Cloud Detections and Properties Retrievals in Infrared Hyperspectral Observations Based on Different Machine-Learning Algorithms

Qikai Hu
wsxgshqk@umich.edu
University of Michigan, Ann Arbor
Ann Arbor, MI, USA

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the “acmart” document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

ACM Reference Format:
Qikai Hu. 2024. Cloud Detections and Properties Retrievals in Infrared Hyperspectral Observations Based on Different Machine-Learning Algorithms. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Abstract

This project focuses on cloud detection and cloud property retrieval. For the cloud detection problem, the project evaluates the performance of traditional machine learning algorithms and compares them with two physical methods. For the cloud property retrieval problem, the project demonstrates that using simulation datasets generated by a radiative transfer model enhances the performance of DNNs by incorporating physical principles. Additionally, the project highlights that applying PCA to the model further improves performance, likely by reducing overfitting and enhancing feature representation.

2 Introduction of the problem

Clouds play a crucial role in Earth’s energy budget and atmospheric circulation [1]. They are responsible for precipitation and contribute significantly to maintaining the climate system’s energy balance [2]. Cloud properties, including cloud fraction and cloud top pressure refers to the proportion of cloud cover within satellite pixels or within grid cells in weather and climate models. It stands as one of the most critical parameters in modeling the downward radiation at the Earth’s surface and the top of the atmosphere [3], representing one of the largest sources of uncertainty in global climate models. Given its essential role in climate models, obtaining accurate and dependable estimates of cloud cover and cloud top height is paramount for climate research.

Hyperspectral instruments can provide abundant information of clouds, offering high spectral resolution that enhances our understanding of cloud properties. Traditional cloud retrieval methods, such as FRESCO+ (Fast Retrieval Scheme for Cloud Observations), primarily leverage the absorption features of O to estimate cloud parameters. These retrieval algorithms relies on the principle that clouds can change the radiative transfer path, which is detectable in spectral observations. However, traditional approaches, which are often based on look-up tables (LUTs) and optimal estimation techniques, face limitations in retrieval efficiency when dealing with the increasingly high spectral and spatial resolution of modern hyperspectral instruments.

Recently, deep learning-based algorithms for cloud and aerosol retrieval have gained traction, demonstrating superior accuracy compared to LUT-based methods. These approaches establish a direct relationship between the satellite-observed top-of-atmosphere reflectance (TOA) and the desired retrieval parameters. While deep learning methods excel due to their robust modeling capabilities, they often lack a unified physical framework for optimizing the integration of diverse satellite measurements. This presents a significant challenge for advancing the retrieval of cloud and aerosol properties, as fully leveraging the potential of hyperspectral observations requires a deeper synthesis of physical principles and data-driven methodologies.

To overcome the limitations mentioned above, this research aims to propose a novel physics-based DNN retrieval method on the cloud properties (e.g.: cloud fraction) assessment using imaging sounder instruments. The proposed phase algorithms will be compared to currently operational infrared-based retrievals.

3 Mathematical background

For a plane-parallel, homogeneous cloud that produces the same radiance I_ν observed at the top of the atmosphere, the measured radiance I_ν at wavenumber ν depends on the cloud fraction η and cloud top pressure p_c and is given by the following equation (Liou, 2002):

$$I_\nu = (1 - \eta\epsilon_\nu)(I_s + I_b)t_\nu(p_c, 0) + I_c + I_a, \tag{1}$$

where:

- The surface contribution is described by

$$I_s = B_\nu(T_s)t_\nu(p_s, p_c), \tag{2}$$

Conference’17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- The contribution of the atmosphere below the cloud is

$$I_b = \int_{p_s}^{p_c} B_\nu(T(p)) \frac{\partial t_\nu(p, p_c)}{\partial p} \, dp, \tag{3}$$

- The cloud's own contribution is

$$I_c = \eta \epsilon_\nu B_\nu(T_c) t_\nu(p_c, 0), \tag{4}$$

- The contribution of the atmosphere above the cloud is

$$I_a = \int_{p_c}^{0} B_\nu(T(p)) \frac{\partial t_\nu(p, 0)}{\partial p} \, dp. \tag{5}$$

In these equations:

- ϵ_ν is the spectral emissivity of the cloud,
- B_ν is the Planck radiation,
- $t_\nu(p_1, p_2)$ is the transmissivity between pressures p_1 and p_2,
- T_c and T_s are the temperatures of the cloud and the surface, respectively, while p_c and p_s are their respective pressures.

4 Related prior work

Over the past two decades, several infrared-based cloud detection algorithms have been developed for various purposes. Notably, a widely adopted channel-dependent cloud retrieval method [8] determines the clouds by the spectral difference between the satellite observations and the cloud-free model forward simulations. However, nowadays cloud detection is challenging for these methods due to the large number of sounders [4]. Another way to accurately identify clouds is to use information from multiple instrument pairs. For example, a method developed by the Atmospheric Infrared Sounder (AIRS) science team gets cloud information spatially and reserves the spectral information of infrared sounders for profile retrievals [5]. Additionally, the cloud information can also be determined by a collocated high-spatial-resolution imager, such as the AIRS observations [6] . Finally, effective cloud detection is calculated by comparing the radiance observed in selected satellite channels from the two resources through modeling [7]. This approach assumes uniform cloud emissivity across the spectrum and selects channels most sensitive to clouds. The AIRS cloud retrieval method has also been successfully extended to other instruments, including the Infrared Atmospheric Sounding Interferometer (IASI) and Cross-track Infrared Sounder (CrIS) [8]. While significant progress has been made in utilizing imaging instruments to identify cloud contamination in infrared sounder data, most of these advancements have primarily focused on distinguishing between cloudy and clear-sky conditions [10]. The capability of imaging instruments to assess cloud coverage at the sub-pixel level for sounder instruments has not received extensive discussion and exploration.

Machine learning-based retrieval algorithms have been developed in recent years to address the limitations of physical models. For example, Kox et al. (2014) [12] introduced a neural network method that utilizes seven infrared bands from the geostationary Meteosat Second Generation (MSG) satellite, which successfully retrieves cirrus cloud properties, specifically cloud top height (CTH) and cloud optical thickness (COT) for clouds with optical thickness $\tau < 2.5$, comparable to target values from the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO). Neural networks have also been extended to handle opaque ice clouds with $\tau > 8$ (Hong et al., 2010, 2012) [13] [14].

Subsequently, Minnis et al. (2016) [15] developed a COT-retrieval algorithm based on infrared radiances at wavelengths of 3.7, 6.7, 11.0, and 12.0 μm from the Moderate Resolution Imaging Spectroradiometer (MODIS), capable of predicting opaque ice COT values up to 150 with substantially improved accuracy. Most recently, Min et al. (2020) [16] proposed a machine learning-based algorithm for CTH estimation using Himawari Imager (AHI) data. The authors trained four machine learning algorithms, all of which demonstrated improved retrieval accuracy compared to traditional physical algorithms.

5 Algorithms

5.1 Physical Method

Two physical methods on the clouds detection includes thermal contrast and the bispectral algorithm. One physical retrieval method of the cloud properties is the *optimal estimation* (OE) method. It uses several channels and any available prior information with suitable weighting according to errors. Essential diagnostic outputs of the OE method are a measure of the model fit to the observation, that is, the cost function J, and formal error estimates of the retrieved parameters. The cost function J is defined as follows:

$$J(\mathbf{x}) = (\mathbf{y}(\mathbf{x}) - \mathbf{y}_m)^{\mathrm{T}} \mathrm{S_y}^{-1} (\mathbf{y}(\mathbf{x}) - \mathbf{y}_m) + (\mathbf{x} - \mathbf{x_a})^{\mathrm{T}} \mathrm{S_a}^{-1} (\mathbf{x} - \mathbf{x_a}), \tag{6}$$

where $\mathbf{y}_m$ are the measurements and $\mathbf{y}(\mathbf{x})$ are the radiances simulated by assuming the state $\mathbf{x}$. $\mathbf{x_a}$ is an estimate of the state prior to the retrieval, and $\mathrm{S_y}^{-1}$ and $\mathrm{S_a}^{-1}$ are the inverted error covariance matrices of the measurements and the prior state, respectively. The state vector $\mathbf{x}$ is varied to minimize the cost function J, yielding the optimal state given the observation $\mathbf{y}_m$, the prior knowledge, and their respective uncertainties.

In the context of this paper, the state parameter $\mathbf{x}$ contains the physical properties of the cloud. Using a window channel alone leads to very large solution spaces for CTP (cloud top pressure) retrievals for cirrus, while the inclusion of a single absorbing channel greatly decreases the solution space. Iterative techniques can also be used to simultaneously fit more than one parameter to the same number of channels.

5.2 DNN model

This project will first employ a multi-layer neural network for the inference, consisting of an input layer, hidden layers, and one output layer. To enhance the performance of the model and facilitate faster convergence and regression, the highly correlated channels in the original spectra may be first transformed into their principal components (PCs) and only the leading PCs are utilized as model inputs. In this model, field-of-views (FOVs) with a clear-sky probability value greater than some specific value as the threshold will be classified as clear sky, while those falling below this threshold are deemed cloudy. The DNN model will be trained to optimize a categorical cross-entropy loss function using batch gradient descent (BGD), with all training data considered in each iteration step. More advanced architecture may be employed later for better performance.

Additionally, this study will combine the DNN model with simulation-based inference, where the DNN models will act like a surrogate model. Simulators are the modern manifestation of scientific theories. They implement mechanistic models of the underlying natural phenomena of interest as well as models for the instruments used to observe those phenomena. Specifically, the goal of Simulation-based Inference (SBI) is to infer the parameters q, given empirically observed data , i.e., to obtain the posterior distribution . In our research, the cloud fraction will be taken as the parameters and the satellite radiance will be taken as the observed data x. SBI can be divided into two components, including the simulators and the inference. A simulator is defined as an algorithm that, given a parameter vector as input, generates a series of internal states or latent variables, denoted as , and ultimately produces an output data vector . Here the parameters q are responsible for characterizing the fundamental mechanistic model, influencing the transition probabilities accordingly. The latent variables z involved in the data-generation process may have a direct or indirect association with a physically significant system state. Finally, the output data x represents the observations. These observations can vary widely, from a small set of unstructured numerical values to high-dimensional and intricately structured data.

To summarize, the model will first use a simulator to generate synthetic data for various parameter sets, and then train the DNN on this simulated dataset. Once trained, apply the model to real observations, sampling from the posterior to infer parameters, thus leveraging the DNN as a surrogate to approximate intractable posteriors in simulation-based inference.

5.3 Neural Operators

At the moment, there are dozens of different neural operators, among which we can distinguish meta-architectures (e.g., deep operator networks) and operators based on integral transforms (e.g., the Fourier neural operator (FNO)). Based on the mathematical background of the cloud properties retrieval problem, We will focus on the properties of networks with integral operators.

Networks with integral operators are based on a natural generalization of the linear layer. A classical neural network transforms vectors using the operator

$$\mathbf{v} = A\mathbf{u} + \mathbf{b}, \quad \mathbf{v}, \mathbf{b} \in \mathbb{R}^n, \quad \mathbf{u} \in \mathbb{R}^m, \quad A \in \mathbb{R}^{n\times m}, \tag{7}$$

where $\mathbf{v}$ and $\mathbf{u}$ are the input and output vectors of the layer and A and $\mathbf{b}$ are the layer parameters. By analogy, a linear layer of the neural operator has the form

$$\mathbf{v}(x) = \int dy\, A(x, y; \mathbf{w}_A)\mathbf{u}(y) + \mathbf{b}(x; \mathbf{w}_b), \tag{8}$$

where $\mathbf{v}(x)$ and $\mathbf{u}(x)$ are the layer input and output functions, $A(x, y; \mathbf{w}_A)$ is the kernel of the integral operator, and $\mathbf{w}_A \in \mathbb{R}^{N_A}$ and $\mathbf{w}_b \in \mathbb{R}^{N_A}$ are the layer parameters.

For reasons of numerical efficiency, a linear layer (8) is often implemented using the parametrization

$$\int dy\, A(x, y; \mathbf{w}_A)\mathbf{u}(y) = \sum_{ij} \phi_j(x) w_{ji} \int dy\, \phi_i(y)\mathbf{u}(y), \tag{9}$$

along with fast transforms that allow one to efficiently calculate the coefficients $c_i = \int dy\, \phi_i(y)\mathbf{u}(y)$ and the sum $\sum_j \phi_j(x)\tilde{c}_j$.

In this project we will implement the neural operators by both applying PCA and not applying PCA to the input vector to examine the performance of the dimensioanlity reduction methods.

6 Experimental Setup

6.1 Data

Data collected from AIRS (Atmospheric Infrared Sounder) and MODIS (Moderate Resolution Imaging Spectroradiometer) will be included in the study as the observation data. AIRS and MODIS are two key satellite instruments aboard NASA's Aqua satellite, each providing valuable data on radiances and also cloud properties. MODIS cloud properties (including Cloud Top Pressure) are derived from 36 spectral bands in the visible, near infrared and infrared regions at high spatial resolution (1 – 5 km), while the retrieval of AIRS is based on eigenvector regression, coming with a lower spatial resolution than the products of MODIS. The dataset can be obtained from the website: https://airs.jpl.nasa.gov/data/get-data/standard-data/.

Specifically, AIRS observations over tropical oceans in June 2018 are used, with collocated MODIS cloud flags as target labels, totaling 1,760,943 samples. The original input features include brightness temperatures from 1,551 selected mid-infrared AIRS channels and monthly sea surface temperatures. Cloud properties, including cloud fractions, are used as target for the algorithms. Both input and target features are stored together as a 1554 dimension vector for each sample, requiring 200GB of storage.

6.2 Structure of Models

The experiments will be divided into two parts. In the first part, we will develop a cloud detection classifier using traditional machine learning models, including LinearSVC, Random Forest, Gradient Boosting, and FNN, alongside two physical methods: thermal contrast and the bispectral algorithm. In the second part, we will address a cloud fraction retrieval problem, applying both a traditional DNN and a simulation-based DNN (physics-based DNN).

The DNN model employed in this study features five hidden layers, with the number of neurons per layer set to 512, 125, 128, 64, and 16, respectively. The activation function selected for this architecture is Leaky Relu. Then the observation dataset is set as the training dataset. The training dataset is divided into training and testing subsets at a ratio of 8:2. The training pattern is adopted to optimize the weights and biases of each neuron iteratively.

DNN models with SBI has the same structure of the neural network as the DNN model in the previous, but it takes simulation dataset as the training dataset. The training dataset is generated based on the SCIATRAN radiative transfer model.

Both the standard DNN models and the simulation-based DNN models will be integrated with a feature extraction method utilizing Principal Component Analysis (PCA) to test if the feature extraction method can help to reduce the over-fitting.

After the implementation of the models, the results will be compared by the running time and performance, including the Root Mean Square Error (RMSE) of the cloud fraction and the cloud top

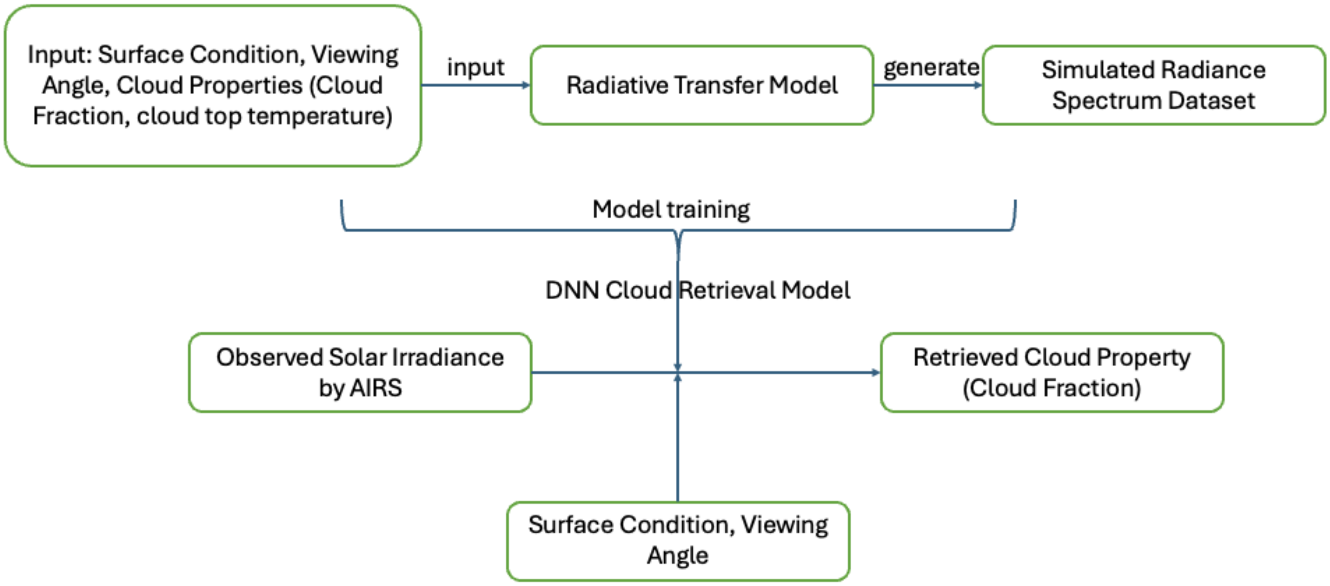

Input: Surface Condition, Viewing Angle, Cloud Properties (Cloud Fraction, cloud top temperature)
input
Radiative Transfer Model
generate
Simulated Radiance Spectrum Dataset
Model training
DNN Cloud Retrieval Model
Observed Solar Irradiance by AIRS
Retrieved Cloud Property (Cloud Fraction)
Surface Condition, Viewing Angle

Figure 1: Structure of Simulation-based DNN

pressure as following:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

where:

- n is the number of data points,
- y_i is the actual value,
- $\hat{y}_i$ is the predicted value.

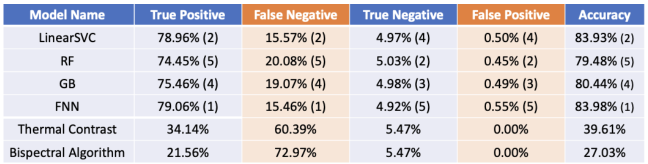
After that, a case study will be performed and the bias of the results may be extensively examined to gain physical understandings of the advantages and disadvantages of the models.

7 Results

The results of the models, including the physical algorithm, traditional machine learning models, DNN (FNN), and simulation-based DNN (physics-based DNN), are presented in Figures 2 and 3. From these results, we can conclude the following:

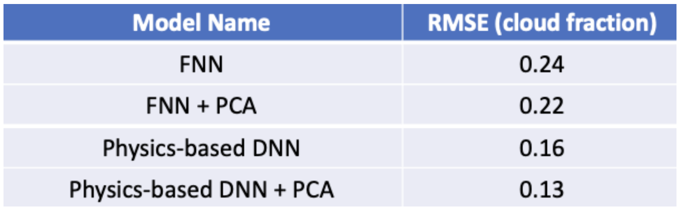
In the cloud detection problem, all traditional machine learning models outperform the physical methods. Among them, the LinearSVC and FNN models demonstrate slightly better performance compared to the others.

In the cloud fraction retrieval problem, the physics-based DNN (simulation-based DNN) outperforms the standard DNN model. Additionally, the application of PCA further reduces the RMSE, likely due to its ability to mitigate overfitting by reducing the dimensionality of the input features.



Model Name	True Positive	False Negative	True Negative	False Positive	Accuracy
LinearSVC	78.96% (2)	15.57% (2)	4.97% (4)	0.50% (4)	83.93% (2)
RF	74.45% (5)	20.08% (5)	5.03% (2)	0.45% (2)	79.48% (5)
GB	75.46% (4)	19.07% (4)	4.98% (3)	0.49% (3)	80.44% (4)
FNN	79.06% (1)	15.46% (1)	4.92% (5)	0.55% (5)	83.98% (1)
Thermal Contrast	34.14%	60.39%	5.47%	0.00%	39.61%
Bispectral Algorithm	21.56%	72.97%	5.47%	0.00%	27.03%

Figure 2: Results of Cloud Detection.



Model Name	RMSE (cloud fraction)
FNN	0.24
FNN + PCA	0.22
Physics-based DNN	0.16
Physics-based DNN + PCA	0.13

Figure 3: Results of Cloud Fraction

8 Discussion

8.1 Differences between simulation and real world data

To evaluate the differences between simulated and observational data, we compared the distribution of cloud fractions derived from both datasets. The simulation data exhibits an even distribution across all cloud fraction values. In contrast, Figure 4 highlights a highly skewed distribution in the observational data, with the majority of values concentrated near 0 and 1. This discrepancy underscores potential challenges in accurately capturing real-world cloud fraction variability in simulations, which may arise from limitations in the model's representation of cloud processes or its underlying parameterizations.

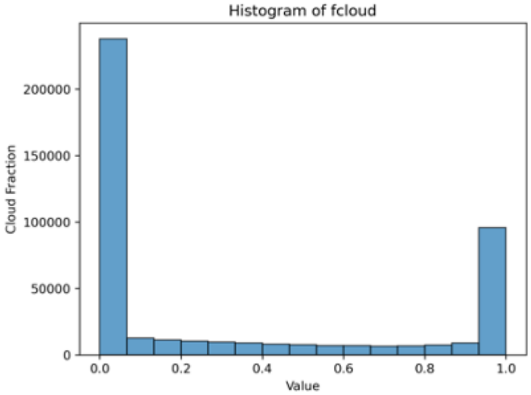


Figure 4: Histgram of Cloud Fraction

8.2 Case Study

To investigate the sources of prediction errors, we utilized the EOS Worldview API and developed a simple interface for conducting case studies. Figure 5 presents true-color images of selected samples where all models performed poorly. The red circle at the center of each panel highlights the AIRS footprint. From the images, it is evident that these cases involve broken cloud cover. This issue may stem from the limitations of non-graph-based algorithms, which do not account for spatial information in their predictions.

8.3 Other Methods

Neural operators, especially the one with integral operators here, are designed to learn complex mappings between input and output spaces in a flexible, data-driven way. They can generalize well across a wide range of conditions without needing explicit physical integration equations, making them adaptable to various scenarios. This flexibility is particularly useful when dealing with nonlinear relationships that are difficult to capture through physics-based equations alone. Also the inverse of the clouds fraction and clouds top pressure is not limited to one solution here according to the physical laws. The flexibility of the neural operator may enable the performance of the model to be more robust. Thus, neural operators can be applied to the retrieval problem in the future.

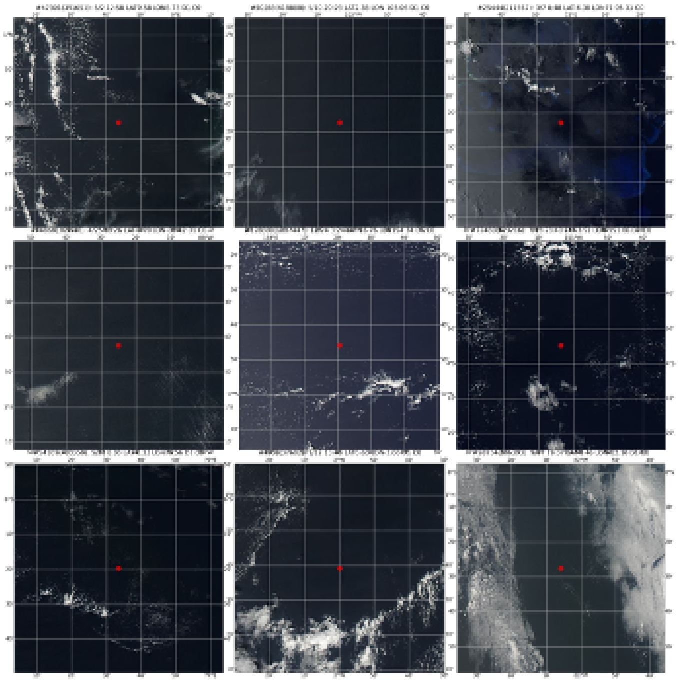


Figure 5: Case Study

9 References

1. Liu, Qian, et al. "Cross-track infrared sounder cloud fraction retrieval using a deep neural network." Computers Geosciences 170 (2023): 105268.

2. Liu, Qian, et al. "Hyperspectral infrared sounder cloud detection using deep neural network model." IEEE Geoscience and Remote Sensing Letters 19 (2020): 1-5.

3. Basurto-Hurtado, Jesus A., et al. "Diagnostic strategies for breast cancer detection: from image generation to classification strategies using artificial intelligence algorithms." Cancers 14.14 (2022): 3442.

4. Boucher, Eulalie, and Filipe Aires. "Towards a new generation of artificial-intelligence-based infrared atmospheric sounding interferometer retrievals of surface temperature: Part II–Assessment." Quarterly Journal of the Royal Meteorological Society 149.754 (2023): 1593-1611.

5. Bai, Wenguang, et al. "A fast piecewise-defined neural network method to retrieve temperature and humidity profile for the vertical atmospheric sounding system of FengYun-3E satellite." IEEE Transactions on Geoscience and Remote Sensing 61 (2023): 1-10.

6. Milstein, Adam B., Joseph A. Santanello, and William J. Blackwell. "Detail Enhancement of AIRS/AMSU Temperature and Moisture Profiles Using a 3D Deep Neural Network." Artificial Intelligence for the Earth Systems 2.2 (2023): 220037.

7. Milstein, Adam B., and William J. Blackwell. "Neural network temperature and moisture retrieval algorithm validation for AIRS/AMSU and CrIS/ATMS." Journal of Geophysical Research: Atmospheres 121.4 (2016): 1414-1430.

8. Ramesh, Poornima, et al. "GATSBI: Generative adversarial training for simulation-based inference." arXiv preprint arXiv:2203.06481 (2022).

9. Cranmer, Kyle, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference." Proceedings of the National Academy of Sciences 117.48 (2020): 30055-30062.

10. Y. Han and Y. Chen, "Calibration algorithm for cross-track infrared sounder full spectral resolution measurements," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 1008–1016, Feb. 2018, doi: 10.1109/TGRS.2017.2757940.

11. P. Bauer et al., "Satellite cloud and precipitation assimilation at oper- ational NWP centres," Quart. J. Roy. Meteorological Soc., vol. 137, no. 661, pp. 1934–1951, Oct. 2011, doi: 10.1002/qj.905.

12. A. P. McNally and P. D. Watts, "A cloud detection algorithm for high- spectral-resolution infrared sounders," Quart. J. Roy. Meteorol. Soc., vol. 129, pp. 3411–3423, Apr. 2003, doi: 10.1256/q.j.02.208.

13. L. Lin, X. Zou, and F. Weng, "Combining CrIS double CO2 bands for detecting clouds located in different layers of the atmosphere," J. Geophys. Res. Atmos., vol. 122, no. 3, pp. 1811–1827, Feb. 2017, doi: 0.1002/2016JD025505.

14. J. Susskind, C. D. Barnet, and J. M. Blaisdell, "Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data in the presence of clouds," IEEE Trans. Geosci. Remote Sens., vol. 41, no. 2, pp. 390–409, Feb. 2003.

15. G. Antonia. (Aug. 2017). The NOAA Unique Combined Atmospheric Processing System (NUCAPS) Algorithm Theoretical Basis Document, Version 2.0. [Online].

16. J. Li, W. P. Menzel, F. Sun, T. J. Schmit, and J. Gurka, "AIRS sub- pixel cloud characterization using MODIS cloud products," J. Appl. Meteorol., vol. 43, pp. 1083–1094, Apr. 2004, doi: 10.1175/1520-0450(2004)043<1083:ASCCUM>2.0.CO.2.

17. R. Eresmaa, "Imager-assisted cloud detection for assimilation of infrared atmospheric sounding interferometer radiances: Imager-assisted cloud detection for radiance assimilation," Quart. J. Roy. Meteorological Soc., vol. 140, no. 684, pp. 2342–2352, Oct. 2014, doi: 10.1002/qj.2304.