Supplementary Sections

Contents

Section.S1 Customized CNN InceptionNet model architecture

The workflow implements supervised learning to predict a target vertical profile on a fixed altitude grid (length: 2000 points). Each input sample consists of two paired profiles (two channels) namely, the refractivity vertical profile (N; 0-20 km at 10 m altitude resolution) and its wavelet covariance transform (WCT_N; Same resolution as N). The model outputs a single-channel predicted profile. Two independent models are trained and applied (as shown in Fig.3 in main text): **Model 0** predicts the **dry refractivity** ($\boldsymbol{N_d}$) from the two input profiles and **Model 1** predicts **dry pressure** ($\boldsymbol{P_d}$) from the two input profiles. The idea is to define and develop a data driven relation between the profiles of the N and profiles of N_d & P_d to bypass the under-determined nature of the Equation (1) in main text which has 3 variables.

The core predictive model employed in this study is a specialized Convolutional Neural Network (CNN) incorporating Inception-style modules within an Encoder-Decoder framework, referred to as "InceptionNet" (Szegedy et al., 2015). The customized architecture is designed to handle high-dimensional atmospheric profile data, mapping input profiles with dimensions (2, 2000, 1) to target profiles of (1, 2000, 1). The methodology comprises three primary components:

(i) Data pre-processing and normalization,

(ii) Model architecture design with multi-scale feature extraction,

(iii) Ensemble training with K-fold cross-validation.

S1.1. Data Pre-processing

The atmospheric profile data undergoes several pre-processing steps prior to model training. The Quality control procedure for converting raw measurements of temperature, pressure and relative humidity into evenly gridded model input data from 10 m to 20 km at 10 m interval is explained in Section 2.3. Further pre-processing pipeline removes the lowest 100 meters (10 data points) from each profile to eliminate sharp and often noisy measurement gradients close to the surface, accommodate WCT_N which is computed with a dilation of 150 m (150/2 = 75 m should be ideally the lowest altitude where unclipped WCT_N begins). The removed data points are replaced with padding values extrapolated from the nearest valid measurements at the top to maintain consistent tensor dimensions. However, the final predicted profiles consist of 100 m to 20 km at 10 m vertical resolution (1990 valid points per profile).

Z-score normalization is applied to both input and output profiles independently. For an input profile x with altitude index i, the normalized value x'_i is computed as:

$$x'_i = (x_i - \mu_i) / \sigma_i \quad \text{(S1)}$$

where μ_i and σ_i represent the mean and standard deviation at altitude level *i* computed across the training dataset. This altitude-wise normalization preserves the vertical structure of atmospheric variability while ensuring numerical stability during training.

S1.2. Model architecture

The predictive model is an Inception-style convolutional encoder–decoder (InceptionNet) implemented using Keras & TensorFlow. The profile input is treated as a 2D tensor where one axis represents profile channels and the other represents altitude. The encoder comprises an initial convolution followed by two Inception-style blocks with parallel convolutional pathways. The decoder mirrors this structure using transposed convolutions to reconstruct a full-resolution profile. The architecture is illustrated in Figures S1 – S3 and described in detail below.

Encoder blocks (multi-scale feature extraction)

- **Initial convolution:** Conv2D with kernel (1×7), stride (1×2), 64 filters, ReLU activation.
- **InceptionBlock 1:** four parallel branches: **(a)** Conv2D with kernel (channels×1), stride (channels×5), **(b)** Conv2D with kernel (channels×1) → Conv2D with kernel (channels×9), stride (channels×5), **(c)** Conv2D with kernel (channels×1) → Conv2D with kernel (channels×13), stride (channels×5), and **(d)** MaxPool with kernel (channels×30), stride (channels×5) → Conv2D with kernel (1×1). Branch outputs are concatenated along the filter dimension. The "channels" parameter is set to 1 to process each input altitude profile separately.

- **InceptionBlock 2:** same multi-branch structure with increased filter counts and with the "channels" parameter set to the input stream count to merge input altitude profiles.
- **Global reductions:** Filter axis averaging (reduce_mean) followed by a Dense layer (ReLU) to yield a compact latent representation.
- **Latent bottleneck:** A Dense layer with tanh activation produces the latent vector of size lat_dim (set to 120/80 for *Model-0*/*Model-1* respectively). The latent space serves as a compressed representation of the atmospheric profile, capturing the essential information required for reconstruction. The latent dimension represents a hyperparameter that balances reconstruction fidelity against model complexity.

Decoder blocks (reconstruction)

The decoder reconstructs the output profile from the latent representation using transposed convolutions (also known as deconvolutions). The architecture mirrors the encoder

but employs three-pathway DeInception blocks that omit the max pooling pathway. It consists of the following parts:

- **Dense:** The latent vector is passed through a fully connected layer to make an 80-dimension feature vector. This feature vector is expanded by a factor of 25 through the De-Inception Blocks to predict the 2000-point output.
- **Reshape:** the projected vector is reshaped to (1, decoder_input_size, 1).
- **DeInceptionBlock 1:** three parallel transposed-convolution branches: **(a)** Conv2DTranspose with kernel (channels×1), stride (channels×5), **(b)** 1×1 → kernel (channels×9), stride (channels×5), **(c)** 1×1 → kernel (channels×13, stride (channels×5). Outputs are concatenated along the filter axis; activation tanh.
- **DeInceptionBlock 2:** analogous structure with larger filter counts; activation linear.
- **Final averaging:** Filter axis averaging reduces the concatenated tensor to an output of shape (1, 2000, 1).

Loss Function

The model is trained using a composite loss function combining frequency-domain and spatial-domain components:

$$L = L_{FFT} + \lambda \cdot L_{MAE} \quad \text{(S2)}$$

The frequency-domain component L_{FFT} is computed as the mean absolute difference between the Fast Fourier Transform (FFT) magnitudes of the predicted and actual profiles:

$$L_{FFT} = (1/N)\ \Sigma\ |FFT(\hat{y}) - FFT(y)| \quad \text{(S3)}$$

where $\hat{y}$ and y denote the predicted and actual profiles, respectively, and N is the number of frequency components. This term penalizes discrepancies in the spectral characteristics of the profiles, encouraging the model to capture both fine-scale variations and large-scale trends (Fuoli et al., 2021; Yadav et al., 2021).

The spatial-domain component L_{MAE} represents the standard Mean Absolute Error between predicted and actual profile values. The weighting parameter λ (set to 10) controls the relative importance of spatial accuracy versus spectral fidelity. This design is intended to encourage accurate pointwise reconstruction while constraining the overall spectral characteristics of the profile.

Table S1: Tensor Dimensions Through the Network

Layer	Output Shape	Operation
Input	(N, 2, 2000, 1)	—
Initial Conv2D	(N, 2, 1000, 64)	1×7 conv, stride 2
InceptionBlock-1	(N, 2, 200, 112)	Multi-scale, stride 5
InceptionBlock-2	(N, 1, 40, 224)	Multi-scale, stride 5
Channel Average	(N, 1, 40)	Mean over channels
Dense	(N, 1, 120) - Model 0 (N, 1, 80) - Model 1	Fully Connected Layer, ReLU
Spatial Average	(N, 120) - Model 0 (N, 80) - Model 1	Mean over Spatial Dimension
Latent Vector	(N, 120) – Model 0 (N, 80) – Model 1	Dense, tanh
Dense (projection)	(N, 80)	Dense, tanh
Reshape	(N, 1, 80, 1)	Reshape
DeInceptionBlock-1	(N, 1, 400, 96)	Transpose, stride 5
DeInceptionBlock-2	(N, 1, 2000, 192)	Transpose, stride 5
Final Average	(N, 1, 2000, 1)	Mean over channels
Output	(N, 1, 2000, 1)	—

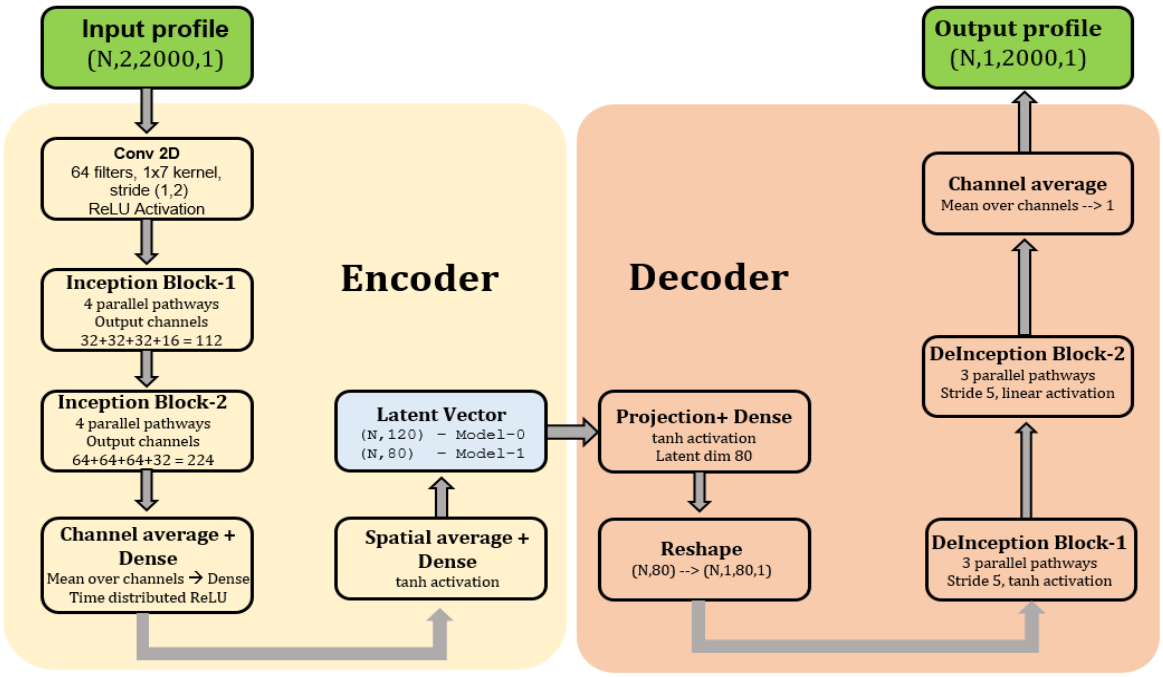

Figure-S1 InceptionNet Encoder-Decoder architecture

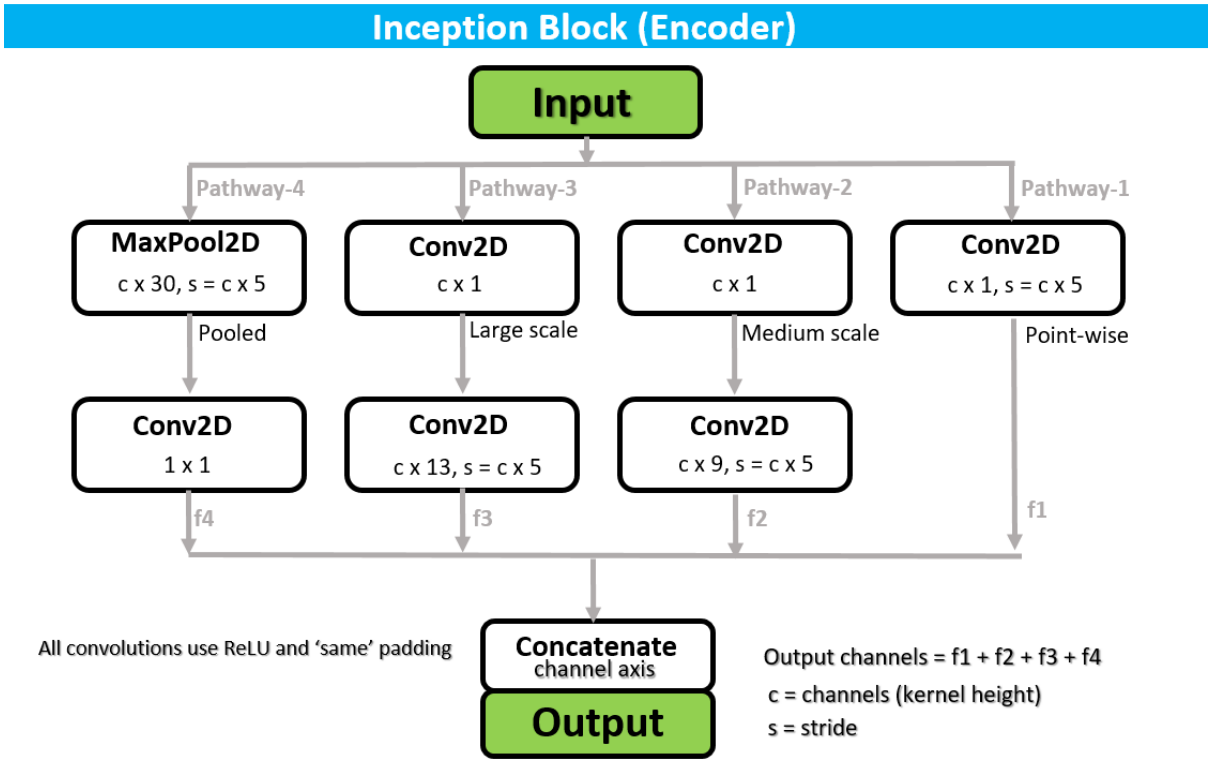

Figure S2 Data flow within the Inception Block Internal structure (Encoder)

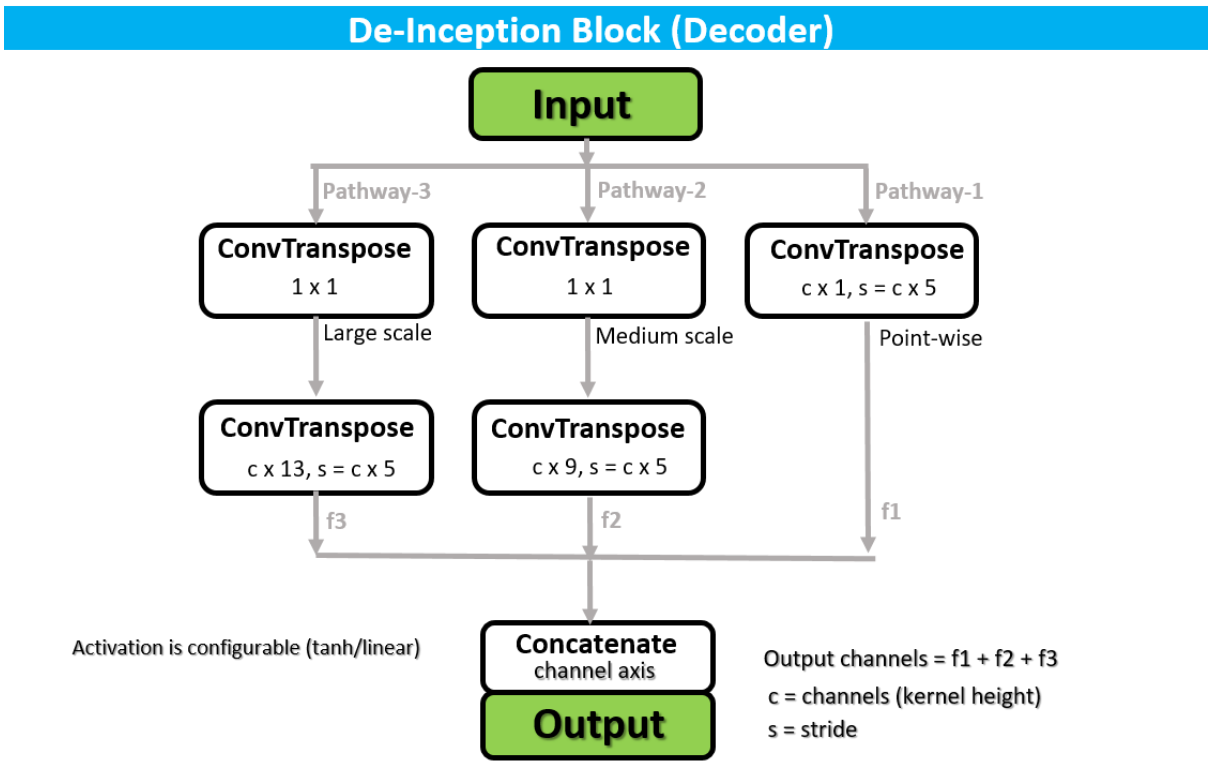

Figure S3 Data flow within the De-Inception Block Internal structure (Decoder)

S1.3. Ensemble training with K-fold cross-validation

To improve prediction robustness, the model is trained using K-fold cross-validation. K is set to 5, thus, partitioning the training dataset into 5 mutually exclusive subsets (folds). For each fold, K−1 (=4; 80% of data) subsets are used for training while the remaining subset (20% of data) serves as validation. This process yields 5 independent models, each trained on a different data partition. For each fold, two models are trained independently resulting in a total ensemble of 10 models. The best-performing model weights for each target variable are retained based on validation loss for each model. Shuffling of data is done before making the folds and a fixed random seed is used to ensure reproducibility.

During the inference phase, predictions are generated by aggregating outputs from the full ensemble (5 models for each target variable). The final prediction is derived via the median of the ensemble outputs to ensure robustness against outliers. This also ensures that the final predictions of this supervised learning exercise are not overtly sensitive to the selection of the training data and the validation data during model construction.

S1.3.1. Optimization Configuration

The Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 5×10^{-4} is employed for model training. Following regularization and optimization techniques are implemented:

1. **Learning rate warmup:** The learning rate linearly increases from 10^{-4} to 10^{-3} over the first 3 epochs to stabilize early training.
2. **Learning rate scheduling:** Upon validation loss plateau (patience: 5 epochs), the learning rate is reduced by a factor of 0.5, with a minimum value of 10^{-7}.
3. **Early stopping:** Training terminates if validation loss does not improve for 10 consecutive epochs, with the best-performing weights restored.
4. **Gradient clipping:** Gradient norms are clipped to 1.0 to prevent exploding gradients.

The training configuration and hyperparameters finally used in the study are summarized in Table S1. These are finalized after the sensitivity tests (described in Sect. S2)

Table S2. Training hyper parameters and configuration used.

Parameter	Value	Description
Batch size	512	Number of samples per gradient update
Maximum epochs	100	Upper limit on training iterations
Initial learning rate	0.0005	Adam optimizer learning rate
Latent dimension	120 Model-0 N_d 80 Model-1 P_d	Size of compressed representation
MAE loss weight (λ)	10	Weight of spatial loss component
Number of folds (K)	5	Cross-validation partitions
Early stopping patience	10	Epochs without improvement
LR reduction factor	0.5	Learning rate decay multiplier

S1.3.2. Implementation Details

The model is implemented using TensorFlow 2.x with the Keras API (tf_keras). Training is conducted on GPU-enabled hardware. Data loading employs the tf.data API with prefetching for efficient pipeline execution. The implementation supports both cloud environments and local computing infrastructure.

Section.S2 Sensitivity of hyper-parameter configurations

S2.1. Experimental Design

A systematic sensitivity analysis was conducted to evaluate the influence of key hyperparameters on the predictive performance and training stability of the dual-model InceptionNet architecture Model-0 predicting N_d and Model-1 predicting P_d. The analysis examined three primary hyperparameter categories: **(i)** learning rate (LR) and batch size, which govern training optimisation dynamics; **(ii)** latent space dimensionality, which controls the information bottleneck within the encoder-decoder structure; and **(iii)** the mean absolute error (MAE) loss weight coefficient (α) in the composite Fourier-MAE loss function. All configurations were evaluated using five-fold cross-validation to ensure robust estimation of performance metrics and to quantify cross-fold variability.

The experimental design comprised ten distinct configurations (Table S3). Cases 01–03 established baseline performance with symmetric latent dimensions (dim = 80) for both models, while varying learning rate, batch size, and number of training epochs to characterise convergence behaviour. Cases 04–06 systematically varied the symmetric latent dimension (60, 80, and 120) to determine the optimal compression ratio for atmospheric profile reconstruction. Cases 07–10 explored asymmetric configurations in which the two models employed different latent dimensions, motivated by the hypothesis that the optimal feature representation may differ between the two prediction targets (N_d and P_d). Additionally, Cases 07 and 10 incorporated modified MAE weight coefficients to examine the trade-off between spectral fidelity (Fourier loss component) and point-wise accuracy (MAE component).

Table S3. Summary of hyperparameter configurations evaluated in the sensitivity analysis. M0 and M1 denote Model 0 (for N_d) and Model 1 (for P_d), respectively. The composite loss function is defined as $L = L_{Fourier} + \alpha \times L_{MAE}$, where α is the MAE loss weight. All experiments employed early stopping with a patience of 10 epochs and learning rate scheduling with a reduction factor of 0.5. All other parameters for model were unchanged and as per Table S2.

Case	Experiment ID	Batch Size	Latent Dim (M0)	Latent Dim (M1)	MAE Weight (M0/M1)	Initial LR	No. of Epochs	Purpose
01	Baseline-1	512	80	80	10/10	1×10^{-4}	50	Reference configuration
02	Baseline-2	512	80	80	10/10	1×10^{-4}	100	Reproducibility verification (extended epochs)
03	Baseline-3	512	80	80	10/10	1×10^{-3}	100	Learning rate sensitivity (10×)
04	LS-60	128	60	60	10/10	5×10^{-4}	100	Reduced latent dimension
05	LS-80	128	80	80	10/10	5×10^{-4}	100	Standard latent dimension
06	LS-120	128	120	120	10/10	5×10^{-4}	100	Expanded latent dimension
07	LS-80/60-MAE12	128	80	60	12/12	5×10^{-4}	100	Asymmetric latent dimensions + modified MAE weight
08	LS-120/80	128	120	80	10/10	5×10^{-4}	110	Asymmetric latent dimensions (high/mid)
09	LS-150/60	128	150	60	10/10	5×10^{-4}	130	Asymmetric latent dimensions (extreme)
10	LS-80/40-MAE20	128	80	40	10/20	1×10^{-3}	100	Asymmetric latent dimensions + differential MAE weight

S2.2. Evaluation Metrics

Six quantitative metrics were employed to comprehensively assess each hyperparameter configuration during each training run. **Mean validation loss** denotes the average of the best

validation losses achieved across the five cross-validation folds, computed using the composite loss function $L = L_{Fourier} + \alpha \times L_{MAE}$. The **coefficient of variation (CV)** quantifies cross-fold stability as the ratio of the standard deviation to the mean validation loss, expressed as a percentage; lower values indicate more consistent performance across data partitions. The **generalisation gap** measures the absolute difference between training and validation loss at the optimal epoch (|train loss − validation loss|), serving as an indicator of overfitting propensity. The **convergence epoch** denotes the mean epoch at which validation loss first falls below 50% of its initial value, characterising training efficiency.

S2.3. Results: Model 0 (N_d)

Table S4. Sensitivity analysis results for Model 0. Cases marked with (*) employ non-standard MAE weights; validation loss values should be normalised prior to cross-experiment comparison. Generalization gap computed at the best validation epoch (lowest validation loss).

Case	Experiment ID	Latent Dim	MAE Wt	Val Loss	CV (%)	Gen Gap	Conv Epoch (mean)	Best Epoch (mean)
01	Baseline-1	80	10	7.034	1.46	0.036	48.0	50
02	Baseline-2	80	10	6.607	1.48	0.069	47.8	98.6
03	Baseline-3	80	10	6.059	0.78	0.256	13.0	98.6
04	LS-60	60	10	6.145	0.46	0.364	11.0	83.2
05	LS-80	80	10	5.932	0.30	0.303	8.0	89.4
06	LS-120	120	10	5.821	1.15	0.303	6.4	92.6
07*	LS-80-MAE12	80	12	6.505	1.28	0.464	7.8	86.2
08	LS-120/80	120	10	5.817	1.17	0.314	6.8	101.6
09	LS-150/60	150	10	5.817	0.36	0.303	7.2	95.4
10	LS-80/40-MAE20	80	10	5.840	0.99	0.416	7.4	71.8

Model 0 demonstrated a clear and systematic sensitivity to latent space dimensionality. The symmetric latent dimension sweep (Cases 04–06) revealed a monotonic improvement in mean validation loss with increasing representational capacity: 6.145 (dim = 60), 5.932 (dim = 80), and 5.821 (dim = 120). The optimal configuration (Case 08; latent dim = 120) achieved the lowest mean validation loss of 5.817, corresponding to a 17.3% improvement relative to the conservative Baseline-1 configuration (Case 01). Further increasing the latent dimension to 150

(Case 09) yielded no additional improvement, indicating that the model had reached a representational saturation point for atmospheric profile reconstruction.

Configurations employing a batch size of 128 and an initial learning rate of 5×10^{-4} exhibited convergence within 6–8 epochs, compared to 47–48 epochs for the baseline configurations (Cases 01–02), indicating substantially enhanced training efficiency. The following additional characteristics were observed across Model 0 configurations:

Training stability. Model 0 maintained a consistently low coefficient of variation across all configurations, reaching a minimum of 0.30% in Case 05 (CV = 0.30%). This indicates that the model converges to a reproducible solution irrespective of the data partitioning strategy employed.

Generalisation behaviour. The generalisation gap remained consistently low across all configurations (range: 0.036–0.464), indicating minimal propensity for overfitting and robust generalisation to held-out validation data.

Training duration. Model 0 consistently utilised the full training duration (mean ≈ 100 epochs), indicating continued extraction of informative features throughout the training cycle without any indication of overfitting.

S2.4. Results: Model 1 (P_d)

Table S5. Sensitivity analysis results for Model 1. Cases marked with (*) employ non-standard MAE weights. Note that Case 10 uses an MAE weight of 20, resulting in substantially elevated loss values that are not directly comparable to other cases without normalisation. Generalization gap computed at the best validation epoch (lowest validation loss).

Case	Config	Latent Dim	MAE Wt	Val Loss	CV (%)	Gen Gap	Conv Epoch (mean)	Best Epoch (mean)
01	Baseline-1	80	10	7.159	0.77	0.088	50.0	49.8
02	Baseline-2	80	10	6.875	0.56	0.172	99.4	97.6
03	Baseline-3	80	10	6.430	1.05	0.467	19.8	56
04	LS-60	60	10	6.491	1.13	0.413	15.4	37
05	LS-80	80	10	6.420	0.88	0.494	14.2	40.2
06	LS-120	120	10	6.293	1.25	0.569	12.0	51

Case	Config	Latent Dim	MAE Wt	Val Loss	CV (%)	Gen Gap	Conv Epoch (mean)	Best Epoch (mean)
07*	LS-60-MAE12	60	12	7.172	0.74	0.570	15.6	39
08	LS-120/80	80	10	6.437	0.98	0.460	13.6	38.2
09	LS-150/60	60	10	6.529	1.40	0.505	16.6	38.4
10*	LS-40-MAE20	40	20	9.942	1.15	0.824	11.8	29.2

Model 1 exhibited a qualitatively similar, but less pronounced, response to latent dimension variations compared to Model 0. The symmetric dimension sweep yielded mean validation losses of 6.491 (dim = 60), 6.420 (dim = 80), and 6.293 (dim = 120), representing a comparatively modest improvement range. Case 06 (symmetric latent dim = 120) achieved the optimal validation performance among all symmetric configurations.

Collectively, the results indicate that Model 1 faces a more challenging optimisation landscape than Model 0. The following characteristics were consistently observed:

Elevated loss baseline and early saturation. Model 1 exhibited systematically higher validation losses and larger generalisation gaps (range: 0.088-0.824) relative to Model 0. However, as demonstrated by the stable validation learning curves (Fig. S5(d)), the loss plateaus rather than progressively degrading after the optimal epoch. This behavior does not indicate classical overfitting, but rather reflects the physical nature of the target variable. Model 1 predicts dry pressure (P_d), a macroscopically smooth profile that physically represents the integrated mass of the overlying atmospheric column. Because the convolutional architecture relies on extracting features from localized refractivity inputs (N and WCT_N), it rapidly extracts the available point-wise and regional statistical signals. The elevated baseline and early plateau indicate that the network quickly reaches the accuracy ceiling of mapping local atmospheric features to a vertically integrated profile, rendering additional representational capacity (e.g., latent dimensions > 80) ineffective

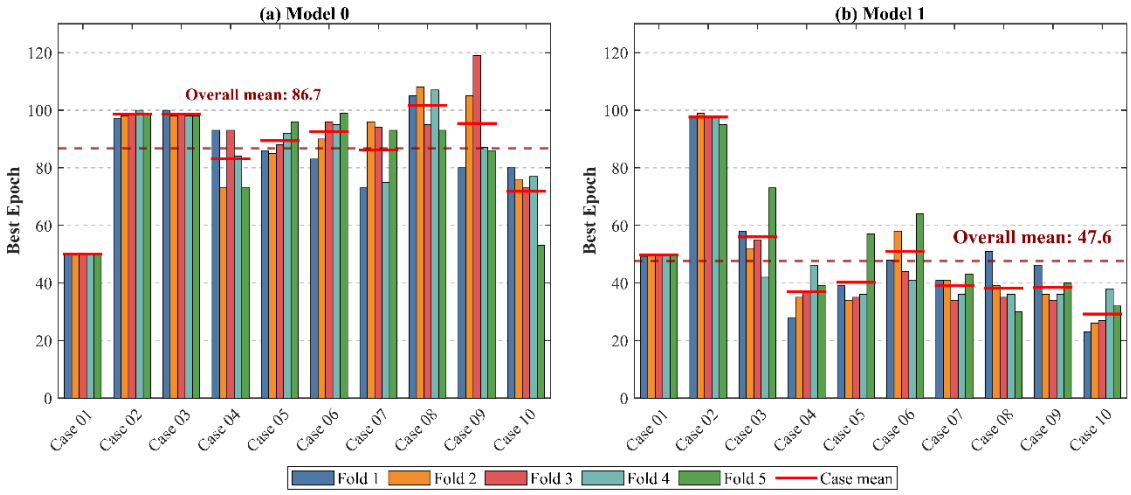

Figure S4. Distribution of optimal training epochs for **Model 0** (predicting N_d) and **Model 1** (predicting P_d) across five cross-validation folds and ten experimental configurations (Table S3), evaluated on 20,826 training atmospheric profiles. The mean optimal epoch for each case is reported alongside individual fold results; the overall mean is indicated for each model.

As illustrated in Fig. S4, Model 0 requires approximately twice as many training epochs to reach optimal performance as Model 1. This observation suggests that Model 0 continues to productively refine its learned representations throughout the training process, benefiting from extended optimisation. In contrast, Model 1 exhibits early convergence, whereby additional training epochs provide no further improvement in generalisation performance, consistent with a more constrained or noisy predictive target.

S2.5. Learning Curves for Model 0 (N_d) & Model 1 (P_d)

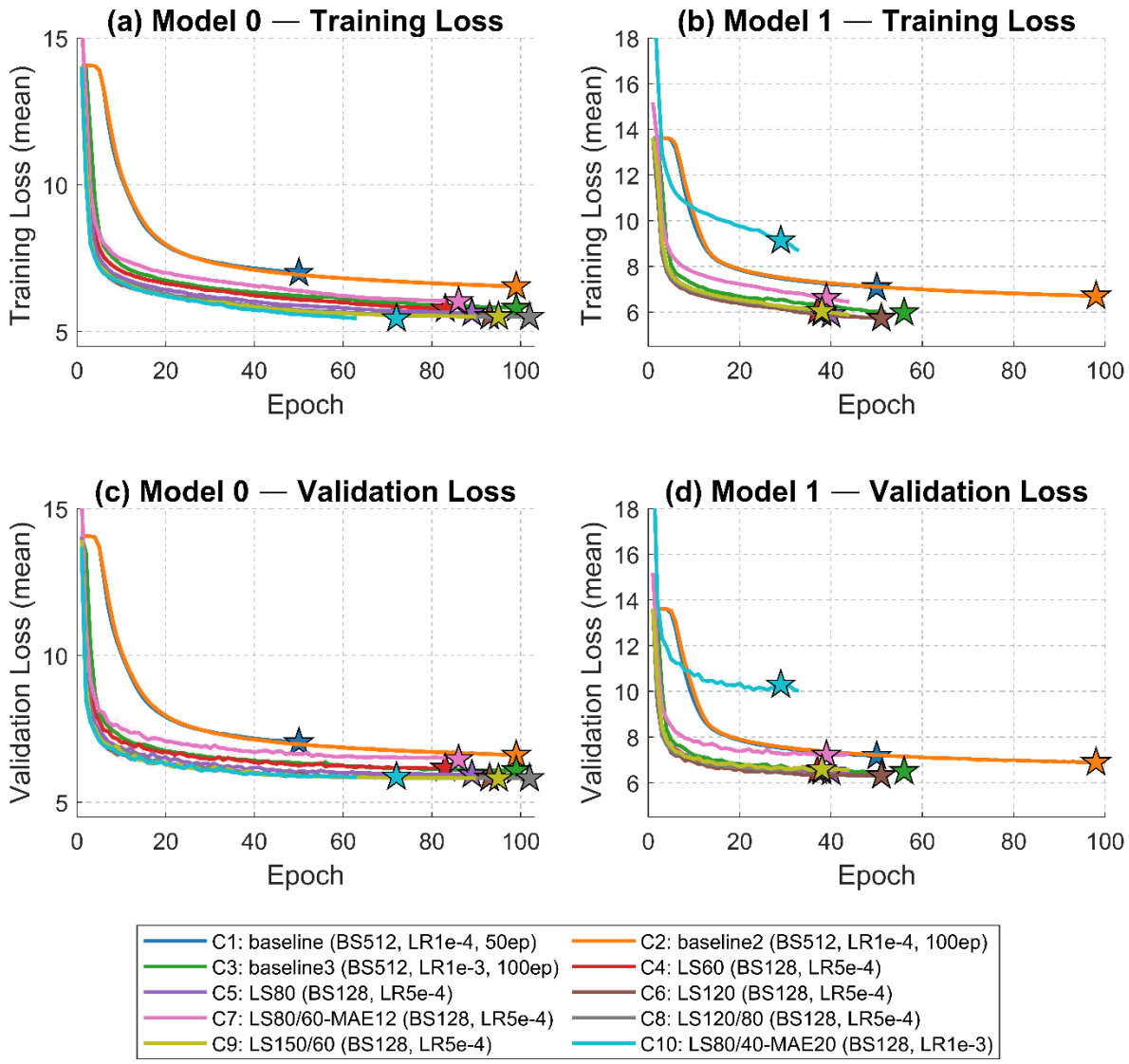

Figure S5. Learning curves for the 10 cases listed in Table S4. Curves represent mean loss across five folds; stars indicate the best epochs reported in Tables S4 and S5.

The training and validation loss curves for all 10 cases are presented in Figure S5. With the exception of Cases 1, 2, and 10, the loss curves converge to comparable values, indicating consistent model performance across varying hyperparameters. The smooth convergence of these curves further suggests that the models are not overfitting. The selected models (Case 8) also show smooth convergence to a solution. This finding is corroborated by the independent test data results presented below.

S2.6. Validation on Independent Test Data

To assess whether differences in validation loss across configurations correspond to meaningful improvements in predictive accuracy of reconstructed target profiles, all trained models were evaluated on an independent test dataset comprising 27,427 atmospheric profiles not used during training or cross-validation. Predictive accuracy was quantified using root-mean-square error (RMSE) and mean bias as a function of altitude. Results for all 10 experimental configurations, for both the Model 0 target (N_d) and the Model 1 target (P_d), are presented in Fig. S6.

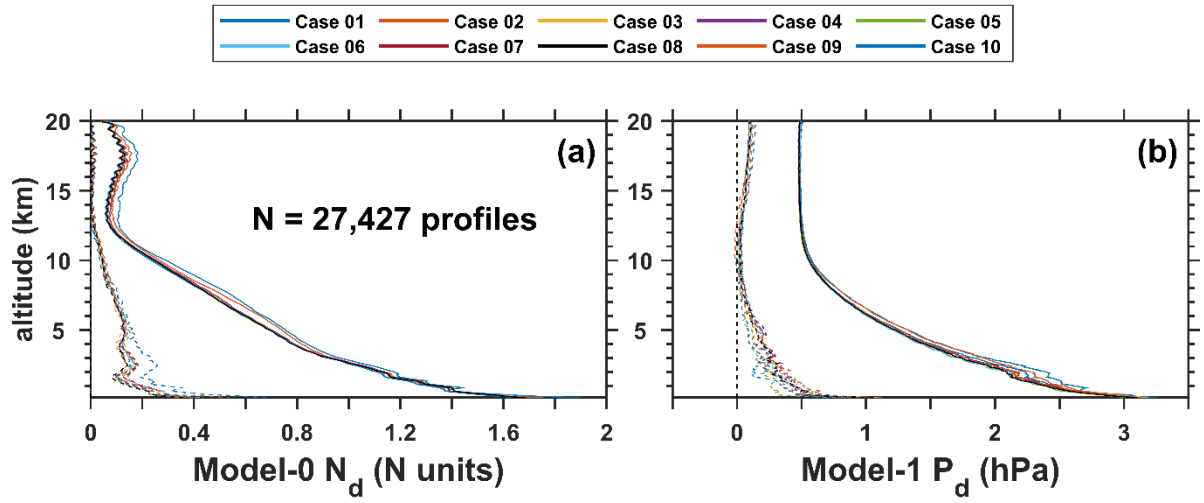

Figure S6. Model performance evaluated against independent in-situ radiosonde observations. Root-mean-square error (RMSE; solid lines) and mean bias (dashed lines) for N_d **(a)** and P_d **(b)**, computed across 27,427 independent radiosonde profiles spanning 100 m to 20 km altitude at 10 m vertical resolution. Profiles were drawn from the independent test dataset (Tables 3 and 4 in the main Section). Results are shown for each of the 10 experimental cases (Table.S2.1).

The vertical profile errors for temperature (T), water vapour pressure (P_w), and relative humidity (RH) are presented in Fig. S7.

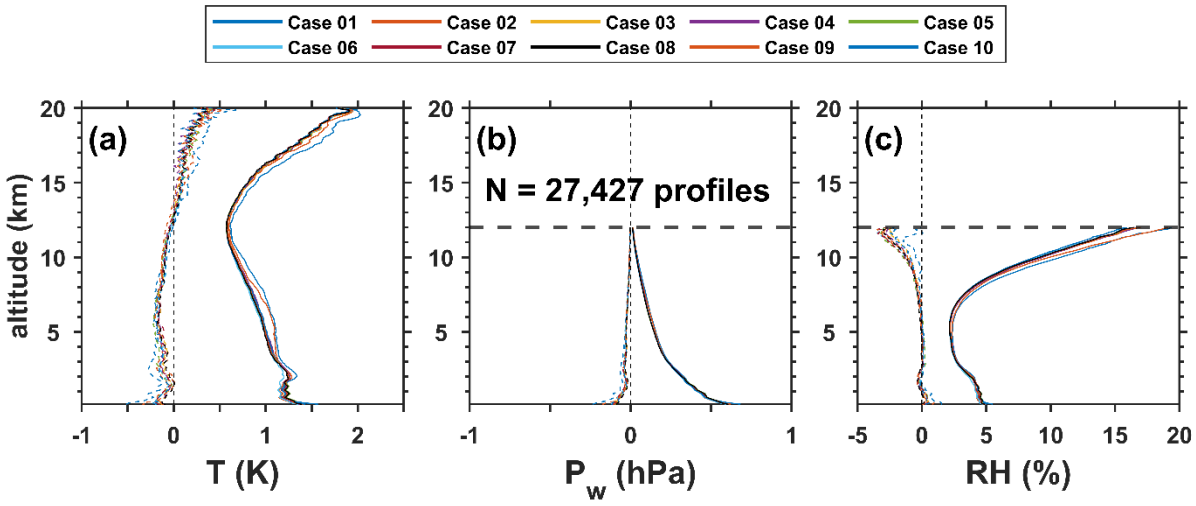

Figure S7. Model performance for derived atmospheric variables evaluated against independent radiosonde measurements. Root-mean-square error (RMSE; solid lines) and mean bias (dashed lines) for temperature (*T*; **a**), water vapour pressure (P_w; **b**), and relative humidity (*RH*; **c**), computed across 27,427 independent profiles spanning 100 m to 20 km altitude at 10 m vertical resolution (Tables 3 and 4 in main manuscript). Results are shown for all 10 experimental cases (Table.S3).

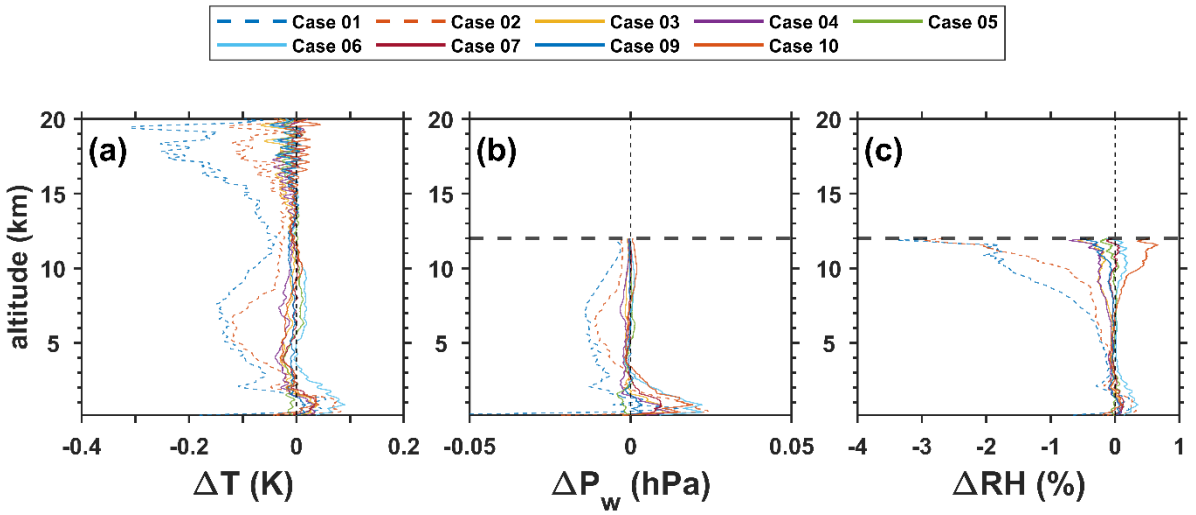

Figure S8. Pairwise RMSE differences relative to Case 08. Differences in RMSE between *Case 08* and each of the nine remaining cases (i.e., *Case 08 – Case 01*, and so on) are shown for temperature, water vapour pressure, and relative humidity, illustrating the insensitivity of predictive skill to the hyperparameter variations explored in this study.

Notably, despite substantial variation in validation loss across configurations, both Model 0 and Model 1 exhibited negligible differences in RMSE and mean bias on the independent test dataset. This result indicates that the sensitivity analysis principally captured differences in

training dynamics and validation-set fitting, rather than fundamental improvements in generalisation capability of profile reconstruction. The insensitivity of test-set metrics to hyperparameter variations suggests that the selected architecture (Table S2) possesses sufficient representational capacity to capture the underlying atmospheric relationships, and that marginal reductions in validation loss do not necessarily translate to enhanced predictive skill on unseen data.

Furthermore, configurations employing modified MAE loss weight coefficients (Cases 07 and 10) yielded no discernible improvement in test-set error metrics. This finding indicates that, within the tested range ($\alpha = 10–20$), the relative weighting of the Fourier spectral loss component and the point-wise MAE component does not significantly influence the model's ultimate predictive accuracy on independent data. The default MAE weight of $\alpha = 10$ is therefore sufficient, and systematic adjustment of the loss function weighting does not represent a viable pathway to performance enhancement for this application.

In aggregate, while the sensitivity analysis revealed meaningful differences in training efficiency and validation performance—particularly with respect to latent space dimensionality and learning rate—these differences converge to statistically equivalent predictive accuracy when evaluated on a large independent test dataset. This insensitivity implies that the predictive accuracy is fundamentally constrained by the inherent under-determined nature of the atmospheric retrieval problem and the physical information limits of the input profiles, rather than the representational capacity of the network architecture. Consequently, computational resources may be more efficiently directed towards alternative research directions, such as input feature engineering or training data augmentation, rather than continuing to scale model complexity.

On the basis of these results, **Case 08** (utilizing an asymmetric latent dimension of **120 for Model 0** and **80 for Model 1**) was selected as the preferred hyperparameter configuration. The selection is empirically justified by the divergent optimization landscapes of the two prediction targets. Model 0 (predicting N_d) must reconstruct a profile characterized by sharp, high-frequency vertical gradients, requiring a prolonged feature extraction phase (~100 epochs) and achieving its minimum validation loss (5.817) only at a higher representational capacity (latent dimension = 120). Conversely, Model 1 predicts dry pressure (P_d), a macroscopic variable characterized by a smooth, exponential vertical decay. Consequently, Model 1 reaches a representational plateau much earlier (~38 epochs); expanding its latent dimension beyond 80 yields no meaningful reduction in validation loss.

Therefore, the asymmetric 120/80 architecture of *Case 08* was selected to provide maximum representational headroom for the highly complex N_d profile, while capping the parameter space for the simpler P_d profile.

While the pairwise RMSE differences on the independent radiosonde test dataset (Fig. S8) reveal near-zero differences across configurations, this parity is a function of the limited variance within the in-situ training/testing datasets. The ultimate deployment target for this architecture is global GNSS-RO satellite observations, which are orders of magnitude larger and encompass a vastly wider distribution of atmospheric states. Selecting a higher-capacity latent space (dim = 120) for the primary refractivity target proactively ensures the network possesses the necessary representational capacity to absorb this expanded real-world variance. Because *Case 08* demonstrated robust generalization without overfitting on the restricted radiosonde dataset, it provides the optimal architectural headroom required to prevent underfitting when scaled to global satellite retrievals. Thus, *Case 08* was selected despite equivalent test performance to provide architectural symmetry with Part 2 satellite applications, acknowledging that this choice reflects anticipated scalability requirements rather than empirical superiority on the current dataset.

Section.S3 Error metrics for individual model evaluation stations

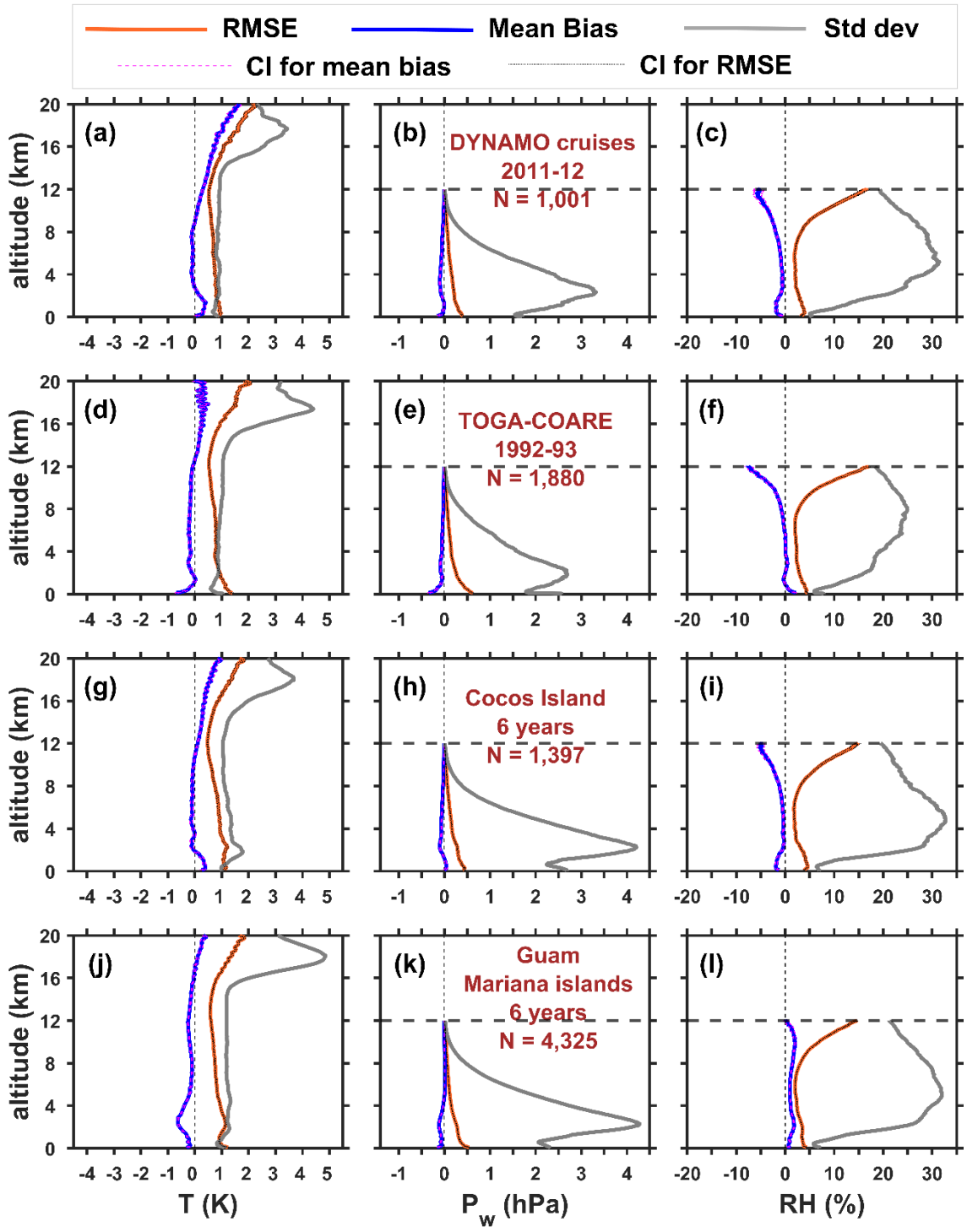

Fig. S9. Vertical profiles of RMSE, mean bias (Retrieved – Reference), and standard deviation (Std Dev) for temperature *(T)* **(a, d)**, water vapor pressure *(P_w)* **(b, e)**, and relative humidity *(RH)* **(c, f)** at 10 m vertical grid size from 100 m to 20 km. Standard deviation (grey lines) represents natural parameter variability in the measured profiles of *T*, P_w & *RH* for the

respective stations, providing context for the magnitude of retrieval errors. The 95% confidence intervals for mean bias and RMSE are plotted as regions bounded by their upper and lower limits (as indicated in the legend); however, owing to the large number of profiles, the separation between the upper and lower bounds is narrower than the line widths used in the plots and is therefore not visually discernible. Panels (**a–c**) represent retrievals from DYNAMO ship cruises (Table-4), (**d-f**) for TOGA-COARE data (Table-4), (**g-i**) for Cocos island data (Table-3), and (**j-l**) show for Guam Mariana Islands (Table-3). Model retrievals are compared against their respective radiosonde measurements. The number of radiosonde ascents used for each evaluation station is indicated.

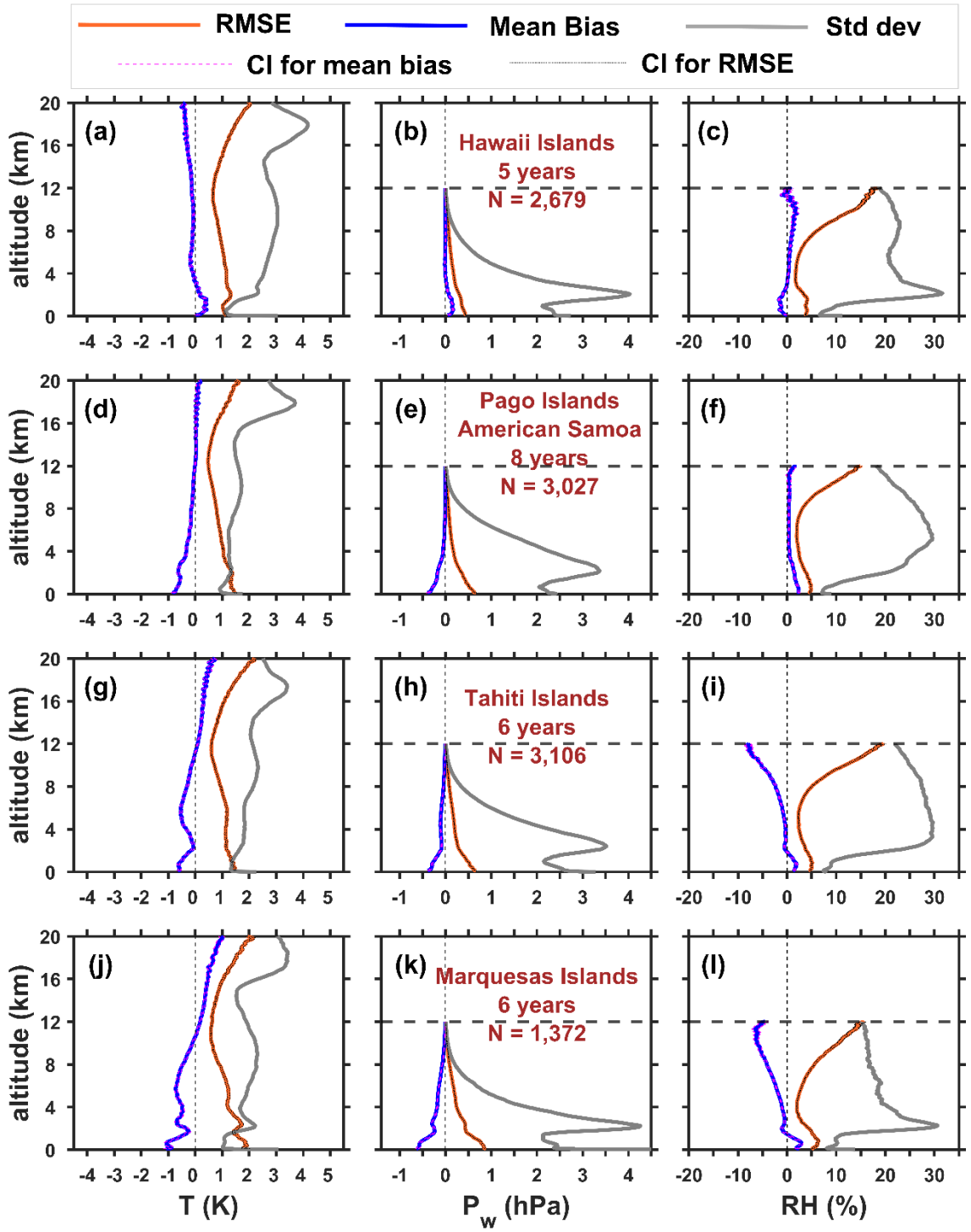

Fig. S10. Same as Fig.S9 but for Hawaii islands (**a-c**), Pago Islands (**d-f**), Tahiti Islands (**g-i**), and for Marquesas Islands (**j-l**). See Table-3 (in main manuscript) for details of each station.

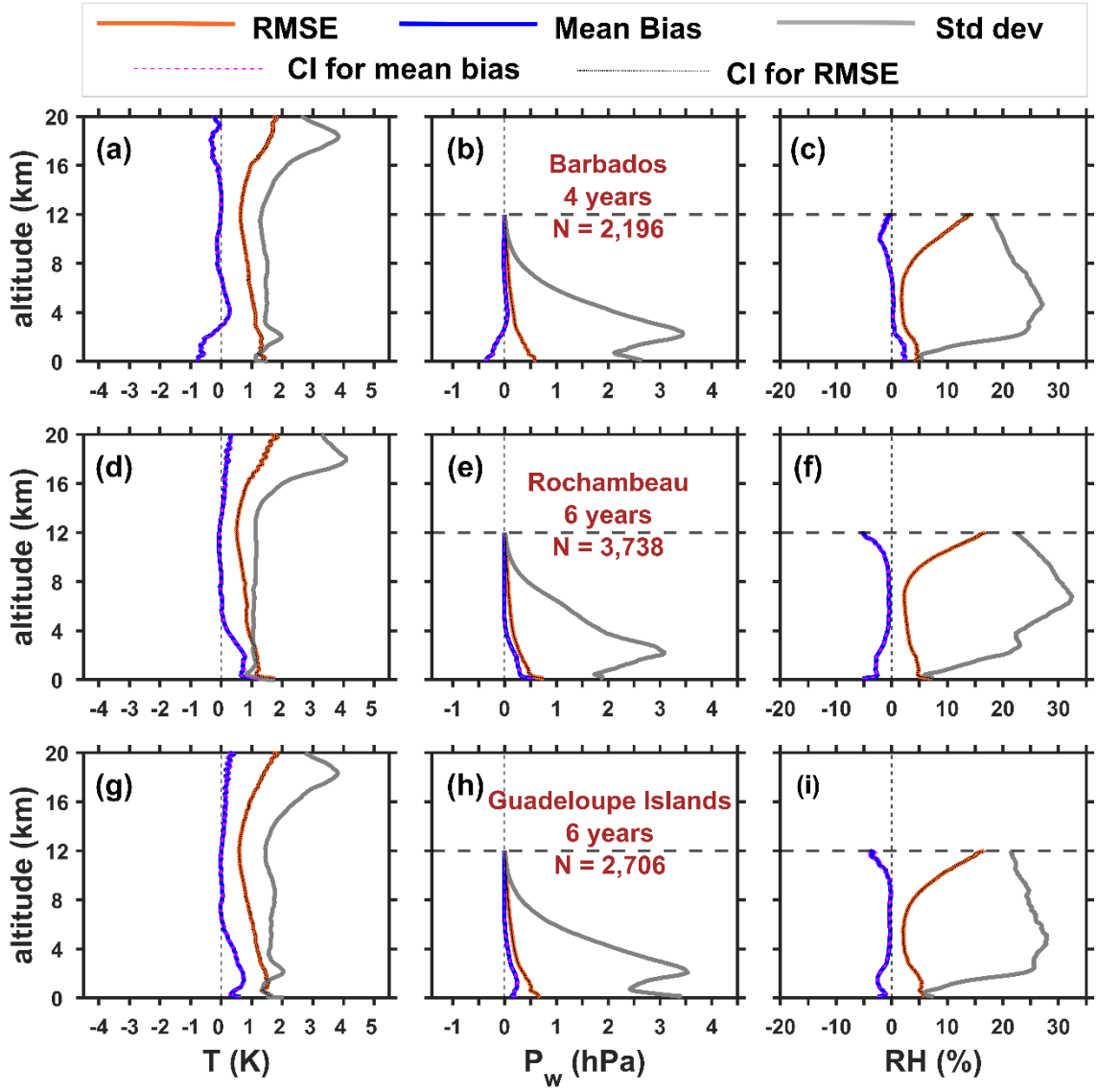

Fig. S11. Same as Fig.S9 but for Barbados (**a-c**), Rochambeau (**d-f**), and Guadeloupe Islands (**g-i**). See Table-3 (in main manuscript) for details of each station.

Section.S4 DYNAMO 2011-12 figures for P_w & T

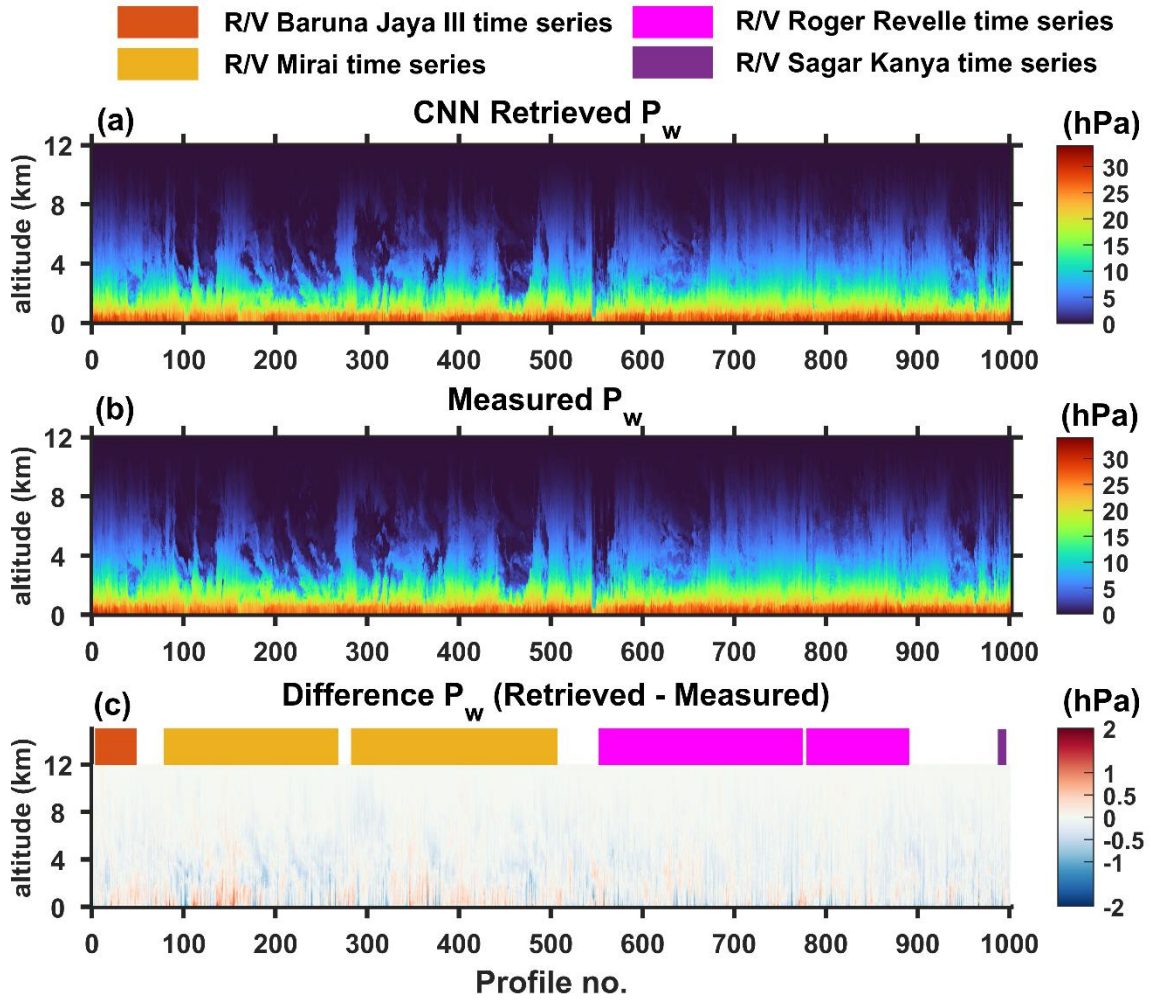

Figure S12 The CNN retrieved P_w (**a**) and the radiosonde measured P_w (**b**), at 10 m vertical resolution from 100 m to 20 km for 1001 profiles chronologically arranged for each ship cruise of DYNAMO 2011-12 campaign as per sequence listed in Table 4. The difference at each height for each profile forms the bottom panel (**c**). The shaded sections at the top of the bottom panel (**c**) show observations conducted as part of the time series for the cruises in the locations indicated in Table 4 & Fig. 2. At other times the observations were taken as ships traversed the tropical Indian Ocean.

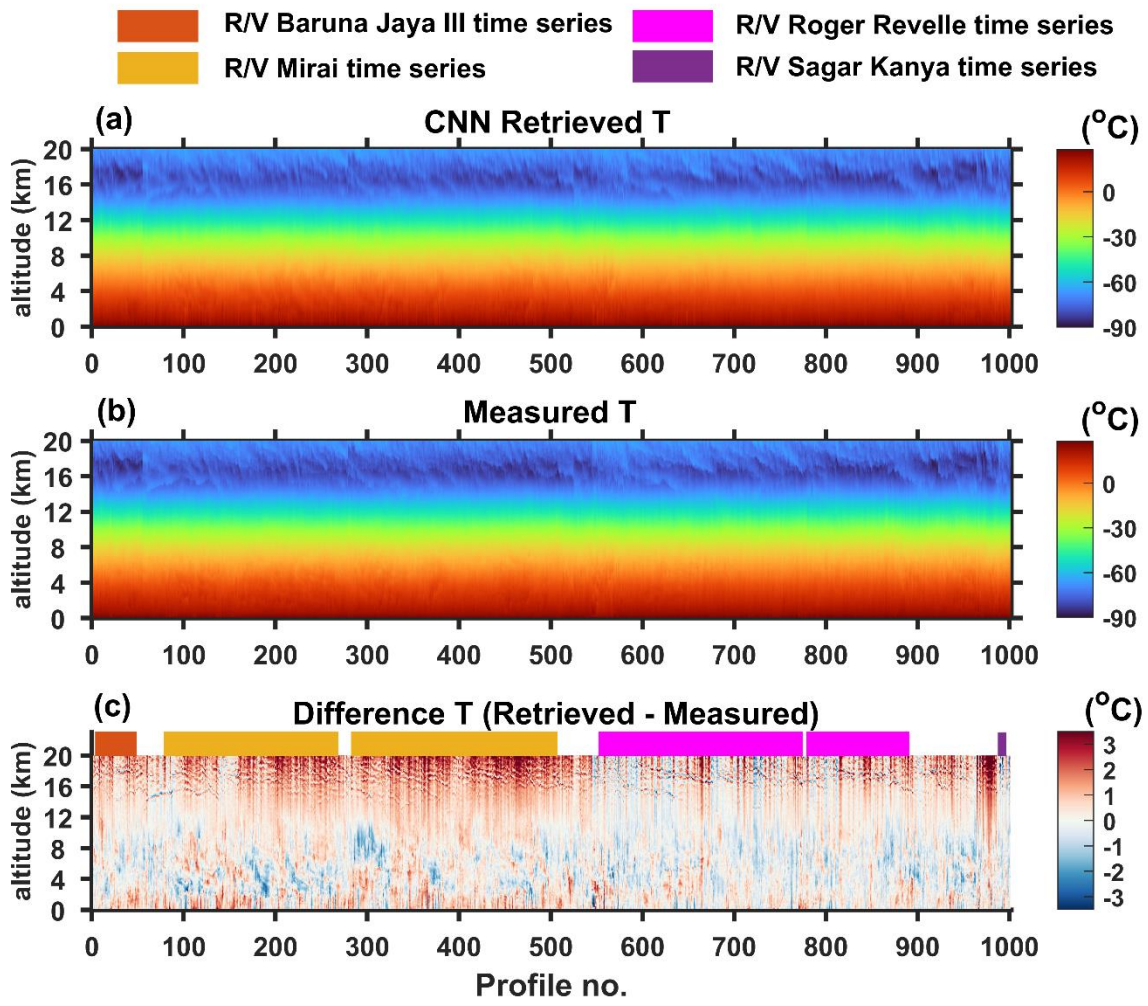

Figure S13 Same as Fig.S12 but for temperature.

Section.S5 Single-level input perturbation experiments on CNN model

S5.1.Overview

Atmospheric refractivity constitutes the sole vertical profile input to the CNN. Observational refractivity — whether derived from GNSS RO or radiosonde measurements — inevitably carries noise whose magnitude varies with measurement conditions. Three diagnostic questions motivate the controlled perturbation experiments described here. First, how robust are CNN retrievals to noise superimposed on the input refractivity profile? Second, when a localised single-level anomaly in *T* or *RH* is introduced and the resulting refractivity profile is supplied to the CNN, does the model correctly attribute the refractivity anomaly to moisture, temperature, or a linear combination thereof — despite the fact that temperature and water vapor contributions to N are inseparably fused within a single refractivity observation? Third, does a perturbation applied at one altitude propagate spuriously into retrievals at adjacent levels? Because each perturbation is applied to a known variable at a precisely defined altitude (Table S6), the experiments directly probe the model's capacity to disentangle superimposed thermodynamic signals within refractivity. Beyond retrieval diagnostics, these experiments serve as a behavioural probe into the internal workings of an otherwise opaque black-box operator, offering interpretable insight into what the model has learned from its training data. Collectively, these single-level sensitivity tests are designed to characterise both the attribution fidelity and the vertical locality of CNN retrievals.

Controlled perturbations to the vertical profile inputs of the CNN form the basis of this diagnostic analysis. Two vertical profile quantities enter the CNN: total refractivity N and its wavelet covariance transform WCT_N.

As outlined in Sect. 2.1 of main manuscript, both N and WCT_N are computed from radiosonde observations of pressure, temperature, and relative humidity. The DL model predicts dry refractivity (N_d) and dry pressure (P_d), from which T, P_w, *and RH* are analytically derived and subsequently compared against their measured counterparts.

A single radiosonde ascent serves as the anchor for all perturbation experiments: the sounding acquired aboard R/V Mirai on 18 October 2011 at 1200 UTC at (80.5°E, 0°) over the tropical Indian Ocean (profile 219, Fig. 6). Measured *T, RH*, and the analytically derived P_w from this ascent, together with the corresponding N and WCT_N computed from it, constitute the unperturbed baseline (*original* observation), hereafter denoted T_{org}, RH_{org}, $P_{w\ org}$, and N_{org} respectively. CNN *retrievals* obtained from these unperturbed inputs — designated T_{ret}, P_{wret}, and

RH_{ret} — define the reference (*original*) retrieval state and are presented in Fig. S14 alongside the pointwise differences between retrieved and measured quantities (e.g., $\Delta T_{org} = T_{ret} - T_{org}$; bottom panels).

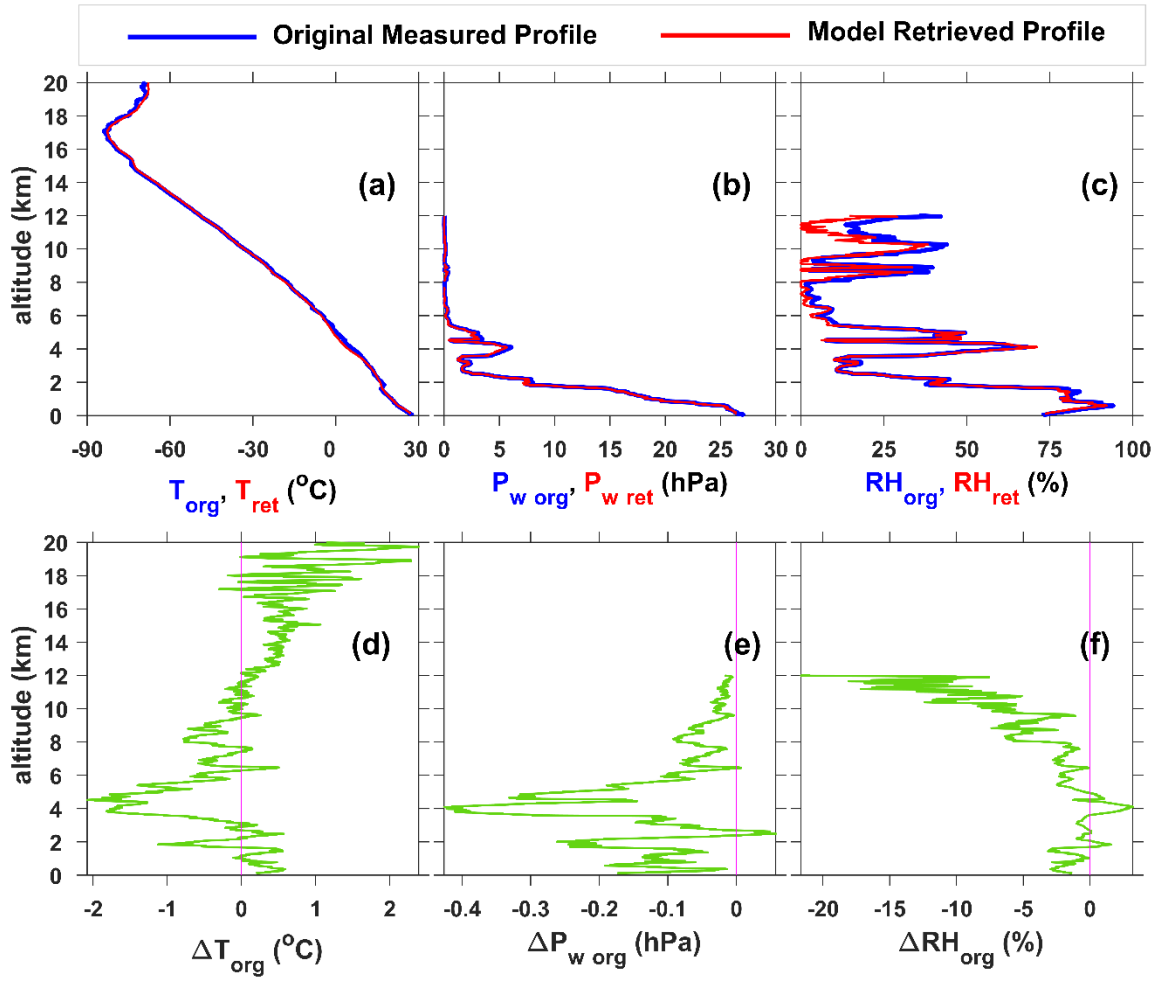

Figure S14. Unperturbed baseline profiles for the R/V Mirai ascent (18 October 2011, 1200 UTC; 80.5°E, 0°). Panels **(a–c)**: radiosonde-measured T_{org}, $P_{w\,org}$, and RH_{org} (blue) overlaid with CNN retrievals T_{ret}, P_{wret}, and RH_{ret} (red). Panels **(d–f)**: pointwise departure of each retrieved quantity from its measured counterpart (ΔT_{org}, $\Delta P_{w\,org}$, ΔRH_{org}), serving as the reference error against which perturbation-induced changes are assessed. All profiles span 100 m–20 km at a 10 m vertical grid interval.

A targeted anomaly is subsequently superimposed on either T or RH at a prescribed altitude, yielding the modified profiles T_{pert} and RH_{pert}. Refractivity and its wavelet covariance transform are recomputed from these modified thermodynamic profiles, producing N_{pert} and WCT_{Npert}; the corresponding water vapor pressure $P_{w\,pert}$ is obtained simultaneously. The pair (N_{pert}, $WCT_{N\,pert}$) is then passed to the CNN in place of the unperturbed inputs. CNN outputs from this perturbed forward pass — denoted T_{ret2}, RH_{ret2}, and $P_{w\,ret2}$ — are evaluated against T_{pert}, RH_{pert}, and $P_{w\,pert}$ to quantify how faithfully the model recovers the imposed anomaly

S5.2. Sensitivity test cases

The perturbation tests are categorized into four independent cases to evaluate effects in the lower and upper troposphere. The magnitudes of the perturbations are kept unrealistically high to test the limits of their effects on DL model retrievals. The DL model was trained only on unperturbed profiles (from Table-S6). All input and output profiles for the DL model used in these sensitivity tests span an altitude range of 100 m to 20 km at a vertical grid spacing of 10 m as described in Sect.2.3.

Four distinct experimental scenarios probe the CNN response across both the lower and upper troposphere. Anomaly amplitudes are deliberately set well beyond physically realistic bounds, ensuring that the diagnostic exercises stress-test the outer limits of model behaviour rather than representing atmospherically plausible states. Crucially, the CNN was exposed exclusively to unperturbed radiosonde profiles during training (Table S6), so the perturbed inputs constitute genuinely out-of-distribution stimuli. Consistent with the preprocessing protocol of Sect. 2.3, all vertical vectors supplied to and extracted from the CNN — in both the baseline and perturbed configurations — are defined on the standard 100 m–20 km domain at 10 m intervals

Table S6: Sensitivity test cases:

Case 1	A Delta increase of +20% added to RH_{org} at exactly 10000 m altitude (Fig.S15(c))
Case 2	A Delta increase of +20% added to RH_{org} at exactly 3000 m altitude (Fig.S15(f))
Case 3	A Delta increase of +2^{o}C added to T_{org} at exactly 10000 m altitude (Fig.S15(g))
Case 4	A Delta increase of +2^{o}C added to T_{org} at exactly 3000 m altitude (Fig.S15(j))

Each experiment modifies only one thermodynamic variable — either T or RH — with the other held fixed. Because P_w is a function of both, any single-variable perturbation necessarily propagates into P_w; CNN-retrieved P_w is therefore evaluated against this analytically consistent perturbed counterpart. Taken together, **Cases 1–4** characterise the CNN response to a localised, single-altitude anomaly in the input refractivity.

S5.3. Results

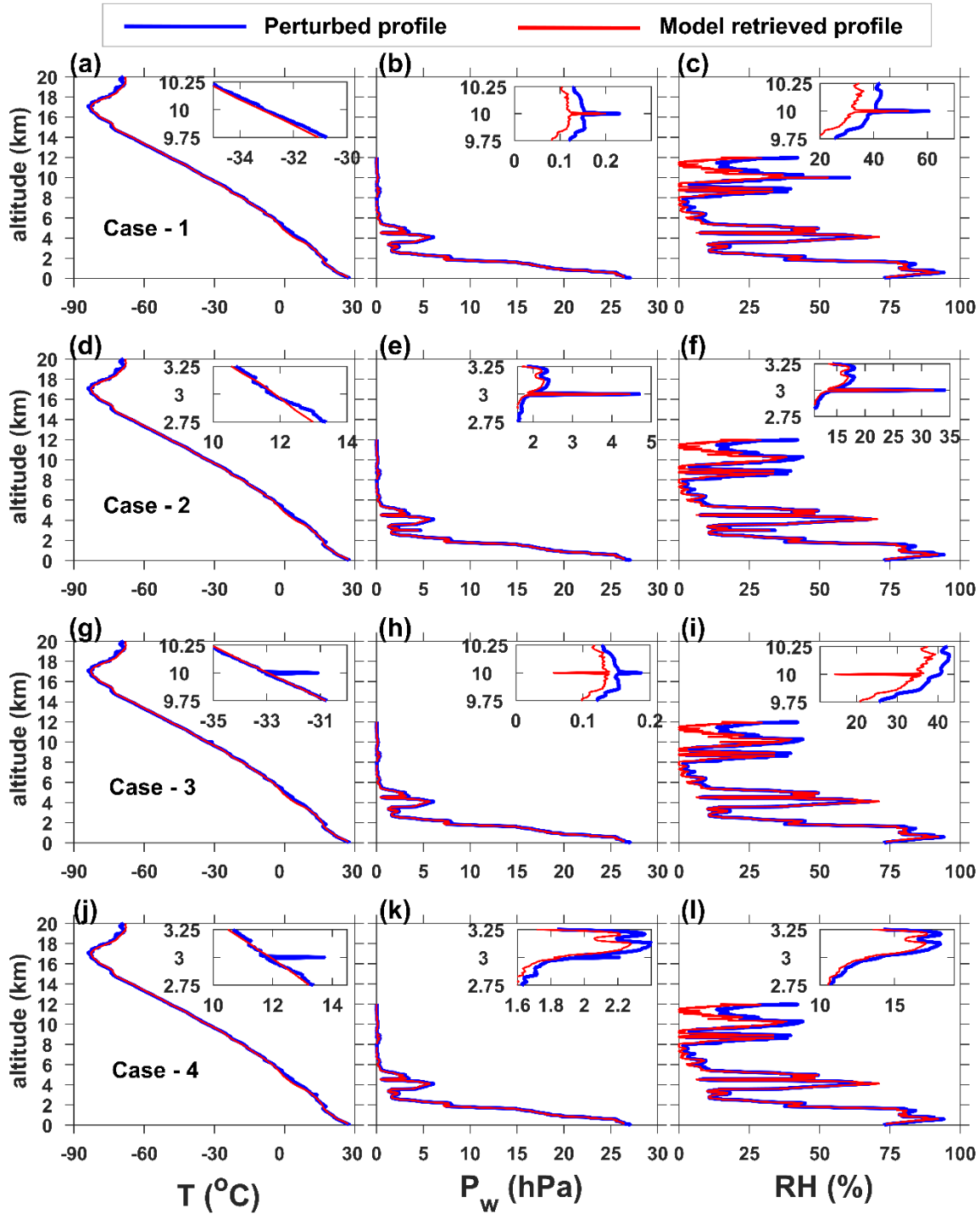

Figure S15. Side-by-side comparison of analytically constructed perturbation targets (T_{pert} , $P_{w\ pert}$ & RH_{pert}) and CNN-retrieved quantities (T_{ret2} $P_{w\ ret2}$ & RH_{ret2}) for all four experimental scenarios: Case 1 (panels **a–c**), Case 2 (panels **d–f**), Case 3 (panels **g–i**), and Case 4 (panels **j–l**). An expanded view of the perturbed altitude neighbourhood is provided as an inset within each panel; inset and main-panel ordinate and abscissa scales are identical.

Two contrasting behaviours emerge from the four cases. For moisture perturbations (Cases 1 and 2), the imposed *RH* anomaly is faithfully captured in both CNN-retrieved *RH* (Figs. S15(c), (f)) and the analytically consistent P_w (Figs. S15(b), (e)), while *T* retrievals remain unaffected (Figs. S15(a), (d)). Conversely, for thermal perturbations (Cases 3 and 4), the CNN largely disregards the imposed *T* anomaly (Figs. S15(g), (j)), instead manifesting the refractivity signal as a co-located spike in *RH* (Figs. S15(i), (l)) and P_w (Figs. S15(h), (k)). Two inferences follow: the CNN does not recover sharp, isolated discontinuities in temperature, yet it successfully resolves 10 m scale anomalies in moisture. Across all four cases, both perturbed and CNN-retrieved profiles remain close to the unperturbed baseline of Fig. S14.

The inter-retrieval difference — quantifying how much the CNN output shifts between the unperturbed and perturbed forward passes — is examined in Fig. S16, which additionally reveals whether a single-level perturbation propagates spuriously into retrievals at neighbouring altitudes. Defining $\Delta T_{org} = T_{ret} - T_{org}$ *and* $\Delta T_{pert} = T_{ret2} - T_{pert}$, their difference $\Delta T_{org} - \Delta T_{pert} = T_{ret} - T_{ret2} + (T_{pert} - T_{org})$. Because $(T_{pert} - T_{org})$ is non-zero only at the single perturbed level, $\Delta T_{org} - \Delta T_{pert}$ reduces to the pure inter-retrieval contrast ($T_{ret} - T_{ret2}$) at all other altitudes, isolating the CNN response to the perturbation from pre-existing retrieval error and thereby directly exposing any vertical leakage in the retrievals. Analogous expressions of subscripts apply to P_w and *RH* in Fig. S16.

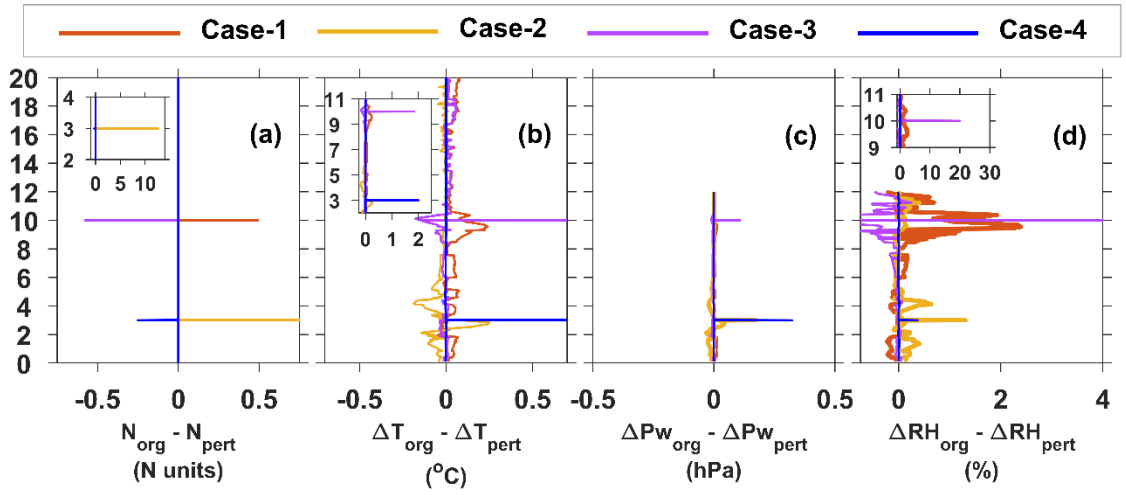

Figure S16. (a) Difference between perturbed and original refractivity (Δ*N*) profile for *Cases 1-4*. RH increase manifests as positive difference, and temperature increase manifests as negative difference. Difference between the CNN model retrievals of the perturbed and original temperature **(b)**, water vapor pressure **(c)**, and relative humidity **(d)** profiles are shown. $\Delta T_{org} = T_{ret} - T_{org}$ *and* $\Delta T_{pert} = T_{ret2} - T_{pert}$. The differences for P_w *and RH* are also defined similarly. The inset

in each plot shows the region of interest. The axes of the inset plots have the same units as the main axes.

Each of the four experiments introduces a localised anomaly into the refractivity input (Fig. S16(a)). A benchmark for ideal CNN behaviour is that retrievals at altitudes other than the perturbed level should remain entirely unaffected; any residual signal elsewhere constitutes undesirable vertical leakage. Panels b–d of Fig. S16 demonstrate that this leakage is negligible: away from the perturbed level, T deviates by no more than 0.1 K, P_w by no more than 0.05 hPa, and RH by no more than 1% — all well within the respective measurement uncertainties of Fig. 1 of the main text. The moisture signal itself is recovered with high fidelity: the imposed 20% delta function anomaly in RH is reproduced to within 2% in CNN-retrieved RH (Fig. S16(d)), below the one-sigma radiosonde RH uncertainty, and this holds for both the lower-tropospheric scenario (Case 2) and its upper-tropospheric counterpart (Case 1). Vertical confinement of the perturbation response is largely maintained, though upper-tropospheric cases exhibit marginal leakage over a ±1 km window; the associated magnitudes remain a small fraction of the measurement uncertainty and are therefore inconsequential. Collectively, the four experiments confirm that the CNN exhibits near-ideal vertical locality in its response to sharp, single-level anomalies.

S5.3.1. Model Response for RH Perturbations

- For delta-function increases in RH (cases 1 and 2), the model precisely detected the altitude-specific changes. Retrievals showed an accurate reproduction of the delta increase in RH with less than 3% deviation (Figs. S15 (c), & (f)). The effects were confined to the perturbed altitude, with largely no spill over to adjacent altitudes.
- The model demonstrated a strong capability to accurately reproduce RH perturbations. Model response is largely localized to perturbation altitudes.

S5.3.2. Model Response to T perturbations

- For cases 3 and 4, the sharp T increase was ignored, with RH & P_w retrievals showing a sharp decrease at the same altitude. This behaviour suggests the model attributes refractivity changes to water vapor rather than T increases which are not observed in real atmosphere with the model response largely confined to perturbing altitude within limits shown in Fig.S16.

Summary

Abrupt changes in the vertical profile of *RH* frequently occur in the tropical troposphere due to moisture advection, stratification, and phase changes, leading to sharp vertical gradients. In contrast, temperature profiles generally vary much more smoothly over small vertical scales. Consequently, within the radiosonde training manifold, high-frequency (e.g., 10 m scale) variance in refractivity (*N*) is statistically dominated by moisture fluctuations rather than temperature fluctuations. The sensitivity tests demonstrate that the deep learning model successfully maps this dominant statistical covariance; it acts effectively as a low-pass filter for temperature, attributing sharp, high-frequency perturbations in *N* almost exclusively to the moisture output channels (*RH* and P_w)

Although the model directly predicts only N_d and P_d, the resulting vertical structures of *T,* P_w, and *RH* are physically consistent, owing to the direct thermodynamic relationship between the model targets (N_d, P_d) and the retrieved atmospheric variables (*T,* P_w, and *RH*) as shown in Sect.2.1. Single-level perturbations to *T* or *RH* show only negligible vertical propagation (which are a fraction of the measurement uncertainty shown in Fig.1(b) & 1(c)) beyond the altitude of perturbation (as demonstrated in cases 1 to 4 Fig.S16).

In Cases 1 and 2, the DL model accurately reproduced the delta increases in RH, demonstrating sensitivity to moisture structures on the order of 10 m in the vertical. However, in Cases 3 and 4, the model largely smoothed over the temperature perturbations, wrongly attributing the resulting local drop in N into an artificial decrease in retrieved moisture. Rather than incorporating physical plausibility, this behavior highlights a fundamental characteristic and limitation of the statistical retrieval: because the network aggressively partitions high-frequency N signals to moisture, it risks aliasing genuine, sharp temperature inversions (such as those occasionally found at the top of the marine boundary layer) into erroneous moisture anomalies. Recognizing this differential response — where the network captures moisture at fine vertical scales but temperature only at macroscopic scales — is critical for correctly interpreting the model outputs. Nonetheless, the results for Cases 1 and 2 (as well as those in Sect. 3.2.2 of the main text) demonstrate the model's robust capability to extract localized moisture features from refractivity inputs within the troposphere.

Section S6: Justification for WCT_N as an Explicit Model Input

S6.1 Rationale

While convolutional neural networks can theoretically learn gradient-sensitive features through convolutional filters, we demonstrate that explicit provision of the wavelet covariance transform of refractivity (WCT_N) is necessary for robust retrievals from limited training data (~20,800 radiosonde ascents) in the current architecture.

The wavelet covariance transform using the Haar wavelet function with a dilation of 150 m (Eq. 7–8 in main text) captures sharp vertical transitions in refractivity. This scale corresponds to characteristic features of the tropical troposphere: boundary layer tops, trade wind inversions, humidity discontinuities at cloud tops, and entrainment zones (Santosh (2022) and references therein) —all of which affect the vertical structure of refractivity and its relationship to temperature and water vapor.

S6.2 Ablation Experiments

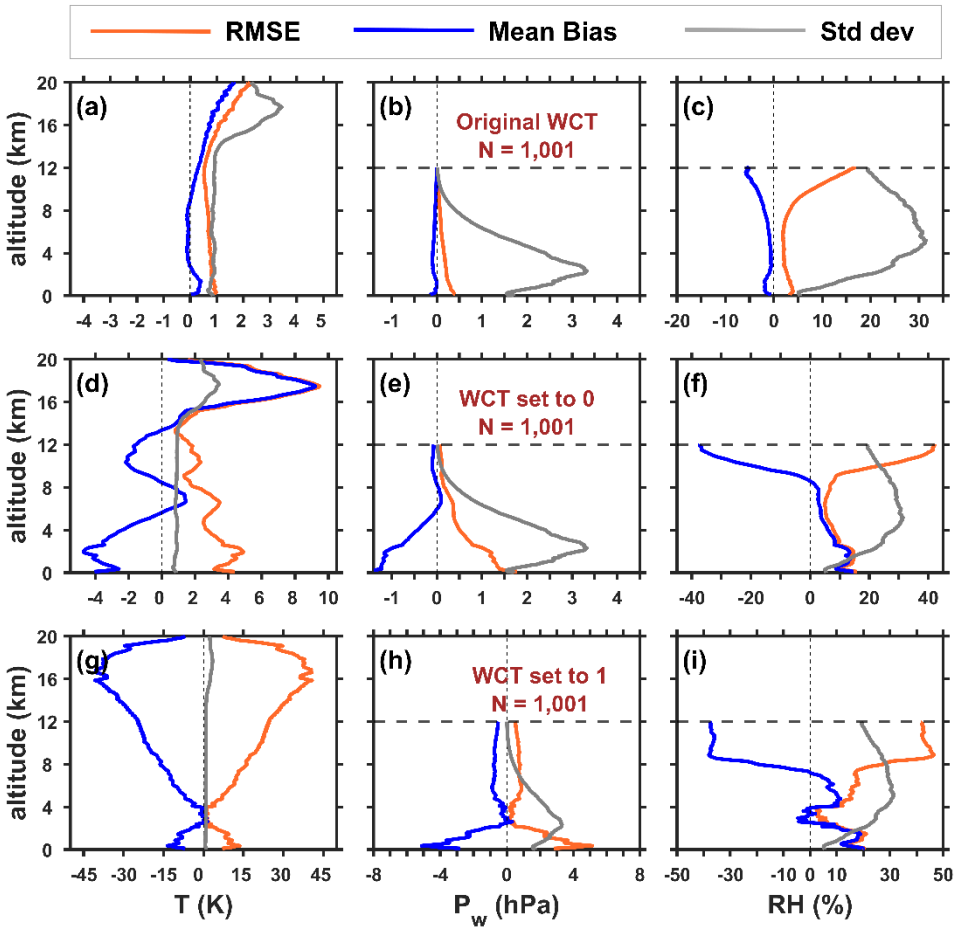

Figure S17. Vertical profiles of RMSE, mean bias (Retrieved – Reference), and standard deviation for retrieved parameters temperature (T), water vapor pressure (P_w), and relative humidity (RH) at 10 m vertical resolution from 100 m to 20 km using DYNAMO cruise data (Table 3; N = 1,001 profiles). **Panels (a–c)**: retrievals using correct WCT_N values. **Panels (d–f)**: WCT_N set to 0 at all altitudes (removal ablation). **Panels (g–i)**: WCT_N set to 1 at all altitudes (saturation ablation). Removal of WCT_N increases RMSE by >300%; saturation increases errors by >1000% (for T), confirming that explicit WCT_N is essential for accurate retrievals.

The necessity of explicit WCT_N inclusion is validated through two ablation experiments on the 1,001 DYNAMO radiosonde profiles:

Removal ablation ($WCT_N = 0$): Setting WCT_N to zero at all altitudes for all profiles eliminates explicit gradient information. WCT_N has a small but non-zero value above 10 km. This produces a **>300% increase in RMSE** for temperature (from ∼1 K to ∼4 K), and > 200% increase in RMSE for water vapor pressure (up to 1.5 hPa), and comparable RMSE increase in relative humidity in the lower troposphere. Strong altitude-dependent biases emerge: temperature bias ranges from −4 K to +9 K; water vapor pressure bias reaches −1.5 hPa; relative humidity bias reaches 10% (Fig. S17 (d–f)).

Saturation ablation (adversarial perturbation; $WCT_N = 1$): Setting WCT_N to unity at all altitudes for all profiles creates a physically implausible uniform high-gradient state absent in training data. This produces **errors exceeding 1000% (for T) and substantial increases in errors for P_w and RH at all altitudes**, rendering retrievals unusable (Fig. S17(g–i)).

The scaling of degradation (300% → 1000%) with degree of physical implausibility suggests that WCT_N constrains the model to valid atmospheric states, not merely provides supplementary information.

We acknowledge that these ablation experiments demonstrate learned dependency rather than fundamental necessity. Thus, for operational purposes, the demonstrated dependency establishes that explicit WCT_N is required for accurate retrievals with the current architecture.

S6.3 Summary

The ablation experiments establish that:

1. WCT_N provides non-redundant information for both the models.

2. Removal of WCT_N degrades retrieval accuracy by >300% for T, P_w & RH.

3. Providing physically inconsistent WCT_N values renders retrievals unusable.

Explicit inclusion of WCT_N with 150 m dilation encodes domain knowledge about tropical tropospheric structure that the CNN cannot reliably discover from limited training data. A CNN learning gradients *ab initio* would need to identify this optimal scale from the data alone. By providing WCT_N explicitly, the model receives gradient information at the physically appropriate scale for tropical tropospheric retrieval, reducing the hypothesis space and ensuring robust performance of the model during training (see learning curves in Fig.S5). Together, these findings demonstrate that explicit inclusion of WCT_N profile as model input is not merely beneficial but essential for robust atmospheric retrieval in the current architecture with the available training data.

Section.S7 References Supplementary Sections

Fuoli, D., Van Gool, L., and Timofte, R.: Fourier Space Losses for Efficient Perceptual Image Super-Resolution, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2340–2349, https://doi.org/10.1109/ICCV48922.2021.00236, 2021.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/ARXIV.1412.6980, 2014.

Santosh, M.: Estimation of daytime planetary boundary layer height (PBLH) over the tropics and subtropics using COSMIC-2/FORMOSAT-7 GNSS–RO measurements, Atmospheric Research, 279, 106361, https://doi.org/10.1016/j.atmosres.2022.106361, 2022.

Yadav, O., Ghosal, K., Lutz, S., and Smolic, A.: Frequency-domain loss function for deep exposure correction of dark images, SIViP, 15, 1829–1836, https://doi.org/10.1007/s11760-021-01915-4, 2021.