remote sensing

MDPI

Article

Dual Perception of Space and Spectrum: Reconstructing 3D Temperature and Humidity Fields via Heterogeneous Sensor Fusion

Lei Tang [1,2], Xiangang Zhao [3,4,5,*], Yikai Li [1,2], Cunqun Fan [3,4,5], Manyun Lin [3,4,5], Xi Zhang [3,4,5], Xiangyu Chang [3,4,5], Baoquan Zhang [1,2] and Yunming Ye [1,2]

1 Shenzhen Key Laboratory of Internet Information Collaboration, Harbin Institute of Technology, Shenzhen 518055, China; 25b351007@stu.hit.edu.cn (L.T.); 23s151148@stu.hit.edu.cn (Y.L.); baoquanzhang@hit.edu.cn (B.Z.); yeyunming@hit.edu.cn (Y.Y.)
2 School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
3 National Satellite Meteorological Center (National Center for Space Weather), Beijing 100081, China; fancq@cma.cn (C.F.); linmy@cma.gov.cn (M.L.); zhangx@cma.gov.cn (X.Z.); changxy@cma.gov.cn (X.C.)
4 Innovation Center for FengYun Meteorological Satellite (FYSIC), Beijing 100081, China
5 Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites, China Meteorological Administration (CMA), Beijing 100081, China
* Correspondence: xgzhao@cma.gov.cn

Highlights

What are the main findings?

- MS3D integrates adaptive patch embedding and gated cross-modal fusion for multi-sensor 3D atmospheric field retrieval.
- MS3D reduces temperature and specific humidity MAEs by 10.8% and 7.9%, respectively, compared with ViT.

What are the implications of the main findings?

- Infrared–microwave complementarity improves atmospheric retrieval robustness under cloudy conditions.
- Spatial-context modeling enables coherent reconstruction of three-dimensional atmospheric temperature and humidity fields.

Abstract

The accurate acquisition of meteorological variables, such as temperature and humidity, is critical for diverse societal and industrial applications. Meteorological retrieval, defined as the process of deriving atmospheric variables from satellite-observed radiances, remains a pivotal research domain. Existing meteorological retrieval generally relies on single-sensor or point-wise approaches. Single-sensor systems lack full vertical coverage due to spectral limits, and point-wise retrieval ignores the spatial context required to capture horizontal atmospheric flow. To overcome these dual limitations, we propose a framework that integrates multi-sensor fusion with direct 3D spatial retrieval. Our approach not only leverages the complementarity of infrared hyperspectral and microwave data to penetrate clouds and resolve vertical structures, but also utilizes 3D modeling to explicitly capture spatial correlations. This synergy is particularly effective for dynamic variables like humidity. By simultaneously expanding the spectral and spatial fields of view, our model successfully reconstructs a robust and continuous 3D atmospheric field. Quantitatively, under the ERA5-based evaluation, MS3D reduces the MAE of temperature and specific humidity by 18.9% and 23.9%, respectively, compared with the pixel-wise IReA baseline, and by 10.8% and 7.9%, respectively, compared with the 3D-field ViT baseline. Furthermore,

Check for updates

Academic Editors: Qian Liu and Manzhu Yu

Received: 11 September 2026
Revised: 02 November 2026
Accepted: 06 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

experiments conducted under cloudy conditions validate that the 3D meteorological fields constructed via multi-sensor fusion possess the robustness to effectively handle complex meteorological scenarios.

Keywords: temperature profile retrieval; humidity profile retrieval; Hyperspectral Infrared Atmospheric Sounder-II (HIRAS-II); Micro-Wave Humidity Sounder-II (MWHS-II); Micro-Wave Temperature Sounder-III (MWTS-III)

1. Introduction

Meteorological elements denote the atmospheric parameters that characterize its physical state and dynamic processes, primarily including temperature, pressure, wind, and humidity. Satellite observations serve as the primary means for acquiring meteorological data, enabling continuous, all-weather global monitoring. Specifically, satellites measure atmospheric radiances; the subsequent process of decoupling meteorological state variables from these observed radiances is referred to as meteorological retrieval. Accurate retrieval of these elements provides precise initial conditions for numerical weather prediction (NWP) models, thereby playing a critical role in enhancing forecast accuracy [1–3].

With the rapid advancement of artificial intelligence, employing deep learning for meteorological retrieval has emerged as a prominent trend. Unlike statistical retrieval methods, which are often constrained by simplistic linear or non-linear models and frequently suffer from insufficient accuracy [4], and distinct from physical retrieval methods that require explicit consideration of complex physical processes—leading to high computational complexity and instability due to sensitivity to initial conditions—neural network-based approaches operate in a data-driven manner. Deep learning retrieval methods directly learn the intrinsic mappings between satellite radiances and meteorological variables. However, existing deep learning retrieval methods primarily face two limitations: (i) Neglect of 3D Spatial Structure: Pixel-wise retrieval methods ignore the inherent 3D structure of atmospheric fields [5]. Satellite radiance observations should be modeled as a unified 3D input rather than processed in isolation, thereby preserving spatial continuity. (ii) Underutilization of Multi-Sensor Data: Despite the abundance of satellite sensors, retrieval methods rely predominantly on single sensors [6], failing to leverage complementary multi-source information. Furthermore, integration is often limited to basic pixel-level concatenation [7], which struggles to capture complex non-linear features, leaving the latent value of multi-sensor data untapped. Consequently, these methods perform poorly in complex scenarios, particularly under cloudy conditions.

In response to these limitations, we introduce the MS3D (Multi-Sensor 3D fusion) retrieval framework, which integrates infrared and microwave sensors for 3D-field retrieval, as shown in Figure 1. The network utilizes 3D image modeling to capture spatial correlations, leveraging information from adjacent points to assist in retrieving fluid dynamics, particularly for humidity. To address spectral deficiencies in single-sensor data, we exploit the rich information from hyperspectral infrared sensors to finely model the atmospheric vertical structure, while incorporating the cloud-penetrating ability of microwave sensors to complete spectral data beneath cloud layers. This design equips the model with a robust field of view, guaranteeing the overall integrity and structural consistency of the atmospheric retrieval. A key challenge in implementing this architecture lies in effectively aligning sensors with varying spatial resolutions and fusing spectral information across different sensors. To overcome this, we employ adaptive patch embedding to align data from different sensors in physical space, maintaining spatial consistency. Given the physical

mechanism differences among HIRAS, MWHS, and MWTS, we utilize a Gated Cross-Fusion Module (GCFM) to control interactions between feature maps at different scales. This addresses challenges related to feature misalignment and noise conflict, preventing the over-homogenization of multi-modal features and ensuring the decoder receives diverse information. Distinct from previous pixel-wise retrieval methods and single-sensor 3D-field retrieval methods, we realize 3D atmospheric reconstruction by fusing multi-sensor data within a 3D-field retrieval framework, utilizing multi-modal techniques to further enhance retrieval accuracy.

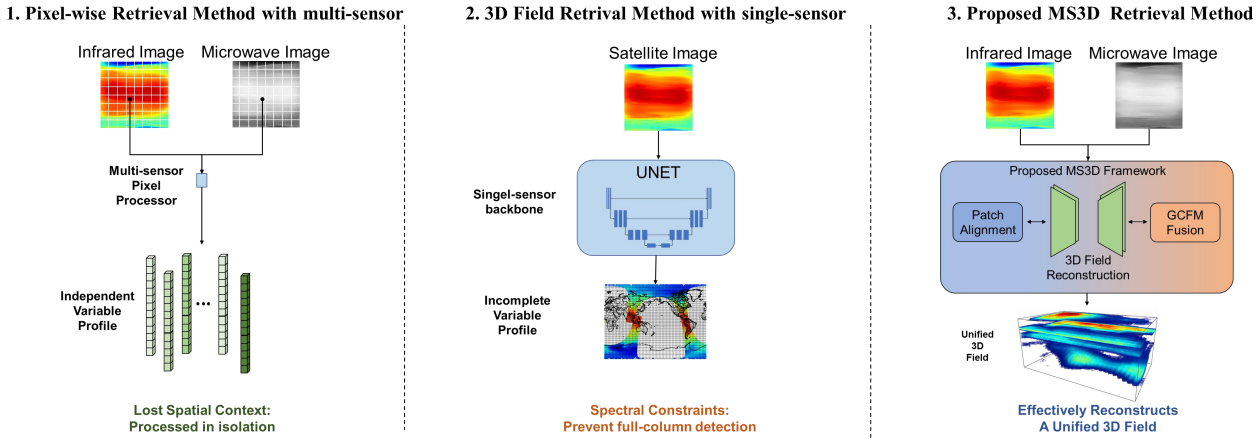


Figure 1. Schematic comparison of (**a**) pixel-wise retrieval, (**b**) single-sensor 3D-field retrieval, and (**c**) the proposed MS3D multi-sensor retrieval framework. Here, 3D-field retrieval refers to the reconstruction of vertical atmospheric profiles across a two-dimensional spatial domain.

The main contributions of this work are as follows:

1. We propose MS3D, a multi-sensor 3D fusion retrieval framework. Unlike existing methods restricted by single-sensor spectral limits and pixel-wise spatial isolation, MS3D integrates multi-modal fusion with 3D spatial modeling. This dual-enhancement strategy enables the reconstruction of a complete 3D atmospheric field with both full vertical coverage and horizontal spatial continuity.
2. We design a specialized deep architecture for heterogeneous data fusion. To bridge the significant modality gap between infrared and microwave data, we introduce adaptive patch embedding and the GCFM. These modules explicitly address spatial misalignment and feature disparity, ensuring a consistent and robust 3D representation across different atmospheric layers.
3. Extensive experiments demonstrate the effectiveness of MS3D in three-dimensional atmospheric field retrieval. Under the ERA5-based evaluation, MS3D reduces the MAE of temperature and specific humidity by 18.9% and 23.9%, respectively, compared with the pixel-wise IReA baseline, and by 10.8% and 7.9%, respectively, compared with the 3D-field ViT baseline. Furthermore, independent validation against IGRA radiosonde observations and experiments under cloudy conditions provide additional evidence of the reliability and robustness of the proposed framework.

2. Related Work

Statistical retrieval methods generally include eigenvector regression, dual regression, and improved first-guess approaches [8–12]. Among these, Principal Component Analysis (PCA) is widely adopted. Tan et al. [13] combined PCA with stepwise regression, while Duma et al. [14] advanced this with Kernel PCA to better handle collinearity. Distinct from standard regression, Deep Gaussian Processes (DGPs) have also been applied, showing

superior capability in capturing hierarchical relationships for temperature and humidity estimation [15].

Physical retrieval relies on radiative transfer models. Foundational works by Chahine, Rodgers, and Lorenc established the iterative and Bayesian optimal estimation frameworks that underpin modern algorithms [16–18]. Recent developments focus on efficiency and precision, such as the adaptive AERIoe algorithm [19] and variational methods for greenhouse gas profiling [20–23]. To further reduce uncertainty, hybrid approaches have been developed that combine statistical initializations or deep learning constraints with physical iterations [24–26].

Machine learning has increasingly been applied to atmospheric retrieval. Early studies utilized standard NNs or combined them with 1D-Var algorithms [27–30]. Later, CNNs were introduced to capture nonlinear dependencies [6,31]. However, recent paradigms have shifted toward deep information fusion to handle data heterogeneity. However, recent paradigms have shifted toward deep information fusion to handle data heterogeneity. To address the redundancy in multi-source data, Transformer-based architectures with advanced feature aggregation mechanisms have been extensively explored to extract singular and common features while dynamically calibrating interaction weights [32]. Furthermore, effective cross-modal interaction has proven critical for fusing complementary sensors, where locality-guided strategies enable robust pixel-level fusion [33,34]. Most relevant to atmospheric challenges, deep fusion strategies are increasingly employed to reconstruct missing information, utilizing dual-domain attention or historical constraints to effectively penetrate cloud cover and mitigate noise interferences [35]. Inspired by these developments, our MS3D framework leverages a multi-sensor fusion strategy to reconstruct 3D atmospheric fields.

3. Materials and Methods

3.1. Data and Instruments

Microwave Sensor: The Micro-Wave Humidity Sounder-II (MWHS; National Space Science Center, Chinese Academy of Sciences, Beijing, China) and Micro-Wave Temperature Sounder-III (MWTS; Xi'an Branch of China Academy of Space Technology, Xi'an, China) onboard FY-3 are passive radiometers designed to measure vertical temperature and humidity distributions up to 6 hPa. Crucially, their ability to penetrate non-precipitating clouds enables essential all-weather atmospheric observation, providing robust data even in cloudy regions.

Hyperspectral Infrared Sensor: Hyperspectral Infrared Atmospheric Sounder-II (HIRAS; Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China) provides high-precision observations with 3053 channels and a spectral resolution of 0.625 cm^{-1}. To balance computational efficiency with information retention, we select a subset of 457 representative channels. As outlined in Table 1, this selection employs a two-step strategy: defining key spectral windows and applying uniform sampling to minimize redundancy while preserving essential spectral features.

ERA5: This study adopts the ERA5 reanalysis dataset (ERA5 hourly data on pressure levels from 1940 to the present) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) as the reference "ground truth." ERA5 is one of the most authoritative and physically consistent global reanalysis products and is widely regarded as the dataset closest to real atmospheric conditions. Its high temporal and spatial resolution, together with the assimilation of a wide range of observations, makes it well suited to serve as a benchmark for global-scale retrieval tasks. By using ERA5 as the reference during training, the model is expected to achieve realistic and physically consistent retrieval performance.

Table 1. Spectral bands sensitive to temperature and humidity [36].

Parameter	Values
Temperature-sensitive spectral bands (cm^{-1})	648.75∼975 1125∼1375 1700∼2000 2160∼2375
Humidity-sensitive spectral bands (cm^{-1})	750∼900 1200∼1250 1300∼1750
Spectral intervals selected for retrieval (cm^{-1})	648.75∼750 (82) 1300∼1650 (281) 2160∼2250 (73) 2350∼2375 (21)

The three instruments exhibit distinct spatial sampling characteristics and viewing geometries. HIRAS-II employs cross-track scanning with a 3 × 3 FOV arrangement and a nominal nadir spatial resolution of 14 km, covering a swath of approximately 2300 km. MWHS-II also performs cross-track scanning, with nominal spatial resolutions of approximately 16 km for the 183 GHz channels and 32 km for the remaining channels. MWTS-III has nominal nadir FOV sizes of approximately 33 km for channels 3–17 and 60 km for channels 1–2. The corresponding swath widths of MWHS-II and MWTS-III are approximately 2600 km and 2800 km, respectively. This study utilizes Level-1 data from the HIRAS, MWTS, and MWHS instruments onboard the FY-3E satellite, covering the global domain from 1 March 2023 to 1 March 2025. Specifically, data from 1 March 2023 to 30 November 2024 are used for training, data from 1 December to 31 December 2024 are used for validation, and data from 1 January to 1 March 2025 are used for testing. All datasets employed in this study are publicly available at https://satellite.nsmc.org.cn/ (accessed on 2 July 2025).

3.2. Multi-Sensor 3D Fusion Framework

To address the challenges posed by three sensors with distinct spatial resolutions and spectral ranges, we propose a multi-sensor 3D fusion (MS3D) framework, as illustrated in Figure 2. In this framework, each sensor is treated as an independent modality for feature extraction, while a GCFM is employed to facilitate effective feature fusion across modalities. Furthermore, separate decoder branches are employed to enable variable-specific reconstruction of temperature and specific humidity, while the shared encoder and bottleneck representations preserve interactions between the two retrieval tasks. This framework exhibits high scalability; incorporating additional sensors or retrieving new meteorological variables can be achieved by simply extending the encoder or decoder branches, respectively.

To facilitate multi-sensor fusion for temperature and humidity retrieval, the proposed framework adopts a U-shaped architecture characterized by downsampling and upsampling paths. Specifically, intermediate feature maps from the three encoders are integrated via concatenation followed by linear projection. These fused features are then transferred to the corresponding layers of the two decoders through skip connections. This mechanism provides the decoders with multi-scale fused features, thereby effectively preserving spatial details during the reconstruction process. Internally, we employ stacked Swin Transformer blocks as the fundamental building units. By leveraging the shifted window-based self-attention mechanism, this design ensures high computational efficiency on high-resolution imagery without compromising the model's ability to capture global dependencies.

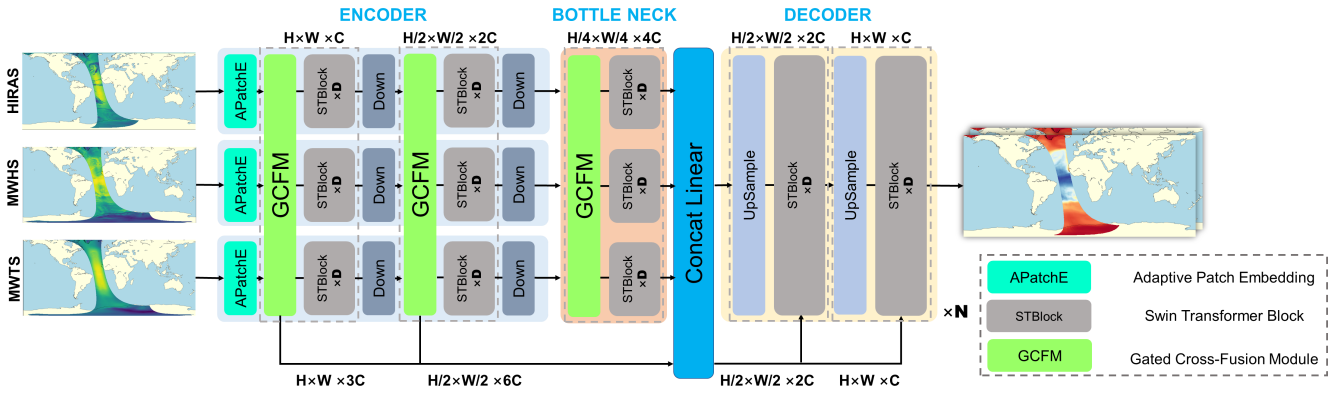


Figure 2. Overall architecture of the proposed MS3D framework. The framework consists of three sensor-specific encoders, cross-modal feature fusion modules, and variable-specific decoders. The notation "×N" denotes N parallel decoder branches (N = 2 for temperature and specific humidity), while the arrows along the bottom represent multi-scale skip connections between the encoders and decoders.

The input dimensions of HIRAS, MWHS, and MWTS are 466 × 1280 × 96, 15 × 1440 × 128, and 17 × 640 × 128, respectively, following the channel × height × width convention. To accommodate their different spatial sampling characteristics, sensor-specific patch sizes of 8 × 3, 9 × 4, and 4 × 4 are adopted, respectively, yielding an identical token grid of 160 × 32 for all three encoder branches. The initial embedding dimension is set to 402, and the attention window size is 2 × 2. Each encoder consists of three Swin Transformer stages containing 4, 6, and 8 blocks, with 3, 6, and 6 attention heads, respectively. The corresponding feature dimensions are 402, 804, and 1608. Two successive 2 × 2 patch-merging operations progressively halve the spatial dimensions while doubling the feature dimension. Each decoder employs patch-expanding operations for upsampling and contains two Swin Transformer stages with 6 and 4 blocks and 6 and 3 attention heads, respectively. The two decoder stages contain 6 and 4 Swin Transformer blocks, respectively, while the notation "×N" in Figure 2 denotes the number of parallel decoder branches (N = 2). Finally, an 8 × 3 patch-expanding operation restores the spatial dimensions to those of the HIRAS observation array, with the two decoder branches independently producing temperature and specific humidity profiles at 37 pressure levels.

The training dataset consists of 17,408 paired multi-instrument observation samples. Z-score normalization is applied independently to each channel of the three instruments, while the target temperature and specific humidity fields are normalized separately for each variable and pressure level. The model is optimized using a weighted L1 loss, with equal weights of 1 assigned to temperature and specific humidity. Specifically, the overall loss is calculated as the average of the mean absolute errors of the two variables in the normalized space.

3.3. Adaptive Patch Embedding

Given the varying spatial resolutions of the sensors, consistent spatial organization of observations is important for multi-modal feature fusion. Specifically, the Level-1 observations from HIRAS, MWHS, and MWTS are organized into two-dimensional arrays according to their native scanning and detector arrangements, preserving the relative spatial ordering of the observations. No spatial resampling, interpolation, or explicit geographical co-registration is performed across the three instruments. Since the three sensors are onboard the same satellite, their observations generally cover overlapping geographical regions, although their spatial resolutions and FOV footprints differ. To accommodate these differences, we carefully design sensor-specific patch sizes for the

embedding process. The adaptive patch embedding strategy enables the generation of an identical number of tokens from different sensors, allowing their features to be integrated within a unified framework. It should be noted that each patch represents a group of observations in the original scanning arrangement rather than a fixed geographical grid cell. Therefore, corresponding tokens represent approximately overlapping geographical regions rather than exactly identical physical footprints. The original FOV sampling characteristics, including potential overlaps, gaps, and viewing-angle-dependent geometric variations, are retained without additional geometric correction. As shown in Figure 3 and Equation (1), adaptive patch embedding produces token sequences of equal length from the three sensors despite their different input dimensions.

$$N = H_1 W_1 / P_1^2 = H_2 W_2 / P_2^2 = H_3 W_3 / P_3^2 \quad (1)$$

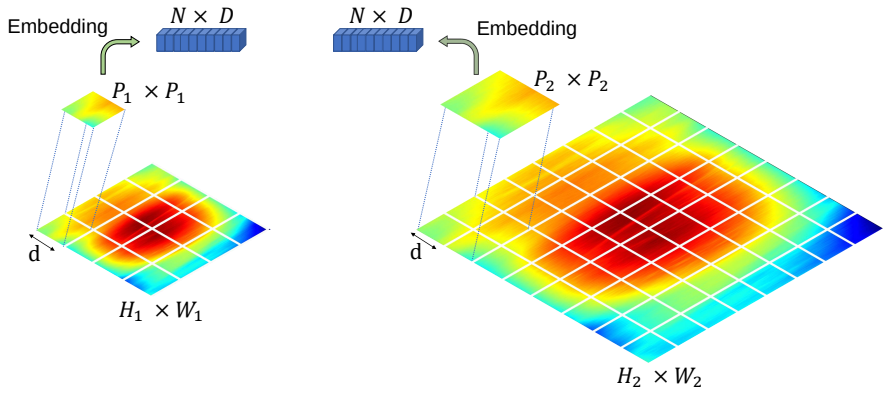


Figure 3. Adaptive patch partitioning. Sensor-specific patch sizes are selected based on the spatial sampling characteristics of different instruments to accommodate differences in their spatial resolutions. $H_1 \times W_1$ and $H_2 \times W_2$ denote the dimensions of two satellite observation arrays, while d represents the relative spatial scaling factor between them. By applying corresponding patch sizes P_1 and P_2, the inputs are partitioned and projected into a unified feature space, resulting in an equal number of N tokens with a feature dimension of D for each instrument. This design enables consistent token representations for subsequent multi-modal feature fusion.

3.4. Gated Cross-Fusion Module

Due to their distinct physical sensing mechanisms, MWHS, MWTS, and HIRAS exhibit significant variations in noise and information distributions. Consequently, naive concatenation of these heterogeneous modalities fails to achieve effective complementarity. Instead, it often leads to feature interference, where data from one sensor acts as noise to another. This challenge is further exacerbated by the substantial dimensional imbalance among the sensors. After spectral channel selection, HIRAS retains 457 spectral channels, considerably more than the 15 and 17 channels of MWHS and MWTS, respectively. In addition, nine auxiliary variables from the HIRAS Level-1 data are incorporated into the HIRAS input, including latitude, longitude, satellite azimuth angle, satellite zenith angle, solar azimuth angle, solar zenith angle, land cover type, land–sea mask, and observation time. These auxiliary variables, together with the 457 selected spectral channels, constitute a total of 466 input channels for HIRAS.

Specifically, the GCFM receives output tokens from the three encoders at the corresponding layer, as shown in Figure 4. These inputs are designated as one primary vector ($\mathbf{X}_M$) and two auxiliary vectors ($\mathbf{X}_{Ai}$). The module performs cross-attention (CA) between

the ($\mathbf{X}_M$) and each ($\mathbf{X}_{Ai}$) in parallel. The resulting vectors ($\mathbf{F}_i$) are then aggregated and processed by an MLP block for deep fusion, yielding the final output representation ($\mathbf{F}_{\text{fuse}}$).

$$\begin{aligned} \mathbf{F}_i &= \text{CA}\big((\mathbf{X}_M\mathbf{W}_Q), (\mathbf{X}_{Ai}\mathbf{W}_K), (\mathbf{X}_{Ai}\mathbf{W}_V)\big), \quad i \in \{1,2\} \\ \mathbf{F}_{\text{fuse}} &= \text{MLP}(\text{Concat}(\mathbf{F}_1, \mathbf{F}_2)] \end{aligned} \tag{2}$$

where $\mathbf{W}$ denotes the corresponding projection matrix.

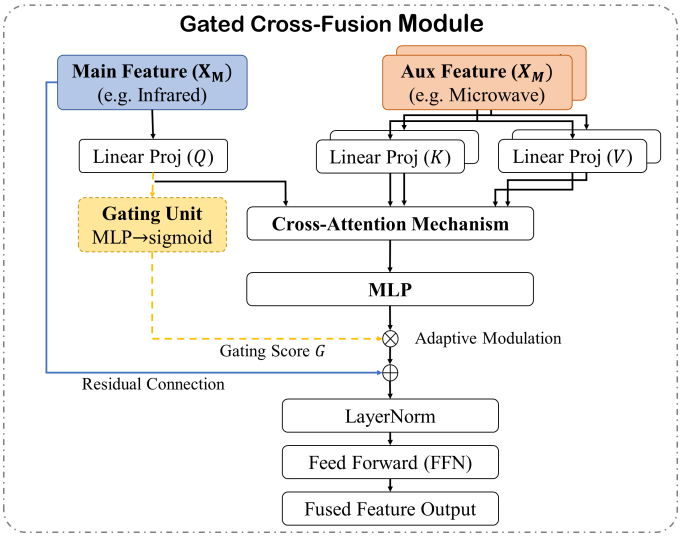


Figure 4. Architecture of the Gated Cross-Fusion Module (GCFM), illustrating cross-attention between the primary and two auxiliary sensor features and the adaptive gating mechanism for feature fusion.

To generate gating information intrinsically aligned with the primary sensor, we utilize the projected query ($\mathbf{X}_M\mathbf{W}_Q$) from the cross-attention mechanism. This feature vector is processed through an MLP layer followed by a Sigmoid activation, constraining values to the range $[0, 1]$ to serve as the gating signal ($\mathbf{G}$). The gating signal modulates the fused features, effectively filtering out noise introduced by other sensors, particularly in shallow network layers (as validated in subsequent experiments). Finally, the module concludes with standard residual connections, followed by Layer Normalization (LN) and a Feed-Forward Network (FFN).

$$\begin{aligned} \mathbf{G} &= \sigma\big(\text{MLP}\big(\mathbf{X}_M\mathbf{W}_Q\big)\big) \\ \mathbf{Output} &= \text{FFN}(\text{LN}(\mathbf{X}_M + \mathbf{G} \odot \mathbf{F}_{\text{fuse}})) \end{aligned} \tag{3}$$

where σ is the sigmoid activation.

3.5. Evaluation Metrics

The performance of the retrieval model is evaluated using two commonly applied metrics: mean absolute error (MAE) and root mean square error (RMSE). ERA5 reanalysis data are interpolated based on latitude and longitude to match the spatial coverage of the retrieval results, providing ground truth over the corresponding regions. Both metrics are then computed on the test set to obtain the average retrieval errors for temperature and specific humidity across the entire swath. The equations of these two metrics are as follows:

$$\text{RMSE}(l) = \sqrt{\frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(X_{\text{era5}}(i,j) - X_{\text{retrieval}}(i,j)\right)^2} \tag{4}$$

$$\mathrm{MAE}(l) = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} |X_{\mathrm{era5}}(i,j) - X_{\mathrm{retrieval}}(i,j)| \quad (5)$$

The evaluation is performed at different pressure levels, where l denotes the pressure level index. X_{era5} represents the reference temperature or humidity values obtained from ERA5 reanalysis data, and $X_{\mathrm{retrieval}}$ denotes the corresponding values retrieved by the proposed model. The indices i and j indicate the latitude and longitude positions within the study region, respectively. Consequently, H and W correspond to the spatial height and width of the retrieval domain, determined by the satellite scanning swath and spatial resolution.

4. Results

To evaluate the effectiveness of the proposed multi-sensor retrieval framework for atmospheric profile inversion and to further demonstrate the advantages of the MS3D network, a series of comprehensive experiments were conducted. These experiments included comparisons with alternative networks for satellite swath retrieval, aiming to assess the accuracy of MS3D under real-world conditions. In addition, ablation studies were performed to investigate the individual and combined contributions of infrared and microwave sensors, highlighting the benefits of multi-sensor fusion. Furthermore, the model was tested under cloudy conditions to examine the robustness and effectiveness of multi-sensor retrieval in scenarios where traditional single-sensor methods typically underperform.

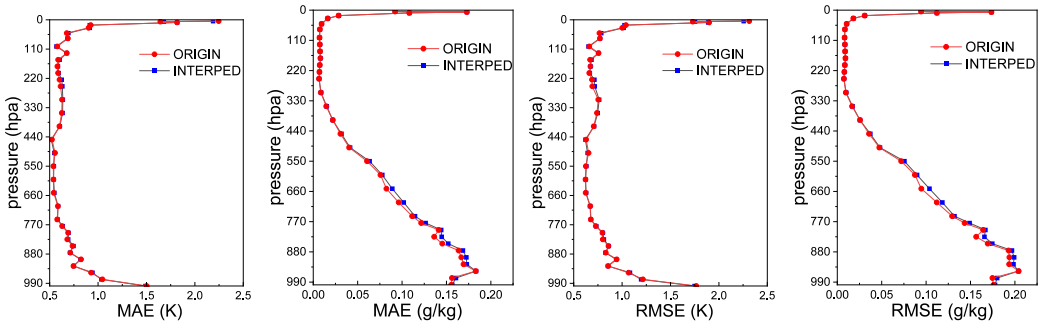
4.1. Ground Truth Validation

Conventional meteorological retrieval tasks typically employ either radiosonde observations or ERA5 reanalysis data as the ground truth. However, radiosonde stations are spatially sparse; for instance, the Integrated Global Radiosonde Archive (IGRA) comprises only 885 available stations globally. Furthermore, radiosonde data are characterized by irregular pressure levels and frequent missing values, particularly in the upper atmosphere. Nevertheless, representing in-situ measurements, they remain the absolute ground truth for meteorological applications.

In contrast, ERA5—the fifth-generation reanalysis dataset produced by ECMWF—assimilates multi-source observations, including data from satellites, aircraft, radar, and weather balloons. It is widely acknowledged as the dataset that best approximates the true atmospheric state. ERA5 offers continuous global coverage with a spatial resolution of 0.25° and spans 37 pressure levels from 1 to 1000 hPa. Given its structured format and spatiotemporal continuity—as opposed to the sparsity and instability of radiosonde data—this study selects ERA5 as the ground truth for training the neural network.

Specifically, temperature and specific humidity from the ERA5 reanalysis dataset are selected as the target variables. Given the hourly temporal resolution of ERA5, temporal collocation is performed by assigning each satellite observation to the ERA5 analysis field corresponding to the beginning of the hour in which the satellite scan was acquired. For instance, a satellite observation acquired at 09:35 UTC is paired with the ERA5 analysis field at 09:00 UTC. Meanwhile, since satellite observation footprints do not spatially align with the ERA5 grid points, bilinear interpolation is employed to derive the true meteorological states at the specific satellite observation locations. To verify that this interpolation process does not introduce unreasonable bias, we designed a validation experiment using radiosonde data as the independent benchmark. We calculated the mean absolute error (MAE) and root mean square error (RMSE) for both the original gridded ERA5 data (matched via nearest neighbor) and the interpolated ERA5 data, relative to the radiosonde observations. As

shown in Figure 5, the close alignment of the error curves indicates that the interpolation process preserves data integrity without introducing significant bias, thereby validating the rationality of our ground truth configuration.



(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

Figure 5. Error comparison between the original and interpolated ERA5 data relative to radiosonde observations. This verification aims to confirm that the interpolation process does not introduce unreasonable bias that would compromise the validity of the ground truth configuration.

4.2. Comparative Experiment

This section presents a comparative study of two categories of networks with a consistent dataset and partitioning strategy to control for confounding variables, namely pixel-wise retrieval methods and 3D-field retrieval methods. The maximum number of training epochs was set to 100. For each model, the top three checkpoints exhibiting the lowest validation loss during training were retained, and the model corresponding to the earliest epoch was selected to mitigate the risk of overfitting. The training strategy was configured as follows: the initial learning rate was set to 5.0×10^{-5} and decayed by a factor of 0.3 every 30 epochs. The AdamW optimizer was employed, with ϵ set to 1.0×10^{-8}, and β_1 and β_2 set to 0.9 and 0.999, respectively. Figure 6 illustrates the evolution of the training and validation losses of MS3D.

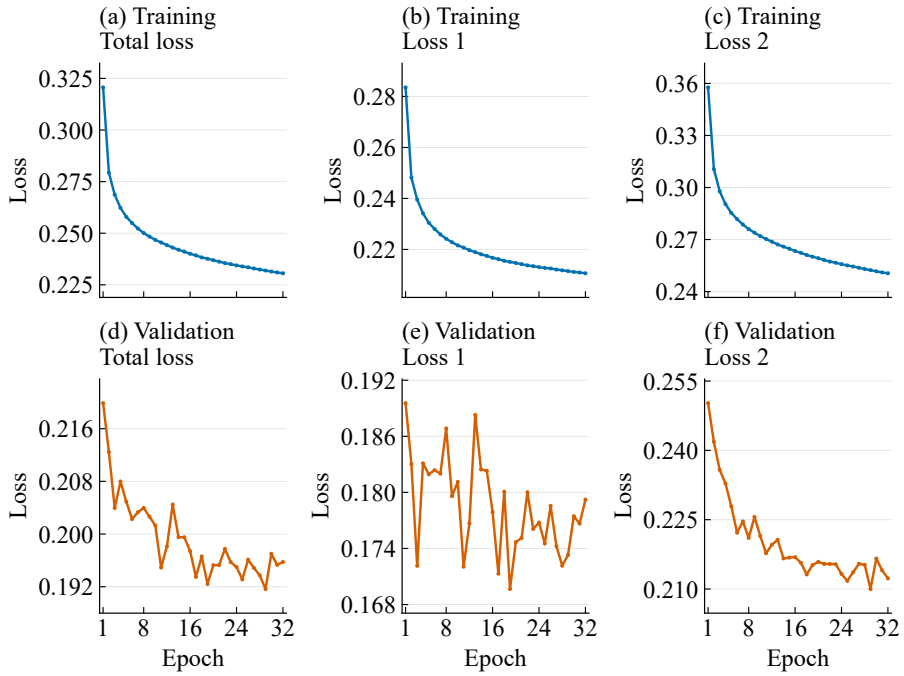
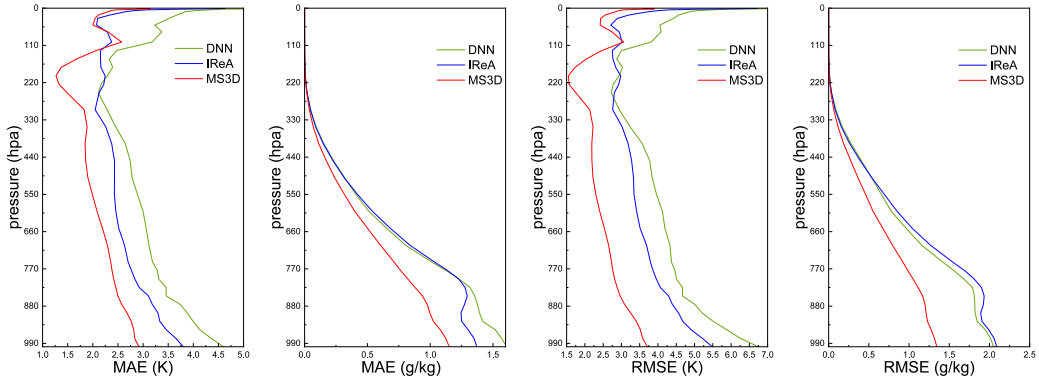


Figure 6. Evolution of training and validation losses for MS3D, including the total loss and its two components.

Pixel-Wise Retrieval Methods. This study employs commonly used deep neural network (DNN) for temperature and humidity profile retrieval, consisting of sequential linear and activation layers that directly map channel radiances to atmospheric profiles. In addition, the recently developed IReA Xiao et al. [7] is considered, which integrates one-dimensional convolutional networks with MLP-based feedforward layers and incorporates multi-sensor data to support the retrieval process.

The retrieval performance of satellite data was evaluated using MAE and RMSE metrics, as shown in Figure 7. Examining the temperature performance first, it is evident that below the 110 hPa level, the pixel-wise retrieval method shows a significant performance gap compared to MS3D. Generally, errors trend upward as altitude decreases, primarily due to the increasing complexity of atmospheric conditions near the surface, where meteorological variables are influenced by a multitude of factors. However, the pixel-wise retrieval method exhibits a much sharper degradation in accuracy, with near-surface errors approaching 5 K (MAE) and 7 K (RMSE). This is fundamentally attributed to the pixel-wise retrieval method's inability to model 3D fields or leverage spatial context, which hinders its capacity to handle complex atmospheric conditions.



(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

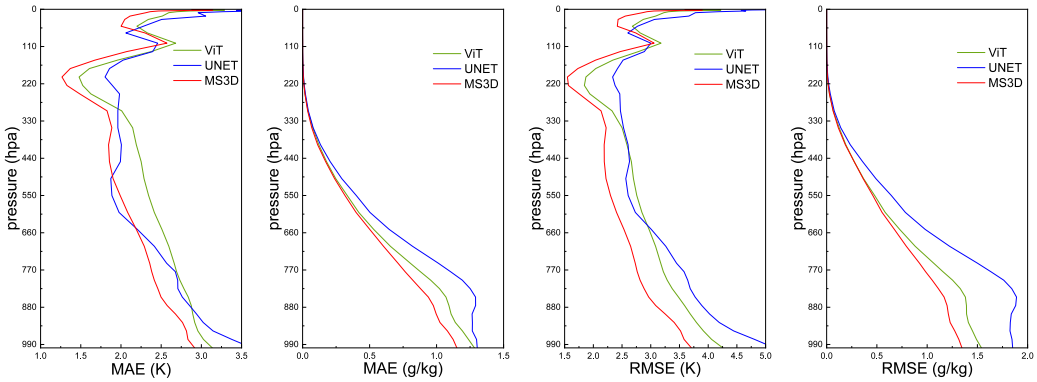
Figure 7. Vertical profiles of MAE and RMSE for temperature and specific humidity retrievals, comparing MS3D with the pixel-wise DNN and IReA baselines.

Regarding humidity, the pixel-wise retrieval method significantly underperforms MS3D across all 37 pressure levels. As humidity is a complex variable highly sensitive to wind and cloud cover, it necessitates the capture of data coupling within the 3D field. Furthermore, the notably poorer RMSE performance of the pixel-level method highlights the instability of point-wise retrieval models and their susceptibility to outliers (or 'bad pixels'). In contrast, by explicitly modeling the 3D meteorological field, the proposed method effectively captures macroscopic atmospheric flow features, demonstrating superior stability and precision.

3D-Field Retrieval Methods. U-Net, a classical neural network widely used for image-to-image translation, was applied to temperature and humidity profile retrieval (Yao and Guan [6]), with HIRAS data serving as input following previous studies. The Vision Transformer (ViT) (Dosovitskiy et al. [37]), a representative transformer architecture in computer vision, was employed to evaluate the effectiveness of self-attention mechanisms in profile inversion, also using HIRAS data as input. All networks were trained with the same hyperparameters as in prior experiments, allowing a systematic comparison of their performance in 3D-field retrieval.

The retrieval performance of satellite data was evaluated using MAE and RMSE metrics, as shown in Figure 8. All three temperature error curves exhibit a characteristic U-shaped pattern (higher errors at the boundaries, lower in the middle), which can be

attributed to two main factors: (1) the scarcity of detection channels for the upper atmosphere (<20 hPa) results in a lack of observational constraints for these altitudes; (2) the near-surface layers are subject to complex influences, such as surface topography and emissivity, making accurate retrieval challenging.



(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

Figure 8. Vertical profiles of MAE and RMSE for temperature and specific humidity retrievals, comparing MS3D with the 3D-field U-Net and ViT baselines.

A comparative analysis of the three 3D-field retrieval methods reveals that UNet (Yao and Guan [6]), by employing multi-scale convolutions, effectively captures accurate temperature values in the middle atmosphere. However, due to the inherent locality of convolution operations, UNet fails to model long-range dependencies, resulting in suboptimal performance for humidity retrieval. In contrast, ViT, as a representative attention-based architecture, generally outperforms UNet. However, it struggles to effectively integrate multi-modal data from heterogeneous sensors. Specifically, a significant performance gap exists compared to MS3D in the lower troposphere (<300 hPa). This region critically requires complementary data from microwave sensors capable of cloud penetration. Yet, ViT lacks a robust multi-modal fusion mechanism to achieve effective information complementarity. Addressing this limitation was a primary motivation for proposing MS3D, which is designed to synergistically fuse data from diverse sensors for the cooperative retrieval of temperature and humidity profiles.

The quantitative evaluation using MAE and RMSE, averaged over the 37 pressure levels, is summarized in Table 2, where MS3D achieves the best performance for both temperature and humidity retrieval. This improvement can be attributed to several architectural characteristics: (i) the multi-scale structure for 3D fields, which enriches spatial feature representation; (ii) the self-attention mechanism, which effectively models long-range dependencies in image; and (iii) the use of multiple encoders and the GCFM, which facilitates the integration of multi-source observations and ensures complementary information across different sensing windows. Compared with pixel-wise retrieval, the proposed method leverages 3D-field retrieval to capture spatial representation, thereby improving humidity estimation. Compared to other 3D-field retrieval baselines, MS3D employs a meticulously designed multi-encoder-decoder architecture. This framework allows for the full exploitation of cross-modal information while simultaneously preserving the independence of single-modal features. Furthermore, given that the hyperspectral sensor possesses significantly more channels and higher information density than the microwave sensors, there is a risk that microwave features could be diluted (or overshadowed) by the dominant hyperspectral signals. To address this, the GCFM incorporates a gating mechanism to selectively fuse information from auxiliary modalities, ensuring that critical microwave signatures are retained within the microwave branch. Compared with the ViT baseline, MS3D

reduces the MAE and RMSE of temperature retrieval by 10.8% and 12.3%, respectively, while achieving corresponding reductions of 7.8% and 12.0% for specific humidity. These results demonstrate the effectiveness of the proposed multi-sensor retrieval framework.

Table 2. Quantitative comparison of MAE and RMSE averaged over 37 pressure levels for different neural networks. The best results are highlighted in bold.

Model	Params (M)	FLOPs (G)	Time (ms)	Temperature (K)		Specific Humidity (g/kg)	
				MAE	RMSE	MAE	RMSE
DNN (pixel-wise)	2.947	722.94	18.02	3.37	4.45	0.49	0.67
IReA (pixel-wise)	0.186	199.39	85.86	2.75	3.79	0.46	0.70
ViT (3D-field)	51.8	1367.86	122.68	2.50	3.07	0.38	0.50
U-Net (3D-field)	1.862	460.24	52.97	2.59	3.41	0.45	0.66
MS3D (3D-field)	1127.16	2545.37	130.53	**2.23**	**2.69**	**0.35**	**0.44**

4.3. Comparison with Radiosonde Observations

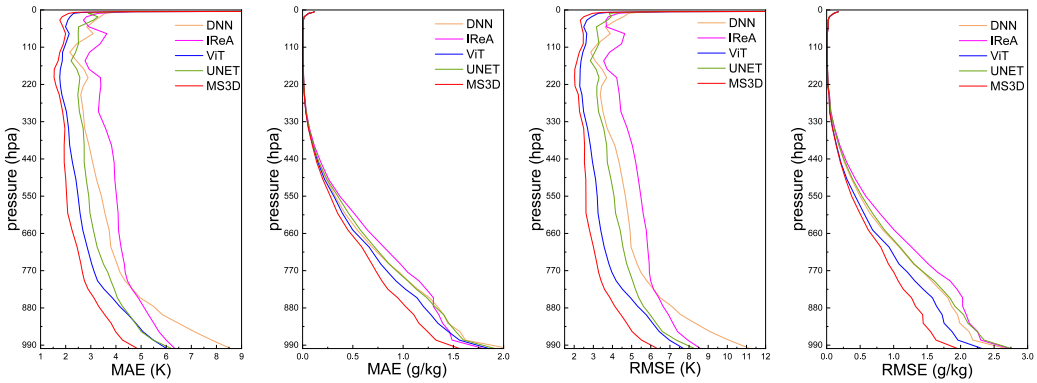
To independently assess the agreement between the atmospheric profiles retrieved by different models and in situ measurements, we conducted a spatiotemporally collocated evaluation using radiosonde observations from the Integrated Global Radiosonde Archive (IGRA) as the reference. Taking the radiosonde station locations and nominal UTC observation times as references, we selected the nearest satellite pixel within a temporal difference of ± 5 min and a great-circle distance of 50 km. Only one matched pair was retained for each station time combination, and the same satellite pixels were used to evaluate all five models.

After quality control to remove missing values and physically implausible observations, the radiosonde profiles were linearly interpolated in logarithmic pressure coordinates onto the 37 target pressure levels. Interpolation was restricted to the pressure range covered by valid radiosonde observations, without extrapolation. At each pressure level, evaluation was performed using a common set of valid samples shared by the radiosonde observations and predictions from all five models. The root mean square error (RMSE) and mean absolute error (MAE) were calculated separately for temperature and specific humidity. A total of 1540 valid profile pairs for temperature and 1537 for specific humidity were obtained. The number of valid samples at each pressure level varied depending on the vertical coverage of the radiosonde observations.

As shown in Table 3 and Figure 9, MS3D consistently outperforms the four baseline models in both temperature and specific humidity retrievals when evaluated against IGRA radiosonde observations. Specifically, MS3D achieves MAE and RMSE values of 3.47 K and 4.35 K for temperature, and 0.41 g/kg and 0.54 g/kg for specific humidity, respectively. The vertical error profiles further demonstrate that MS3D generally maintains lower retrieval errors across different pressure levels, with more pronounced improvements in the lower atmosphere, particularly for specific humidity. These results provide observational evidence of the reliability of the retrieved atmospheric profiles, complementing the ERA5-based evaluation. It should be noted that the models are trained using ERA5 as the reference target rather than radiosonde observations. Therefore, discrepancies relative to radiosonde measurements reflect not only retrieval errors but also differences between ERA5 and actual atmospheric observations. Nevertheless, the consistently superior performance of MS3D against radiosonde measurements further supports the effectiveness of the proposed framework.

Table 3. Comparison of temperature and specific humidity retrieval performance against IGRA radiosonde observations. The best results are highlighted in bold.

Model	Params (M)	FLOPs (G)	Time (ms)	Temperature (K)		Specific Humidity (g/kg)	
				MAE	RMSE	MAE	RMSE
DNN (pixel-wise)	2.947	722.94	18.02	5.00	6.31	0.51	0.74
IReA (pixel-wise)	0.186	199.39	85.86	5.31	6.71	0.52	0.82
ViT (3D-field)	51.8	1367.86	122.68	3.96	4.99	0.47	0.64
U-Net (3D-field)	1.862	460.24	52.97	4.59	5.84	0.51	0.76
MS3D (3D-field)	1127.16	2545.37	130.53	**3.47**	**4.35**	**0.41**	**0.54**



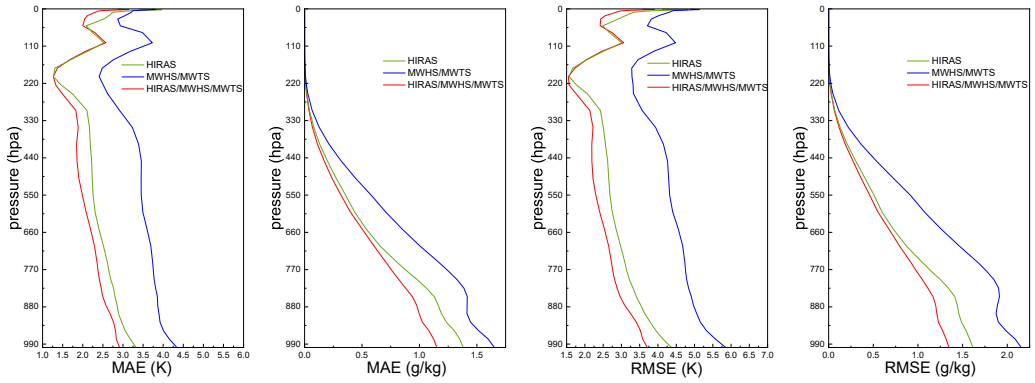
(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

Figure 9. Vertical profiles of MAE and RMSE for temperature and specific humidity retrievals against IGRA radiosonde observations.

4.4. Ablation Experiments

To further assess the contribution of the multi-sensor design, ablation experiments were conducted by dividing the data sources into three groups: (a) a control group using all three sensors (HIRAS, MWHS, and MWTS) as proposed in this study; (b) Experimental Group 1 using only HIRAS; and (c) Experimental Group 2 using MWHS and MWTS. In Experimental Group 2, the two microwave sensors are combined because of their complementary but narrower detection windows with MWTS being a temperature sounder and MWHS a humidity sounder. Given that the retrieval task involves both temperature and humidity profiles, their combination ensures a consistent experimental setup. By contrast, HIRAS, with its high spectral resolution and large number of channels, is evaluated independently to examine its stand-alone contribution.

MAE and RMSE were used as evaluation metrics. Figure 10 presents the results for the 37-level temperature and humidity profiles. Retrieval performance with all three sensors is consistently superior to that obtained with either microwave-only or infrared-only configurations. The most notable finding is the clear performance gap between the microwave sensors (MWHS/MWTS) and the infrared sensor (HIRAS), with the microwave-based results underperforming across all pressure levels. This discrepancy is mainly attributable to the narrow observational windows of microwave sensors, which restrict the information available to the neural network and limit its capacity to learn complex retrieval mappings. In contrast, HIRAS, with its higher spectral resolution and larger number of channels, provides richer feature representations, allowing the network to construct more sophisticated transformations and thereby achieve more accurate mappings from radiances to atmospheric variables. This observation is consistent with the general principle that neural networks benefit from more abundant and higher-quality input data.



(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

Figure 10. MAE and RMSE of atmospheric profile retrievals across 37 pressure levels for HIRAS/MWHS/MWTS, HIRAS, and MWHS/MWTS groups.

When comparing the joint configuration (HIRAS/MWHS/MWTS) with HIRAS alone, the improvement is most pronounced in the mid-to-lower atmosphere (<220hpa). The additional microwave observations effectively complement the infrared channels, enhancing the model's ability to retrieve temperature and humidity profiles in these regions. This advantage is evident in the MAE and RMSE curves, where the three-sensor setup exhibits reduced errors relative to HIRAS alone. Quantitative results averaged across the 37 pressure levels, summarized in Table 4, further confirm that the combined HIRAS/MWHS/MWTS configuration yields the best performance, underscoring the effectiveness of the multi-sensor retrieval strategy.

Table 4. Quantitative comparison of MAE and RMSE averaged over 37 pressure levels for HIRAS/MWHS/MWTS, HIRAS and MWHS/MWTS groups. The best results are highlighted in bold.

Temperature (K)	**MAE**	**RMSE**
HIRAS/MWHS/MWTS	**2.23**	**2.69**
HIRAS	2.54	3.08
MWHS/MWTS	3.42	4.40
Specific humidity (g/kg)	MAE	RMSE
HIRAS/MWHS/MWTS	**0.35**	**0.44**
HIRAS	0.41	0.52
MWHS/MWTS	0.54	0.75

In addition, we conducted a divergence verification experiment. Specifically, 100,000 points were randomly sampled from different spatial locations and pressure levels within a single retrieval scene. The retrieval outputs from the three models were plotted against the corresponding ERA5 ground truth values, with model predictions on the y-axis and ERA5 references on the x-axis. A linear regression was fitted to each scatter plot, and the coefficient of determination (R^2) was computed. The results are shown in Figures 11 and 12 for temperature and humidity, respectively.

For temperature, the MWHS/MWTS retrieval exhibits the largest deviations from the truth values, while HIRAS results also contain a subset of outliers. In contrast, the HIRAS/MWHS/MWTS configuration shows much tighter alignment with the 1:1 reference line, reflecting a stronger fitting capability. This is corroborated by the higher R^2, which indicates greater explanatory power.

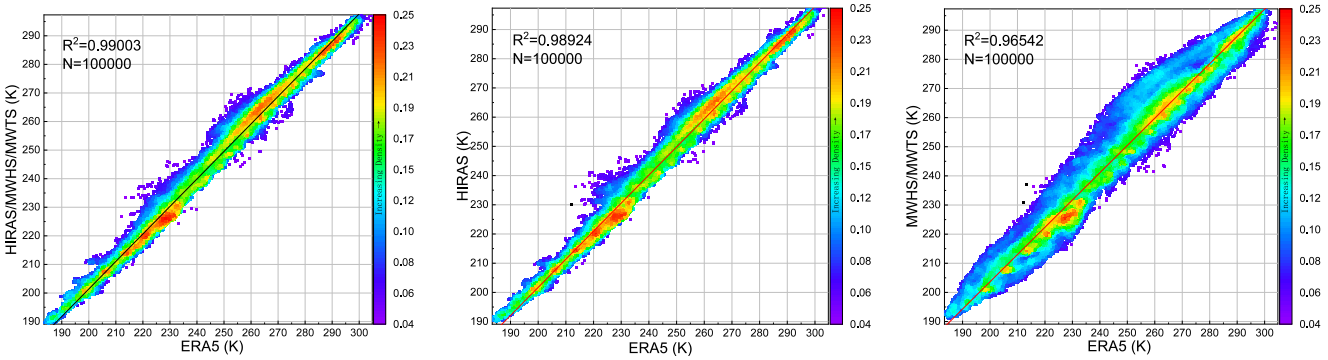


Figure 11. Divergence between retrieved temperature profiles and ERA5 data at 100,000 randomly selected points from (**a**) HIRAS/MWHS/MWTS, (**b**) HIRAS, and (**c**) MWHS/MWTS.

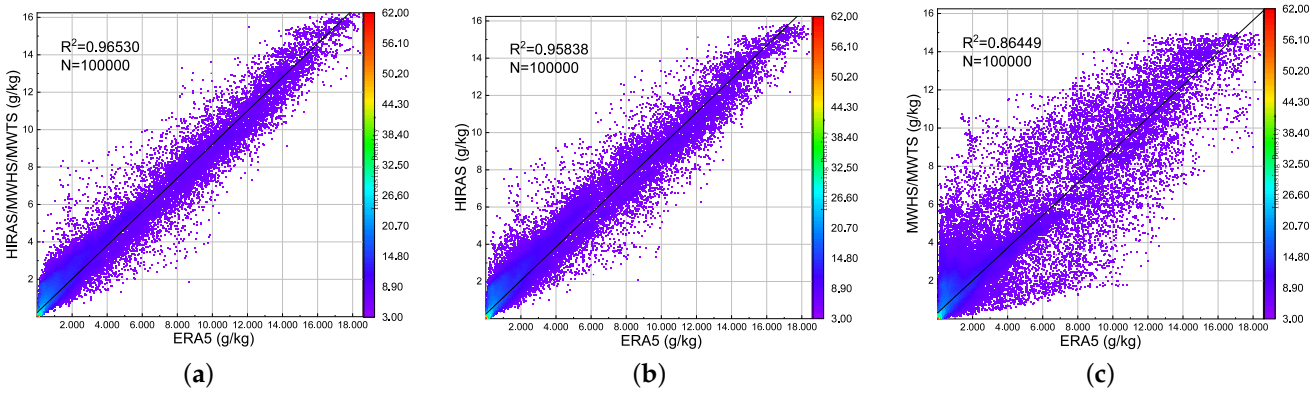


Figure 12. Divergence between retrieved specific humidity profiles and ERA5 data at 100,000 randomly selected points from (**a**) HIRAS/MWHS/MWTS, (**b**) HIRAS, and (**c**) MWHS/MWTS.

For humidity, the scatter distribution is less uniform due to the pronounced variability in both the upper and lower atmosphere, with most values concentrated below 0.002 g/kg. Nevertheless, both the scatter plots and the R^2 values consistently demonstrate that the three-sensor (HIRAS/MWHS/MWTS) configuration outperforms the other two setups, confirming its superior retrieval accuracy.

5. Discussion

The comparative experiments and ablation studies presented above demonstrate the effectiveness of the proposed MS3D framework in atmospheric temperature and specific humidity retrieval. To further understand the mechanisms underlying these improvements, this section provides a more detailed analysis of the proposed framework from several perspectives, including the behavior of the gated cross-fusion mechanism, the reconstruction of horizontal and vertical atmospheric structures, and retrieval performance under cloudy conditions. These analyses offer further insights into how MS3D exploits complementary multi-sensor information and spatial dependencies to improve atmospheric field reconstruction.

5.1. Contribution of the Gating Mechanism

Given the significantly higher information density of the hyperspectral HIRAS sensor and the physical heterogeneity arising from the distinct spectral ranges of the three sensors, the gating mechanism is essential within the cross-attention module. This mechanism serves to adaptively regulate the information flow, ensuring that critical multi-modal features are preserved rather than being misinterpreted as noise. To validate this design,

we sampled gating signals from the GCFMs across three encoder layers (N = 10,000). The gating values range from 0 to 1, where 0 indicates complete rejection and 1 represents full acceptance.

As illustrated in Figure 13, the evolution of gating signals across layers reveals the model's dynamic strategy for information fusion: (i) Layer 1 (Shallow Interaction): The gating signal exhibits a semi-open state. This configuration facilitates limited cross-modal interaction during the initial feature extraction phase, effectively preventing noise propagation from interfering with the specific features of each branch. (ii) Layer 2 (Asymmetric Fusion): A distinct differentiation among sensors emerges. The HIRAS branch maintains a semi-open state due to its inherent high information density. In contrast, the gating values for MWHS and MWTS increase significantly to approximately 0.8. This indicates that these branches are actively assimilating the rich spectral information provided by HIRAS to complement their own representations. (iii) Layer 3 (Independence Preservation): In the final layer, the gating mechanism tends to close. At this stage, where the network processes high-level semantic features, restricting information flow is crucial. This isolation ensures the independence of each sensor's retrieval path and prevents feature homogenization, thereby preserving the diversity of multi-source information and avoiding the convergence of distinct modal features.

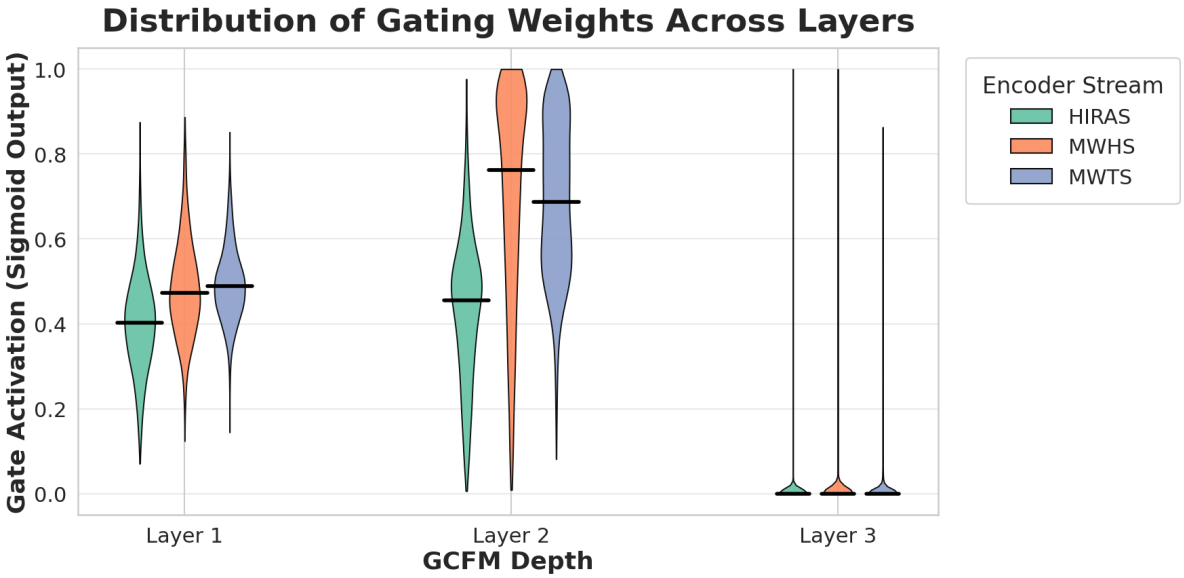
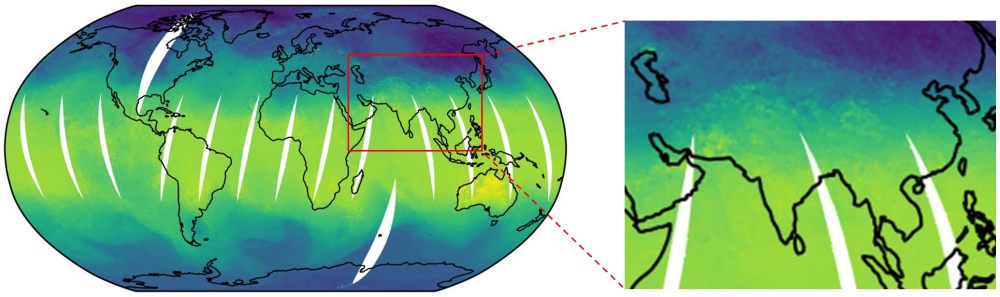


Figure 13. Violin plots showing the distribution of gating signals across different layers of the encoders for the three sensors. The plots illustrate the probability density of sampled gating values ranging from 0 to 1. The black horizontal lines denote the median (50th percentile) of the signals for each corresponding layer, indicating the central tendency of information acceptance at various stages of feature extraction.

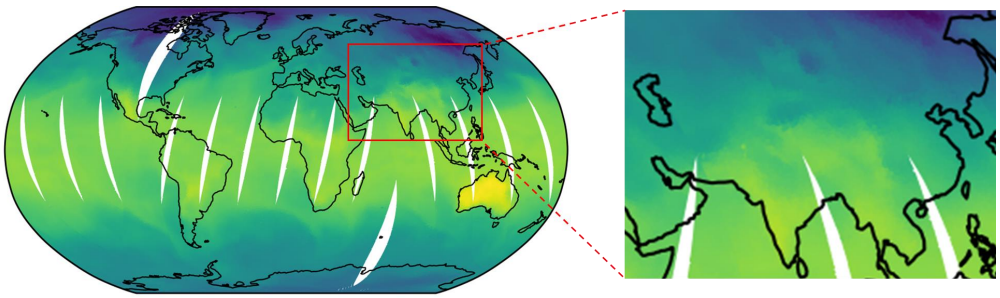
To validate the capability of the proposed 3D-field retrieval framework in capturing atmospheric spatial information, we stitched retrieval results from 15 consecutive time steps to generate global maps of temperature and specific humidity at 1000 hPa, as shown in Figures 14 and 15. By comparing the outputs of MS3D, the pixel-wise retrieval method (IReA), and the ground truth, distinct differences are observed.

In terms of temperature, the IReA exhibits obvious graininess and a lack of detailed texture, highlighting the limitation of methods that disregard spatial context. In contrast, MS3D successfully captures distinct temperature textures over the Eurasian continent; specifically, it accurately identifies the temperature gradients and boundaries associated with the complex terrain of the Tibetan Plateau and Xinjiang region. Furthermore, since

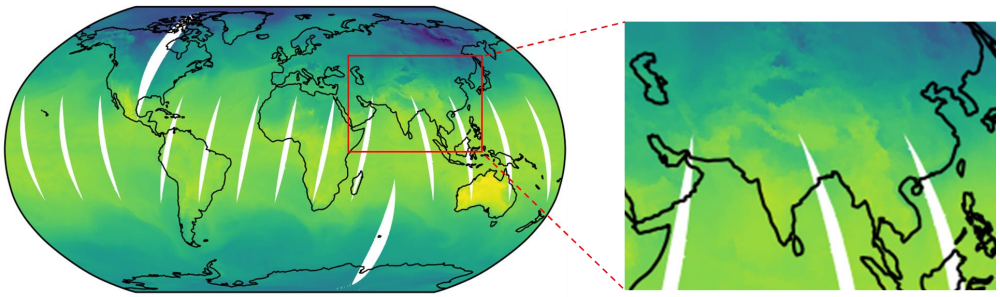
humidity is a meteorological variable governed by fluid dynamics and highly sensitive to spatial features, the pixel-wise method fails to represent the atmospheric flow, resulting in a loss of textural detail. Conversely, MS3D preserves clear atmospheric textures. Notably, the model successfully captures the vortex structure over the Atlantic Ocean, a detail that reveals critical information regarding heavy precipitation and cyclones, which is of significant value for severe convection forecasting.



(**a**) Visualization of temperature from IReA



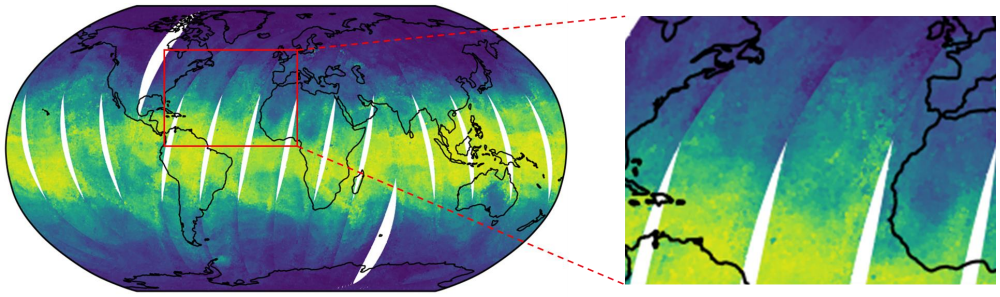
(**b**) Visualization of temperature from MS3D



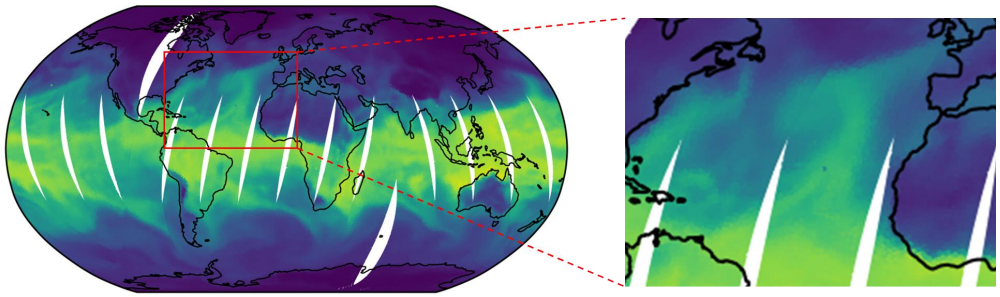
(**c**) Visualization of temperature from ERA5

Figure 14. Comparison of temperature fields at 1000 hPa obtained from (**a**) IReA, (**b**) MS3D, and (**c**) ERA5. The enlarged regions highlight differences in the reconstructed horizontal spatial structures. The colors represent numerical values, with brighter colors indicating higher values.

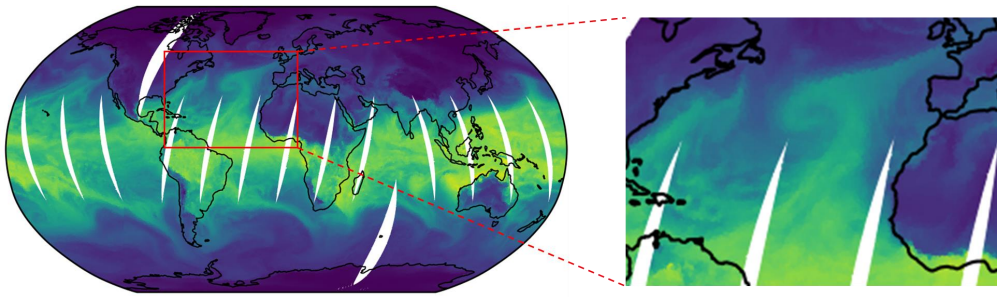
Consequently, this experiment provides compelling evidence that the MS3D framework, with its broader spatial field of view, significantly outperforms pixel-wise methods in capturing spatial information, demonstrating its substantial value for meteorological research.



(**a**) Visualization of specific humidity from IReA



(**b**) Visualization of specific humidity from MS3D



(**c**) Visualization of specific humidity from ERA5

Figure 15. Comparison of specific humidity fields at 1000 hPa obtained from (**a**) IReA, (**b**) MS3D, and (**c**) ERA5. The enlarged regions highlight differences in the reconstructed horizontal spatial structures. The colors represent numerical values, with brighter colors indicating higher values.

5.2. Vertical Structure Reconstruction Analysis

To evaluate the ability of MS3D to reconstruct atmospheric vertical structures, we selected an orbital segment over the Pacific Ocean observed between 01:44 and 01:54 UTC on 1 March 2025. A total of 71 along-track samples were extracted at a fixed scan column and detector position, covering an along-track distance of approximately 3749 km. Vertical cross-sections were constructed from the MS3D predictions and the corresponding ERA5 reference fields at identical locations and pressure levels. Among the 37 output pressure levels, the range from 100 to 900 hPa was selected for visualization, with temperature and specific humidity expressed in K and g/kg, respectively. No cross-track averaging or spatial smoothing was applied, and identical color scales were used for the reference and predicted fields of each variable. As shown in Figure 16, MS3D successfully reproduces the major vertical stratification of temperature and the concentration of atmospheric moisture in the lower layers. The model also captures variations in the vertical extent of moist layers along the orbital track. Although some localized moisture structures appear smoother than those in the ERA5 reference fields, the results demonstrate that MS3D can reconstruct the principal vertical structures of temperature and humidity and their along-track variations.

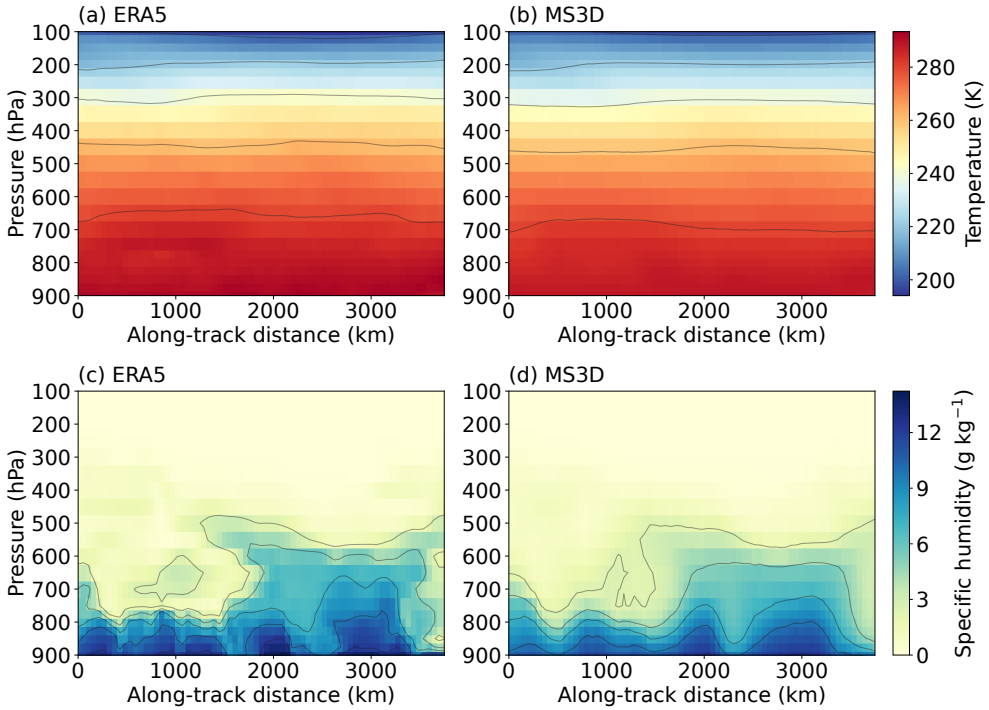


Figure 16. Vertical cross-sections of temperature and humidity from MS3D and ERA5.

5.3. Evaluation Under Cloudy Conditions

To further validate the benefits of multi-sensor input, we performed a comparative experiment under cloudy conditions, evaluating temperature and humidity profile retrieval using either three sensors (HIRAS/MWHS/MWTS) or HIRAS alone. Cloud detection was carried out using the FY-4B AGRI Cloud Mask (Full Disk) product provided by the Fengyun Satellite Remote Sensing Data Service (https://satellite.nsmc.org.cn/). This dataset offers a high spatial resolution of 4 km and has been quality-validated. Since FY-4B is a geostationary satellite, its observations are limited to a fixed region. We therefore adopted its full-disk coverage, spanning latitudes from −80.88° to 80.88° and longitudes from 23.81° to −173.81° (crossing the 180° antimeridian), as illustrated in Figure 17. The meanings of the CLM cloud mask values are summarized in Table 5.

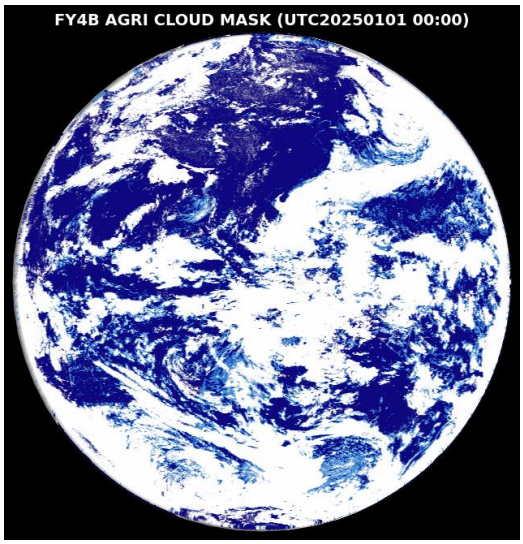
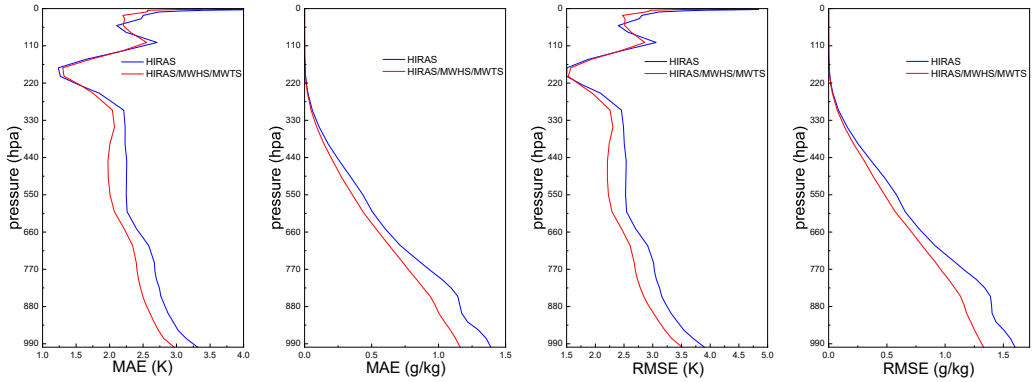


Figure 17. FY-4B AGRI Cloud Mask (Full Disk) product visualization.

Table 5. FY4B AGRI Cloud Mask (Full Disk) product parameters.

	Values	Meaning
Cloud Mask	0	cloud
	1	probably cloud
	2	probably clear
	3	clear

For spatial matching, we applied nearest-neighbor interpolation, in which the cloud classification at a target location was determined by majority voting among the four nearest latitude-longitude grid points. A maximum search radius of 1° was set. Given the 15 min temporal resolution of the cloud product, the temporally nearest available observation was selected and assumed to be representative of the cloud conditions at the target time. If no valid points were found within this range, the result was assigned as NaN and excluded from further analysis. Based on this procedure, we generated a cloud mask for part of the retrieval domain, with pixels labeled 0 (cloud) used to identify cloudy-condition samples. All retrieval samples were filtered using this cloud mask, and invalid interpolation results were discarded. The retrieval performance under cloudy conditions was then evaluated for the two sensor configurations, with MAE and RMSE as metrics. The results are presented in Figure 18.



(**a**) Temperature MAE (**b**) Humidity MAE (**c**) Temperature RMSE (**d**) Humidity RMSE

Figure 18. MAE and RMSE of 37-Level temperature and humidity profiles under cloudy conditions.

The comparison between the two experiments is particularly pronounced in the atmosphere below 220 hPa, where the HIRAS/MWHS/MWTS triple-sensor configuration exhibits substantially lower errors than HIRAS alone. This indicates that the inclusion of microwave sensors effectively enhances the retrieval of temperature and humidity profiles in the atmosphere below 220 hPa, producing significant improvements from this level down to the surface; this observation is entirely consistent with our initial hypothesis. Given that clouds are predominantly concentrated within the troposphere, the single-sensor HIRAS, operating on the infrared spectrum, faces inherent limitations in retrieving tropospheric temperature and humidity under cloudy conditions due to significant signal attenuation. However, by integrating the microwave sensors MWHS and MWTS, the MS3D framework leverages their superior cloud penetrability. Consequently, compared to the single-sensor model, MS3D effectively mitigates retrieval errors in the troposphere, demonstrating robust performance and improved accuracy even in the presence of cloud cover. Quantitative metrics averaged over the 37 pressure levels under cloudy conditions are summarized in Table 6. Under cloudy conditions, compared with the HIRAS-only configuration, MS3D reduces the MAE and RMSE of temperature retrieval by 10.1% and 10.2%, respectively,

while achieving corresponding reductions of 16.3% and 17.0% for specific humidity. These improvements highlight the benefits of incorporating complementary microwave observations under cloudy conditions.

Table 6. Average quantitative metrics over 37 pressure levels under cloudy conditions for HIRAS/MWHS/MWTS and HIRAS. The best results are highlighted in bold.

Temperature (K)	**MAE**	**RMSE**
HIRAS/MWHS/MWTS	**2.31**	**2.63**
HIRAS	2.57	2.93
Specific humidity (g/kg)	MAE	RMSE
HIRAS/MWHS/MWTS	**0.36**	**0.44**
HIRAS	0.43	0.53

We further evaluated the retrieval performance under clear-sky conditions, using a total of 5,955,710 valid atmospheric profiles. As shown in Table 7, the results demonstrate that MS3D maintains reliable retrieval performance under clear-sky conditions. As expected, the retrieval errors are generally lower than those under cloudy conditions, primarily because infrared observations are less affected by cloud interference. This trend is consistent with the physical characteristics of atmospheric sounding and further supports the reliability of the evaluation results.

Table 7. Average quantitative metrics over 37 pressure levels under sunny conditions for HIRAS/MWHS/MWTS and HIRAS. The best results are highlighted in bold.

Temperature (K)	**MAE**	**RMSE**
HIRAS/MWHS/MWTS	**2.28**	**2.58**
HIRAS	2.46	2.89
Specific humidity (g/kg)	MAE	RMSE
HIRAS/MWHS/MWTS	**0.31**	**0.40**
HIRAS	0.37	0.48

In summary, by modeling both multi-sensor interactions and 3D fields, MS3D realizes a dual perception of space and spectrum, successfully reconstructing a comprehensive 3D atmospheric field for precise temperature and humidity retrieval. Moreover, the proposed framework demonstrates excellent scalability. The multi-encoder multi-decoder architecture allows for the rapid integration of new sensors by adding encoder branches, as well as the inclusion of additional retrieval variables via new decoder branches. Consequently, this framework is well-positioned to integrate novel data sources in the future, helping to bridge data gaps and supporting broad transferability across application scenarios and datasets, such as global retrieval utilizing multi-satellite constellations. These represent promising and forward-looking directions for future exploration.

6. Conclusions

In this paper, we propose the MS3D framework, designed to overcome the spatial limitations of existing retrieval methods and the spectral constraints of single-sensor observations by synergizing broad spatial context with deep spectral penetration. Extensive experiments confirm that MS3D achieves state-of-the-art (SOTA) performance, yielding overall improvements of 10.8% (MAE) and 12.3% (RMSE) for temperature, and 7.8% (MAE) and 12% (RMSE) for humidity. Furthermore, the framework demonstrates exceptional capability in capturing dynamic atmospheric flow features and performing precise me-

teorological retrievals in all-weather scenarios. Notably, under cloudy conditions, the multi-sensor fusion significantly outperforms the single-sensor baseline, with accuracy gains of 10.1% (MAE) and 10.2% (RMSE) for temperature, and 16.3% (MAE) and 17.0% (RMSE) for humidity.

In future work, we aim to explore global retrieval utilizing multi-satellite and multi-sensor constellations. Leveraging the strong scalability of the proposed framework to integrate additional sensors and satellites will be of significant value for providing precise initial fields for global weather forecasting.

Author Contributions: Conceptualization, X.Z. (Xiangang Zhao) and L.T.; methodology, L.T.; software, L.T.; validation, L.T. and Y.Y.; formal analysis, Y.Y.; investigation, X.Z. (Xiangang Zhao); resources, X.Z. (Xi Zhang); data curation, Y.L.; writing—original draft preparation, L.T.; writing—review and editing, C.F., X.Z. (Xiangang Zhao) and M.L.; visualization, B.Z.; supervision, M.L. and X.Z. (Xiangang Zhao); project administration, X.C.; funding acquisition, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Shenzhen Science and Technology Program ZDCYKCX20250901092700001. This work was supported in part by National Nature Science Foundation of China No. 62302127.

Data Availability Statement: The raw satellite data used in this study are publicly available from the Fengyun Satellite Data Center at https://satellite.nsmc.org.cn/ (accessed on 2 July 2025)

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, Q.; Guo, X.; Zhang, R.; Li, J.; Zhang, L.; Jia, J.; Zhou, X. A Comparison of Physical-Based and Statistical-Based Radiative Transfer Models in Retrieving Atmospheric Temperature Profiles from the Microwave Temperature Sounder-II Onboard the Feng-Yun-3 Satellite. *Atmosphere* **2025**, *16*, 44.
2. Chen, L.; Xu, N.; Wang, J.S.; Shang, J.; Shou, Y.X.; Li, B.; Xu, R.H.; Wu, S.L.; Wang, X.; Zheng, W.; et al. FengYun satellites: From observations to quantitative applications. *Natl. Remote Sens. Bull.* **2025**, *29*, 1462–1479. https://doi.org/10.11834/jrs.20254459.
3. Thies, B.; Bendix, J. Satellite based remote sensing of weather and climate: Recent achievements and future perspectives. *Meteorol. Appl.* **2011**, *18*, 262–295.
4. Li, J.; Geer, A.J.; Okamoto, K.; Otkin, J.A.; Liu, Z.; Han, W.; Wang, P. Satellite All-sky Infrared Radiance Assimilation: Recent Progress and Future Perspectives. *Adv. Atmos. Sci.* **2022**, *39*, 9–21.
5. Hang, R.; Cao, S.; Shang, J.; Ge, L.; Shi, C.; Liu, Q. Physical Knowledge Constrained Convolutional Network for Temperature Profile Retrieval from FY-4A/GIIRS Hyperspectral Data. *J. Remote Sens.* **2025**, *5*, 0841.
6. Yao, S.; Guan, L. Comparison of three convolution neural network schemes to retrieve temperature and humidity profiles from the FY4A GIIRS observations. *Remote Sens.* **2022**, *14*, 5112.
7. Xiao, Y.; Hu, S.; Deng, W.; Dang, R.; Liu, L.; Huang, W.; Yang, W. Integrated Retrieval of the Temperature and Humidity Profiles of Atmospheric Boundary Layer by Combining Ground-Based Infrared Hyperspectral Interferometers and Microwave Radiometers. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–13.
8. Qu, Y.; Goldberg, M.D.; Divakarla, M. Ozone profile retrieval from satellite observation using high spectral resolution infrared sounding instrument. *Adv. Atmos. Sci.* **2001**, *18*, 959–971.
9. Chedin, A.; Scott, N.A.; Wahiche, C.; Moulinier, P. The Improved Initialization Inversion Method: A High Resolution Physical Method for Temperature Retrievals from Satellites of the TIROS-N Series. *J. Appl. Meteorol.* **1985**, *24*, 128–143.
10. Smith, W.L.; Woolf, H.M. The Use of Eigenvectors of Statistical Covariance Matrices for Interpreting Satellite Sounding Radiometer Observations. *J. Atmos. Sci.* **1976**, *33*, 1127–1140.
11. Smith, W.L.; Weisz, E.; Kireev, S.V.; Zhou, D.K.; Li, Z.; Borbas, E.E. Dual-Regression Retrieval Algorithm for Real-Time Processing of Satellite Ultraspectral Radiances. *J. Appl. Meteorol. Climatol.* **2012**, *51*, 1455–1476.
12. Li, N.; Guan, L.; Wright, J.S. Cloud detection and sea surface temperature retrieval by HY-1C COCTS observations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19853–19863.
13. Tan, H.; Mao, J.; Chen, H.; Chan, P.W.; Wu, D.; Li, F.; Deng, T. A study of a retrieval method for temperature and humidity profiles from microwave radiometer observations based on principal component analysis and stepwise regression. *J. Atmos. Ocean. Technol.* **2011**, *28*, 378–389.

14. Duma, Z.-S.; Sihvonen, T.; Susiluoto, J.; Lamminpää, O.; Haario, H.; Reinikainen, S.-P. Kernel-based retrieval models for hyperspectral image data optimized with Kernel Flows. In Proceedings of the 2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Helsinki, Finland, 9–11 December 2024; pp. 1–5.
15. Svendsen, D.H.; Morales-Álvarez, P.; Ruescas, A.B.; Molina, R.; Camps-Valls, G. Deep Gaussian processes for biogeophysical parameter retrieval and model inversion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 68–81.
16. Chahine, M.T. Inverse problems in radiative transfer: Determination of atmospheric parameters. *J. Atmos. Sci.* **1970**, *27*, 960–967.
17. Rodgers, C.D. Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys.* **1976**, *14*, 609–624.
18. Lorenc, A.C. Optimal nonlinear objective analysis. *Q. J. R. Meteorol. Soc.* **1988**, *114*, 205–240.
19. Huang, W.; Liu, L.; Yang, B.; Hu, S.; Yang, W.; Li, Z.; Li, W.; Yang, X. Retrieval of temperature and humidity profiles from ground-based high-resolution infrared observations using an adaptive fast iterative algorithm. *Atmos. Meas. Tech.* **2023**, *16*, 4101–4114.
20. Moreau, E.; Lopez, P.; Bauer, P.; Tompkins, A.M.; Janisková, M.; Chevallier, F. Variational retrieval of temperature and humidity profiles using rain rates versus microwave brightness temperatures. *Q. J. R. Meteorol. Soc.* **2004**, *130*, 827–852.
21. Hewison, T.J. 1D-VAR retrieval of temperature and humidity profiles from a ground-based microwave radiometer. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 2163–2168.
22. Benedetti, A.; Lopez, P.; Moreau, E.; Bauer, P.; Venugopal, V. Verification of TMI-adjusted rainfall analyses of tropical cyclones at ECMWF using TRMM precipitation radar. *J. Appl. Meteorol.* **2005**, *44*, 1677–1690.
23. Robert, C.; Durbiano, S.; Blayo, E.; Verron, J.; Blum, J.; Le Dimet, F.-X. A reduced-order strategy for 4D-Var data assimilation. *J. Mar. Syst.* **2005**, *57*, 70–82.
24. Thapliyal, P.K.; Shukla, M.V.; Bisht, J.H.; Pal, P.K.; Navalgund, R.R. Improvement in the retrieval of humidity profiles using hybrid regression technique from infrared sounder data: A simulation study. *Meteorol. Appl.* **2014**, *21*, 301–308.
25. Bi, Y.; Zhang, P.; Yang, Z.; Wang, Q.; Zhang, X.; Liu, C.; Xu, P.; Hou, L.; Ke, J.; Zhang, N. Fast CO2 retrieval using a semi-physical statistical model for the high-resolution spectrometer on the fengyun-3D satellite. *J. Meteorol. Res.* **2022**, *36*, 374–386.
26. Wang, H.; Mao, K.; Shi, J.; Bateni, S.M.; Altantuya, D.; Sainbuyan, B.; Bao, Y. A normal form for synchronous land surface temperature and emissivity retrieval using deep learning coupled physical and statistical methods. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103704.
27. Singh, D.; Bhatia, R.C. Study of temperature and moisture profiles retrieved from microwave and hyperspectral infrared sounder data over Indian regions. *Proc. SPIE* **2006**, *6408*, 640818.
28. Huang, P.; Guo, Q.; Han, C.; Zhang, C.; Yang, T.; Huang, S. An Improved Method Combining ANN and 1D-Var for the Retrieval of Atmospheric Temperature Profiles from FY-4A/GIIRS Hyperspectral Data. *Remote Sens.* **2021**, *13*, 481.
29. Milstein, A.B.; Blackwell, W.J. Neural network temperature and moisture retrieval algorithm validation for AIRS/AMSU and CrIS/ATMS. *J. Geophys. Res. Atmos.* **2016**, *121*, 1414–1430.
30. Liu, S.; Grassotti, C.; Liu, Q.; Zhou, Y.; Lee, Y.-K. Improvement of MiRS sea surface temperature retrievals using a machine learning approach. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1857–1868.
31. Malmgren-Hansen, D.; Laparra, V.; Nielsen, A.A.; Camps-Valls, G. Statistical retrieval of atmospheric profiles with deep convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 231–240.
32. Ma, M.; Ma, W.; Jiao, L.; Liu, X.; Li, L.; Feng, Z.; Liu, F.; Yang, S. A multimodal hyper-fusion transformer for remote sensing image classification. *Inf. Fusion* **2023**, *96*, 66–79.
33. Cao, Y.; Luo, X.; Yang, J.; Cao, Y.; Yang, M.Y. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Inf. Fusion* **2022**, *88*, 1–11.
34. Li, Z.; Guo, F.; Zhang, Z.; Zhang, X. Correction of the SMAP sea surface brightness temperature and retrieval of sea surface salinity incorporating CYGNSS observables. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3226–3235.
35. Wang, Z.; Wang, X.; Li, G.; Wu, W.; Liu, Y.; Song, Z.; Song, H. Historical information fusion of dense multi-source satellite image time series for flood extent mapping. *Inf. Fusion* **2024**, *109*, 102445.
36. Han, L.H.; Jiang, G.M. FY-3E HIRAS-II Channel Selection for Temperature and Humidity Profile Retrieval using Principal Component Analysis and Weighting Functions. In Proceedings of the 2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece, 31 October 2023–2 November 2023; pp. 1–6.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.