The Floor Is Not the Limit: Representing Physics-Aligned Weak LiDAR Returns

Tianchi Xu
Peking University
School of Earth and Space Sciences
Beijing, China
tianchixu@stu.pku.edu.cn

Yuanpeng He*
Peking University
School of Computer Science
Beijing, China
heyuanpeng@stu.pku.edu.cn

Jintao Xu
Peking University
Department of Atmospheric and Oceanic Sciences, School of Physics
Beijing, China
jintaox@zju.edu.cn

Kang He
Peking University
Institute of Remote Sensing and Geographic Information System
Beijing, China
kang.he9486@gmail.com

Linyu Li
Peking University
School of Computer Science
Beijing, China
linyuli@stu.pku.edu.cn

Huanyao Zhang
Peking University
School of Computer Science
Beijing, China
huanyao.zhy@gmail.com

Dongming Jin
Peking University
School of Computer Science
Beijing, China
dmjin@stu.pku.edu.cn

Bohan Zhou
The Chinese University of Hong Kong
Department of Computer Science and Engineering
Hong Kong, China
zhoubh@link.cuhk.edu.hk

Wenpin Jiao
Peking University
School of Computer Science
Beijing, China
jwp@sei.pku.edu.cn

Zhi Jin
Peking University
School of Computer Science
Beijing, China
zhijin@pku.edu.cn

Abstract

Vertically resolved backscatter reveals cloud and aerosol layers and supports radiative analysis. Elastic lidar measures attenuated backscatter, and pixel-aligned unattenuated profiles are rarely available for training. Clipping weak returns to one floor value discards attenuation-related variation. To address these gaps, **AtmosBench** renders the same atmospheres with and without attenuation: 36,288 paired curtains at 355 and 532 nm from 48 globally distributed events over four years. **Physical Subfloor Deficit Coordinate (PhysicalSDC)** is a single-wavelength retrieval method trained on these pairs. Its parameter-free encoding uses logarithmic distance below the floor, following the lidar equation. On held-out events, mean absolute log error falls 12.45% relative to a raw-logarithm control. Without fine-tuning, log error against operational cloud profiles across 1,270 nighttime orbits drops 6.15% relative to clipping and 3.11% relative to an equal-capacity control, with comparable EarthCARE radar agreement. On airborne observations, layer-base error falls 27.29% relative to the raw-logarithm control, improving layer tracing under attenuation. AtmosBench and code are available at https://github.com/TianchiXu1/AtmosBench.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Earth and atmospheric sciences**.

*Corresponding author. Email: heyuanpeng@stu.pku.edu.cn.

KDD '27, San Jose, CA, USA
2027.

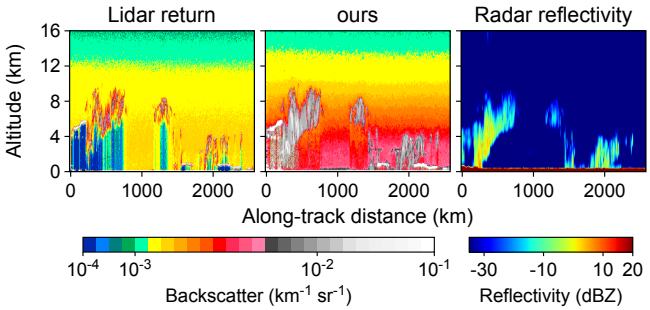


Figure 1: Retrieval beneath cloud. On EarthCARE orbit 02963C, co-orbiting radar provides a structural reference for the lidar-based estimate. The estimate shows structure beneath the cloud deck where the measured lidar return weakens. The radar panel uses a linear reflectivity scale, and Appendix A.1 gives the discrete lidar backscatter intervals.

Keywords

spaceborne lidar, atmospheric backscatter, inverse problems, scientific machine learning

ACM Reference Format:
Tianchi Xu, Yuanpeng He, Jintao Xu, Kang He, Linyu Li, Huanyao Zhang, Dongming Jin, Bohan Zhou, Wenpin Jiao, and Zhi Jin. 2027. The Floor Is Not the Limit: Representing Physics-Aligned Weak LiDAR Returns. In *Proceedings of 33rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '27).* ACM, New York, NY, USA, 12 pages.

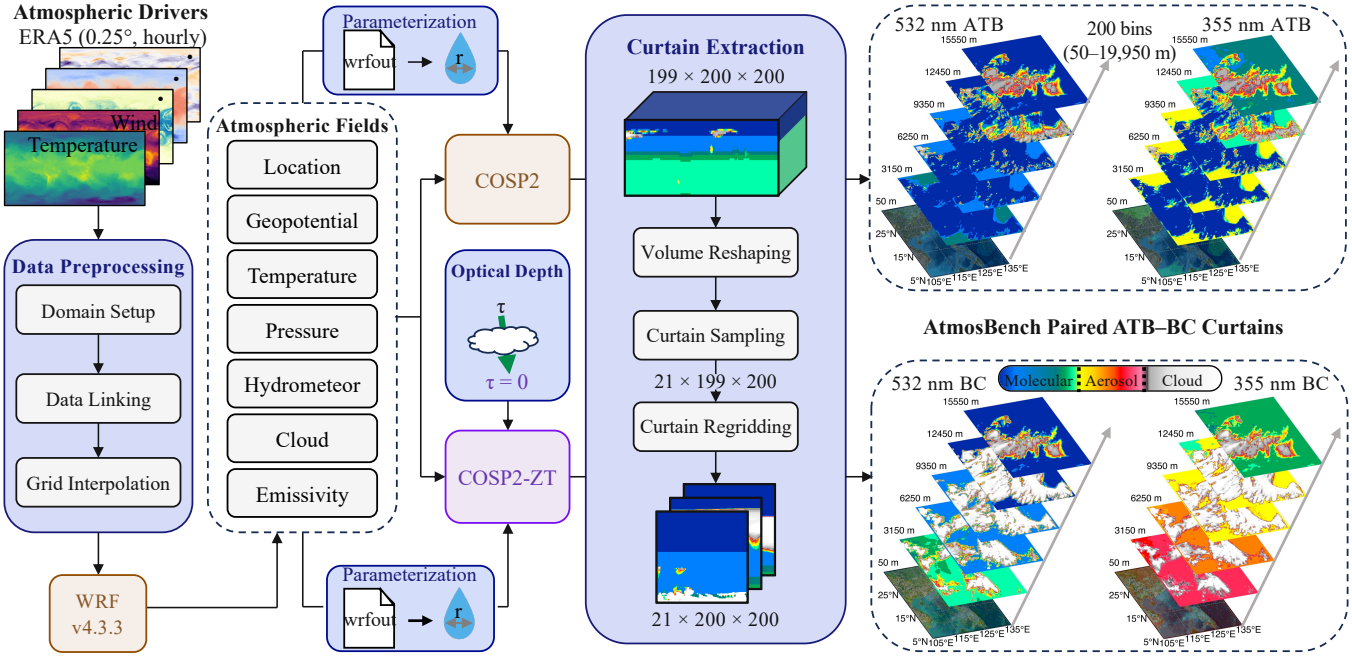


Figure 2: AtmosBench data generation. Two renderings of the same atmosphere produce pixel-aligned attenuated returns and unattenuated backscatter at 355 and 532 nm. The layered views illustrate the two renderings. The square thumbnails show representative 532 nm backscatter curtains.

1 Introduction

Clouds and aerosols regulate atmospheric radiation and precipitation, and these processes depend on vertical structure that passive imagery cannot fully resolve [13]. Spaceborne light detection and ranging (lidar) supplies height-resolved profiles along the satellite track. The retrieved backscatter coefficient (BC) supports cloud and aerosol typing [35], optical-property inversion [37], and analysis of atmospheric cloud layers [12].

Elastic lidar records attenuated total backscatter (ATB), in which local scattering is reduced by two-way transmission loss above each range bin. Classical inversions address this ambiguity through a prescribed lidar ratio [17] or a boundary condition [8]. Operational products add calibration, molecular profiles, and scene-dependent retrieval logic [36]. A learned single-wavelength retrieval faces the same observational gap: a real orbit supplies ATB, while the pixel-aligned unattenuated BC of that atmosphere is rarely available. In the Earth Cloud Aerosol and Radiation Explorer (EarthCARE) scene in Figure 1, the lidar return weakens sharply below the cloud deck. Co-orbiting radar provides a reference for hydrometeor occurrence in the same region [32].

Strong attenuation can push returns below the nominal signal floor even where layers are present. Level-1 lidar products can retain finite measurements in this regime. A single amplitude coordinate mixes weak returns with more reliable measurements. Assigning the entire regime one floor value suppresses that variation, including information related to attenuation. The lidar equation suggests a separate coordinate: on a logarithmic scale, multiplicative attenuation appears as an additive displacement.

We construct **AtmosBench** to supply aligned inputs and targets (Figure 2). The fifth-generation atmospheric reanalysis (ERA5) supplies globally consistent forcing [11], and the Weather Research and Forecasting (WRF) model resolves 48 regional events over four years [29]. The Cloud Feedback Model Intercomparison Project Observational Simulator Package (COSP) renders each state with standard attenuation and at zero optical depth [2, 30]. The two renderings hold clouds, aerosols, and thermodynamic fields fixed. This supplies a target even for layers whose returns are strongly attenuated and avoids registration between separate data products. The two wavelengths and event-level splits let us train retrieval models and compare representations against the same targets on held-out atmospheres.

We introduce **Physical Subfloor Deficit Coordinate (PhysicalSDC)**, a single-wavelength backscatter retrieval method. Its parameter-free *Physics-Aligned Encoding* retains the conventional above-floor coordinate and appends the logarithmic distance below the floor. An AdaIR-derived multiscale network [3] combines these coordinates with the surrounding profile to estimate BC. The matched controls separate the value of retaining weak variation from the value of how it is represented. We test those comparisons on reserved AtmosBench events and across networks with point and probabilistic outputs. We then apply the trained models without fine-tuning to satellite and airborne observations to evaluate backscatter-product agreement, lidar–radar structure, and layer-boundary retrieval. The radar and airborne reference products enter only in evaluation. The observed input remains the attenuated return at one wavelength.

2 Related Work

2.1 Atmospheric Lidar Retrieval

The elastic-lidar equation couples local backscatter to cumulative optical depth. Klett [17] and Fernald [8] use lidar-ratio or boundary assumptions. Retrievals from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) draw on calibration, layer detection, and quality control [35–37]. EarthCARE combines a 355-nm Atmospheric Lidar (ATLID) and a 94-GHz Cloud Profiling Radar (CPR) [13, 32]. Its Level-1 products support single- and multi-sensor retrievals and radiative closure [1, 5, 6, 18, 24, 27]. PhysicalSDC estimates BC in one forward pass from ATB using relationships learned on AtmosBench.

2.2 Paired Atmospheric Rendering

ERA5 supplies atmospheric forcing [11], WRF resolves regional atmospheric evolution [29], and COSP maps model states to satellite-like observables [2, 30]. PRE-Net uses multimodal knowledge expansion for precipitation retrieval [31]. AtmosBench uses paired COSP renderings to supply pixel-aligned targets for evaluating single-wavelength input representations.

2.3 Physics-Aligned Representations and Validity Masks

Physical structure can enter scientific machine learning through data, representation, objective, or architecture [10, 15, 34]. PhysicalSDC uses the lidar equation to design an input coordinate for finite below-floor measurements. The input-validity mask distinguishes these measurements from non-finite and fill cells, following mask-aware restoration [19].

2.4 Restoration Models and Vertical Retrieval

Restormer [38] provides a multiscale Transformer architecture for image restoration. PromptIR [25], AdaIR [3], and DATPRL [4] are evaluated restoration baselines. Recent vertical-retrieval systems such as C3DIR evaluate cloud occurrence and boundary geometry against EarthCARE products [33] using passive imagery, forecast fields, and surface elevation. Prior deep learning studies for spaceborne lidar have focused on signal denoising, feature detection, and cloud-property retrieval [12, 28]. Our models use single-wavelength lidar ATB and fixed spatial coordinates. Our convolutional baselines use the U-Net encoder–decoder family [26], with a Gaussian variant adding an input-dependent variance head [16].

3 AtmosBench

3.1 Task Definition

For wavelength $\lambda \in \{355, 532\}$ nm, each sample is a height-by-track curtain pairing an ATB array $A_\lambda \in \mathbb{R}^{200\times200}$ with its zero-attenuation BC target $B_\lambda \in \mathbb{R}^{200\times200}$. Rows span 200 vertical bins at 100-m spacing from 0.05 to 19.95 km, and columns span approximately 200 km along track at 1-km spacing. We learn the mapping

$$\widehat{B}_\lambda = f_\theta(A_\lambda, M_A, H, X, \lambda), \tag{1}$$

Here M_A marks valid ATB cells, and H and X give the height and along-track coordinates.

Each retrieval uses ATB from one wavelength. The representations receive the same validity and position fields and the same wavelength indicator. This isolates how the measured return is encoded in the comparisons in Section 4.

3.2 Paired Data Generation

Figure 2 outlines the AtmosBench data generation. Each of the 48 regional events starts from hourly ERA5 pressure-level and single-level fields on their native 0.25° grid over a $10° \times 10°$ domain [11], supplying lateral boundary conditions every 6 h. WRF version 4.3.3 [29] uses a 221×221 Mercator domain at 1-km horizontal spacing with 200 vertical interfaces and a model top of 50 hPa. The first 12 h of a 30-hour run are spin-up, leaving 18 hourly states.

Events span 2022–2025 across a 4×12 geographic grid covering four latitude bands between 60°S and 60°N and twelve 30° longitude sectors. The year–month pairs map one-to-one onto the grid, with one event per month and twelve per season. Initialization uses synoptic hours. Within selected sectors, an ERA5 relative- and specific-humidity score refines the WRF domain center. The production manifest records the date, initialization hour, and final domain center for each event, so both renderings can use the same event specification.

AtmosBench uses the COSP2 satellite lidar simulator [2, 30] from the `COSPv2.0-2.1.7` source directory. We trim ten boundary cells on each side of the WRF domain before rendering. For each atmospheric state we run COSP2 with standard and zero optical depth, holding clouds, aerosols, and thermodynamics fixed. The standard branch applies path transmission to local backscatter. The zero-optical-depth branch returns the same backscatter before attenuation. Thus the BC target can retain local layer structure even under strong path attenuation.

Both renderings use the same interpolation and track coordinates. The event manifest selects the along-track direction on the horizontal tile, and each hourly state yields 21 curtains per wavelength. AtmosBench contains 36,288 paired curtains in total. An ATB return and its target share the same atmospheric state, wavelength, height and along-track position. The same-input rendering check in Appendix A verifies the alignment and the expected ordering of ATB and BC. This paired construction also supports an attenuation-stratified evaluation using the corresponding BC after inference (Section 7.1).

3.3 Event-Level Splitting and Evaluation

Curtains from the same WRF event share spatial and temporal structure, so we split by event. The training set contains 40 events (30,240 curtain pairs). Validation and reserved test each contain four events (3,024 pairs per split). The validation events span all four latitude bands and determine checkpoint selection. Both wavelengths and all times from an event stay in the same partition. Over valid cells, we report mean absolute error (MAE) in log backscatter and in physical BC units, and root mean square error (RMSE) in physical units. We pool valid cells within each event and wavelength before averaging the resulting scores. This gives each event–wavelength combination equal weight even when valid-pixel counts differ.

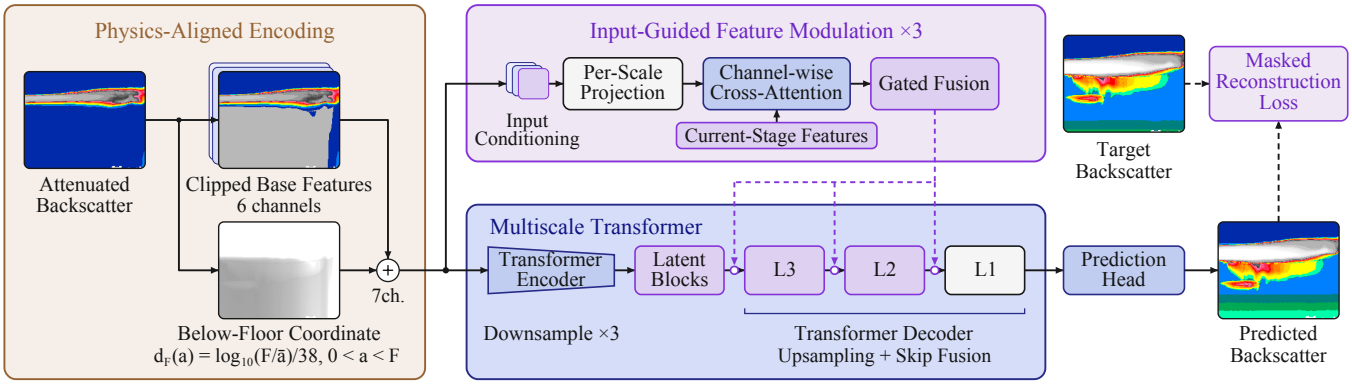


Figure 3: PhysicalSDC architecture. A parameter-free coordinate for weak returns augments the input to a multiscale retrieval network. The paired backscatter target supervises the prediction through the masked reconstruction loss.

3.4 Valid Inputs and Targets

AtmosBench retains finite below-floor ATB. The nominal signal floor defines the encoding regime, while validity is determined from ATB alone. Supervision also requires finite, positive, non-fill BC:

$$M_A = \mathbf{1}\{\text{finite}(A),\ |A| < 10^{30}\},$$
$$M_{\text{sup}} = M_A \wedge \text{finite}(B) \wedge (|B| < 10^{30}) \wedge (B > 0). \tag{2}$$

Only M_A is provided as an input. We use M_{sup} for loss and evaluation, retaining matched supervision for 98.92% of the corpus cells. Finite negative numerical-noise cells remain input-valid. RawLog and floor-collapsed representations map them to their respective guards, while Asinh preserves their sign.

4 PhysicalSDC

4.1 Input Representations

Let a denote an ATB return in $\text{m}^{-1}\,\text{sr}^{-1}$ and $F = 10^{-7}\,\text{m}^{-1}\,\text{sr}^{-1}$ denote the nominal signal floor. The shared six-channel input is

$$[\, r(a),\ O(A),\ M_A,\ H,\ X,\ \lambda\,], \tag{3}$$

where $r(a)$ is the encoded ATB coordinate, $O(A)$ is a deterministic ATB-only field constructed from train-only statistics (Appendix B), M_A is the input-validity mask, and H, X, and λ denote normalized height, along-track position, and wavelength. The input representations differ in $r(a)$ and the additional channel: **Clip** applies $\log_{10}\max(a, F)$ and collapses all returns below F to one value; **RawLog** applies a single log coordinate down to the numerical guard 10^{-45}; **Asinh** preserves finite signed values through an inverse hyperbolic sine transform; **Null** appends a copy of Clip's normalized ATB channel, constant below the nominal floor, matching the seven-channel input size and parameter count of PhysicalSDC.

The observability field summarizes smoothed log ATB and accumulates evidence along the curtain from the top down. Keeping this field, the validity mask, and the coordinates fixed gives each model the same context for interpreting weak returns. Clip and Null then test the effect of collapsing below-floor variation while preserving that context.

4.2 Physics-Aligned Encoding

Physics-Aligned Encoding pairs the clipped log amplitude $r(a)$ with a Subfloor Deficit Coordinate (SDC). For finite positive returns below the floor, the SDC measures logarithmic distance from it:

$$d_F(a) = \begin{cases} \text{clip}\left(\dfrac{\log_{10} F - \log_{10} \bar{a}}{38},\ 0,\ 1\right), & 0 < a < F, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\bar{a} = \text{clip}(a, 10^{-45}, F)$. The denominator spans the 38 decades between the nominal signal floor $F = 10^{-7}$ and the numerical guard 10^{-45}; it is set by the floor-to-guard ratio in the sensitivity tests. Above the floor, the primary channel retains the log amplitude and $d_F(a) = 0$. Below the floor, the primary channel holds its floor reference and $d_F(a)$ represents the logarithmic distance below F. Finite nonpositive values remain input-valid with $d_F(a) = 0$. The encoding has no learned parameters. The nominal signal floor and numerical guard are fixed across sensors.

4.3 Physical Basis of the Encoding

Under the elastic-lidar equation [8, 17],

$$a(z) = \beta(z)T^2(z) = \beta(z)\exp[-2\tau(z)], \tag{5}$$

where $\beta(z)$ is the backscatter coefficient and $\tau(z)$ is the one-way optical depth accumulated above height z. Within the unsaturated branch ($10^{-45} \le a < F$) of Eq. (4),

$$d_F(a(z)) = \frac{\log_{10}[F/\beta(z)]}{38} + \frac{2\tau(z)}{38\ln 10}. \tag{6}$$

For fixed local backscatter, d_F is affine in cumulative optical depth. Within the unsaturated range, every decade of attenuation shifts the coordinate by the same amount. Local backscatter also enters Eq. (6), and optical depth is unknown at inference. A single measured return can therefore be consistent with different combinations of local backscatter and attenuation. The coordinate re-expresses the recorded return, and the network learns to combine it with neighboring structure and relationships from the paired WRF–COSP data. Inference uses measured ATB and its derived fields, together with position and wavelength. The simulator's optical depth and BC are never supplied as inputs.

4.4 Architecture and Experimental Setup

Figure 3 shows how the encoded returns enter the retrieval network. PhysicalSDC uses a multiscale encoder–decoder with input-guided feature modulation following AdaIR [3]. We train the network from scratch on AtmosBench. The modulation feeds features derived from the encoded input into the latent and decoder stages at their respective scales. The prediction head produces a single normalized BC curtain. The paired BC curtain supervises the output through the masked reconstruction loss, and the same input representation is used for training and inference.

We adapt the restoration backbones to the same ATB conditioning and a single BC output and train them from random initialization. The matched representation comparisons vary the input encoding within a fixed backbone. The Gaussian U-Net pair shares its convolutional trunk and tests the same encoding with a probabilistic mean-and-variance output.

The matched representations share the event split, train-only normalization, AdamW [20] with constant learning rate $4{\times}10^{-4}$ and weight decay 10^{-4}, effective batch size 12, ten epochs, gradient-norm clipping at 1.0, and the same evaluator. The checkpoint with the lowest validation macro log MAE is used for the reserved test events. BC targets are log-normalized using wavelength-specific bounds estimated on the training events. For log scoring we use the targets clipped to those bounds, expressing the error in decades. For scoring in physical units, we decode predictions and compare them with the original positive BC targets on the same supervised cells. Thus the physical-unit metrics also test the magnitude recovered by a model trained on normalized log backscatter.

Point-estimate models optimize a masked loss on normalized log BC. The cellwise terms constrain the retrieved amplitude, and the gradient term constrains how it changes between neighboring cells. A weak ATB return still contributes to supervision whenever the pair meets the validity conditions in Section 3.4:

$$\mathcal{L} = \mathcal{L}_1 + 0.25\mathcal{L}_{\mathrm{MSE}} + 0.05\mathcal{L}_{\mathrm{grad}}, \tag{7}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\mathrm{MSE}}$ average absolute and squared errors over supervised cells in each curtain. The gradient term averages absolute differences between predicted and target gradients along height and track. Each direction uses only adjacent pairs with two supervised cells, and we weight the available directions equally. We then average the losses across curtains. Gaussian U-Net uses a Gaussian output head and masked negative log likelihood [16]. Appendix B defines the input fields.

The matched representation and cross-backbone experiments use single-precision floating point (FP32). For the reserved-test comparisons with AdaIR and Log-domain U-Net, we report the mean and sample standard deviation over three matched seeds. Gaussian U-Net and the individual perturbation experiments use one validation-selected run per configuration. The broader restoration comparison includes mixed-precision runs.

We time the learned models in the same precision on an NVIDIA RTX 3090, processing one 200×200 curtain at a time. End-to-end timing includes input encoding, data transfer, inference, and backscatter decoding. We count floating-point operations (FLOPs) for supported operators, excluding Fourier transforms.

Table 1: Input representations on reserved AtmosBench events. Values are means ± sample standard deviations. Log MAE is in 10^{-3} dex. Linear MAE and RMSE are in $10^{-6}\,\mathrm{m}^{-1}\,\mathrm{sr}^{-1}$. Bold and underlining mark the best and second-best comparable results, respectively.

Representation	Log MAE↓	Lin. MAE↓	RMSE↓
Clip	242.87 ± 11.01	74.79 ± 1.28	325.04 ± 3.76
Null	233.66 ± 14.29	73.87 ± 0.78	325.02 ± 0.51
Asinh	105.39 ± 8.16	57.63 ± 3.75	271.22 ± 6.75
RawLog	<u>45.90 ± 0.62</u>	<u>35.14 ± 2.61</u>	<u>199.66 ± 6.97</u>
PhysicalSDC (Ours)	**40.18 ± 0.32**	**34.12 ± 2.24**	**190.07 ± 2.76**

Table 2: RawLog and PhysicalSDC on reserved AtmosBench events with two U-Net models. Log MAE is in 10^{-3} dex. Linear MAE and RMSE are in $10^{-6}\,\mathrm{m}^{-1}\,\mathrm{sr}^{-1}$.

Encoding	Log MAE↓	MAE↓	RMSE↓
Gaussian U-Net			
RawLog	66.65	40.13	241.50
PhysicalSDC (Ours)	**59.16**	**39.07**	**239.14**
Log-domain U-Net			
RawLog	54.39 ± 0.74	39.00 ± 2.52	213.97 ± 2.73
PhysicalSDC (Ours)	**48.23 ± 1.70**	**36.81 ± 1.55**	**211.64 ± 2.22**

5 Evaluation on AtmosBench

5.1 Representation Ablation on Held-Out Events

We apply the validation-selected models to the reserved test events. Table 1 holds the retrieval backbone fixed and varies how the input represents weak returns.

PhysicalSDC reduces log MAE by 83.46% relative to Clip and 82.80% relative to Null. Both controls retain the network and spatial context while collapsing below-floor variation. Null also matches the seven-channel input size and parameter count. RawLog retains finite positive measurements above the numerical guard and closes most of the gap left by flooring. Asinh preserves signed finite measurements through another global transform, with higher error than RawLog under the fixed scaling. The remaining gain over RawLog compares one scale shared by weak and strong returns with a coordinate dedicated to the below-floor range. These controls separate the larger effect of retaining weak variation from the effect of how it is encoded.

Table 3: Log MAE reduction relative to RawLog across validation events. Positive values favor PhysicalSDC (Ours).

Validation event	355 nm	532 nm
Event 1	−13.63%	+12.76%
Event 2	+5.53%	+30.29%
Event 3	+1.12%	+1.31%
Event 4	+25.30%	+25.93%

Table 4: Comparison with restoration models on AtmosBench validation events (top) and on the EarthCARE and CALIOP subsets (bottom). Log errors are in dex. Linear errors are in 10^{-6} m^{-1} sr^{-1}.

AtmosBench (Validation)	355 nm			532 nm		
Method	Log MAE↓	MAE↓	RMSE↓	Log MAE↓	MAE↓	RMSE↓
Gaussian U-Net	0.1072	35.94	226.9	0.1470	36.74	230.6
Log-domain U-Net	0.1047	36.18	232.9	0.1471	36.52	229.4
DATPRL	0.0969	33.57	214.7	0.1328	34.50	217.9
PromptIR	0.0925	35.16	223.7	0.1252	36.61	228.4
AdaIR	0.0922	33.22	213.0	0.1250	34.17	216.9
PhysicalSDC (Ours)	**0.0273**	**17.93**	**156.7**	**0.0341**	**19.45**	**162.0**
Satellite observations	EarthCARE ATLID–CPR (355 nm)			CALIOP CPro (532 nm)		
Method	AP↑	AUROC↑		Median abs. log error↓	Log MAE↓	MAE↓
Climatology	0.180	0.500		1.125	1.225	73.57
ATB binned mean	0.349	0.676		0.880	0.955	47.38
Raw ATB (direct)	0.373	0.518		0.253	0.452	29.24
Gaussian U-Net	0.606	0.800		0.340	0.441	30.41
Log-domain U-Net	0.585	0.781		0.286	0.377	28.34
DATPRL	0.590	0.788		0.333	0.417	27.74
PromptIR	0.578	0.786		0.281	0.378	26.75
AdaIR	0.591	0.794		0.287	0.375	25.85
PhysicalSDC (Ours)	**0.625**	**0.803**		**0.162**	**0.274**	**24.69**

PhysicalSDC gives the lowest error on all three metrics. Relative to RawLog, log MAE falls by 12.45%, MAE in physical units by 2.92%, and RMSE by 4.80%. The log MAE reduction is 5.71 in table units, with a paired-event 95% confidence interval of [1.75, 10.84]. For the bootstrap, we average runs and wavelengths within each event and resample whole events. The standard deviations in Table 1 describe variation across training runs, while the interval treats the held-out event as the unit of comparison.

Log error evaluates relative discrepancies across backscatter magnitudes, while physical-unit error gives larger absolute discrepancies more weight. On the validation events, log MAE and RMSE decrease by 12.17% and 1.15%, and physical-unit MAE differs by less than 1% from RawLog (Table A1). The relative log-error reduction is similar across the two event sets. On the reserved events the improvement also appears in both physical-unit metrics.

5.2 Cross-Backbone Evaluation

The gains extend to convolutional networks with point and probabilistic outputs (Table 2). Relative to RawLog, log MAE and physical-unit MAE fall by 11.33% and 5.62% for Log-domain U-Net. Its paired-event 95% interval for the log-error reduction is [3.93, 8.39] in table units. Gaussian U-Net gives reductions of 11.23% and 2.65%, respectively. The two comparisons preserve the training and evaluation setup within each network. Together with AdaIR, they show improvement across different backbones and with both the masked reconstruction loss and the Gaussian likelihood.

5.3 Evaluation Across Wavelengths and Events

Table 4 compares the full retrieval system with standard restoration models and task-specific U-Nets. Log error is higher at 532 nm for each model in the table, and PhysicalSDC gives the lowest validation error at both wavelengths. Its log MAE is 70% lower than standard AdaIR at 355 nm and 73% lower at 532 nm. Averaged across wavelengths, the physical-unit MAE and RMSE decrease by about 45% and 26% relative to AdaIR. Figure A1 shows selected predictions and their paired BC targets on a common scale.

In the matched comparison with RawLog, log MAE falls in seven of eight event–wavelength combinations (Table 3). The highest-error event improves at both wavelengths. The fourth event gives the largest reductions at both wavelengths taken together and accounts for about 82% of the aggregate decrease. The gain is widespread, with the largest reductions concentrated in one event.

6 Real-Data Evaluation

We apply the AtmosBench-trained models without real-data fine-tuning. Three references test complementary properties: backscatter agreement with the operational CALIOP product, layer structure against co-orbiting radar, and layer boundaries from airborne high-spectral-resolution lidar.

6.1 Spaceborne Lidar Observations

On CALIOP [35], we use 532-nm Level-1 ATB [21] and compare retrieved backscatter with the operational 5-km cloud-profile product (CPro, V4-51) [22, 36, 37]. CPro derives backscatter from measurements and retrieval assumptions. We aggregate the Level-1 profiles by CPro profile identifier so that the amplitude comparison uses the same along-track footprint. All representations are scored against the same registered cells under the nighttime sampling protocol.

On EarthCARE [6, 13], we use 355-nm ATLID Level-1 ATB [7]. Reflectivity from the co-orbiting CPR [14, 18] provides a structural reference for hydrometeor occurrence beneath cloud (Figure 1). We rank cells by estimated BC and compare their ordering with radar detections. Average precision (AP) and area under the receiver

Table 5: Comparison of input representations on 1,270 nighttime CALIOP CPro orbits and 1,298 EarthCARE CPR orbits. CPro log MAE is in 10^{-3} dex and linear MAE is in 10^{-6} $\mathbf{m^{-1}\,sr^{-1}}$. CPR AP and AUROC are scaled by 100.

	CALIOP CPro		EarthCARE CPR	
Representation	Log MAE↓	Lin. MAE↓	AP↑	AUROC↑
Clip	294.52	21.21	67.35	82.15
Null	285.27	20.25	**70.83**	**84.07**
Asinh	392.89	21.35	65.10	82.93
RawLog	465.75	23.05	64.71	81.17
PhysicalSDC (Ours)	**276.40**	**19.75**	69.21	83.48

operating characteristic curve (AUROC) summarize the resulting structural agreement (Appendix C).

On nighttime CALIOP orbits, PhysicalSDC gives the closest agreement with CPro in log and absolute backscatter error (Table 5). Relative to RawLog and Clip, log MAE decreases by 40.65% and 6.15%, and physical-unit MAE by 14.33% and 6.87%, respectively. The paired-orbit 95% intervals are positive for all four reductions. Relative to equal-capacity Null, log and physical-unit MAE decrease by 3.11% and 2.47%. Agreement in both metrics shows that the benefit on recorded lidar measurements extends beyond the simulated log-error comparison. Figure A2 shows the measured curtain and CPro alongside the estimates on a shared scale.

On EarthCARE, PhysicalSDC improves both structural scores over RawLog, Clip, and Asinh. Its scores are close to the higher scores of Null (Table 5). Null suppresses noise and weak-return variation together by assigning both the same below-floor code. The two satellite comparisons therefore distinguish closer backscatter-product agreement on CPro from comparable structural agreement with radar. CPR reflectivity is subject to different detection sensitivity, and the radar scores compare the location of detected structure.

PhysicalSDC leads the evaluated restoration architectures in the satellite comparisons in Table 4. On CALIOP, the log-MAE reduction over AdaIR is 27%, with lower median absolute log error and linear MAE. On the EarthCARE architecture cohort, CPR detections occupy 18% of valid pixels. AP focuses the comparison on this minority class, and the gain over Gaussian U-Net is more pronounced in AP than in AUROC. For AP and AUROC, the paired-orbit 95% intervals over Gaussian U-Net are $[0.0048, 0.0342]$ and $[0.0008, 0.0062]$, respectively.

The multilayer scene in Figure 1 gives the highest AP in the architecture cohort (Table A2). In that scene, the radar reference shows hydrometeor layers in the region of weak lidar returns beneath cloud. Figure 4 compares the same broad, vertically continuous curtain across every learned baseline on a separate EarthCARE orbit. Its Level-2 extinction, backscatter, and depolarization product (EBD) is a qualitative backscatter reference [5]. The aggregate EarthCARE structural scores use the co-orbiting CPR detections.

6.2 Airborne Layer-Boundary Retrieval

We test layer-boundary estimates against the 532-nm particulate backscatter measured by the second-generation High Spectral Resolution Lidar (HSRL-2) aboard the ER-2 during the Plankton, Aerosol,

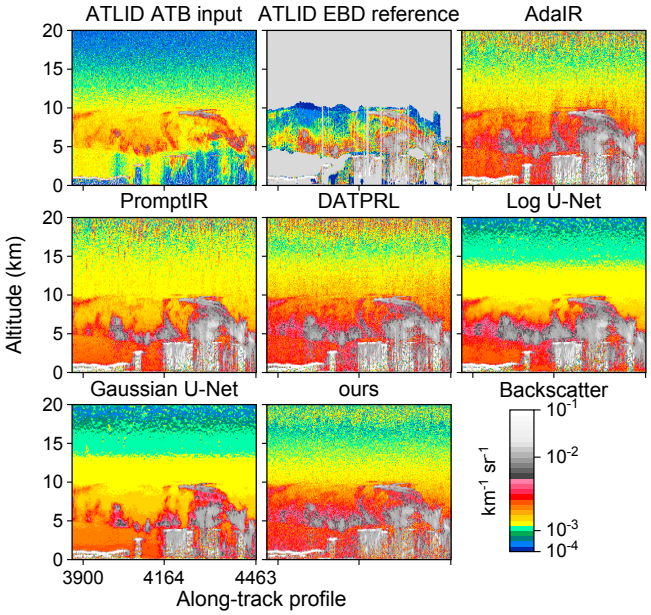


Figure 4: EarthCARE backscatter retrieval on orbit 04458C. The lidar input, operational retrieval, and model estimates share a 600-profile window and a common backscatter scale. Gray marks invalid cells.

Table 6: Layer-boundary retrieval on five PACE-PAX flights. Coverage is the share of profiles with a reference layer for which the method returns a boundary. MAE is averaged across flights on profiles with both reference and predicted boundaries, in meters.

Method	Cov. ↑	Top↓	Base↓	Thick. ↓
Fernald ($S = 50$ sr)	59.6%	804	1,310	941
Direct ATB	100.0%	1,281	310	1,224
Clip	100.0%	1,211	272	1,192
RawLog	100.0%	1,395	395	1,253
PhysicalSDC (Ours)	100.0%	1,204	265	1,192

Cloud, ocean Ecosystem Postlaunch Airborne Experiment (PACE-PAX) [9, 23]. We apply the same boundary rule to each prediction and the airborne reference to estimate the dominant layer's top, base, and thickness (Appendix C.2).

Across all 12 available flight-days, the clearest gain over RawLog is in layer base, where MAE falls 27.29% to 273.68 m. Top and thickness errors decrease by 11.79% and 4.00%, respectively. Appendix C.2 gives the paired comparison with Clip.

The five-flight comparison with classical inversion tests coverage as well as accuracy (Table 6). Klett and Fernald rely on lidar-ratio or boundary assumptions [8, 17]. Fernald returns boundaries for approximately 55–65% of profiles with a reference layer across the tested lidar ratios, with 59.6% at 50 sr. Its reported errors are calculated on the subset with a returned boundary. PhysicalSDC uses one ATB wavelength and relationships learned from WRF–COSP, yielding an estimate without a prescribed lidar ratio at inference. It

returns boundaries for all profiles with reference layers and gives the lowest base error. On profiles where both methods return a boundary, PhysicalSDC lowers base MAE by 1,034.18 m relative to the 50-sr Fernald inversion.

Direct ATB thresholding also provides full coverage. With the same layer operator, PhysicalSDC reduces its top and base errors by 6% and 15%, respectively. Coverage and accuracy together matter for tracing a layer along a curtain: each profile needs a usable boundary at a height consistent with the airborne reference.

7 Robustness and Computational Cost

7.1 Strong-Attenuation Analysis

We use the paired simulator targets to group predictions by ATB/BC after inference. For positive returns in the clean simulator, this ratio gives two-way transmission. We compare the same frozen models and targets within each group. In the strongly attenuated group ($0 < \mathrm{ATB/BC} < 0.1$), which covers about two fifths of the valid pixels, PhysicalSDC reduces log MAE by 6.35% and MAE in physical units by 11.21% relative to RawLog. The paired-event intervals favor PhysicalSDC for both metrics on these simulated events. A weak return can correspond to substantial local backscatter, and the improvement appears in both log-scale error and the estimated magnitude of backscatter.

In the strongly attenuated group, Clip and Null each give more than an order of magnitude higher log MAE than PhysicalSDC. Their errors are much closer in the weakly attenuated group ($0.5 \leq \mathrm{ATB/BC} \leq 1$). The difference is consistent with the role of the extra coordinate, which preserves variation that Clip and Null collapse. Across the positive transmission groups, the strongly attenuated group also gives the largest absolute reduction in physical-unit MAE over RawLog. These groups differ in target magnitude, so the paired comparisons within each group are the basis for attributing that gain. For finite nonpositive returns, the additional coordinate is inactive, and the paired differences with RawLog include zero.

7.2 Instrument Perturbations

The clean simulation isolates attenuation. To test instrument perturbations, we modify ATB before recomputing the input channels and keep the BC targets fixed. All representations receive the same perturbed ATB. The Gaussian scales are based on ATLID and CALIOP products, with $\sigma = 3.89\times10^{-7}$ and 3.45×10^{-7} m^{-1} sr^{-1} at 355 and 532 nm, respectively. We also test calibration errors, background offsets, quantization and clipping.

With clean-trained models, PhysicalSDC gives about half the log error of RawLog at 1σ and remains close to the floor-collapsed controls (Table 7). Null's constant below-floor code removes both weak-return variation and noise in that range. The gain over RawLog holds at all three Gaussian amplitudes.

The calibration and background-offset tests change the measurements in different ways. A multiplicative calibration shift preserves the ordering of positive returns, and PhysicalSDC gives the lowest log error under both tested shifts. An additive background offset can move weak returns across the encoding's positive-return boundary. Quantization and clipping can map distinct returns to the same value before encoding. The relative gain changes across these perturbations, consistent with the method's use of variation retained by the input product.

Table 7: Retrieval under input perturbations. Log MAE is in 10^{-3} dex. Dashes indicate untested configurations.

Test input	Clip	Null	Asinh	RawLog	PhysicalSDC (Ours)
Models trained on clean inputs					
Clean	230.34	229.56	103.12	45.43	39.90
0.5σ Gaussian	484.46	457.12	989.34	865.81	476.26
1σ Gaussian	522.85	504.08	1226.63	1049.80	531.60
2σ Gaussian	590.98	574.18	1443.24	1193.98	611.86
-0.5σ offset	337.68	301.39	796.94	1168.64	478.18
$+0.5\sigma$ offset	465.28	481.38	459.91	455.05	491.04
-2% calibration	232.37	232.84	103.98	46.69	42.58
$+2\%$ calibration	237.40	228.18	103.55	45.87	39.16
10^{-9} quantization	230.33	229.57	301.46	581.04	348.84
Clipped at F	230.34	229.56	431.96	439.69	424.18
Models trained with 1σ Gaussian noise					
Clean	308.23	–	–	311.40	302.02
1σ Gaussian	265.19	–	–	282.90	239.25

Table 8: Sensitivity to the nominal signal floor and numerical guard. Log MAE is in 10^{-3} dex. Dashes identify the input settings left untested in the sensitivity sweep.

Configuration	RawLog	PhysicalSDC (Ours)
$F = 10^{-8}$	–	51.34
$F = 10^{-7}$ (default)	45.43	**39.90**
$F = 10^{-6}$	–	41.04
Guard 10^{-30}	52.96	**43.12**

Training PhysicalSDC with the test noise scale reduces its noisy-test log MAE from 531.60 to 239.25 in table units. In the matched noise-training comparison, its mean error is 15.43% lower than RawLog and 9.78% lower than Clip. Paired-event intervals for those comparisons include zero. With the same noise-trained checkpoints on clean inputs, PhysicalSDC still has the lowest error of the three, while all three noise-trained models give higher clean-input error than their clean-trained counterparts. The results characterize adaptation to the tested noise distribution.

The encoding uses a nominal signal floor to separate the two input coordinates and a numerical guard to bound the logarithmic range. Table 8 varies these constants in matched training. The standard floor of 10^{-7} gives the lowest error among the tested floors. Increasing the guard to 10^{-30} compresses the range of the below-floor coordinate from 38 to 23 decades and retains an 18.58% log-MAE reduction over RawLog. The gain therefore extends beyond the default numerical range.

7.3 Computational Cost and Training Dynamics

The extra coordinate uses elementwise operations and increases the network's parameter count by 0.046%. With the same backbone, PhysicalSDC and RawLog have similar end-to-end latency, at 74.86

Table 9: Model size and inference cost. Latency is the median end-to-end measurement.

Model	Params (M)	FLOPs (G)	Latency (ms)
Log-domain U-Net	31.14	207.08	15.76
DATPRL	42.57	228.71	75.06
PromptIR	35.60	211.66	67.45
RawLog	28.83	198.36	75.69
PhysicalSDC (Ours)	28.84	198.50	74.86

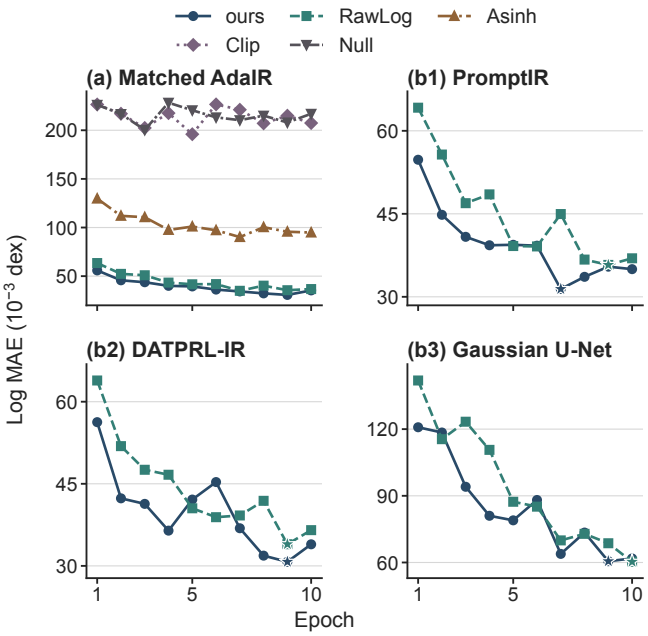


Figure 5: Validation error during training. (a) Five encodings in the same AdaIR network. (b1–b3) PhysicalSDC and RawLog paired within each network. Stars mark checkpoints selected by validation log MAE.

and 75.69 ms respectively (Table 9). The supported operation counts differ by 0.14 billion. The classical Fernald retrieval takes 0.59 ms on a central processing unit with different inputs.

In the matched AdaIR runs (Figure 5a), the two logarithmic encodings that retain below-floor variation already give validation log error at least 71.9% below the floor-collapsed controls at epoch 1. Null follows Clip despite having the additional channel. The separation persists across training, including at the validation-selected checkpoints used for test evaluation.

PromptIR and DATPRL give lower validation error with PhysicalSDC at epoch 1 and at the selected checkpoints (Figure 5b1–b2). At their validation minima, the encoding reduces log MAE from 0.0358 to 0.0314 dex for PromptIR and from 0.0340 to 0.0307 dex for DATPRL.

Gaussian U-Net has similar validation minima with both encodings. On the reserved events, the Gaussian model with PhysicalSDC gives lower errors than its RawLog counterpart (Table 2).

8 Limitations and Ethical Considerations

The paired renderings share an atmospheric state and the assumptions of WRF, the aerosol fields, and COSP. This construction isolates attenuation inside the simulator. The event grid was designed for controlled geographic and seasonal diversity, with each event–wavelength combination weighted equally in the reported scores. The retained COSP configuration documents the rendering procedure (Appendix A).

In measured products, noise, calibration, quantization, and fill-value processing determine which weak returns can be used. The encoding operates on finite measurements retained by the sensor product. If processing removes a return or maps all weak values to the same floor, those measurements no longer provide distinct coordinates. The real-data comparisons therefore use the available reference from each observing system. A co-registered corpus of satellite observations would make it possible to train or fine-tune on measured data and test transfer to the atmospheric conditions encountered by the sensor.

The data comprise atmospheric simulations and public Earth observations. Operational use of the retrieved profiles requires domain quality flags, uncertainty estimates, and source provenance. Non-finite and fill returns are explicitly masked in the retrieval.

Artifact Availability. The AtmosBench release includes the original paired arrays, fixed event splits, code, and three trained PhysicalSDC checkpoints.[1] The generation package includes the WRF and COSP2 configurations, the same-input check and the curtain-extraction code. A standalone scorer applies the same valid-cell mask and event–wavelength aggregation to predicted backscatter in physical units. Observational reference products are available from their instrument data providers.

9 Generative AI Usage

Generative AI tools assisted with literature organization, manuscript editing, and figure formatting. The authors defined the experimental protocols, physical formulations, and implementations and verified the reported experimental results. No generative AI output served as experimental evidence.

10 Conclusion

PhysicalSDC gives finite weak returns a separate input coordinate derived from the lidar equation. Matched comparisons on held-out AtmosBench events show the benefit of retaining below-floor variation and of how it is represented. Gains extend across backbones and to backscatter error in strongly attenuated layers. The paired renderings also let future retrieval methods train and evaluate on aligned targets from the same atmospheric states, at both wavelengths, under fixed event splits.

The satellite results show closer agreement with the operational CALIOP cloud-profile product and comparable structural agreement with EarthCARE radar relative to the equal-capacity control. In the airborne evaluation, the retrieved profiles give more accurate layer bases than RawLog. Encoding weak returns relative to the measurement physics lets a single lidar wavelength support vertical structure retrieval and layer tracing under attenuation.

[1] https://github.com/TianchiXu1/AtmosBench

References

[1] Howard W. Barker, Jason N. S. Cole, Najda Villefranque, Zhipeng Qu, Almudena Velázquez Blázquez, Carlos Domenech, Shannon L. Mason, and Robin J. Hogan. 2025. Radiative Closure Assessment of Retrieved Cloud and Aerosol Properties for the EarthCARE Mission: The ACMB-DF Product. *Atmospheric Measurement Techniques* 18 (2025), 3095–3107. doi:10.5194/amt-18-3095-2025

[2] M. Bonazzola, H. Chepfer, P.-L. Ma, J. Quaas, D. M. Winker, A. Feofilov, and N. Schutgens. 2023. Incorporation of Aerosol into the COSPv2 Satellite Lidar Simulator for Climate Model Evaluation. *Geoscientific Model Development* 16 (2023), 1359–1377. doi:10.5194/gmd-16-1359-2023

[3] Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. 2025. AdaIR: Adaptive All-in-One Image Restoration via Frequency Mining and Modulation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=M5t0WvjfCg

[4] Guanglu Dong, Chunlei Li, Chao Ren, Jingliang Hu, Yilei Shi, Xiao Xiang Zhu, and Lichao Mou. 2026. Learning Domain-Aware Task Prompt Representations for Multi-Domain All-in-One Image Restoration. In *The Fourteenth International Conference on Learning Representations*. arXiv:2603.01725 [cs.CV] https://openreview.net/forum?id=CzVlgDOF7L

[5] David Patrick Donovan, Gerd-Jan van Zadelhoff, and Ping Wang. 2024. The EarthCARE Lidar Cloud and Aerosol Profile Processor (A-PRO): The A-AER, A-EBD, A-TC, and A-ICE Products. *Atmospheric Measurement Techniques* 17, 17 (2024), 5301–5340. doi:10.5194/amt-17-5301-2024

[6] Michael Eisinger, Fabien Marnas, Kotska Wallace, Takuji Kubota, Nobuhiro Tomiyama, Yuichi Ohno, Toshiyuki Tanaka, Eichi Tomita, Tobias Wehr, and Dirk Bernaerts. 2024. The EarthCARE Mission: Science Data Processing Chain Overview. *Atmospheric Measurement Techniques* 17 (2024), 839–862. doi:10.5194/amt-17-839-2024

[7] European Space Agency. 2025. EarthCARE ATLID NOM Level 1B. Data set. doi:10.57780/ECA-BAD0EF3

[8] Frederick G. Fernald. 1984. Analysis of Atmospheric Lidar Observations: Some Comments. *Applied Optics* 23, 5 (1984), 652–653. doi:10.1364/AO.23.000652

[9] Johnathan W. Hair, Chris A. Hostetler, Anthony L. Cook, David B. Harper, Richard A. Ferrare, Terry L. Mack, Wayne Welch, Luis Ramos Izquierdo, and Floyd E. Hovis. 2008. Airborne High Spectral Resolution Lidar for Profiling Aerosol Optical Properties. *Applied Optics* 47, 36 (2008), 6734–6752. doi:10.1364/AO.47.006734

[10] Erhu He, Yiqun Xie, Licheng Liu, Zhenong Jin, Dajun Zhang, and Xiaowei Jia. 2024. Knowledge Guided Machine Learning for Extracting, Preserving, and Adapting Physics-aware Features. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 715–723. doi:10.1137/1.9781611978032.82

[11] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146, 730 (2020), 1999–2049. doi:10.1002/qj.3803

[12] Karen Hu and Xiaomei Lu. 2024. Spaceborne Lidar Measurement of Global Cloud Properties through Machine Learning. *Frontiers in Remote Sensing* 5 (2024), 1477503. doi:10.3389/frsen.2024.1477503

[13] Anthony J. Illingworth, H. W. Barker, A. Beljaars, M. Ceccaldi, H. Chepfer, N. Clerbaux, J. Cole, J. Delanoë, C. Domenech, D. P. Donovan, et al. 2015. The EarthCARE Satellite: The Next Step Forward in Global Measurements of Clouds, Aerosols, Precipitation, and Radiation. *Bulletin of the American Meteorological Society* 96, 8 (2015), 1311–1332. doi:10.1175/BAMS-D-12-00227.1

[14] Japan Aerospace Exploration Agency. 2024. EarthCARE/CPR L1B CPR One-Sensor Received Echo Power Products and Doppler Product. Data set. doi:10.57746/EO.01jdvcydwjf63vpdbjp0vz6v64

[15] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-Informed Machine Learning. *Nature Reviews Physics* 3, 6 (2021), 422–440. doi:10.1038/s42254-021-00314-5

[16] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, CA, USA, 5574–5584. https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[17] James D. Klett. 1981. Stable Analytical Inversion Solution for Processing Lidar Returns. *Applied Optics* 20, 2 (1981), 211–220. doi:10.1364/AO.20.000211

[18] Pavlos Kollias, Bernat Puidgomènech Treserras, Alessandro Battaglia, Paloma C. Borque, and Aleksandra Tatarevic. 2023. Processing Reflectivity and Doppler Velocity from EarthCARE's Cloud-Profiling Radar: The C-FMR, C-CD and C-APC Products. *Atmospheric Measurement Techniques* 16 (2023), 1901–1914. doi:10.5194/amt-16-1901-2023

[19] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Computer Vision – ECCV 2018 (Lecture Notes in Computer Science, Vol. 11215)*. Springer, Cham, Switzerland, 89–105. doi:10.1007/978-3-030-01252-6_6

[20] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[21] NASA/LARC/SD/ASDC. 2023. CALIPSO Lidar Level 1B Profile Data, V4-51. Data set, Version 4-51. doi:10.5067/CALIOP/CALIPSO/CAL_LID_L1-Standard-V4-51

[22] NASA/LARC/SD/ASDC. 2023. CALIPSO Lidar Level 2 Cloud Profile, V4-51. Data set, Version 4-51. doi:10.5067/CALIOP/CALIPSO/CAL_LID_L2_05kmCPro-Standard-V4-51

[23] NASA/LARC/SD/ASDC. 2025. PACE-PAX ER-2 High Spectral Resolution Lidar-2 Data. Data set, Version 1. doi:10.5067/ASDC/SUBORBITAL/PACE-PAX/DATA001/ER2/AircraftRemoteSensing/HSRL2_1

[24] Tomoaki Nishizawa, Rei Kudo, Eiji Oikawa, Akiko Higurashi, Yoshitaka Jin, Nobuo Sugimoto, Kaori Sato, and Hajime Okamoto. 2026. Algorithms to Retrieve Aerosol Optical Properties Using Lidar Measurements on Board the EarthCARE Satellite. *Atmospheric Measurement Techniques* 19 (2026), 729–744. doi:10.5194/amt-19-729-2026

[25] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. 2023. PromptIR: Prompting for All-in-One Image Restoration. In *Advances in Neural Information Processing Systems*, Vol. 36. doi:10.52202/075280-3121

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351)*. Springer, Cham, Switzerland, 234–241. doi:10.1007/978-3-319-24574-4_28

[27] Kaori Sato, Hajime Okamoto, Tomoaki Nishizawa, Yoshitaka Jin, Takashi Y. Nakajima, Minrui Wang, Masaki Satoh, Woosub Roh, Hiroshi Ishimoto, and Rei Kudo. 2025. JAXA Level 2 Cloud and Precipitation Microphysics Retrievals Based on EarthCARE Radar, Lidar, and Imager: The CPR_CLP, AC_CLP, and ACM_CLP Products. *Atmospheric Measurement Techniques* 18 (2025), 1325–1338. doi:10.5194/amt-18-1325-2025

[28] Patrick Selmer, John E. Yorks, Edward P. Nowottnick, Amanda Cresanti, and Kenneth E. Christian. 2024. A Deep Learning Lidar Denoising Approach for Improving Atmospheric Feature Detection. *Remote Sensing* 16, 15 (2024), 2735. doi:10.3390/rs16152735

[29] William C. Skamarock, Joseph B. Klemp, Jimy Dudhia, David O. Gill, Zhiquan Liu, Judith Berner, Wei Wang, Jordan G. Powers, Michael G. Duda, Dale M. Barker, and Xiang-Yu Huang. 2019. *A Description of the Advanced Research WRF Model Version 4*. Technical Report NCAR/TN-556+STR. National Center for Atmospheric Research. doi:10.5065/1dfh-6p97

[30] D. J. Swales, R. Pincus, and A. Bodas-Salcedo. 2018. The Cloud Feedback Model Intercomparison Project Observational Simulator Package: Version 2. *Geoscientific Model Development* 11 (2018), 77–81. doi:10.5194/gmd-11-77-2018

[31] Zheng Wang, Kai Ying, Bin Xu, Chunjiao Wang, and Cong Bai. 2025. From Swath to Full-Disc: Advancing Precipitation Retrieval with Multimodal Knowledge Expansion. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. 3091–3101. doi:10.1145/3711896.3736971

[32] Tobias Wehr, Takuji Kubota, Georgios Tzeremes, Kotska Wallace, Hirotaka Nakatsuka, Yuichi Ohno, Rob Koopman, Stephanie Rusli, Maki Kikuchi, Michael Eisinger, Toshiyuki Tanaka, Masatoshi Taga, Patrick Deghaye, Eichi Tomita, and Dirk Bernaerts. 2023. The EarthCARE Mission–Science and System Overview. *Atmospheric Measurement Techniques* 16 (2023), 3581–3608. doi:10.5194/amt-16-3581-2023

[33] Charles H. White, Yoo-Jeong Noh, John M. Haynes, and Imme Ebert-Uphoff. 2026. C3DIR: A Deep Learning 3-Dimensional Cloud Property Retrieval Scheme for Passive Satellite Imagers. *arXiv preprint arXiv:2607.16929* (2026). arXiv:2607.16929 [physics.ao-ph] doi:10.48550/arXiv.2607.16929

[34] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2023. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *Comput. Surveys* 55, 4 (2023), 66:1–66:37. doi:10.1145/3514228

[35] David M. Winker, Mark A. Vaughan, Ali Omar, Yongxiang Hu, Kathleen A. Powell, Zhaoyan Liu, William H. Hunt, and Stuart A. Young. 2009. Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms. *Journal of Atmospheric and Oceanic Technology* 26, 11 (2009), 2310–2323. doi:10.1175/2009JTECHA1281.1

[36] Stuart A. Young and Mark A. Vaughan. 2009. The Retrieval of Profiles of Particulate Extinction from Cloud-Aerosol Lidar Infrared Pathfinder Satellite Observations (CALIPSO) Data: Algorithm Description. *Journal of Atmospheric and Oceanic Technology* 26, 6 (2009), 1105–1119. doi:10.1175/2008JTECHA1221.1

[37] Stuart A. Young, Mark A. Vaughan, Anne Garnier, Jason L. Tackett, James D. Lambeth, and Kathleen A. Powell. 2018. Extinction and Optical Depth Retrievals for CALIPSO's Version 4 Data Release. *Atmospheric Measurement Techniques* 11, 10 (2018), 5701–5727. doi:10.5194/amt-11-5701-2018

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5728–5739. doi:10.1109/CVPR52688.2022.00564

A AtmosBench Data and Construction

The rendering configuration accompanying Figure 2 uses the retained `COSPv2.0-2.1.7` source. The reproduction check renders both branches from one atmospheric state and verifies matching output grids, finite returns and ATB $\leq$ BC at both wavelengths. The check uses six columns from the UKMO fixture. The released arrays are the benchmark reference because the historical build and input mapping are incomplete.

A.1 Curtain Color Scale

The lidar panels in Figures 1, 4, A1, and A2 use the same 27 backscatter bins. Equal-width swatches represent unequal numerical intervals. In units of 10^{-3} km^{-1} sr^{-1}, the bin edges are 0.1, 0.3, 0.8, 0.9, 1, 1.5, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The bars label the four decade boundaries. The radar reflectivity scale in Figure 1 is linear in dBZ.

B Conditioning Fields and Validation Metrics

B.1 Input Normalization and Observability

Let u_λ denote the train-only 99.9th percentile of positive ATB and $\tilde{u}_\lambda = \max(u_\lambda, 10F)$. For positive valid a, RawLog uses the following normalized log coordinate:

$$r_{\mathrm{raw}}(a) = \mathrm{clip}\left(\frac{\log_{10}\mathrm{clip}(a, 10^{-45}, \tilde{u}_\lambda) + 45}{\log_{10}\tilde{u}_\lambda + 45}, 0, 1\right). \tag{8}$$

Finite nonpositive values map to the 10^{-45} guard. Clip, PhysicalSDC, and Null apply the same linear scaling over $[\log_{10} F, \log_{10}\tilde{u}_\lambda]$ after mapping $a \leq F$ to F. Asinh applies $\mathrm{asinh}(a/s_\lambda)/\mathrm{asinh}(u_\lambda/s_\lambda)$ while preserving finite signed values. The frozen train-only constants are $s_{355} = s_{532} = 10^{-12}$, $u_{355} = 4.06625\times10^{-5}$, and $u_{532} = 4.77817\times10^{-5}$ m^{-1} sr^{-1}.

The five matched input representations share the deterministic observability field $O(A)$. Invalid or below-floor cells are mapped to the fixed 10^{-7} for this field, including in the floor-sensitivity tests. The log ATB is clipped at the train-only 99.9th percentile and smoothed with the separable 3×3 kernel $[1,2,1]^\top[1,2,1]/16$, giving L. We compute

$$\begin{aligned} p_s &= \sigma((L - n_\lambda)/m_\lambda - 3), \\ p_a &= \sigma\left(\frac{L - q_{90,\lambda}}{(q_{99.9,\lambda} - q_{50,\lambda})/8}\right), \\ t_i &= \Delta z \sum_{j\leq i} p_{a,j}, \qquad O_i = M_{A,i}\,\mathrm{clip}\left(p_{s,i}e^{-t_i/t_{95,\lambda}}, 0, 1\right). \end{aligned} \tag{9}$$

Here n_λ and m_λ are the median and median absolute deviation of high-altitude log ATB. We fit $q_{\cdot,\lambda}$ and $t_{95,\lambda}$ on the training events. For N_z vertical bins, $\Delta z = 20/\max(N_z - 1, 1)$ km. The cumulative sum starts at the top of the curtain. The normalized height index H runs from 0 at the top to 1 at the bottom, and along-track position X spans $[0, 1]$. The wavelength indicator distinguishes the two sensing wavelengths: -1 for 355 nm and $+1$ for 532 nm.

B.2 Validation-Set Representation Ablation

Table A1 summarizes validation log and linear metrics across all five input representations, complementing Table 1 and the per-event validation results in Table 3.

Table A1: Representation ablation on the validation set. Log MAE is in dex. Linear MAE and RMSE are in 10^{-6} m^{-1} sr^{-1}.

Representation	Log MAE↓	Linear MAE↓	RMSE↓
Clip	0.1959	42.28	239.1
Null	0.2004	43.16	245.4
Asinh	0.0902	31.90	213.2
RawLog	<u>0.0349</u>	**18.52**	<u>161.2</u>
PhysicalSDC (Ours)	**0.0307**	<u>18.69</u>	**159.3**

C Registration and Layer-Boundary Evaluation

C.1 Satellite Registration and Per-Orbit EarthCARE Scores

For the multi-architecture CALIOP cohort, we aggregate Level-1 profiles by the Level-2 profile identifier before comparing backscatter at the 5-km CPro footprint. The registered cohort covers 29 granules with 106,144 profile columns and 2,435,744 valid cells. For EarthCARE, 18 neighboring ATLID profiles and 11 neighboring CPR profiles are grouped before vertical interpolation onto the common 100-m grid. The multi-architecture cohort contains 5,957 registered columns and 1,129,149 valid pixels across six orbits. Table A2 lists the per-orbit structural scores on this cohort. The wider matched-representation cohorts are reported in Table 5.

Table A2: Per-orbit EarthCARE ATLID–CPR structural scores for PhysicalSDC (Ours) on the reference cohort in Table 4. Orbit 02963C is shown in Figure 1.

Orbit ID	AP↑	AUROC↑
01337E	0.606	0.773
01490B	0.545	0.755
01551E	0.387	0.745
01684D	0.675	0.811
02923H	0.703	0.818
02963C	0.834	0.919

C.2 Airborne Layer-Boundary Evaluation

HSRL-2 profiles are aggregated linearly into 100-m vertical bins using cell overlap (excluding bins with < 80% coverage) and resampled onto a 1-km navigation grid. Each flight track is segmented at missing positions, time gaps exceeding 30 s, or geodesic jumps exceeding 6 km, retaining complete non-overlapping 200-km windows (103 segments in the Fernald comparison). Every method and the cloud-screened 532-nm particulate reference share a 10^{-7} m^{-1} sr^{-1} threshold, 0.3–12 km altitude range, three-bin minimum contiguous run, and at least 12 common valid bins. When multiple candidate runs exist, the run with the largest integrated backscatter defines the layer, whose highest and lowest above-ground altitudes give layer top and base. Fernald inversions are evaluated at $S \in \{30, 50, 70\}$ sr. Table 6 reports $S = 50$ sr.

On the full airborne set, the paired-flight 95% intervals for the reduction relative to Clip are $[-41.30, 40.15]$ m for top and $[-24.14, 30.63]$ m for base.

D Additional Qualitative Results

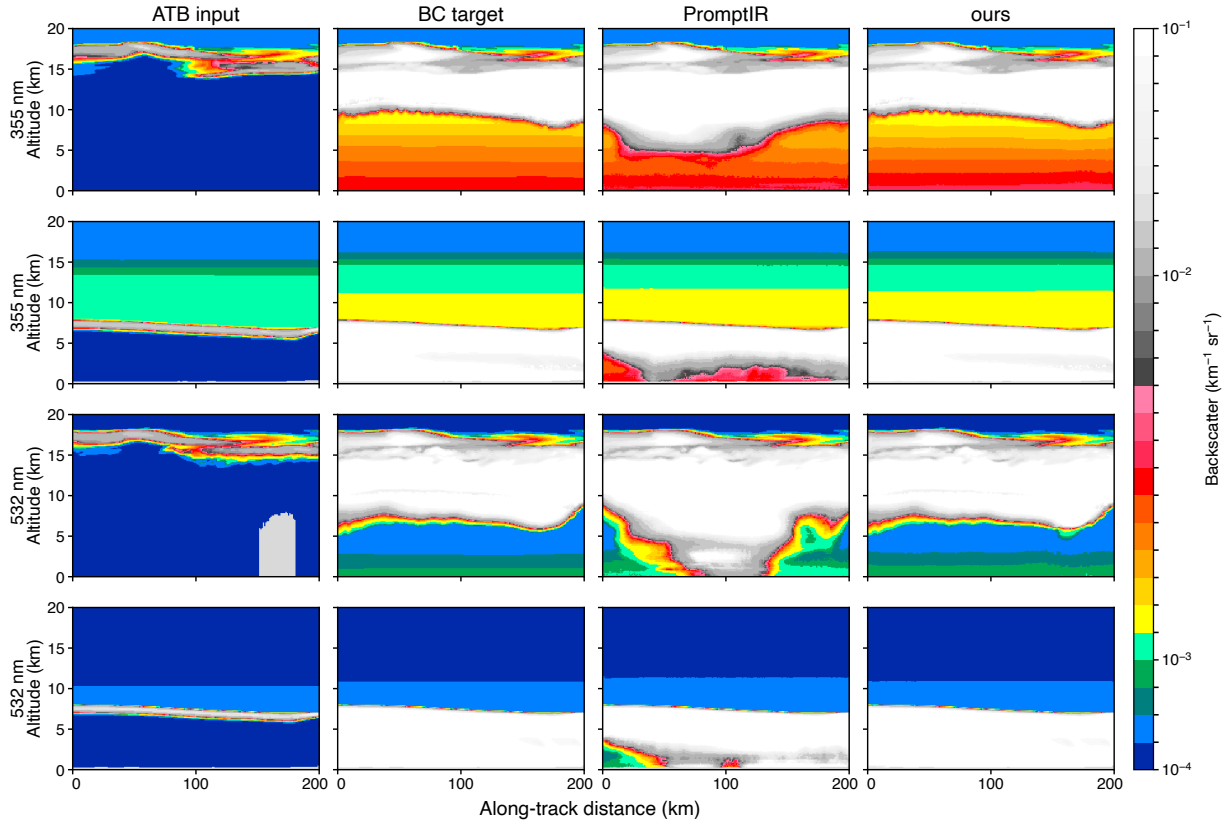


Figure A1: Example retrievals on AtmosBench. Among curtains with PhysicalSDC log MAE at most 0.02 dex and lower than PromptIR's, we show the events with the largest within-event gains at each wavelength. Gray marks invalid cells.

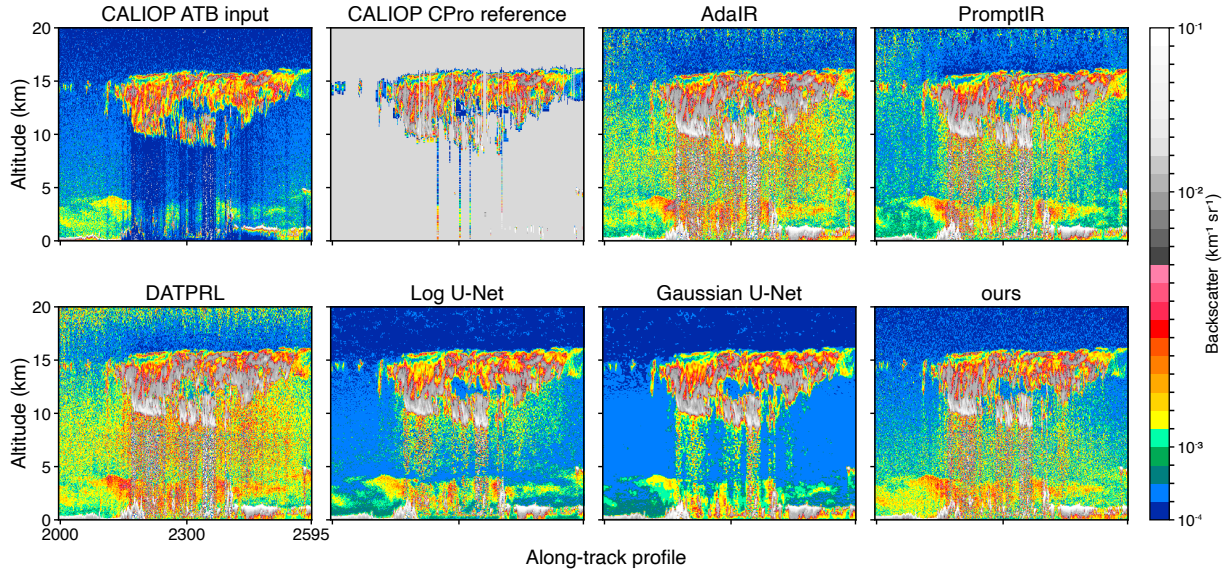


Figure A2: Example retrievals on CALIOP. The lidar input, operational CPro retrieval, and model estimates share the same 600-profile window from nighttime granule 2021-01-01T01-06-51ZN. Gray marks invalid cells.