathbb{E}_{\mathcal{D}_{\mathrm{CS}\cap\mathrm{GB}}}\big[\, S \mid v_{\mathrm{CS}} \,\big]$$

Although inference uses v_{CS} from the full CloudSat archive, model parameters are learned only from the collocated subset $\mathcal{D}_{\mathrm{CS}\cap\mathrm{GB}} \subset \mathcal{D}_{\mathrm{CS}}$. Restricting the training distribution in this way introduces domain shift we introduced in Section 1.2 and additional error stemming from space–time mismatches between satellite and ground-based observations.

A notable extension of this framework is the CNN-based U-Net architecture developed by King et al. [32], which uses upper-profile reflectivity to inpaint the CZ and lowest bins of CS. The model is trained on a simulated KaZR VAP product from the ARM program that mimics CloudSat reflectivity in Arctic regions. To address device mismatch, Kollias et al. aligned KaZR and CloudSat profiles via statistical collocation [34], assuming

the resulting CloudSat-like reflectivity structure was sufficient for training. The model is trained exclusively on $\mathcal{D}_{\text{GB}}$ but applied using v_{CS}, based on the assumption that $v_{\text{CS}} \approx v_{\text{GB}}$. Its general form is:

$$\hat{s} = f_{\text{Z-S}}\left(\mathbb{E}_{D_{\text{GB}}}(S \mid v_{\text{CS}})\right)$$

where $f_{\text{Z-S}}$ represents a Z–S conversion function that transforms inpainted reflectivity into snowfall rate. While the model is widely deployable in principle, which requires only CloudSat and densely-covered model data at inference, the training set remains highly localized and sparsely distributed. This limits its exposure to the atmospheric variability present along the horizontal dimensions in global CS and CZ domains. Moreover, the additional Z–S conversion introduces further modeling uncertainty, particularly under varying climatic regimes [36, 31, 5]. This further motivates the development of a snowfall-rate-specific model that directly estimates snowfall end-to-end without relying on intermediate reflectivity conversion.

In summary, supervised self-contained models, including both regression- and CNN-based variants, offer practical advantages by relying solely on satellite input (and other equally available data) at inference, which enables the models' use anywhere in CS, not limited to a narrow horizontal coverage only. However, they suffer from three critical limitations: (1) their training data are spatially restricted due to the need for ground-based collocation, which limits exposure to global atmospheric variability; so (2) they exhibit significant domain shift when applied globally, particularly in remote or climatologically distinct regions such as the open ocean or tropics; and (3) they are prone to collocation and device mismatch errors, which can degrade model accuracy even within the training domain.

1.3.4 Summary

Therefore, all of the reviewed methods, including replication, assimilation, and self-contained, suffer from limitations that hinder their ability to generalize across the broader snowfall retrieval community. For self-contained models, the need to collocate CloudSat observations with sparse ground-based data restricts the available training samples and introduces geographic bias. The collocation process itself is operationally challenging and remains biased toward local site characteristics, especially when corrections rely on ground-based inputs rather than spaceborne proxies such as other A-Train instruments. Assimilation models face similar issues: site-specific correction schemes tend to favor horizontal locality and are difficult to generalize to regions lacking dense auxiliary data. Replication methods offer

better applicability due to their simplicity and independence from training data, but they rely on a vertical homogeneity assumption that often breaks down in regions with complex near-surface meteorology.

1.4 Research Gaps and Objectives

In summary, strong surface clutter in current CPR-based remote sensing systems prevents reliable observation of near-surface hydrometeors on a global scale, particularly within the boundary layer. This limitation significantly impacts our ability to characterize and model hydroclimatological processes. Existing hardware configurations are unable to recover the clutter-contaminated zone effectively, especially over land where surface reflectivity is most severe. As a result, there is a need to leverage other well-observed portions of the atmospheric column, possibly in combination with auxiliary data from independent sources, to infer or calibrate snowfall conditions near the surface.

Having reviewed existing approaches through the lens of statistical learning, we observe that despite decades of development, current models continue to face major challenges in reconstructing near-surface hydrometeor profiles globally, particularly in the lowest 1 km of the boundary layer. These challenges stem from three primary issues: (1) reliance on ground-based inputs that are geographically sparse, (2) dependence on overly simplified assumptions that are often violated in practice, and (3) vulnerability to domain shift when applied beyond the training region.

Among existing methods, supervised autoregressive models come closest to meeting our reconstruction goals, as they offer the potential to overcome the first two limitations. However, they still suffer from significant domain shift, especially across the horizontal dimension. To address this, we require a new training paradigm that fully exploits the data available and explicitly targets domain generalization across all spatial and vertical dimensions. To our knowledge, no such approach has yet been proposed within the CloudSat snowfall retrieval community.

This study sets out to design and evaluate such a model. Specifically, our research objectives are to:

1. Develop a method to reconstruct near-surface snowfall rates in the clutter zone ($\sim$1 km) globally using spaceborne Cloud Profiling Radar data.
2. Evaluate the performance of the proposed method against the current replication-based baseline.

3. Assess the generalizability of the model by identifying the conditions under which the inpainting results can be trusted.

In addition, we examine the model's residual limitations, with the goal of informing future efforts in both algorithm design and satellite mission development. Upon achieving the research objectives, we introduce a new autoregressive training paradigm that maximally leverages available data to mitigate domain shift across both vertical and horizontal dimensions. This paradigm enables accurate reconstruction of snowfall profiles within the clutter zone at a global scale, even under severe training data limitations. The resulting inpainting model produces more accurate near-surface snowfall estimates with acceptable computational efficiency and has the potential to improve upon the current DPC official SnowProf product by reducing biases and uncertainties in downstream scientific analyses related to boundary-layer snowfall. Beyond snowfall, the proposed training framework may be transferable to other globally distributed meteorological retrievals, such as microphysical or optical properties of shallow boundary-layer clouds, thereby offering a broader impact on hydrometeorological research across the Earth system.

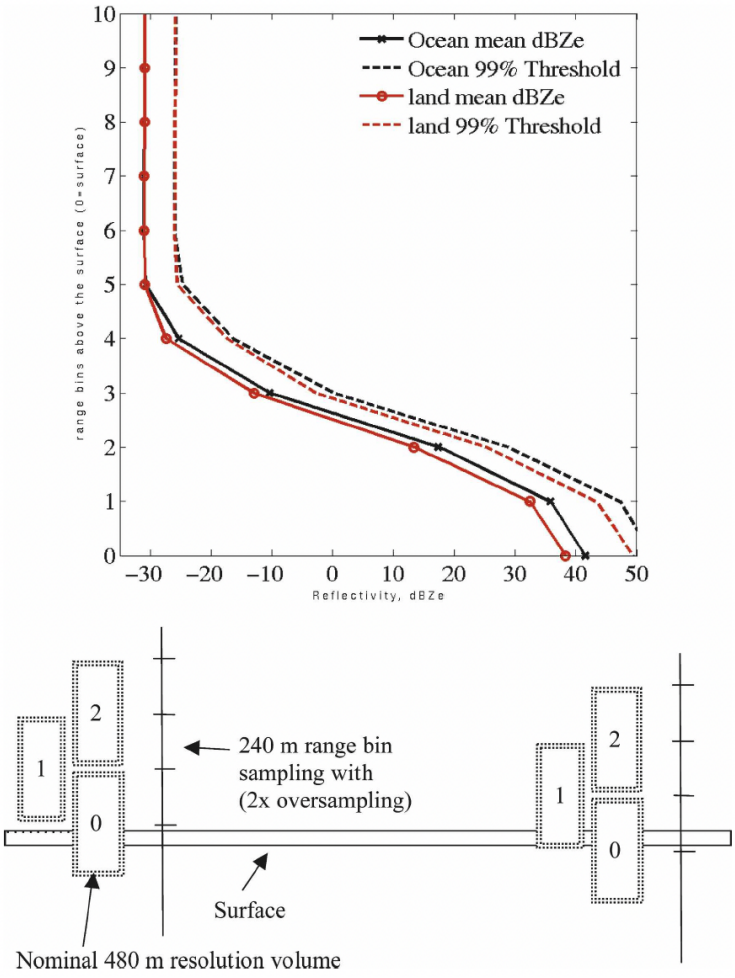

Figure 1.2: Surface responses to CloudSat's radar observations (top) and the oversampling process near the surface (bottom). Source: Adapted from Marchand et al. (2008) [45].

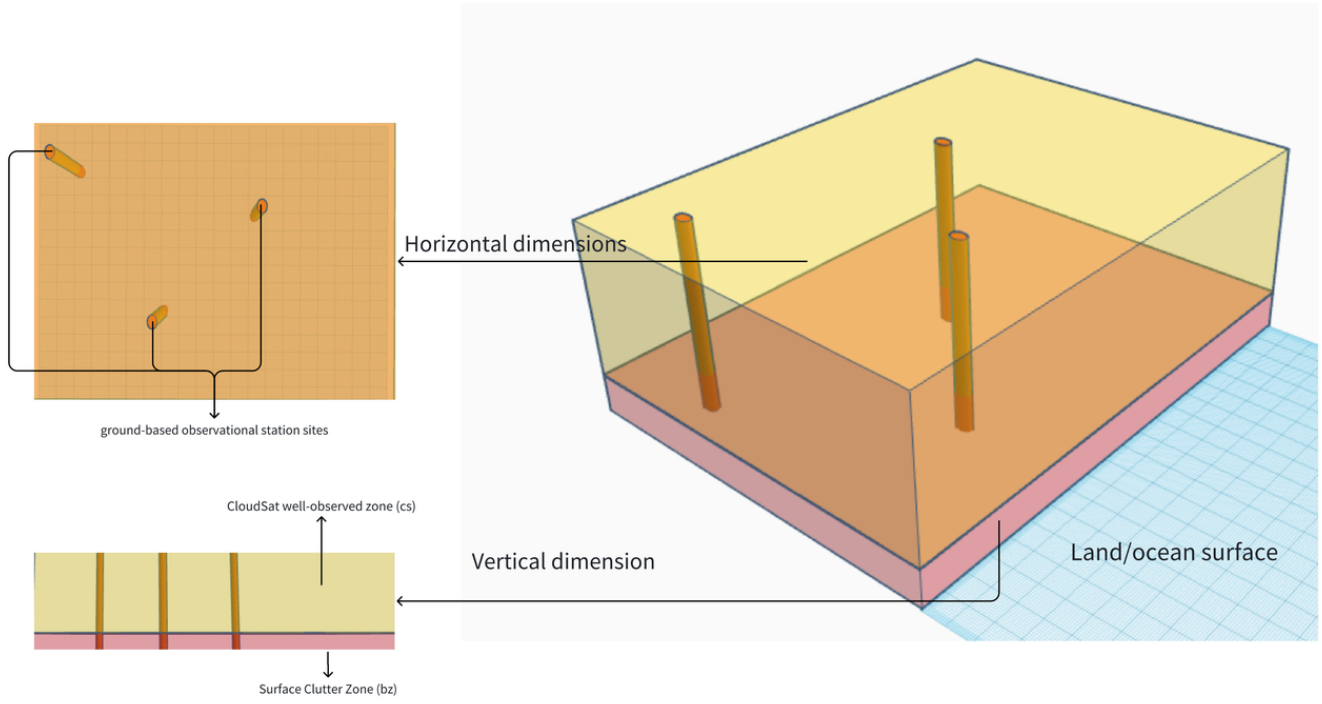

Figure 1.3: Conceptual illustration of CloudSat observational domains and domain shift. **Top left:** Horizontal view showing the global extent of CloudSat coverage, with three representative ground-based snowfall observation sites (e.g., ARM KaZR) marked as dots (GB). **Bottom left:** Vertical cross-sectional view along latitude and height, representing the CloudSat profile geometry. The lowest pink layer indicates the clutter zone (CZ); the yellow region above represents the well-observed CloudSat domain (CS). **Right:** Combined 3D view integrating both horizontal and vertical perspectives. The blue surface layer represents land and ocean boundaries.

Chapter 2

Data

To construct the dataset used in this study, we extract three-dimensional vertical profiles (e.g., snowfall rate) and two-dimensional footprint-level attributes (e.g., surface type) from a suite of CloudSat data products. For conceptual clarity, the three-dimensional profiles can be visualized as vertical cross sections, as shown in Fig. 1.1, while the footprint attributes correspond to geolocated surface-level metadata, typically one value per spatial position along the satellite track.

We organize the extracted data as a time-ordered sequence of georeferenced footprints, each consisting of a two-dimensional attribute vector and a three-dimensional multichannel profile, where each channel corresponds to a distinct retrieved geophysical quantity. The formal structure of the dataset is defined as:

$$X_i \in \mathbb{R}^{125\times c}, \quad x_i \in \mathbb{R}^d, \quad \forall (X_i, x_i) \in \mathcal{D} = \{\text{footprint } i \mid i = 1, 2, \ldots, N\}$$

Here, i denotes the footprint index ordered by time, X_i is the vertical profile consisting of 125 height bins and c physical channels, and x_i is a vector of d footprint-level planar attributes. The dataset $\mathcal{D}$ includes all N CloudSat footprints collected between April 2016 and October 2017. Fig. 2.1 provides a visual summary of the spatial and temporal distribution of the test dataset, which spans a broad range of geolocations and meteorological conditions. Nonetheless, due to CloudSat's sun-synchronous orbit and its operation in Daytime-Only mode during the selected study period, the sampled local hours are heavily concentrated between approximately 11:00 and 14:00 local time [56, 49, 64].

In the following subsections, we detail the selection of profile channels, footprint attributes, and temporal coverage criteria, all of which were guided by the scientific goals of this study.

2.1 Snowfall Profiles and Auxiliary Data from CloudSat

In this study, we utilize snowfall-rate profiles and ancillary retrievals from the CloudSat mission. These data serve two primary purposes: (1) to explore physical relationships among variables observable by the Cloud Profiling Radar (CPR), and (2) to construct training and test sets for the proposed inpainting model, in conjunction with additional observations from the A-Train constellation and model-derived datasets. The study period spans from April 1, 2016, to October 1, 2017, during which CloudSat remained part of the A-Train formation and had not yet reached the end of its operational stability [64, 6].

Below, we briefly outline the core specifications of the CloudSat data products used in this study. CloudSat is a sun-synchronous, near-polar-orbiting satellite with an inclination angle of approximately 98°, designed to collect high-resolution hydrometeorological observations, particularly over remote polar regions that are often undersampled by ground-based networks. Its onboard CPR provides horizontal spatial resolutions of approximately 1.1 km (along-track) and 1.4 km (across-track) per footprint. The radar's native vertical resolution is 500 m, which is oversampled to 240 m in the released data products [56].

CloudSat observations are segmented into granules, each corresponding to a single orbit and packaged as an HDF4 binary file containing approximately 37,000 footprints. During the daytime-only mode applied in our study period, approximately 20,000 valid footprints are retained per granule after filtering out instrument anomalies. The official CloudSat Data Processing Center (DPC) releases a hierarchy of data products, ranging from raw radar backscatter (Level 1B) to high-level geophysical and application-specific variables (Level 3). Typically, lower-level products are used as inputs to generate derived higher-level datasets [3].

For this study, we integrate four key CloudSat data products: 1B-CPR, 2B-GEOPROF, 2C-PRECIP-COLUMN, and 2C-SNOW-PROFILE [65, 66, 23]. Together, these products provide the core radar reflectivity fields and derived snowfall-related geophysical variables. The data are joined granule-wise and separated into two components for each granule: a two-dimensional dataset containing footprint-level attributes and a three-dimensional profile containing vertical structure. The two-dimensional dataset includes georeferenced attributes, such as latitude, longitude, and timestamp, as well as footprint-specific metadata. For example, surface type and precipitation flags are footprint-level (2D) attributes, assigning one value per footprint or ray. In contrast, quantities with vertical structure, such as bin height, radar reflectivity, and snowfall rate, are organized into three-dimensional matrices. In these, the first dimension corresponds to the along-track footprint index and the

second to the vertical bin height. The snowfall-rate profiles used in this study are extracted from the 2C-SNOW-PROFILE product spanning the period from April 1, 2016, to October 1, 2017. This product derives snowfall rates directly from the radar reflectivity profiles provided in 2B-GEOPROF. To ensure data quality, we also extract flags from 2B-GEOPROF to construct a validity mask for each vertical bin. Specifically, we mark as invalid all bins classified as cluttered or located below the detected surface height. In addition, we remove any footprint flagged as missing, anomalous, or otherwise unusable in 2B-GEOPROF. In addition, we incorporate several auxiliary variables from other CloudSat data products to support exploratory analysis and model diagnostics, including the generalizability assessments described in Section 3.5. These variables include the bin-centered height for each vertical level, digital elevation model (DEM) elevation, surface radar backscatter (σ_0), near-surface reflectivity, surface type, and the top-bin index of the lowest significant hydrometeor layer. These ancillary fields are extracted from 1B-CPR, 2B-GEOPROF, 2C-SNOW-PROFILE, and 2C-PRECIP-COLUMN. Several of these variables are themselves derived from external sources. For example, the surface type field in 2C-PRECIP-COLUMN is originally based on sea ice detection from AMSR-E, which has since been replaced by the daily sea ice product from SSM/I [23, 11]. Similarly, the near-surface reflectivity field incorporates ECMWF-AUX meteorological profiles to estimate attenuation effects from air and hydrometeors on the W-band CPR signal [23, 1].

Together, these three-dimensional profiles and two-dimensional footprint-level attributes, referenced by timestamp and geolocation, enable us to: (1) explore both physical and empirical relationships between snowfall rate and other meteorological variables, (2) train and validate the proposed inpainting model, and (3) assess its overall reconstruction performance.

2.2 Auxiliary Meteorological Profiles from ECMWF

As demonstrated by King et al. in 2024, meteorological context around a CPR-detectable scene can significantly improve reconstruction accuracy by providing physically relevant information [32]. Motivated by this, we incorporate a set of auxiliary meteorological profiles from the CloudSat ECMWF-AUX data product into our dataset.

The ECMWF-AUX product is derived from the operational AN-ECMWF analysis fields produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It includes five key meteorological state variables:

- Pressure

- Temperature
- Specific humidity
- Zonal wind component
- Meridional wind component

These profiles are used in our study to enrich the physical representation of atmospheric conditions associated with each CloudSat footprint. As a model-derived product, AN-ECMWF differs from CloudSat in both spatial and temporal resolution. It typically offers more complete spatio-temporal coverage and higher temporal frequency, but at the cost of coarser spatial resolution and increased uncertainty due to model-based assumptions.

To reconcile these differences, the CloudSat Data Processing Center (DPC) subsets and collocates AN-ECMWF fields to align with CloudSat's observational grid. The ECMWF variables are first interpolated linearly and then bilinearly in three-dimensional space to match CloudSat's vertical and horizontal bins, followed by linear interpolation in time. The result is a temporally and spatially collocated meteorological profile for each CloudSat footprint, aligned with the rest of the retrieval variables in the dataset [1].

To enable efficient and consistent spatio-temporal management of the large-scale dataset used in this study, the meteorological profiles are appended as five additional channels to the snowfall-rate profiles. These auxiliary channels are referenced by three-dimensional geolocation and timestamp, aligned with each column of the multi-channel vertical profiles shown in Fig. 3.2.

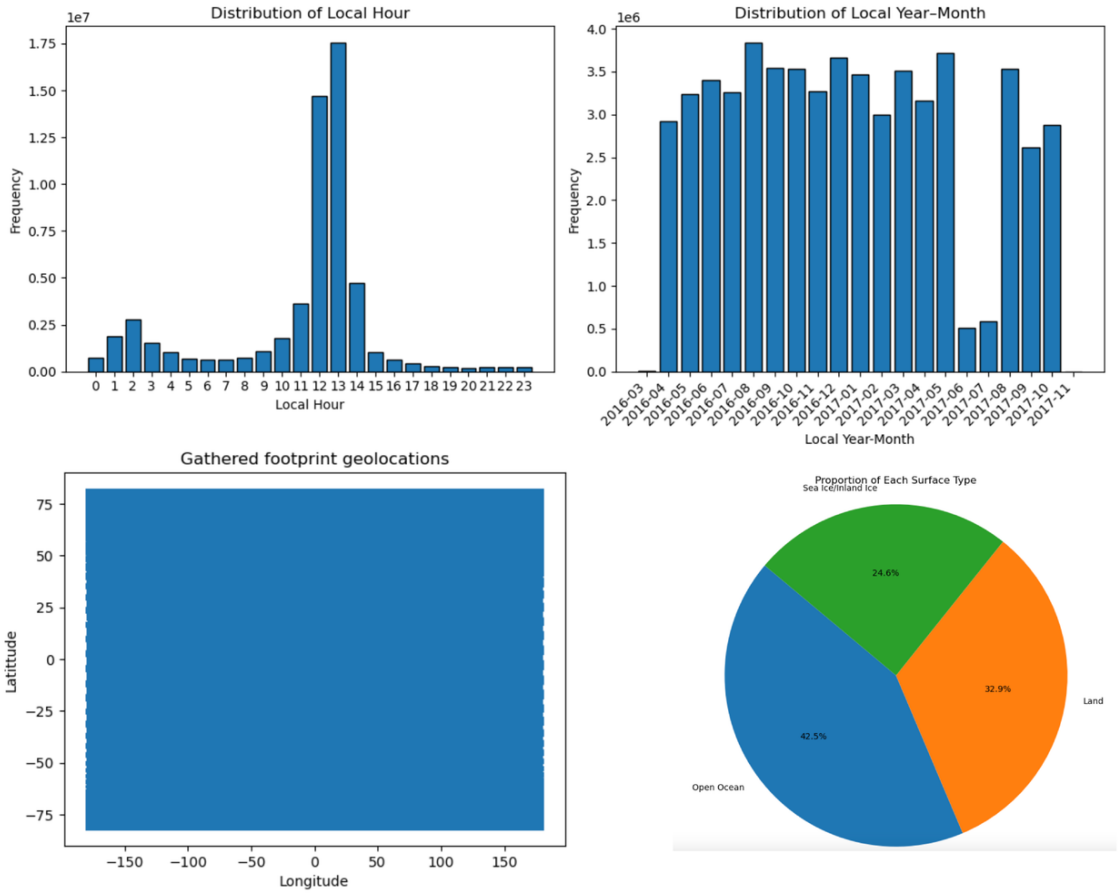

Figure 2.1: Spatio-temporal overview of the sampled test dataset. Local times are converted from the original TAI/UTC timestamps using the time zones inferred from geolocation.

Chapter 3

Methodology

In this section, we present the methodology designed to achieve the research objectives outlined in Section 1.4. We begin by identifying a physically grounded underlying model for reconstructing snowfall rates in the lowest radar bins based on known observations above. Building on this foundation, we introduce *SnowProfGPT*, a new autoregressive model tailored for snowfall-rate profile reconstruction.

To address the domain shift challenges identified in prior methods, we then propose a transfer learning training paradigm adapted from the field of text generation. This paradigm is designed to enable generalization to unseen geographic and meteorological conditions. To support quantitative evaluation of the reconstruction, we define a key concept—*Synthetic Clutter Zone* (SCZ), which provides dense, high-quality proxy data to serve as ground-truth references for the real clutter zone. Finally, we outline our generalizability evaluation framework, which assesses the performance of the proposed model across different spatial, temporal, and meteorological regimes.

3.1 Inductive Bias: An Identified Underlying Relationship

To develop a reconstruction model for snowfall profiles in the clutter zone (CZ), we must first identify a plausible form of the underlying relationship between observable and missing data, a formulation commonly referred to as the *inductive bias* in statistical learning [4]. This bias defines the assumed model family in which the true data-generating process is

believed to reside. Only by selecting an appropriate inductive bias can we constrain the model search space in a way that aligns with physical intuition and empirical structure.

In this study, we propose that the underlying process governing snowfall rate in lower atmospheric bins can be reasonably approximated as a Markov model [9]: that is, the snowfall rate at a given bin is conditionally dependent on the rates observed in the few bins directly above it, as well as on the local meteorological conditions. This assumption reflects both physical intuition and prior evidence of spatio-vertical autocorrelation in snowfall fields.

Specifically, the snowfall rate at a given radar range R can be described by the physical model used in the 2C-SNOW-PROFILE product, which relates snowfall rate to radar-observable microphysical quantities. It is given by:

$$S(R) = \frac{1}{\rho_{\text{liq}}} \int_{D_{\min}}^{D_{\max}} N(D, R)\, m(D, R)\, V(D, R)\, dD$$

Here, $S(R)$ is the snowfall rate in mm/hr at range R, and D, N, m, and V denote the snowfall particle diameter, particle size distribution (PSD), particle mass, and fall speed, respectively. The integral represents the expected mass flux of particles at range R, normalized by the liquid water density ρ_{liq}. In practice, this physical relationship is inferred indirectly through radar reflectivity, which is a function of the backscattered power associated with the PSD at each bin [66, 65].

However, below the near-surface bin (NSB), clutter-free radar reflectivity is no longer available. As a result, the snowfall rate in lower bins must be inferred based on information from the overlying, well-observed portion of the profile. One common approach to such extrapolation is to apply a Taylor expansion, assuming that the snowfall rate is sufficiently smooth and differentiable with respect to radar range R:

$$S(R + dR) = S(R) + dR \cdot \frac{dS}{dR}(R) + \mathcal{O}\left((dR)^2\right) \tag{3.1}$$

This expansion rewrites the undetectable snowfall rate $S(R + dR)$ as a function of the closest well-observed value $S(R)$, along with its local rate of change in the vertical direction. The $\mathcal{O}\left((dR)^2\right)$ term represents a higher-order remainder that becomes negligible when the bin spacing dR is sufficiently small. To estimate $S(R+dR)$, we therefore require knowledge of the vertical snowfall rate gradient $\frac{dS}{dR}$ at position R, which describes how the snowfall rate evolves from the observed level downward into the clutter-contaminated zone.

Assuming continuity in the snowfall rate scalar field, we can express the vertical gradient of the snowfall rate using the chain rule and the physical model introduced earlier. Applying Leibniz's rule for differentiation under the integral sign yields:

$$\frac{dS}{dR} = \frac{1}{\rho_{\text{liq}}} \int_{D_{\min}}^{D_{\max}} \left[\frac{dN}{dR} \cdot m \cdot V + N \cdot \frac{dm}{dR} \cdot V + N \cdot m \cdot \frac{dV}{dR} \right] dD$$

To gain more physical insight, we simplify this expression in terms of the first-order moment (i.e., the expectation) of the particle mass flux. The snowfall rate gradient can then be approximated as:

$$\frac{dS}{dR} = \frac{1}{\rho_{\text{liq}}} \left[\mathbb{E}_{D|R} \left[\frac{d(mV)}{dR} \right] + \frac{\mathbb{E}_{D|R+dR}[m(R)V(R)] - \mathbb{E}_{D|R}[m(R)V(R)]}{dR} \right], \quad \text{as } dR \to 0$$

Here, $\mathbb{E}_{D|R}[\cdot]$ and $\mathbb{E}_{D|R+dR}[\cdot]$ denote expectation operators over the particle size distribution (PSD) at ranges R and $R+dR$, respectively, corresponding to $N(\cdot, R)$ and $N(\cdot, R+dR)$.

Recognizing that mV represents the particle mass flux Φ, the expression simplifies further to:

$$\frac{dS}{dR} = \frac{1}{\rho_{\text{liq}}} \left[\mathbb{E}_{D|R} \left[\frac{d\Phi}{dR} \right] + \frac{d}{dR} \mathbb{E}_{D|R}[R, \Phi(R)] \right], \quad \text{as } dR \to 0$$

This formulation reveals that the vertical rate of change of snowfall rate at range R can be decomposed into two physically interpretable components:

1. The expected rate of change of the snow mass flux at range R, due to the physical process of snow falling (e.g., gravitational settling and aggregation).

2. The rate of change of the expected snow mass flux, capturing variability in the snow particle size distribution with altitude (i.e., snow-particle microphysical spatio-vertical heterogeneity).

For clarity, we refer to the two terms composing the linear rate of change of snowfall rate $\frac{dS}{dR}$ as the *motion effect* and the *microphysical effect*, respectively. The motion effect reflects the change in snowfall mass flux due to vertical dynamics, which, under a Eulerian perspective (as adopted by CloudSat's CPR), depends primarily on mechanically driven

meteorological variables such as vertical wind velocity, air pressure, and large-scale atmospheric motion. In contrast, the microphysical effect captures variability in the particle size distribution (PSD) and is driven by thermodynamic factors such as temperature, humidity, and phase transitions.

We assume that both effects can be encapsulated by a function $g(\vec{x}(R))$, which maps local meteorological conditions at range R to the vertical rate of change of snowfall rate:

$$g(\vec{x}(R)) \approx \frac{1}{\rho_{\text{liq}}} \left[\mathbb{E}_{D|R} \left[\frac{d\Phi}{dR} \right] + \frac{d}{dR} \mathbb{E}_{D|R}[R, \Phi(R)] \right] = \frac{dS}{dR}, \quad \exists\, g : \mathbb{R}^c \to \mathbb{R}$$

Here, $\vec{x}(R) \in \mathbb{R}^c$ denotes a vector of local meteorological variables at range R, and g is an unknown functional that maps these inputs to the vertical snowfall gradient.

Substituting this functional form into the Taylor expansion, we obtain a first-order autoregressive approximation for the snowfall rate at an undetectable bin in the clutter zone:

$$S(R + dR) \approx S(R) + dR \cdot g(\vec{x}(R)) + \mathcal{O}\left((dR)^2\right), \quad \exists\, g : \mathbb{R}^c \to \mathbb{R}$$

This expression constitutes the most basic autoregressive model, grounded in physical reasoning. It expresses the snowfall rate in an unobservable bin as a function of the nearest observed snowfall rate above, modulated by local meteorological conditions that are not directly measurable by CloudSat's CPR.

Before presenting the proposed model, we first revisit the replication model 1.2 employed by the official DPC product for handling bins within the clutter zone. This method can be interpreted as the **zero-order** approximation of Equation 3.1, wherein the snowfall rate at the next vertical bin is assumed to be identical to the current one, and the error term is the first-order term $\mathcal{O}(dR)$ of the distance between the known bin and the predicted bin. While simple, this strategy is physically grounded and generally reasonable if the expectation of its first-order remainder $\mathcal{O}(dR)$ is zero at some scales.

To empirically support this interpretation, we analyze a randomly selected CloudSat granule containing only footprints with positive snowfall rates. Figure 3.1 shows two empirical distributions derived from this sample: the right panel plots the marginal distribution of snowfall rates $S(R)$, while the left panel shows the conditional distribution of the difference $S(R + dR) - S(R)$. The former follows an approximately exponential distribution, consistent with the long-tailed behavior of PSD-derived snowfall rates. In contrast, the

distribution of differences $S(R+dR)-S(R)$ closely resembles a narrow normal distribution centered near zero.

This empirical behavior implies that the snowfall rate in any given bin is typically very similar to that in the bin directly above, with only small random fluctuations. In other words, the expected value of the first-order Taylor expansion term, $dR \cdot g(\vec{x}(R)) + \mathcal{O}((dR)^2)$, is close to zero. Consequently, a simple Markov assumption of the form:

$$S(R + dR) \mid S(R) \sim \mathcal{N}(S(R), \sigma^2)$$

for some unknown variance σ^2, provides a reasonable approximation. This assumption directly underpins the replication strategy used by the DPC. However, when this process is applied recursively, e.g., to estimate the snowfall rate at the fourth bin below the NSB, the variance accumulates over time. Under a random-walk interpretation [9], the uncertainty at depth increases linearly with the number of steps:

$$S(R + 4dR) \mid S(R) \sim \mathcal{N}(S(R), 4\sigma^2)$$

In Section 4, we empirically observe this uncertainty accumulation in the baseline model's performance. As shown in the R-square plot in Figure 4.1, the degradation is even more rapid than predicted by the random-walk variance, likely due to a small but persistent bias in the model (approximately 0.01, as seen in Figure 3.1) that compounds with depth, further deteriorating predictive accuracy in lower bins.

In summary, while the replication model offers robustness through its simplicity and minimal parameterization, it corresponds to a zero-order approximation of the underlying vertical process. Significant improvements can be achieved by addressing two key limitations: (1) reducing the bias inherent in the Markov assumption, and (2) shrinking predictive uncertainty. These improvements are enabled by incorporating additional meteorological context, allowing the model to debias and enrich the first-order structure.

This leads to a probabilistic autoregressive formulation in discrete form:

$$S_{i+1} \mid S_i, \vec{x}_i \sim \mathcal{N}\left(S_i + \Delta R \cdot g(\vec{x}_i) + \mathcal{O}(\Delta R^2),\, \sigma^2\right) \tag{3.2}$$

Here, S_i denotes the snowfall rate at bin i, $\vec{x}_i \in \mathbb{R}^c$ represents local meteorological inputs at that bin, ΔR is the vertical bin spacing, which is 240 meters for the CPR of CloudSat, and g is a learnable function mapping auxiliary inputs to the expected vertical gradient in snowfall rate.

Accordingly, our autoregressive forward model takes the form:

$$\hat{S}_{i+1} = S_i + \Delta R \cdot g(\vec{x}_i) + \mathcal{O}(\Delta R^2)$$

with residual uncertainty characterized by standard deviation σ. This autoregressive inductive bias provides the physical foundation for the model architecture developed in Section 3.4, designed to capture sequential dependencies between the unobserved snowfall rates in the CZ and surrounding meteorological variables.

3.2 SnowProfGPT: A Self-Attention-Based Autoregressive Model for Snowfall-Rate Inpainting

In this section, we introduce *SnowProfGPT*, a self-attention–based autoregressive model designed to reconstruct cluttered meteorological profiles on a global scale. The model is trained using a generative pretraining (GPT-style) paradigm [28], enabling it to generalize across domain shifts and unseen geophysical regimes. Its architecture is built to satisfy two core objectives: (1) to align with the GPT paradigm for scalable domain adaptation, as further detailed in Section 3.3, and (2) to adhere to the autoregressive inductive bias established in Section 3.1, thereby capturing physically meaningful patterns in snowfall-rate profiles.

To maintain clarity and focus, we present only the core design components of the architecture in this section. Secondary implementation details, such as layer normalization, input standardization, and regularization techniques, are provided in the Table 3.1 and the accompanying GitHub repository for reproducibility.

3.2.1 Forward model definition: Inputs and outputs

To move beyond the simplest Markov model 3.2 identified in Section 3.1, we expand the model's input space to better reflect the nonlinear, spatially structured nature of real-world atmospheric systems. In particular, the chaotic and vertically heterogeneous structure of snowfall profiles cannot be adequately captured by strictly local or memoryless assumptions [50]. Therefore, rather than conditioning only on the local snowfall rate and meteorological variables, we introduce a larger spatiotemporal input window that includes neighboring vertical bins and adjacent horizontal footprints.

Table 3.1: Core architectural components and hyperparameter configurations of the SnowProfGPT model. Embeddings are applied separately for each channel, footprint, and auxiliary metadata feature. The model uses a multi-head transformer encoder–decoder architecture, with autoregressive decoding over four vertical bins. All embeddings operate in a shared latent space of 512 dimensions.

Component	Configuration
Architecture Type	Transformer-based autoregressive encoder-decoder
Embedding Dimensionality	512 (shared across all tokens and layers)
Input Channels	6 (each meteorological channel embedded separately)
Input Tensor Shape	$5 \times 9 \times 6$ (width $\times$ height $\times$ channels)
Encoder	8-layer Transformer encoder
Decoder	8-layer Transformer decoder (autoregressive)
Multi-Head Attention	8 heads per encoder/decoder layer
Feedforward Dimensionality	2048 units per layer
Encoder Dropout	0.3
Decoder Dropout	0.1
Activation Function	ReLU (per layer)
Normalization	LayerNorm applied before each encoder/decoder block
Autoregressive Target Sequence	4 lowest snowfall-rate bins in SCZ
Teacher Forcing	Enabled only in pretraining stage 1
Output Layer	Linear projection from $512 \to 1$ (snowfall rate)
Logarithmic Feature Transformation	Applied to all non-snowfall channels
Bin Positional Encoding	Sinusoidal encoding (length $9 + 4$)
Footprint Positional Encoding	Learnable scalars per footprint (5 total)
Channel Positional Encoding	Learnable scalars per input channel (6 total)
Metadata Embeddings	`surface type` (5 categories, including missing)
Output Bias Correction	Predicted snowfall < 0.2 mm/hr is clipped to 0.0

For sufficiently smooth fields, linear combinations of prior vertical values, such as $S(R)$, $S(R - \Delta R)$, $S(R - 2\Delta R), \ldots$ can yield improved approximations of $\frac{dS}{dR}$, as derived in Taylor-series–based finite-difference schemes [7, 39]. By incorporating such a neighborhood, the model can reduce predictive bias and uncertainty, provided the information is well structured and reliable. However, increasing the context window also incurs higher computational costs and may introduce redundant inputs, especially for our computationally expensive Transformer-based model [59], and in light of the quasi-Markov property 3.2 we identified for snowfall-rate profiles. To balance information richness with efficiency and tractability, we fix the input scene to a size of 9×5 in vertical and horizontal dimensions, respectively, with 6 total channels (one snowfall-rate channel and five meteorological channels), as shown in Fig. 3.2. The model's regression task is to learn a function G such that:

$$\hat{s}_{ij} = \mathbb{E}\left[S_{ij} \mid S_{I'J},\, X_{IJC},\, \vec{v}_I\right] = G(S_{I'J},\, X_{IJC},\, \vec{v}_I), \quad \exists\, G : \left(\mathbb{R}^{5\times5\times1},\, \mathbb{R}^{9\times5\times5},\, \{0,1,7,8,-1\}\right) \to \mathbb{R}$$

Here: - $S_{I'J} \in \mathbb{R}^{5\times5\times1}$ contains snowfall-rate values in the five vertical bins immediately above the target bin S_{ij}, across a 5×5 spatial window. - $X_{IJC} \in \mathbb{R}^{9\times5\times5}$ is the full meteorological tensor of five channels over nine vertical bins and five horizontal footprints. - $\vec{v}_I \in \{0, 1, 7, 8, -1\}$ encodes surface types across the horizontal window, representing open ocean, inland water, sea ice, land, and missing, respectively.

The index sets $I', I, J \subset \mathbb{N}$ refer to neighboring vertical and horizontal indices, and the Cartesian product $I \times J = \{(i, j) \mid i \in I,\, j \in J\}$ defines the spatial extent of the scene. The model does not constrain the target bin (i, j) to be in any fixed position within the window; instead, it is expected to adapt to varying relative positions within $I \times J$. This flexibility encourages the model to have a broader generalization and more flexible handling of scenes. Fixing the window size is also essential for managing computational cost, especially in transformer-based models, where attention mechanisms scale quadratically with input length [59].

With this model formulation, our goal is to approximate an unknown black-box function $G(S_{I'J}, X_{IJC}, \vec{v}_I) = \mathbb{E}[S_{ij} \mid S_{I'J}, X_{IJC}, \vec{v}_I]$, which estimates the snowfall rate in any CZ bin given its surrounding context. The model must simultaneously capture:

- **In-channel autocorrelation**, including the Markov dependency $S_{ij} \mid S_{I'J}$, such as the aforementioned *motion effect*, which provides the dominant predictive signal;
- **Cross-channel relationships** between auxiliary meteorological fields and snowfall, reflecting the aforementioned *microphysical effect* and helping to reduce systematic biases and uncertainty.

As demonstrated in recent studies, self-attention mechanisms are well suited for modeling both in-channel and cross-channel dependencies. This operation, often referred to as *contextualization*, allows the model to form a high-dimensional latent representation of the meteorological scene that reflects both the local physical state and its global context. Once this context is encoded, an autoregressive decoder can use it to reconstruct missing snowfall-rate values within the clutter zone. In the following subsections, we describe how SnowProfGPT operationalizes this autoregressive architecture, beginning with scene tokenization and embedding, followed by self-attention encoding, and concluding with autoregressive decoding. An overview of the model architecture is illustrated in Fig. 3.2, which will be covered in the following subsections

3.2.2 Tokenization and Embedding: Unifying heterogeneous meteorological inputs into a shared contextual feature space

We now describe the architecture that transforms multi-channel meteorological scenes into a unified token representation for downstream attention-based modeling. This step is foundational: it converts structured physical inputs into a form that enables semantic reasoning and context-dependent learning, tailored to the autoregressive model defined in Section 3.2.1.

Tokenization. We begin by flattening the structured input into a one-dimensional sequence using a 3D tokenizer, as illustrated in Fig. 3.2. The tokenizer takes as input: - $S_{I'J} \in \mathbb{R}^{5\times5\times1}$: the snowfall-rate values in the 5 vertical bins immediately above the target bin, - $X_{IJC} \in \mathbb{R}^{9\times5\times5}$: the auxiliary meteorological channels across the full 9×5 scene, - $\vec{v}_I \in \mathbb{R}^5$: the 1D footprint-level surface type encoding.

The tokenization function is defined as:

$$\vec{x} = T_m(S_{I'J},\, X_{IJC},\, \vec{v}_I), \quad T_m : \left(\mathbb{R}^{5\times5\times1},\, \mathbb{R}^{9\times5\times5},\, \mathbb{R}^5\right) \to \mathbb{R}^{225}$$

with a bijective mapping $m(i, j, c) = k$, such that $\vec{x}_k = X_{ijc}$.

Each entry $\vec{x}_k$ is referred to as a *token*, representing an atomic unit of information in the meteorological scene.

Why context matters. In NLP, the meaning of a token is inherently contextual [55, 42]. For example, the word "dog" can refer to an animal in the phrase "family dog," but to food in "hot dog." The same principle applies in meteorology: two tokens with the same scalar value, e.g., snowfall rate = 1 mm/hr, may represent entirely different phenomena depending on their location in the profile, the surrounding dynamics, or surface conditions. A single snowfall rate could correspond to a part of light stratiform snow in one case and pre-squall convective bands in another.

Embedding. To allow the model to differentiate these contexts, we linearly embed each token into a shared high-dimensional vector space:

$$\vec{z}_k = \vec{x}_k W_e^l, \quad \text{where } l = [m^{-1}(k)]_3$$

Here, l denotes the channel from which token $\vec{x}_k$ originated, and $W_e^l \in \mathbb{R}^{1\times d}$ is a channel-specific learned embedding matrix lifting up the channel of tokens to the contextual feature space. We implement 7 such embedding layers:

- 5 for the auxiliary meteorological channels in X_{IJC},
- 1 for the snowfall-rate channel in $S_{I'J}$,
- 1 for the surface-type vector $\vec{v}_I$.

These embeddings are expected to project all tokens, regardless of source or physical unit, into a unified semantic space where they can be compared via well-defined similarity measures. For example, inner products between embedded tokens can reflect relevance or contextual similarity within the scene.

Positional Encoding. Because Transformer models are inherently permutation-invariant [59], we inject spatial awareness into the token embeddings via learnable positional encodings:

$$\mathrm{PE}_k = \vec{p}_i^{\,w} + \vec{p}_j^{\,h} + \vec{p}_l^{\,c}, \quad \text{where } (i,j,l) = m^{-1}(k), \quad \vec{p}_i^{\,w},\, \vec{p}_j^{\,h},\, \vec{p}_l^{\,c} \in \mathbb{R}^d$$

This encoding provides relative positioning in width, height, and channel space analogous to the three axes in our spatiophysical input. These are implemented as periodic encodings, as visualized in Fig. 3.2. Without these, the model would lack awareness of token order or structure that are necessary to define a meteorological scene, unlike CNNs, which encode spatial locality via convolutional kernels.

The final embedded and position-aware token representation is:

$$\vec{z}_k = \vec{x}_k W_e^l + \mathrm{PE}_k, \quad \text{for } l = [m^{-1}(k)]_3$$

Each token is thus embedded into a high-dimensional contextual space, where heterogeneous meteorological inputs, originating from distinct devices or even with different physical quantities, may be well interpreted jointly in a shared, semantically meaningful scene representation.

3.2.3 Self-Attention Encoder: Contextual Representation of the Meteorological Scene

In addition to embedding meteorological tokens into a shared semantic space, the model must reason over complex spatial and cross-channel relationships. As we mentioned before, these relationships should involve latent autocorrelations along vertical or horizontal dimensions, cross-variable dependencies, or higher-order physical patterns. To support this,

we introduce a self-attention mechanism [59], adapted from the Transformer architecture, which forms the core of SnowProfGPT's contextual encoding.

Query-Key-Value transformation. To enable attention-based reasoning, each token embedding $\vec{z}_k \in \mathbb{R}^d$ is first linearly projected into three distinct representations: - A *Query* vector, which requests contextual information, - A *Key* vector, which responds to those queries, - A *Value* vector, which contains the information to be passed if selected.

These projections are learned through h independent attention heads, with each head applying its own set of transformation matrices:

$$H_k = \left(W_k^Q,\, W_k^K,\, W_k^V\right) \in \mathbb{R}^{d/h \times d}, \quad k = 1, 2, \ldots, h$$

Each head is designed to capture a different class of relationships: some heads may learn vertical continuity, while others may attend to surface effects, thermodynamic transitions, or other meteorological patterns. This multi-head architecture is necessary: without these projections, each token would default to attending most strongly to itself under the standard inner-product metric, which would prevent the model from learning true contextual dependencies.

Self-attention mechanism. For each head k, we compute attention scores between every pair of tokens based on the similarity between their Query and Key vectors. These scores are used to compute weighted combinations of the Value vectors, allowing each token to selectively aggregate information from other tokens in the scene:

$$\text{head}_k = \text{Attention}_k(\vec{z}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad \text{where} \quad Q = \vec{z}W_k^Q,\ K = \vec{z}W_k^K,\ V = \vec{z}W_k^V$$

The output of this attention layer is a sequence of contextualized embeddings, where each token's new representation reflects a learned combination of all other tokens' Value vectors weighted by their relevance in that specific head. These relevance scores capture how strongly each meteorological token relates to others in the scene, based on both spatial structure and physical variable type.

Multi-head fusion and feedforward transformation. The outputs of all heads are concatenated and linearly projected to merge information from different pattern classes:

$$\text{MultiHead}(\vec{z}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

To further refine the representations, we apply the standard Transformer encoder stack: a residual connection and layer normalization, followed by a feedforward network (FFN), and another residual + normalization block. The feedforward layers operate independently on each token and help the model learn interactions between dimensions of a single token vector. For example, as we mentioned that each dimension of a token embedding is expected to represent a distinct aspect of the token in the scene context, this may allow the model to explore dependencies between a snowfall rate's absolute value and its derived indicators (e.g., likelihood of snow squall, CPR detectability), if those are considered important in light of the real-world data.

The final output of the encoder is referred to as the *contextualized embedding* of the input meteorological scene. It integrates:

- **Global contextual information**, derived from attention-based relationships across all tokens, and
- **Local token-specific structure**, enhanced by the feedforward layers that capture intra-token feature interactions.

This contextualized representation forms the latent "knowledge base" of the scene analogous to "memory" in language models that the decoder will use to perform autoregressive inpainting in the next stage.

3.2.4 Self- and Cross-Attention Decoder: Autoregressive Meteorological Scene Inpainter

With the contextualized representation of the meteorological scene produced by the encoder, we now introduce the decoder block, which performs autoregressive prediction of the snowfall rates in the clutter zone. Specifically, the decoder sequentially estimates the snowfall rate in each of the four lowest bins, using the Markovian structure defined by the inductive bias in Section 3.1.

The decoder models the conditional probability distribution over the snowfall-rate sequence as:

$$P(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T \mid \vec{z}_{\text{enc}}) = \prod_{t=1}^{T} P(\hat{s}_t \mid \hat{s}_{<t}, \vec{z}_{\text{enc}})$$

where $\vec{z}_{\text{enc}}$ denotes the encoder output and $\hat{s}_t$ represents the t-th snowfall token to be predicted. This formulation captures both temporal (i.e., autoregressive) and spatial dependencies, consistent with the scene's 2D structure.

To do so, the decoder first applies a masked self-attention block over previously predicted snowfall tokens, allowing it to learn intra-sequence dependencies. This mechanism is structurally identical to that in the encoder but uses masking to prevent each token from attending to future ones. Next, a cross-attention layer fuses information from the encoded scene context, enabling the decoder to condition on the full meteorological scene while predicting each token. The cross-attention operation is defined as:

$$\text{CrossAtt}(S, \vec{z}_{\text{enc}}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad \text{where}$$

$$Q = SW^Q, \quad K = \vec{z}_{\text{enc}}W^K, \quad V = \vec{z}_{\text{enc}}W^V$$

Here, S denotes the sequence of already predicted (or initially available) snowfall-rate embeddings. The attention matrix should reflect the alignment between the decoder query tokens and the encoder's contextualized memory.

After the attention layers, the decoder applies the same residual connections, layer normalization, and position-wise feedforward network as in the encoder. These further refine the token representations and capture intra-token interactions across dimensions, e.g., the relationship between absolute snowfall rate and its contribution to snow squall likelihood for the same token.

Autoregressive decoding order. To reflect the spatiophysical structure of the scene, we autoregressively traverse the 2D snowfall-rate grid using a fixed order: first top-to-bottom along the vertical (height-bin) dimension, then left-to-right along the horizontal (footprint) dimension. This ordering mirrors the causal structure assumed in our autoregressive framework.

Formally, the decoder produces the next snowfall token as:

$$\hat{s}_t = \text{softmax}\left(W_{\text{out}} \cdot \text{DecoderBlock}(s_{<t}, \vec{z}_{\text{enc}}) + b\right)$$

Each newly predicted token $\hat{s}_t$ is appended to the decoder's input and becomes available for subsequent predictions. This iterative process continues until all missing snowfall-rate tokens in the scene have been reconstructed. By following the fixed 2D traversal order

along the flattened token sequence, the decoder effectively learns and renders vertical and horizontal dependencies within the clutter zone.

The resulting inpainting process thus combines:

- Local autoregressive structure aligned with vertical atmospheric physics, and
- Global conditioning on multi-channel scene context via cross-attention.

As shown in the last "place-back" of the output Value, the final output layer of SnowProfGPT is a linear projection that maps each predicted snowfall-rate token embedding back into its original physical dimensionality and position in the scene. Each reconstructed snowfall-rate value is then positioned in its corresponding spatial location within the scene by applying the inverse of the 3D tokenizer, m^{-1}, restoring the reconstructed values to their original vertical bin indices. Once all snowfall tokens in a scene have been generated, the model completes the inpainting of the clutter zone.

3.2.5 Overview

We have now fully defined the main structure and workflow of the inpainting model. During training, the reconstructed snowfall-rate tensor is compared to the labeled ground truth using a supervised loss function. This difference is then used to update the model parameters through backpropagation [38]. The goal is to minimize reconstruction error while preserving the inductive bias described in Section 3.1, ultimately converging to a function that generalizes well across meteorological regimes.

Why such complexity? Compared to traditional machine learning models, or even convolutional neural networks (CNNs) [40], SnowProfGPT adopts a notably more complex architecture. However, this complexity is not unnecessary. It is driven by the demands of transfer learning, where the goal is to minimize domain shift and support generalization to remote or unseen observational regimes.

First, SnowProfGPT is built on a Transformer backbone and trained under a GPT-style paradigm (detailed in Section 3.3). This design enables the model to benefit from *emergent intelligence* [63, 68], which is an empirically observed property of large attention-based models, where increasing contextual depth and capacity leads to the learning of surprisingly rich and generalizable patterns. As the model scales, it becomes increasingly resilient to domain-specific biases and more capable of constructing robust internal representations of snowfall-related meteorological scenes. This is essential for addressing severe

domain shift, especially in cases where ground-truth profiles are only available for narrow or geographically sparse regions.

Second, the embedding design enables flexible handling of diverse inputs. Regardless of origin, semantics, or physical units, each input variable is treated as a token and embedded into a shared latent space [59]. Within this space, variables of different dimensionalities and meanings, such as surface-type indicators, wind speed in meters per second, and temperature in degrees Celsius, can be jointly reasoned about as part of a coherent meteorological scene. This embedding strategy also allows the model to tolerate spatial mismatches, missing data, variable input structure, or even potential input from different missions that may be adopted in the future. For example, both vertically resolved meteorological profiles and footprint-level attributes can be interpreted in a unified space. The result is a semantically rich, flexible input format that supports reasoning across heterogeneous features.

Finally, the self-attention mechanism plays a critical role by enabling pairwise interactions across all tokens, regardless of their physical variable type or spatial location. This allows the model to learn both local and non-local dependencies across vertical levels, horizontal footprints, and meteorological channels. Unlike most CNNs, which typically transform local features through a sequence of convolutional layers before applying regression to spatially aggregated representations [40], self-attention maintains a global view throughout the encoding process. It does not "resolve" features locally before regression, but instead enables each token to directly attend to all others in the scene. As a result, SnowProfGPT can capture a broader range of physical and statistical patterns, including long-range cross-channel relationships, that are often difficult to represent with convolutional models.

In summary, SnowProfGPT brings together three architectural core designs:

- A GPT-based autoregressive transformer with emergent generalization capability,
- A semantic embedding system for handling heterogeneous and multi-scale meteorological data, and
- A self-attention mechanism capable of capturing non-local dependencies across all meteorological dimensions.

These elements jointly may enable SnowProfGPT to perform accurate, generalizable inpainting of snowfall-rate profiles in the clutter zone, making full use of the globally distributed CloudSat dataset while maintaining robustness under domain shift.

3.3 Generative Pre-training: A Transfer Learning Approach

Having defined the SnowProfGPT architecture according to the physical inductive bias established in Section 3.1, the next step is to design a training paradigm that effectively enables the model to learn the desired patterns from data. In this section, we describe the training scheme used to learn and transfer meteorological knowledge for snowfall-rate inpainting in the clutter zone (CZ).

Importantly, as an overview, this model adheres to the "CS Pretraining + GB Finetuning" paradigm identified in Table 1.1. Unlike prior methods that rely on a very spatially limited subset $\mathcal{D}_{\text{CS}\cap\text{GB}} \subset \mathcal{D}_{\text{CS}}$, collocated with $\mathcal{D}_{\text{GB}}$, our model can be pre-trained on the full CloudSat-observed domain $\mathcal{D}_{\text{CS}}$. This domain spans all well-observed vertical bins across the globe, offering a much larger and richer sample base than the sparsely distributed, GB-collocated datasets typically used in existing methods. This distinction is critical: Prior models suffer from limited generalizability due to their dependence on small, geographically clustered training sets (e.g., $|\mathcal{D}_{\text{CS}\cap\text{GB}}| \approx |\mathcal{D}_{\text{GB}}| \ll |\mathcal{D}_{\text{CS}}|$). In contrast, SnowProfGPT leverages the full extent of CloudSat's spatial coverage during pretraining, enabling it to learn from diverse scenes across a wide range of meteorological regimes. This design offers a potentially principled strategy for mitigating domain shift, particularly in the horizontal dimension, by utilizing all available high-quality CloudSat data, rather than discarding most of it due to the spatial coverage constraints from having to collocate with ground-based sites as almost all existing methods do.

As discussed in the Introduction, one of the central challenges in this task is the presence of severe domain shift: reliable CloudSat observations above the CZ are available globally, while ground-truth snowfall rates within the CZ are sparse, localized, and limited to a small number of ground-based radar sites. Previous work [32] has suggested that transfer learning is a promising direction for this setting, particularly to adapt models across horizontally diverse regions. To address this challenge, we adopt a generative pretraining (GPT) [28, 52] strategy originally developed for natural language processing. This paradigm is designed for scenarios in which large volumes of **unlabeled data** are available, but only limited labeled data exist for a specific downstream task. In such cases, the model is first *pretrained* on the unlabeled data using a self-supervised task, typically by learning to recover artificially corrupted sequences given surrounding context. Through this process, the model internalizes structural relationships among tokens, effectively "understanding" the data intrinsic patterns in a domain-agnostic way (i.e., general but not specific to a specific narrow domain). The pretrained model is then *fine-tuned* [25] on a

small labeled dataset, which allows it to specialize in a task-specific domain with minimal updates, preserving generalization.

We propose that this GPT paradigm is well-suited to our task of snowfall-rate profile reconstruction. Specifically:

1. CloudSat provides a massive archive of high-quality snowfall **unlabeled** observations above the CZ, spanning 17 years and covering diverse geolocations and meteorological regimes.

2. Ground-truth snowfall rates within the CZ are extremely scarce and geographically limited to a handful of well-instrumented ground-based radar sites (i.e., narrow GB in Fig. 1.3).

3. There exists a strong domain mismatch between CloudSat's global observational coverage (CS) and the localized nature of available in-CZ truth data (GB).

Given these constraints, we adapt the GPT training paradigm for meteorological profile reconstruction. The model is first pretrained on CloudSat-observed data from well-measured regions above the CZ, where snowfall rates can be reliably inferred. Then, the pretrained model is fine-tuned using the sparse and spatially constrained ground-based GB data to specialize in reconstructing the vertical layers that were previously unobserved. The overall scheme is illustrated in Fig. 3.3, and further elaborated in the following subsections.

3.3.1 Self-Supervised Pretraining on Unlabeled CloudSat Observations

To construct the pretraining dataset, we extract the lowest 20 reliable vertical bins above the defined Synthetic Clutter Zone (SCZ) from CloudSat observations between April 2016 and October 2017. Inspired by the pretraining tasks used in generative language models, we apply a corruption-based strategy in which portions of the input data, despite being fully observable, are masked, and the model is trained to reconstruct the missing values using the surrounding context. The goal of this phase is to expose the model to as much vertical and horizontal variability as possible, allowing it to learn generalizable spatiophysical patterns inherent in the meteorological data. To achieve this, we design a dynamic scene sampling procedure that includes both a horizontal sampler and a vertical sampler. As shown in Fig. 3.3, the full sampling pipeline proceeds as follows:

1. Extract the well-observed vertical bins (yellow layer) from all available CloudSat profiles.
2. Group every five adjacent footprints to form a horizontal "scene" unit.
3. Randomly select 80% of the available scenes for pretraining.
4. For each selected scene, sample 9 adjacent vertical bins to form a multi-channel 2D meteorological scene.
5. Mask the lowest 4 bins in the snowfall-rate channel of each scene, which serve as the reconstruction targets for the model.

To maximize diversity, we implement real-time dynamic chunking during training. The horizontal and vertical samplers generate new combinations of scene patches for each epoch, exposing the model to a wide range of spatial and vertical conditions. This sampling strategy serves as a form of data augmentation and is critical for reducing both model variance and inductive bias. From a statistical learning perspective, increasing diversity in the training samples helps lower generalization error by improving exposure to rare or extreme conditions across the vertical and horizontal domains [22]. The corrupted snowfall-rate bins act as self-supervised labels. By learning to reconstruct them from the surrounding context, the model internalizes spatial autocorrelation patterns and meteorological structure, tailored from the GPT-style training process used in language models.

To monitor model performance and guide optimization during pretraining, we define a validation set with three objectives: (1) to detect overfitting or spurious pattern learning, (2) to evaluate the model's ability to reconstruct snowfall rates in bins closest to the surface, and (3) to control the learning rate schedule and early stopping behavior. To satisfy the first goal, we randomly sample the remaining 20% of CloudSat scenes not used for training. These are held out entirely from the pretraining phase to provide an unbiased performance evaluation. To achieve the second goal, we constrain the vertical sampler to select only the lowest 9 bins from each scene and mask only the lowest 4 bins in the snowfall-rate channel. This ensures that the validation set targets the near-surface region of the CloudSat domain, providing a proxy for evaluating performance in the clutter zone, which remains unobservable in CloudSat data. For the third goal, we implement a learning-rate scheduler and early stopping mechanism [2, 12]. If the model's validation loss degrades for more than two consecutive evaluations, we halve the learning rate. If the degradation persists for five evaluations and the learning rate drops below a minimum threshold, training is terminated. Detailed pretraining hyperparameters and scheduling rules are documented in Table 3.2 for reproducibility.

Table 3.2: Training strategy and hyperparameter configurations used for pretraining and fine-tuning SnowProfGPT. The model is pretrained on CloudSat observations above the SCZ using a self-supervised autoregressive task, then fine-tuned on a small synthetic ground-based subset. Pretraining is performed in three stages with gradually increasing task complexity. Fine-tuning disables all regularization and uses fixed weights and learning rate.

Category	Configuration Details
Hardware and Device	Nvidia RTX 4090 (CUDA)
Model Architecture	Transformer-based autoregressive decoder (Section 3.2)
Input Tensor Dimensions	$5 \times 9 \times 6$ (scene width $\times$ height $\times$ channels)
Masking Target	Lowest 4 bins in snowfall-rate channel
Training Epochs	100 (early stopping based on validation)
Batch Sampling Strategy	Dynamic scene chunking with random vertical/horizontal samplers
Optimizer	Adam
Initial Learning Rate	1×10^{-7}
Learning Rate Scheduler	ReduceLROnPlateau (factor: 0.5, patience: 3)
Early Stopping	Stop after 5 degraded validations or LR $< 10^{-7}$
Loss Function	Mean Absolute Error (MAE) [24] + regularization terms
Stage 1 (Warm-up)	Mixed teacher-forced training (20% batches), no regularization
Stage 2 (Vertical Generalization)	Add correlation penalty with bin height ($c_1 = 10$)
Stage 3 (Horizontal Generalization)	Add correlation penalty with latitude ($c_2 = 2$)
Fine-Tuning Mode	Single batch; fixed learning rate (1×10^{-7}); Stage-3 regularization
Teacher Forcing (Fine-Tune)	Disabled (fully autoregressive inference)
Validation Strategy	Static subset of lowest 9 bins from held-out CloudSat scenes
Checkpoints	Saved per epoch with MAE and R^2 in filename

3.3.2 Multi-Stage Pretraining and Generalization Regularizations

Pretraining a model of SnowProfGPT's scale and complexity presents substantial challenges, particularly in achieving stable convergence and generalizable performance. To address this, we implement a multi-stage pretraining strategy that gradually increases task difficulty and introduces regularization to support vertical and horizontal generalization, which is shown as Algorithm 1. This approach not only improves convergence but also helps the model internalize physically meaningful patterns rather than overfitting to local data statistics.

Stage 1: Mixed teacher-forced and autoregressive training. In the first stage, the model is trained on a simplified version of the reconstruction task that combines teacher-forcing [20] and fully autoregressive prediction. Specifically, 20% of the training batches are provided with ground-truth snowfall rates from previous vertical bins when predicting

Algorithm 1 Multi-Stage Pretraining of SnowProfGPT on CloudSat Data

Initialize model weights θ
Set learning rate $\eta \leftarrow \eta_0$
Define three-stage loss functions: $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$
Fix validation set from lowest 9 bins in CloudSat profiles
for each training stage $s = 1, 2, 3$ **do**
    for each epoch **do**
        for each batch **do**
            Dynamically sample a 5×9 spatial scene with all 6 channels
            Mask lowest 4 snowfall-rate bins as targets to recover
            Compute loss $\mathcal{L}_s$ based on current stage
            Backpropagate gradients and update θ using learning rate η
        end for
        if validation loss degrades for 3 consecutive checks **then**
            $\eta \leftarrow \eta/2$
        end if
        if validation loss degrades for 5 checks or $\eta < 10^{-7}$ **then**
            break (stage converged)
        end if
    end for
end for
Return pretrained weights θ^*

the next bin in sequence. The remaining 80% of batches follow a purely autoregressive setup, in which each prediction is conditioned only on the model's previous outputs.

This strategy serves two purposes. First, teacher-forced batches help the model rapidly learn the Markovian autocorrelation structure present in vertical snowfall profiles. Second, the autoregressive batches challenge the model to learn more robust spatial dependencies and mitigate error accumulation from recursive prediction, which otherwise leads to random-walk–like error propagation [9].

Stage 2: Vertical generalization regularization. After convergence on the mixed training regime, we introduce an additional regularization term to encourage generalizability across vertical layers. This is implemented by modifying the loss function to penalize correlations between residual errors and altitude:

$$\mathcal{L}_1(y, \hat{y}) = \mathcal{L}_0(y, \hat{y}) + c_1 \cdot \sigma_{h,\, y-\hat{y}}$$

Here, $\mathcal{L}_0$ is the original reconstruction loss, and $\sigma_{h,\, y-\hat{y}}$ denotes the Pearson correlation coefficient [10] between altitude h and the prediction residual $y - \hat{y}$. This regularization term discourages the model from developing altitude-specific biases during training.

The motivation behind this constraint is rooted in the core goal of pretraining: to learn general meteorological patterns that hold across altitudes. Without this regularization, the model risks overfitting to height-dependent structures present in CloudSat data, which would lead to catastrophic degradation when it encounters previously unseen CZ heights during fine-tuning.

medskip

Stage 3: Horizontal generalization regularization. In the final stage of pretraining, we introduce a second regularization term to improve generalization across horizontal locations. Although the model is exposed to globally distributed CloudSat data, the spatial density of observations is uneven, favoring polar regions over equatorial zones due to the satellite's orbit. To prevent the model from developing latitude-dependent biases, we augment the loss function as follows:

$$\mathcal{L}_2(y, \hat{y}) = \mathcal{L}_1(y, \hat{y}) + c_2 \cdot \sigma_{\text{lat},\, y-\hat{y}}$$

Here, $\sigma_{\text{lat},\, y-\hat{y}}$ is the Pearson correlation coefficient between latitude and the snowfall-rate prediction residuals. This term penalizes any systematic dependence of the residuals on latitude, helping the model avoid overfitting to region-specific features.

Each stage of training is monitored using a fixed validation set. While the training set is dynamically constructed through random scene sampling in a $\sim$5 km altitude window, the validation set remains static and is composed of the lowest nine vertical bins across all footprints. We apply a learning-rate scheduler and early stopping to each stage [2]. The learning rate begins at a high value to support warm-up, and validation metrics are recorded every one-tenth of an epoch. If validation loss worsens for three consecutive evaluations, the learning rate is halved. If degradation persists for five rounds or the learning rate falls below 10^{-7}, training is stopped and the stage is considered converged.

3.3.3 Transfer Learning: From CS to GB

Once pretraining is complete, we fine-tune the model on the ground-based (GB) dataset following Algorithm 2, which contains snowfall-rate observations within the clutter zone. This dataset is small, spatially sparse, and limited to a handful of ground radar sites.

However, it contains critical information from vertical bins within the CZ that were unseen during pretraining.

Algorithm 2 Fine-Tuning SnowProfGPT on Ground-Based CZ Dataset

```
Initialize model with pretrained weights θ*
Set minimal learning rate η ← 10^−7
Fix validation set from GB-CZ (lowest bins only)
for each epoch do
    Load entire GB dataset in one batch
    Construct 5 × 9 × 6 scenes using only CZ bins
    Compute loss L_2 with regularizations
    Update weights θ using learning rate η
    if validation loss degrades for 2 consecutive rounds then
        break (fine-tuning converged)
    end if
end for
Return final model θ** (fine-tuned)
```

As illustrated in Fig. 1.3 and Fig. 3.3, the fine-tuning set spans only a few geographic regions and is orders of magnitude smaller than the CloudSat dataset. To preserve the general patterns learned during pretraining, we initialize fine-tuning with the final weights from the pretrained model and apply a minimal learning rate (1e-7) to prevent overfitting or catastrophic forgetting [67]. The vertical sampler is disabled, and each scene is constructed by chunking only the lowest vertical bins in the GB dataset. A validation set is again excluded from the GB data to monitor model behavior during fine-tuning, using the same stopping criteria and learning-rate control as in pretraining. This strategy enables the model to refine its representations for clutter zone conditions using a minimal and targeted weight update. By doing so, we allow the model to specialize in the unseen CZ domain while preserving the robust spatial generalization acquired from global-scale CloudSat observations.

3.3.4 Summary

In summary, as illustrated from the top to the bottom in Fig. 3.3, our GPT-based training scheme for snowfall profile reconstruction consists of three key phases. First, the model is pretrained on the CloudSat-observed domain (CS), which provides a vast and diverse archive of snowfall profiles across a wide range of geolocations and altitudes. This phase

encourages the model to learn generalizable patterns for reconstructing snowfall rates based on surrounding meteorological information: patterns that are valid along both the vertical and horizontal dimensions; Second, the model is fine-tuned on a small, geographically sparse set of ground-based (GB) observations. Although narrow in horizontal coverage, this dataset contains snowfall profiles within the clutter zone (CZ) that were entirely unseen during pretraining. Fine-tuning allows the model to adapt to the vertical zone of interest while preserving the general horizontal patterns acquired from CS. This multi-stage training strategy enables the model to first internalize broad physical relationships from abundant data, and then specialize using a minimal and targeted adjustment. In doing so, it is encouraged to address the central domain-shift problem in snowfall profile reconstruction, transferring generalizable knowledge from a globally sampled source domain (CS) to a sparse, vertically novel target domain (CZ).

3.4 Synthetic Clutter Zone: A Proxy for Model Assessment

In this subsection, we introduce the concept of a Synthetic Clutter Zone (SCZ), along with its associated synthetic CS and synthetic GB domains, to enable global-scale training and evaluation in the absence of genuine snowfall-rate observations within the clutter zone (CZ). The SCZ provides a practical proxy for model validation, designed as the lowest four bins of reliable snowfall-rate observations immediately above the official CZ defined in CloudSat data products. Its vertical thickness is fixed and matches that of the genuine CZ, enabling a consistent substitution for the purpose of training and assessment.

Reliable snowfall-rate profiles within the CZ are only available from upward-looking in-situ radar sites, which are sparse and highly localized. Indeed, ground-based radar sites cover less than 2% of the Arctic domain, and in-situ snowfall profiles over the open ocean are virtually nonexistent [32, 47, 17]. This extreme data sparsity motivates the need for a global proxy. To enable model training and evaluation under these constraints, we define the SCZ as the four lowest reliable bins above the CZ, and we withhold this region from model exposure during pretraining. In this framework, the SCZ functions as a proxy for the true CZ, which is accessible for our evaluation, but not for training, mirroring the inaccessibility of the CZ in real-world applications.

To simulate fine-tuning conditions, we define a synthetic GB dataset by selecting CloudSat profiles in the SCZ located within a 1-degree radius ($\sim$111 km) of two known ground-based radar sites: the ARM NSA site (71.323, -156.615) and the OLI site (70.495, -149.886).

These coordinates are derived from the KaZR-VAP product [34] as showin in the map for the fine-tune set in Fig. 3.3, which statistically aligns ARM Ka-band radar reflectivity with CloudSat observations and served as the ground-truth for CZ profiles for the inpainting model developed by King et al. in 2024 [32]. Since CloudSat's orbital coverage does not match the continuous temporal resolution of ground-based radar systems, we use a larger search radius to collect a sufficient number of SCZ profiles for synthetic fine-tuning. This "shift-up" configuration where the SCZ mimics the position of the CZ while remaining within the well-observed CloudSat domain allows us to train and evaluate a model using the same architecture and transfer learning framework developed in Section 3.3. The synthetic GB provides unseen vertical bins for fine-tuning, while the SCZ serves as an accessible but unseen target for assessing reconstruction accuracy and generalizability. In this setting, SCZ evaluation acts as a surrogate for CZ performance, with the understanding that some unquantified bias may remain due to differences between the SCZ and the actual clutter zone.

Following the official definitions in the 2C-PRECIP-COLUMN product [23], the near-surface bin (NSB) is the fifth lowest bin over land and sea ice, or the third lowest bin over open water. Bins below the NSB are considered unreliable and constitute the genuine clutter zone. For the SCZ, we extract the four vertical bins immediately above the genuine CZ, including the NSB. To simplify data handling and model configuration, we use a consistent SCZ thickness of four bins across all surface types. These SCZ profiles serve as the ground-truth targets for evaluating model performance across a range of metrics: R-squared [16], mean absolute error, and Q-sigma (defined in Section 4), as well as for generalizability assessments discussed in Section 3.5.

3.4.1 Generalizability Evaluation: Through Residual Modeling

Generalizability is a critical aspect of performance for any inpainting model, particularly in geophysical applications where data conditions vary widely across space and time [30]. Deep learning models, while powerful, are often less robust than physically based or simple empirical methods, and may degrade unpredictably when applied outside their training domain [53, 19]. To evaluate the generalization behavior of SnowProfGPT, we conduct a residual analysis designed to identify the key factors that influence the model's performance and quantify its sensitivity to varying environmental conditions.

Rather than evaluating performance strictly in aggregate across fixed dimensions (e.g., latitude, local season) with a strong assumption about the most significant factors influencing the performance of the inpainting model, we train a secondary regression model,

referred to as the Residual Regression Model (RRM), to model the inpainting residuals $\hat{s} - s$ as a function of known footprint/ray-level attributes. This model serves as a proxy for the expected behavior of the inpainting model under different conditions, and allows us to identify:

- Is the performance of SnowProfGPT dependent on the external conditions where it inpaints,
- If so, along which dimensions the model is stable or unstable, and
- Under which conditions the model tends to underpredict or overpredict snowfall rates.

The RRM is trained on a dataset of residuals co-generated during the global SCZ reconstruction phase. For each footprint, we extract a set of predictor variables, including:

- Geolocation (latitude and longitude),
- Predicted bin height,
- Surface elevation,
- Surface-layer wind speed,
- Surface radar backscatter (σ_0),
- Surface type (categorical),
- Top height of the lowest detected hydrometeor layer,
- Local time (hour, month, and season).

Once trained, the RRM predicts the residual error of the inpainting model for any given scene. If the optimally trained RRM still can't capture the variability of the prediction residuals of SnowProfGPT (i.e., an R-sqaure of $\sim 0\%$ or even worse), then SnowProfGPT is perfectly generalizable under any condition, as there is little explorable dependency between the prediction performance of SnowProfGPT and the external conditions when/where it makes a prediction. Otherwise, while the predictions are not exact, they provide insight into both the direction (overestimation or underestimation) and approximate magnitude of model error, conditioned on footprint attributes. The reliability of the generalization assessment of SnowProfGPT through the RRM is evaluated using an unseen residual test set that was not used during RRM training.

To capture nonlinear feature interactions, we train the RRM using an XGBoost model [8] with cross-validated hyperparameter tuning [22]. We then apply a SHAP-based importance analysis [60] to assess which factors most influence the residuals:

SHAP-based analysis. We compute SHAP (SHapley Additive exPlanations) [60] values for each feature in the RRM. This provides a model-aware quantification of how much each feature contributes to residual error of SnowProfGPT. Unlike conventional feature importance metrics, SHAP values account for all interactions among features and do not rely on linear assumptions.

Modified SHAP values. Because the mean residual in the SCZ is not necessarily zero, we redefine SHAP values with respect to a zero baseline rather than the empirical mean. This re-centering allows us to interpret each SHAP value as the direction and magnitude by which a given feature value shifts the model toward overestimation or underestimation. These modified SHAP values are visualized via violin plots, enabling detailed diagnostic analysis of model behavior under specific conditions.

In contrast to correlation-based diagnostics, SHAP captures nonlinear, context-aware feature effects [60]. Applied to the RRM, this framework reveals which footprint attributes most strongly affect the model's prediction errors, and under what conditions these errors are likely to occur.

Summary. By modeling and interpreting residuals through the RRM, we evaluate:

1. Which environmental factors most strongly influence the inpainting model's performance, and
2. Under which conditions the model tends to systematically underpredict or overpredict snowfall rates.

This residual modeling framework provides a robust and interpretable method for generalization analysis, and serves as a key component in understanding the model's operational reliability and uncertainty across diverse geophysical scenarios.

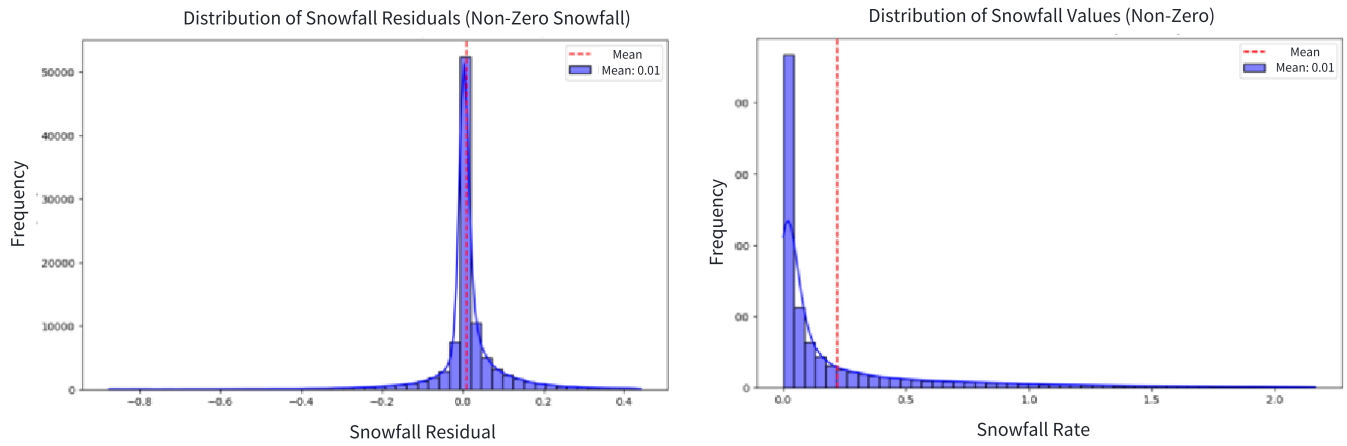

Figure 3.1: Empirical distributions sampled from a random granule of CloudSat snowfall profiles (snowfall events only). **Left:** Residual distribution of snowfall-rate differences between adjacent vertical bins, $S(R + dR) - S(R)$, approximating the first-order local gradient. **Right:** Marginal distribution of snowfall rates $S(R)$ across all valid bins. The left panel supports a Markov assumption by showing that local vertical changes in snowfall rate are approximately zero-mean and narrowly distributed.

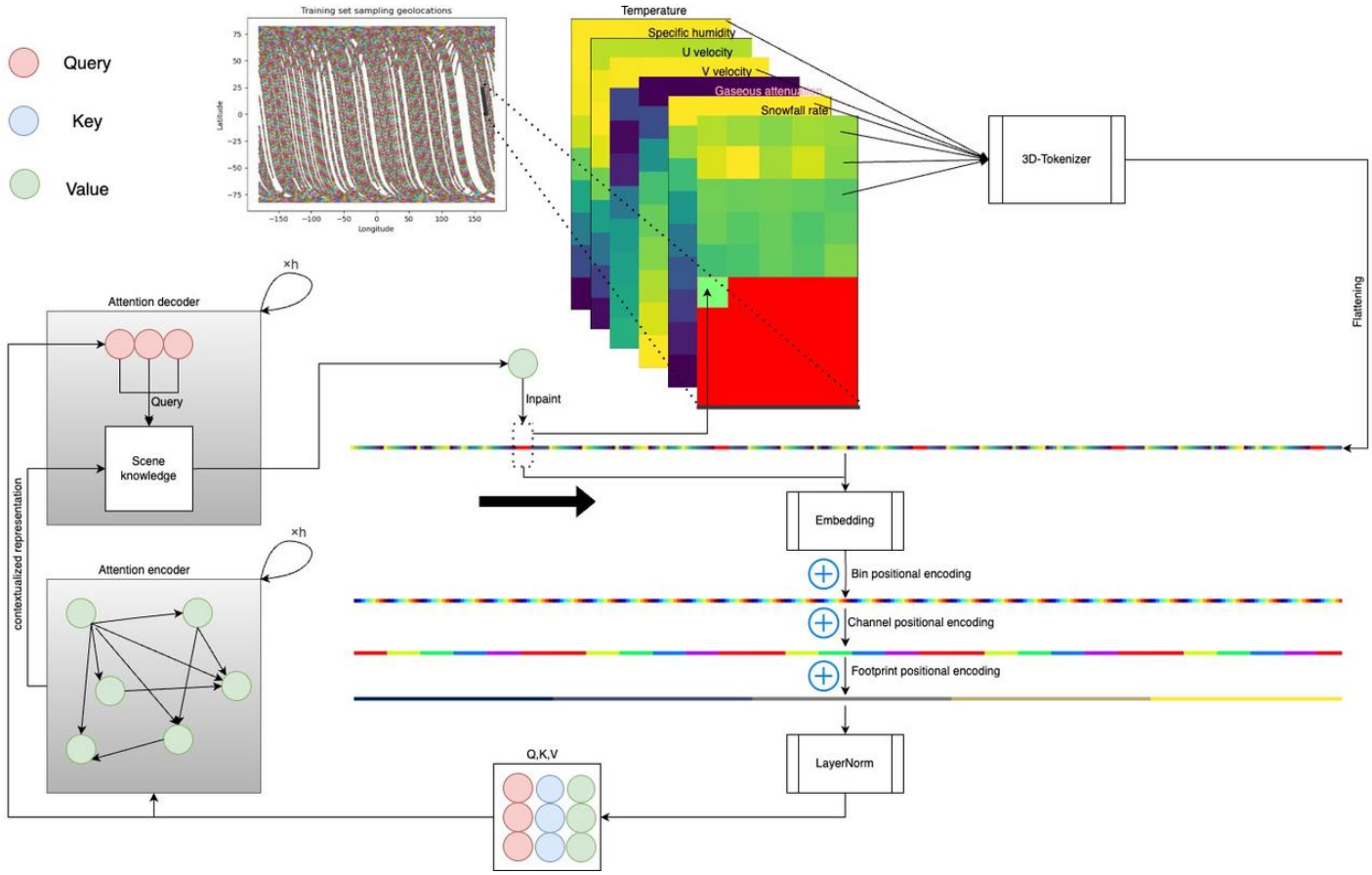

Figure 3.2: Overview of the SnowProfGPT model architecture. The diagram shows the core design components. For each scene, a 5-by-5 window of adjacent footprints is sampled from a CloudSat granule. The scene is tokenized using a 3D tokenizer along the footprint (width), variable (channel), and vertical bin dimensions. Positional encodings for width, height, and channel are added in the same order. Each token is linearly embedded into a vector space and transformed into Query, Key, and Value representations through multi-head projection. The encoder applies multi-head self-attention to compute contextualized token embeddings. The decoder uses masked self-attention to process previously predicted snowfall tokens and applies cross-attention to incorporate encoder outputs. Tokens are inpainted autoregressively from top to bottom and left to right. Each predicted token is projected back into snowfall-rate space and placed into its original location via the inverse tokenizer. The spatial window is fixed at 5-by-5. Encoder memory and decoder inputs are dynamically updated throughout decoding.

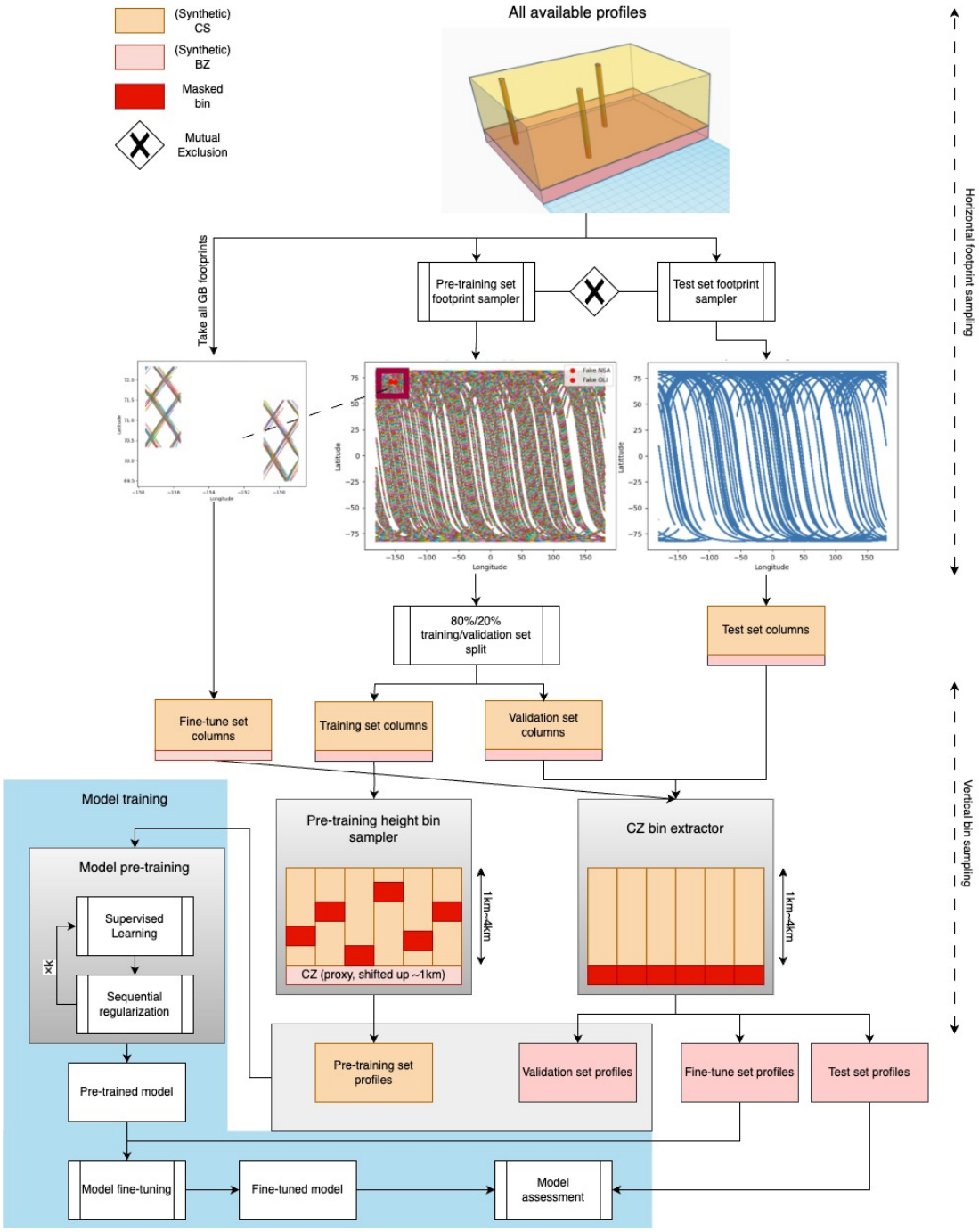

Figure 3.3: Overview of the SnowProfGPT training paradigm. Yellow regions represent well-observed vertical bins in the CloudSat-observed domain (CS), located immediately above the clutter zone (CZ). Red regions denote masked snowfall-rate bins used as autoregressive training targets. Pink regions represent the clutter zone, which is the focus of reconstruction. The top panel shows the horizontal footprint sampling scheme, where adjacent CloudSat footprints are grouped to form each scene. The middle panel shows the vertical bin sampling strategy used to define the scene depth and mask placement. The bottom panel illustrates how the scene is reassembled in three dimensions. The blue pipeline indicates the sequential pretraining and fine-tuning process used in the transfer learning scheme.

Chapter 4

Results

In this section, we present a comprehensive evaluation of SnowProfGPT, focusing on three primary objectives: (1) validating the Markov structure assumed in our inductive bias, (2) assessing the accuracy of the proposed inpainting model across the SCZ, and (3) analyzing the model's generalizability under diverse spatiophysical conditions. To benchmark performance, we compare SnowProfGPT with a simple replication strategy that estimates snowfall rates in the clutter zone (CZ) by copying the last reliable value immediately above it. This replication-based approach serves as the baseline throughout our evaluation.

The baseline is not arbitrary: it reflects the default strategy implemented in CloudSat's official DPC data products and remains one of the most widely used methods in the snowfall retrieval community. Thus, any improvement over this baseline carries meaningful operational implications. If SnowProfGPT demonstrates superior performance, it suggests that current data products relying on the replication method could be enhanced using our inpainting approach or by future models trained using the SnowProfGPT framework introduced in Section 3.

The benchmarking will be operated on the previously held-out test dataset processed by the right pipeline in Figure 3.3. The overview of the sampling distribution of the test dataset has been demonstrated in Figure 2.1 in the Data section.

4.1 Model global performance

We begin by evaluating the global accuracy of SnowProfGPT following fine-tuning, compared with the baseline replication model. Figure 4.1 summarizes three key metrics: co-

efficient of determination (R^2), mean absolute error (MAE), and bias, computed over all snowfall footprints in the SCZ in Figure 2.1.

In the left panel, SnowProfGPT demonstrates consistently higher R^2 values than the baseline across all four SCZ bins. The overall global R^2 of the inpainting model is 50%, compared to 35% for the baseline. Notably, for the lowest bin in the SCZ that is closest to the clutter zone, the inpainting model achieves an R^2 of 28%, substantially outperforming the baseline's 2%. This reflects a 26 percentage point gain in explained variance for the most challenging bin. The middle panel compares MAE across SCZ height levels. While the inpainting model performs slightly worse than the baseline in the uppermost bins (Bins 1 and 2), it yields significantly lower error in the lower bins (Bins 3 and 4). At the lowest bin (Bin 4), SnowProfGPT achieves an average error of 0.160 mm/hr, compared to 0.175 mm/hr for the baseline, which is an important improvement near the surface where accurate snowfall estimates are most critical. In the right panel, we examine bias. The baseline model exhibits a persistent underestimation across all bins, while the inpainting model tends to overestimate, particularly in the upper bins. However, SnowProfGPT substantially reduces bias in the lower bins, leading to more balanced estimates near the surface.

While the inpainting model does not fully preserve vertical smoothness in snowfall profiles, it provides markedly more accurate predictions in the SCZ, especially in bins closest to the surface that are considered most critical, resulting in a clear performance advantage over the baseline.

4.2 Generalizability analysis

We now evaluate the generalizability of SnowProfGPT by analyzing its performance under diverse spatial conditions and environmental contexts. As discussed in Section 3.3, horizontal domain shift is a critical challenge in snowfall retrieval. We begin by evaluating spatial variability in the lowest bin of the SCZ, followed by surface-type analysis, and finally perform a detailed residual-based diagnostic using the Residual Regression Model (RRM).

4.2.1 Spatial heterogeneity of SnowProfGPT predictions

Figure 4.2 shows the mean absolute error (MAE) at each geolocation for the lowest SCZ bin, computed only over footprints where snowfall was detected above the SCZ. Each value

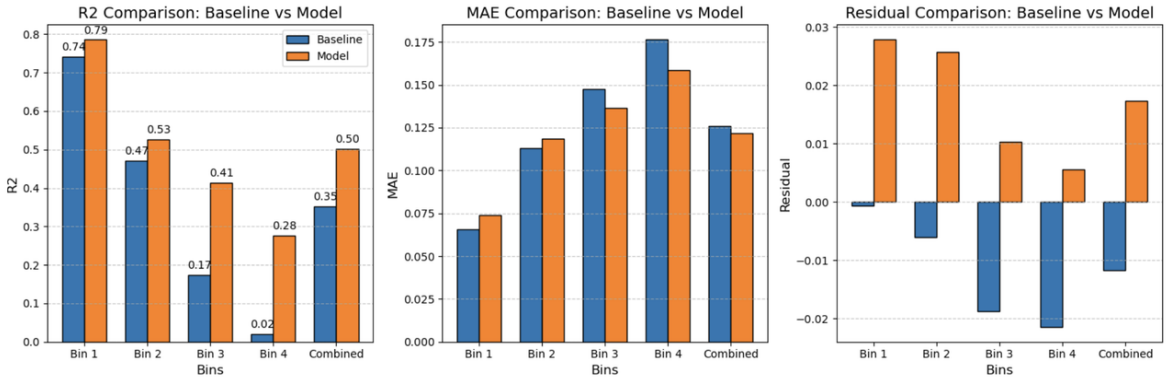

Figure 4.1: Global performance comparison between the baseline (replication) model and the proposed inpainting model (SnowProfGPT), evaluated over snowfall footprints only. **Left:** Global coefficient of determination (R^2) for each height bin in the inpainted clutter zone (SCZ). **Middle:** Global mean absolute error (MAE) in snowfall-rate prediction. **Right:** Global bias (mean signed error) relative to reference snowfall rates. "Bin 1" through "Bin 4" correspond to the four lowest bins in the SCZ, listed in descending order from the surface. "Combined" denotes aggregate metrics across all SCZ bins.

represents the MAE aggregated within a 2-km spatial window. While both models perform well over large areas, the baseline model exhibits a higher frequency of anomalous "spark patches" (i.e., localized regions of extreme error) particularly across the Arctic. In contrast, SnowProfGPT produces a smoother and more spatially coherent error field, indicating higher robustness. Although both models degrade near topographic boundaries, likely due to unresolved elevation mismatches in the DEM (as previously observed by Bennartz et al., 2019), SnowProfGPT mitigates error patterns in flatter regions and reduces non-DEM-related anomalies in polar domains.

To account for heterogeneity in local snowfall variability, we introduce a normalized error metric, Q-sigma, defined as:

$$q = \frac{\hat{y} - y}{\sigma_{\text{local}}}$$

where σ_{local} is the local standard deviation of snowfall rates estimated from the 200 nearest neighbors in space-time. This metric is conceptually similar to a pseudo–z-score and enables fairer comparisons across climatologically complex and homogeneous regions. Figure 4.3 shows that SnowProfGPT yields a more uniform Q-sigma field than the baseline replication model, particularly in polar regions. High-latitude spark patches in the baseline model are reduced or eliminated, and Q-sigma is more stable across open ocean, where snowfall vari-

ability is inherently high. However, both models show greater sensitivity to overestimation than underestimation, especially in transitional climate zones. This suggests a physical discontinuity in snowfall processes between higher altitudes and the boundary layer, where hydrometeors often dissipate before reaching the surface. An exception occurs at the highest latitudes, where the baseline replication model systematically underestimates snowfall, suggesting that snowfall can intensify at lower altitudes in extreme polar conditions.

4.2.2 SnowProfGPT Performance Across Surface Types

We next compare model performance across different surface types (Fig. 4.4). Both models perform best over ice-covered surfaces and degrade significantly over open ocean. For the lowest SCZ bin, the baseline model yields negative R^2 over ocean, indicating performance worse than a constant predictor. SnowProfGPT improves R^2 in this region to over 20%, capturing more than 30% additional variability. Performance over land surfaces is intermediate. These results confirm that the inpainting model better handles surface-type–driven meteorological differences that the baseline fails to resolve.

4.2.3 Interpretable Generalization Insights via SHAP and Residual Modeling

To complement our spatial and surface-type evaluations, we further assess the generalization behavior of SnowProfGPT using the Residual Regression Model (RRM) introduced in Section 3.5. The RRM was trained to estimate inpainting residuals based on footprint-level attributes, enabling model-aware analysis of how external conditions influence performance. The RRM achieves an R^2 of 59% on a held-out residual test set, indicating that SnowProfGPT's errors are partially predictable from external features, implying that the model is not fully generalizable under all conditions, but degrades in structured, learnable ways.

Figure 4.5 presents SHAP-based interpretation results from the RRM. In the left panel, we rank the footprint attributes by their estimated importance to residual prediction. The right panel shows modified SHAP violin plots, re-centered around zero residual, to quantify how individual feature values shift the model's predictions toward overestimation or underestimation. Across most features, the SHAP distributions are narrow and symmetric around a positive mean close to zero, indicating that SnowProfGPT is generally stable across footprint conditions. The small positive mean across most SHAP distributions reflects a consistent but slight overestimation bias in the lowest SCZ bin, consistent with the global bias observed in Fig. 4.1.

Top driver: Hydrometeor layer top height. The most influential attribute is the top height of the lowest detected hydrometeor layer. SHAP values are tightly concentrated but positive on average, suggesting that profiles with deeper (i.e., less shallow) hydrometeor columns tend to be modestly overestimated. While no strong directional trend is observed, the dominance of deeper profiles in the dataset likely drives this slight upward bias.

Latitude and horizontal domain shift. Latitude is the second most influential factor. Although its SHAP distribution is roughly symmetric, a clear right-tailed signature indicates that the model significantly overestimates snowfall in some low-latitude regions. This suggests a limitation in generalizing to equatorial climates, potentially due to sparser CloudSat sampling in these zones and the lack of explicit latitude encoding in the model inputs. At high latitudes, the model also tends to overestimate, but to a much lesser degree. These results are consistent with the spatial residual maps, where overestimation anomalies frequently occur in mid-latitude regions. Through the RRM and SHAP framework, we verify that spatial heterogeneity (especially latitude) is indeed one of the top two most influential factors affecting SnowProfGPT's performance.

Vertical prediction bin height. Bin height ranks third in influence but is significantly less important than the top two. Its SHAP distribution is symmetric, with only a slight overestimation trend at higher altitude bins. This result suggests that the model does not exhibit strong bias across vertical levels within the SCZ, but SnowProfGPT most frequently overestimates snowfall intensities at higher altitudes.

Surface type, backscatter, and wind speed. Surface type, surface radar backscatter, and surface-layer wind speed rank next in importance but are all substantially less influential than the top three predictors. Their SHAP distributions are unimodal, tightly concentrated around zero, and show no directional dependence, suggesting that SnowProfGPT is robust to variations in clutter-prone conditions and low-level meteorological forcing. Although surface type and radar backscatter are often cited as key challenges in satellite-based snowfall retrieval, the model appears to have learned stable relationships that generalize across these factors, at least within the SCZ. The success may be attributed to the model's exposure to a broad range of surface types and horizontal-wind conditions during pretraining: diversity that is preserved even after fine-tuning and likely contributes to its stability across these dimensions.

Local time features. Time-related features, including local hour, month, season, and day/night, rank among the least influential in the SHAP-based analysis. This is expected in part due to CloudSat's operational mode: during the 2016–2017 study period, the satellite operated in Daytime-Only mode, leading to a narrow distribution of sampling times with

minimal diurnal variability [49]. Consequently, the model had limited opportunity to learn meaningful patterns related to day versus night or to specific hours.

Similarly, local month and season show low influence, suggesting that SnowProfGPT is less sensitive to the time of year when a prediction is made than to other geophysical conditions, such as surface type, latitude, or hydrometeor structure. This may reflect the stronger predictive signal carried by spatial and environmental features compared to seasonal or among-month ones.

An exception: Surface elevation. While surface elevation ranks low in overall importance, its SHAP distribution shows a strong directional effect. Specifically, predictions over low-altitude surfaces are more likely to underestimate, whereas predictions over higher-elevation surfaces tend to overestimate snowfall intensity. This may reflect both genuine climatological differences (e.g., precipitation enhancement at elevation) and varying severity of surface clutter artifacts.

Summary. This SHAP-based analysis confirms two key insights:

1. SnowProfGPT's performance is not fully invariant to external footprint conditions; residuals remain partially predictable across both spatial and meteorological features.
2. Despite this, the model is highly stable along most input dimensions with slight, consistent overestimation bias, and only a few dimensions (notably latitude and surface elevation) showing stratified/directional effects.

These findings reinforce the robustness of SnowProfGPT while highlighting conditions under which performance may degrade. By modeling generalization behavior through the RRM and interpreting it via SHAP, we provide a transparent, model-aware diagnostic for understanding prediction uncertainty in real-world deployments.

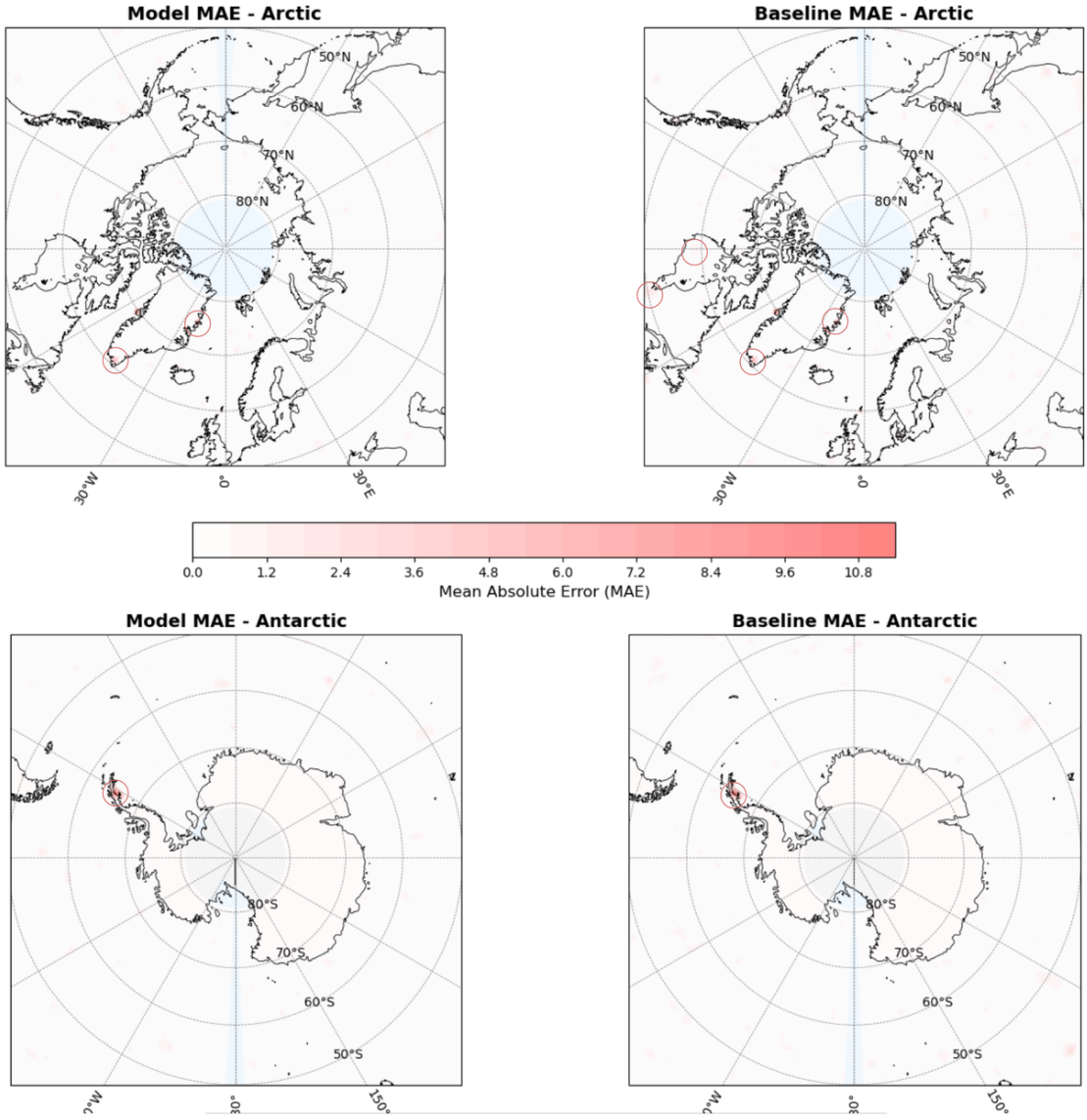

Figure 4.2: Spatial distribution of mean absolute error (MAE) for the inpainting model (SnowProfGPT) and the baseline (replication) model, evaluated over snowfall footprints in the lowest bin of the SCZ. Each point reflects the MAE within a 2-km spatial window aggregated across time. Circled anomalies ("spark patches") indicate regions where performance degrades significantly. Results shown for polar regions only.

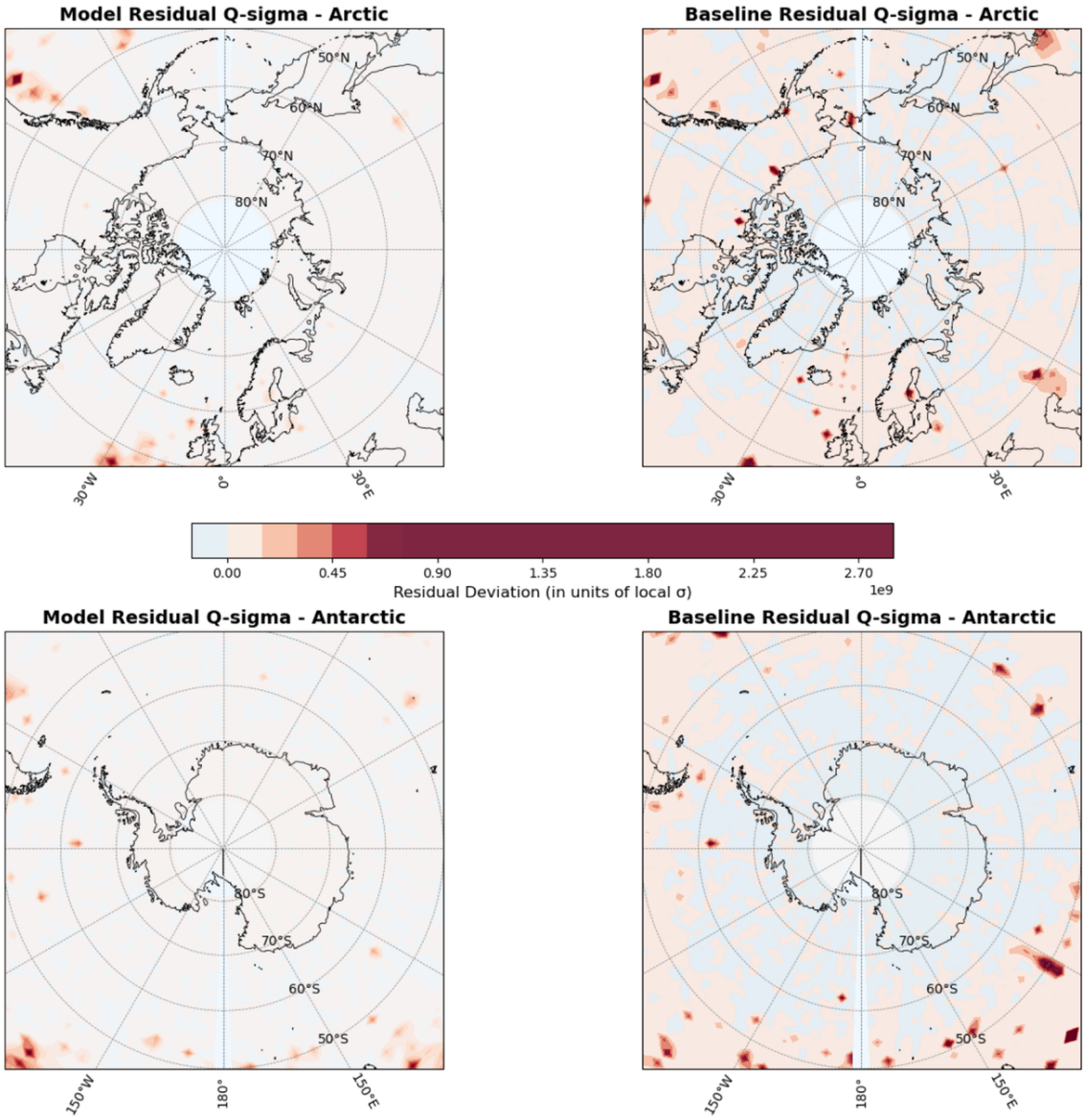

Figure 4.3: Spatial distribution of Q-sigma for the inpainting and baseline models in the lowest SCZ bin. Q-sigma normalizes residuals by local snowfall variability, highlighting error stability across geolocations. Outliers below the 1st percentile and above the 99th percentile are excluded for clarity.

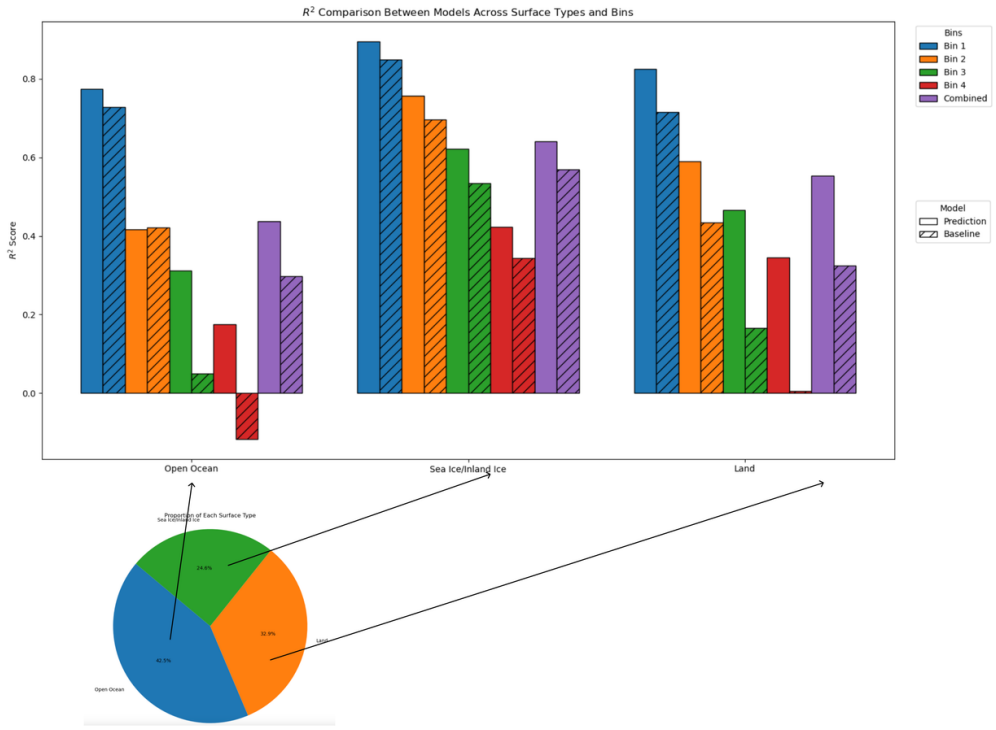

Figure 4.4: Comparison of model performance across surface types. Bars indicate R^2 for the inpainting and baseline models in the lowest SCZ bin, stratified by surface type: open ocean, ice, and land. Surface-type effects reflect both model sensitivity and underlying retrieval challenges.

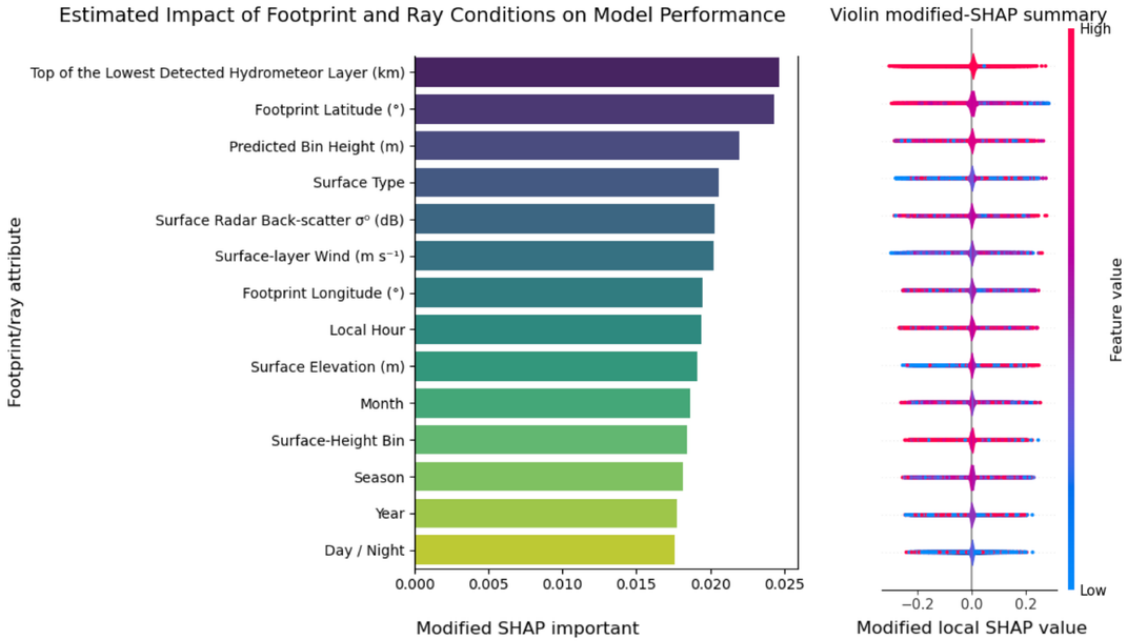

Figure 4.5: Residual sensitivity analysis using the Residual Regression Model (RRM). **Left:** SHAP-based feature importance ranking for footprint-level conditions impacting model performance, based on a residual model trained to predict inpainting errors with 59% R^2 on a held-out test set. **Right:** Adjusted SHAP violin plot showing how each attribute influences residual directionality. SHAP values are re-centered such that zero corresponds to no error, with positive (negative) values indicating a tendency toward overestimation (underestimation).

Chapter 5

Conclusions and Discussion

This study contributes to a longstanding challenge in satellite-based snowfall retrieval: reconstructing near-surface snowfall-rate retrievals within the clutter zone (CZ), where radar measurements are obscured by ground clutter and surface returns. We introduce SnowProfGPT, a transformer-based autoregressive inpainting model trained under a generative pretraining (GPT) paradigm, to reconstruct snowfall-rate profiles in the CZ globally using CloudSat Cloud Profiling Radar (CPR) observations. To our knowledge, this is the first successful application of transfer learning and autoregressive modeling for CZ snowfall reconstruction at a global scale.

To contextualize our contribution, we first review existing approaches within the CloudSat snowfall community under a statistical learning framework. We identify domain shift, especially across the horizontal spatial dimensions, as a core limitation that has remained underrepresented. We also establish a theoretical foundation for snowfall-profile reconstruction by identifying the quasi-Markov structure of vertical snowfall-rate profiles. We show that their variation can be decomposed into motion-driven effects (governed mostly by the physical process of snowfall) and microphysical effects (driven majorly by local meteorological conditions), thereby informing the model's autoregressive inductive bias, facilitating the core designs of SnowProfGPT.

To address the lack of ground truth within the CZ for model assessment and supervision, we introduce a novel proxy: the Synthetic Clutter Zone (SCZ), defined as the lowest reliable bins directly above the official CZ. SnowProfGPT is pretrained on globally distributed CloudSat observations above the SCZ and fine-tuned using a synthetic ground-based (GB) subset extracted near ARM Ka-band radar sites. This transfer learning framework combining large-scale self-supervised learning with targeted finetuning contributes to

the relaxation of domain shift across both vertical and horizontal dimensions by leveraging the massive volume of high-quality observations above the CZ for the first time. This approach may also serve as a foundation for future retrieval tasks in other geophysical domains that lack dense reference labels but require robust generalization across the full spatial and vertical extent of the atmosphere.

5.1 Key Contributions and Findings

Quantitative evaluation confirms substantial performance improvements over the current CloudSat replication baseline. SnowProfGPT improves R^2 from 35% to 50% across all SCZ bins and from 2% to 28% in the most critical lowest bins, capturing 26% more variability of the snowfall intensities than the baseline replication method in those bins. It reduces global MAE in the lowest bin by 0.015 mm/hr and debiases surface-near snowfall estimates, which are systematically underestimated by the baseline model. These gains are especially significant over open oceans, where SnowProfGPT recovers over 30% additional variance compared to the baseline.

Generalizability analysis shows that SnowProfGPT performs consistently well across surface types, particularly over ice and land, and is significantly more stable than the baseline in polar regions. A Residual Regression Model (RRM), combined with SHAP-based interpretation, reveals that most input features, such as wind speed, surface reflectivity, and hydrometeor structure, have only minor, symmetric impacts on model bias. Only a few features, such as latitude and surface elevation, show stratified or directional dependencies, indicating predictable but limited degradation under specific conditions.

Together, these results demonstrate that SnowProfGPT achieves accurate, stable, and interpretable near-surface snowfall reconstruction despite the lack of direct ground-truth data in the CZ. By combining a physically grounded inductive bias with a scalable transfer learning paradigm, it outperforms operational baselines and establishes a generalizable framework for satellite snowfall retrieval under uncertainty.

5.2 Limitations and Future Directions

While SnowProfGPT represents a significant advancement, several limitations remain and offer directions for future research:

First, the SCZ used for training and evaluation is synthetic, and its effectiveness relies on the assumption that performance in the SCZ approximates that in the true clutter zone. Although designed as a proxy, further validation using real-world GB-CZ observations is needed to confirm that the transferability seen in the SCZ generalizes to the genuine CZ.

Second, our study uses a single epoch of CloudSat data spanning only 18 months (2016–2017), during which the satellite operated in Daytime-Only mode and remained in the A-Train [49, 6]. This time span excludes among-year and orbital variability present in other phases of the mission. Future work should incorporate the full 17-year CloudSat record and consider other missions such as EarthCARE [62], which will offer similarly structured CPR observations, to further pretrain and evaluate SnowProfGPT across observational regimes.

Third, the model architecture, while effective, is computationally expensive due to the quadratic cost of full attention. While this complexity may help enable the emergent generalization seen in large autoregressive models, lightweight variants based on efficient attention (e.g., linear attention [29]) or distilled architectures [21] could provide similar performance with reduced resource demands. Nonetheless, maintaining a large, generalist model such as SnowProfGPT offers long-term benefits: it behaves as an adaptive "expert" system that can be fine-tuned to support targeted tasks and missions, and can be upgraded and maintained through long-term sequential training.

Lastly, while we focus here on snowfall retrieval, the SnowProfGPT framework can be extended to other atmospheric retrieval tasks involving spatially structured, partially observed profiles, such as cloud microphysical or optical properties in boundary-layer clouds. With its ability to integrate diverse meteorological inputs into a shared latent space, SnowProfGPT can absorb additional data modalities and bypass traditional collocation pipelines through learned embeddings. This opens new possibilities for end-to-end learning across multi-source geophysical datasets, simplifying and reducing information loss through traditional manual data collocation or assimilation (e.g., over/downsampling) for data-standard mismatching (e.g., spatio-temporal coverages/resolutions).

5.3 Closing Remarks

SnowProfGPT offers a scalable, accurate, and interpretable solution to the problem of reconstructing snowfall rates in the radar clutter zone. It establishes a path forward for training geophysical models under data sparsity, validates the role of large-scale pretraining for atmospheric retrieval, and sets the foundation for future research in cluttered observational domains.

By bridging autoregressive modeling, transfer learning, and atmospheric remote sensing, SnowProfGPT contributes not just a new snowfall product, but a modeling paradigm capable of transforming how we reconstruct and interpret meteorological scenes from space.

References

[1] ECMWF-AUX | CloudSat DPC.

[2] torch.optim — PyTorch 2.6 documentation.

[3] A Users Guide to CloudSat Standard Data Products | Request PDF.

[4] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[5] Ralf Bennartz, Frank Fell, Claire Pettersen, Matthew D. Shupe, and Dirk Schuettemeyer. Spatial and temporal variability of snowfall over Greenland from CloudSat observations. *Atmospheric Chemistry and Physics*, 19(12):8101–8121, June 2019. Publisher: Copernicus GmbH.

[6] Barbara Manganis Braun, Theodore H. Sweetser, Clifford Graham, and Joseph Bartsch. Cloudsat's a-train exit and the formation of the c-train: An orbital dynamics perspective. In *2019 IEEE Aerospace Conference*, pages 1–10, 2019.

[7] Richard L Burden and J Douglas Faires. Numerical analysis, brooks, 1997.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[9] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[10] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[11] Josefino C Comiso and Fumihiko Nishio. Trends in the sea ice cover using enhanced and compatible amsr-e, ssm/i, and smmr data. *Journal of Geophysical Research: Oceans*, 113(C2), 2008.

[12] Christian Darken, Joseph Chang, John Moody, et al. Learning rate schedules for faster stochastic gradient search. In *Neural networks for signal processing*, volume 2, pages 3–12. Citeseer, 1992.

[13] Min Deng, Gerald. G. Mace, Zhien Wang, and Elizabeth Berry. CloudSat 2C-ICE product update with a new Z_e parameterization in lidar-only region. *Journal of Geophysical Research: Atmospheres*, 120(23), December 2015.

[14] Min Deng, Gerald G. Mace, Zhien Wang, and R. Paul Lawson. Evaluation of Several A-Train Ice Cloud Retrieval Products with In Situ Measurements Collected during the SPARTICUS Campaign. *Journal of Applied Meteorology and Climatology*, 52(4):1014–1030, April 2013.

[15] Min Deng, Gerald G. Mace, Zhien Wang, and Hajime Okamoto. Tropical Composition, Cloud and Climate Coupling Experiment validation for cirrus cloud profiling retrieval using CloudSat radar and CALIPSO lidar. *Journal of Geophysical Research: Atmospheres*, 115(D10):2009JD013104, May 2010.

[16] Alessandro Di Bucchianico. Coefficient of determination (r 2). *Encyclopedia of statistics in quality and reliability*, 2008.

[17] L. Edel, C. Claud, C. Genthon, C. Palerme, N. Wood, T. L'Ecuyer, and D. Bromwich. Arctic snowfall from cloudsat observations and reanalyses. *Journal of Climate*, 33(6):2093 – 2109, 2020.

[18] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image Inpainting: A Review. *Neural Processing Letters*, 51(2):2007–2028, April 2020.

[19] Veronika Eyring, William D Collins, Pierre Gentine, Elizabeth A Barnes, Marcelo Barreiro, Tom Beucler, Marc Bocquet, Christopher S Bretherton, Hannah M Christensen, Katherine Dagon, et al. Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9):916–928, 2024.

[20] Yang Feng, Shuhao Gu, Dengji Guo, Zhengxin Yang, and Chenze Shao. Guiding teacher forcing with seer forcing for neural machine translation. *arXiv preprint arXiv:2106.06751*, 2021.

[21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[22] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[23] John M. Haynes, Tristan S. L'Ecuyer, Graeme L. Stephens, Steven D. Miller, Cristian Mitrescu, Norman B. Wood, and Simone Tanelli. Rainfall retrieval over the ocean with spaceborne W-band radar. *Journal of Geophysical Research: Atmospheres*, 114(D8):2008JD009973, April 2009.

[24] Timothy O Hodson. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 2022:1–10, 2022.

[25] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[26] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[27] Jinrui Jing, Qian Li, Xuan Peng, Qiang Ma, and Shaoen Tang. Hprnn: A hierarchical sequence prediction model for long-term weather radar echo extrapolation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4142–4146, 2020.

[28] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048, 2024.

[29] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[30] Fraser King, George Duffy, Lisa Milani, Christopher G. Fletcher, Claire Pettersen, and Kerstin Ebell. DeepPrecip: a deep neural network for precipitation retrievals. *Atmospheric Measurement Techniques*, 15(20):6035–6050, October 2022. Publisher: Copernicus GmbH.

[31] Fraser King and Christopher G Fletcher. Using cloudsat-cpr retrievals to estimate snow accumulation in the canadian arctic. *Earth and Space Science*, 7(2):e2019EA000776, 2020.

[32] Fraser King, Claire Pettersen, Christopher G. Fletcher, and Andrew Geiss. Development of a full-scale connected u-net for reflectivity inpainting in spaceborne radar blind zones. *Artificial Intelligence for the Earth Systems*, 3(2):e230063, 2024.

[33] Rithwik Kodamana and Christopher G. Fletcher. Validation of CloudSat-CPR Derived Precipitation Occurrence and Phase Estimates across Canada. *Atmosphere*, 12(3):295, March 2021. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[34] P. Kollias, B. Puigdomènech Treserras, and A. Protat. Calibration of the 2007–2017 record of atmospheric radiation measurements cloud radar observations using cloudsat. *Atmospheric Measurement Techniques*, 12(9):4949–4964, 2019.

[35] Wouter M. Kouw and Marco Loog. An introduction to domain adaptation and transfer learning, January 2019. arXiv:1812.11806 [cs].

[36] Mark S. Kulie and Ralf Bennartz. Utilizing spaceborne radars to retrieve dry snowfall. *Journal of Applied Meteorology and Climatology*, 48(12):2564–2580, 2009. Publisher: American Meteorological Society.

[37] Katia Lamer, Pavlos Kollias, Alessandro Battaglia, and Simon Preval. Mind the gap – Part 1: Accurately locating warm marine boundary layer clouds and precipitation using spaceborne radars. *Atmospheric Measurement Techniques*, 13(5):2363–2379, May 2020. Publisher: Copernicus GmbH.

[38] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28, 1988.

[39] Randall J LeVeque and Randall J Leveque. *Numerical methods for conservation laws*, volume 132. Springer, 1992.

[40] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[41] Guosheng Liu. Deriving snow cloud characteristics from CloudSat observations. *Journal of Geophysical Research: Atmospheres*, 113(D8), 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2007JD009766.

[42] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

[43] Maximilian Maahn, Clara Burgard, Susanne Crewell, Irina V. Gorodetskaya, Stefan Kneifel, Stef Lhermitte, Kristof Van Tricht, and Nicole P. M. van Lipzig. How does the spaceborne radar blind zone affect derived surface snowfall statistics in polar regions? *Journal of Geophysical Research: Atmospheres*, 119(24):13,604–13,620, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2014JD022079.

[44] Roger Marchand. Product: 2B-GEOPROF Product Version: P1_r05 Document Revision: 0 Date: 18 June 2018.

[45] Roger Marchand, Gerald G. Mace, Thomas Ackerman, and Graeme Stephens. Hydrometeor detection using cloudsat—an earth-orbiting 94-ghz cloud radar. *Journal of Atmospheric and Oceanic Technology*, 25(4):519 – 533, 2008.

[46] Roger Marchand, Gerald G. Mace, Thomas Ackerman, and Graeme Stephens. Hydrometeor Detection Using Cloudsat—An Earth-Orbiting 94-GHz Cloud Radar. *Journal of Atmospheric and Oceanic Technology*, 25(4):519–533, April 2008.

[47] Eva Mekis, Norman Donaldson, Janti Reid, Alex Zucconi, Jeffery Hoover, Qian Li, Rodica Nitu, and Stella Melo. An overview of surface-based precipitation observations at environment and climate change canada. *Atmosphere-Ocean*, 56(2):71–95, 2018.

[48] Lisa Milani, Mark S. Kulie, Daniele Casella, Stefano Dietrich, Tristan S. L'Ecuyer, Giulia Panegrossi, Federico Porcù, Paolo Sanò, and Norman B. Wood. CloudSat snowfall estimates over Antarctica and the Southern Ocean: An assessment of independent retrieval methodologies and multi-year snowfall analysis. *Atmospheric Research*, 213:121–135, November 2018.

[49] Lisa Milani and Norman B. Wood. Biases in cloudsat falling snow estimates resulting from daylight-only operations. *Remote Sensing*, 13(11), 2021.

[50] Hugh Morrison, Marcus van Lier-Walqui, Ann M Fridlind, Wojciech W Grabowski, Jerry Y Harrington, Corinna Hoose, Alexei Korolev, Matthew R Kumjian, Jason A Milbrandt, Hanna Pawlowska, et al. Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of advances in modeling earth systems*, 12(8):e2019MS001689, 2020.

[51] Anders Persson and Federico Grazzini. User guide to ecmwf forecast products. *Meteorological Bulletin*, 3(2), 2007.

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

[53] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[54] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation. pages 3752–3761, 2018.

[55] Noah A Smith. Contextual word representations: A contextual introduction. *arXiv preprint arXiv:1902.06006*, 2019.

[56] Simone Tanelli, Stephen L. Durden, Eastwood Im, Kyung S. Pak, Dale G. Reinke, Philip Partain, John M. Haynes, and Roger T. Marchand. CloudSat's Cloud Profiling Radar After Two Years in Orbit: Performance, Calibration, and Processing. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3560–3573, November 2008.

[57] Simone Tanelli, Kyung Pak, Stephen Durden, and Eastwood Im. Reducing Surface Clutter in Cloud Profiling Radar Data. Technical Report NPO-44873, December 2008. NTRS Author Affiliations: California Inst. of Tech. NTRS Document ID: 20080048034 NTRS Research Center: Jet Propulsion Laboratory (JPL).

[58] Douglas Vandemark, Mark Bourassa, Shu-Hua Chen, Heidi Dierssen, Paul Houser, Lyatt Jaegle, Carol Johnson, Guosheng Liu, George Mount, David Mitchell, Jun Wang, Diane Wickland, and Curtis Woodcock. NASA Earth Science Senior Review Subcommittee Report - 2017. Technical report, NASA Earth Science Advisory Committee, June 2017. Submitted to the NASA Earth Science Advisory Committee.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] Huanjing Wang, Qianxin Liang, John T Hancock, and Taghi M Khoshgoftaar. Feature selection strategies: a comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11(1):44, 2024.

[61] Yu Wang, Yalei You, and Mark Kulie. Global Virga Precipitation Distribution Derived From Three Spaceborne Radars and Its Contribution to the False Radiometer Precipitation Detection. *Geophysical Research Letters*, 45(9):4446–4455, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL077891.

[62] Tobias Wehr, Takuji Kubota, Georgios Tzeremes, Kotska Wallace, Hirotaka Nakatsuka, Yuichi Ohno, Rob Koopman, Stephanie Rusli, Maki Kikuchi, Michael Eisinger, et al. The earthcare mission–science and system overview. *Atmospheric Measurement Techniques*, 16(15):3581–3608, 2023.

[63] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[64] Mona M. Witkowski, Deborah Vane, and Thomas Livermore. CloudSat - Life in Daylight Only Operations (DO-Op). In *2018 SpaceOps Conference*, Marseille, France, May 2018. American Institute of Aeronautics and Astronautics.

[65] N. B. Wood, T. S. L'Ecuyer, F. L. Bliven, and G. L. Stephens. Characterization of video disdrometer uncertainties and impacts on estimates of snowfall rate and radar reflectivity. *Atmospheric Measurement Techniques*, 6(12):3635–3648, December 2013.

[66] Norman B. Wood, Tristan S. L'Ecuyer, Andrew J. Heymsfield, Graeme L. Stephens, David R. Hudak, and Peter Rodriguez. Estimating snow microphysical properties using collocated multisensor observations. *Journal of Geophysical Research: Atmospheres*, 119(14):8941–8961, July 2014.

[67] Zeke Xie, Fengxiang He, Shaopeng Fu, Issei Sato, Dacheng Tao, and Masashi Sugiyama. Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting. *Neural computation*, 33(8):2163–2192, 2021.

[68] Douglas C Youvan. Quantum-holographic self-attention: A unified framework for emergent intelligence in ai. 2025.