A New Atmospheric Dataset for GIIRS Sampled from ERA5 Using Shannon Entropy Method

Qi Zhang¹,², Yi Yu¹ and Weimin Zhang¹,³*
¹ College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China
² College of Computer Science, National University of Defense Technology, Changsha 410073, China
³ Key Laboratory of Software Engineering for Complex Systems, Changsha 410073, China
Email: weiminzhang@nudt.edu.cn (W M Zhang), zhangq@nudt.edu.cn (Q. Zhang)

Abstract. Geostationary Interferometric Infrared Sounder (GIIRS) on board the new Chinese geostationary meteorological satellite Fengyun-4A is aimed to obtain three-dimensional atmospheric information regularly at the regional scale. Since the existing atmospheric datasets are global covered and they are not representative of the GIIRS detection area, a diverse and local representative atmospheric dataset is required for GIIRS in training transmittance parameterization scheme of fast radiative transfer model and retrieving atmospheric profiles. The new atmospheric dataset for GIIRS application in this paper is sampled from the European Centre for Medium-Range Weather Forecasts (ECMWF) fifth-generation reanalysis (ERA5) 2019 dataset using Shannon entropy sampling method. The new dataset demonstrates the feature of uniform distribution within multiple atmospheric variables’ (i.e. atmospheric temperature, specific humidity, 2 meter temperature) variation ranges and GIIRS’s spatial coverage. To address the issue derived from the climate difference and the varied vertical pressure structure of regions with great differences in topographic height, the study area is divided into six categories regarding of topographic height categories where atmospheric dataset are sampled separately. Seasonal variation is also taken into account in the process of sampling. Compared with the initial randomly sampled dataset, the atmospheric and 2 meter temperature in the new dataset demonstrate a higher level of uniformity within their variation ranges. The statistical results indicate that the differences between samples in the new dataset are enlarged, meanwhile, the new dataset well retains the statistical characteristics of the original ERA5 dataset. The new dataset has a potential in the study of the utility of GIIRS data in the numerical weather forecast.

Keywords: atmospheric profiles; GIIRS; hyperspectral infrared sounder; temperature; humidity; fast radiative transfer model

1. Introduction
Hyperspectral Infrared (HIR) instruments are able to provide atmospheric information at a higher accuracy and with more vertical resolution than lower-spectral-resolution infrared radiance (IR) sounder for Numerical Weather Prediction (NWP) [1]. In the proceedings of utilizing the HIR data, a representative atmospheric dataset is essential for training the transmittance parameterization scheme in the radiative transfer model (RTM) [2-3] and atmospheric profile retrieval schemes [4-5].
A prerequisite for exploiting the specific HIR data in the NWP model is the development of a fast radiative transfer model (FRTM) [6]. The core of the FRTM is a fast transmittance calculation that is parameterized by a database of accurate line-by-line (LBL) transmittances computed from a set of diverse and representative atmospheric datasets [7]. Besides, the accuracy of the statistical atmospheric profile retrieval method depends on the temporal and spatial representative atmospheric training samples [1].

The Geostationary Interferometric Infrared Sounder (GIIRS) on board the Fengyun-4A satellite is the first instrument to continuously monitor the vertical structure of the atmosphere and capture the atmospheric motions in both horizontal and vertical directions before the happening of high-impact weather events [8, 9]. Different from HIR sounders on board the polar orbiting satellites (i.e. the Atmospheric Infrared Sounder (AIRS), the Infrared Atmospheric Sounding Interferometric (IASI), the Cross-track Infrared Sounder (CrIS)) which detect the global atmosphere, the GIIRS sounder only detects a specific region. A regional representative atmospheric dataset can better represent the atmospheric conditions within GIIRS’s observation coverage, which will also definitely improve the radiances simulation and atmospheric products retrieval accuracy of GIIRS. However, at present, such a regional representative atmospheric dataset is unavailable for GIIRS.

Focusing on different targets, a few atmospheric datasets consisting of different profile numbers and contents are available and widely used. Thermodynamic Initial Guess Retrieval (TIGR), developed by Laboratoire de Météorologie Dynamique. Building on the experience from TIGR, a series of diverse profile datasets from short-range forecasts are set up by ECMWF center. Nevertheless, the sampling method used in TIGR and ECMWF series dataset has limitation in generalization because the dissimilarity distance, the distance threshold and standard deviation of variable normalization, which are needed in measure the dissimilarity between two atmospheric situations, are in fact hard to define. Besides, the variable independent sampling methods in [10, 11] are not optimal where multiple variables are statistically retrieved at the same time. The global covered clear-sky dataset SeeBor consists of 15704 profiles of temperature, moisture, and ozone at 101 pressure levels. This dataset combines profiles from a number of sources (e.g. ECMWF, TIGR-43, ozone sondes, etc.), but the demand for a uniformly distributed dataset is not considered in the process of sampling. Didi has sampled a local training dataset consisting of 70 profiles from Seebor V5 and ECMWF-83, which simply use Euclidean distance as dissimilarity measurement index and different variables’ profiles are sampled independently. Besides, some small datasets used in training fast radiative transfer models, such as University of Maryland Baltimore County (UMBC)-48 and ECMWF-83, are usually re-sampled from the above mentioned TIGR and ECMWF series datasets, and they are not large enough for the statistical retrieval experiments.

None of the atmospheric datasets mentioned above considered the topographic element during the sampling process. Researchers find that the influence of topography on the land surface temperature differs according to the amount of the solar radiation received [12] and the precipitation is popularly correlated with topographic height [13]. Besides, it should be noted that topographic height differences of different regions will cause the varied pressure structures. Thus, we considered atmospheric datasets of significant topographic height differences to be sampled separately in this paper.

Taking the above discussed aspects into account, a new regional representative atmospheric dataset is built for GIIRS using Shannon entropy method. Samples’ diversity and topographic elements are taken into account at the same time in the sampling process. The organization of this paper is as follows. Section 2 provides a detailed introduction to the data source, the sampling method and the evaluation considerations. Statistical characteristics and spatial distribution of the new dataset are analyzed in Section 3. Advantages of the new dataset over the existing widely-used datasets as well as it’s potential use for GIIRS data’s application in NWP are discussed in Section 4, where the future improvements of the atmospheric dataset are also discussed.
2. Data and Method

2.1. Data Description
Atmospheric datasets which are composed of surface variables (e.g., surface pressure, surface pressure, 10 meter wind speed, etc.), and multi-level atmospheric variables (e.g., temperature, water vapor, ozone, etc.) are mainly derived from atmospheric sounding observations, satellite retrieval and numerical weather forecast products. In March 2019, ECMWF released the fifth generation of its reanalysis (ERA5), which provides hourly estimates of atmospheric, land and oceanic variables with a 30KM horizontal resolution, and 137 pressure levels from the surface to top level at 0.05hPa (~80KM) in the vertical direction. In this paper, the original large atmospheric database is taken from 2019 ERA5 dataset, and only clear sky situations are retained.

The study region is set to within the observation area of GIIRS (10°N-53°N, 70°E-144°E). Considering that the climate difference and the varied vertical pressure structure of the different topographic type and the difference of land/sea surface properties within the region, the original ERA5 dataset of 2019 is split into six categories. Table 1 provides a detailed description of the considered six categories.

Table 1. The detailed information of topographic height classification.

| category | category-1 | category-2 | category-3 | category-4 | category-5 | category-6 |
|----------|------------|------------|------------|------------|------------|------------|
| surface type | sea | land | land | land | land | land |
| Height(m) | h<=500 | h<=500 | 500<h<=1500 | 1500<h<=3000 | 3000<h<=5000 | h>5000 |

2.2. Sampling Method
Shannon entropy measures the amount of information caused by the randomness of a certain system, it is defined for a given discrete probability distribution. Assuming that variable X has a value sequence \((x_1, x_2, \ldots, x_m)\) and an interval sequence \((b_1, b_2, \ldots, b_n, b_{n+1})\). The interval sequence of length \(n+1\) is divided into \(n\) bins \((b_{i-1}, b_i, \ldots, b_n)\), where \(b_i = [b_j, b_{j+1}]\), then the proportion of samples included in every bin by Eq. (1), then Shannon entropy of X is obtained by Eq. (2).

\[
p_j = \frac{\text{Number}_{x \in b_i}}{m}, i = 1,2 \ldots m; j = 1,2 \ldots n
\]

\[
E(X) = \sum_{j=1}^{n} -p_j \log p_j
\]

The original atmospheric dataset, \(S\), is derived from ERA5 and presented on fixed 137 pressure levels at 00:00UTC, and 12:00UTC in 2019. In consideration of the seasonal variation, \(m\) situations are sampled every month from \(S\) independently, the final dataset, \(X\), consisting of totally \(12 \times m\) situations. The value of \(m\) differs for six categories because the amount of original dataset in six categories is decreased by the topographic height. The details of \(m\) are provided in table 2. With the sum entropy of three atmospheric variables (Eq. (3)) as the uniformity measurement index, the final new dataset \(B\) is obtained by retaining the combination of atmospheric samples which has the maximum sum entropy of three variables. As is seen in Eq. (3), the calculation of entropy combines quite different variables (atmospheric temperature, T, atmospheric humidity, WV, and 2 meter temperature, TS) at the same time.

\[
E(B) = -W_T \times \sum_{i=1}^{T} \sum_{n=1}^{N} p_n^T_i \times \log(p_n^T_i) - W_WV \times \sum_{i=1}^{WV} \sum_{n=1}^{N} p_n^{WV}_i \times \log(p_n^{WV}_i) - \sum_{n=1}^{N} p_n^{TS} \times \log(p_n^{TS})
\]

\(L_{var}\) is the selected pressure levels of each variable, which varies with six category since that different category have different vertical pressure structure. \(N\) is the number of bins, which equals to 20, and the bins are calculated by dividing the interval between the minimum and the maximum values of each
variable of every pressure level in S to 20 equally spaced intervals. $W_{\text{var}}$ is the weight of each variable, and it is obtained by computing the ratio of the number of principal components ($n_{\text{PCA, components}}$) in Principal Component Analysis (PCA) when conserving 99.9% of the variable’s cumulative explained variance ratio to the $L_{\text{var}}$ (Eq. (4)).

$$W_{\text{var}} = \frac{n_{\text{PCA, components}}}{n_{\text{levels}}}$$

The iterative process is designed as the following. The initial dataset $O$ is composed of m randomly selected situations from S. At each iteration, a new situation will replace every situation in the current dataset which means the entropy needs to be re-calculated for m times, the positive maximum entropy increment position is then recorded, so that the final replacement occurs between the new situation and the recorded position at each iteration. The iterative process is stopped if the entropy increment (the current entropy minus the initial entropy) exceeds 10 out of the following considerations: (1) avoid too much computational cost; (2) the situation numbers in the original dataset for every category varies; (3) the initial randomly selected dataset reached a higher entropy compared with the dataset without the random selection operation.

The quality control step is carried out before the iteration starts for each new situation: the situations where the value of the variables (atmospheric temperature, specific humidity and 2 meter temperature) differs from the field mean value by more than 25 standard deviation are rejected.

| Category | Monthly amount | Total amount |
|----------|----------------|--------------|
| 1        | 1000           | 12000        |
| 2        | 1000           | 12000        |
| 3        | 1000           | 12000        |
| 4        | 500            | 6000         |
| 5        | 500            | 6000         |
| 6        | 100            | 1200         |

2.3. Evaluation

The sampling result, that is, the new dataset, is evaluated from the following aspects:

1) The distribution uniformity degree of variables within their variation ranges in the new dataset. As seen in Eq. (5), variable’s distribution uniformity is measured by the samples proportion in every bin at every selected pressure level, where ‘i’ refers to the ‘i’th pressure level, ‘j’ refers to the ‘j’th bin. The uniformity distribution comparison is made between the initial randomly sampled dataset and the new dataset.

$$p_{i,j}^{\text{var}} = \frac{\text{Number}_{\text{var},i,j,\text{bin},j}}{m}, \quad i = 1, 2, \ldots, n_{\text{levels}}; \quad j = 1, 2, \ldots, 20$$

2) The commonly used statistical results of each pressure level in the new dataset, including mean, maximum, minimum, standard deviation (std) values are presented. The minimum and the maximum reflects the variation range of the variable, and std is able to represent the dissimilarities between samples in the dataset, the mean and the std together represent variables’ scatter characteristics within the variation range. Statistical results comparisons are made between the new dataset and another two datasets, the initial randomly selected dataset and the original dataset respectively.

3) Examination of whether the new dataset reflects the seasonal variation and the meteorological characteristics of each topographic height category is made.

4) Examination of whether the samples in the new dataset is spatial uniformly distributed is made.
3. Results
In this section, in order to explain clearly the characteristics of the new data set and reduce the length of the article simultaneously, part of the visualization results are presented.

Figure 1 presents the samples’ distribution proportion differences between the initial dataset S and the new dataset B of category-3 in February. It is clearly shown that three variables (atmospheric temperature (figure 1 (a, b), specific humidity (figure 1 (c, d) and 2 meter temperature (figure 1 (e)) have all reached higher uniform distributions within their variation ranges after Shannon entropy sampling, especially that the proportion of values at both ends of atmospheric temperature and 2 meter temperature has increased. Compared with the atmospheric temperature, the specific humidity (figure 1 (c, d)) shows relatively slight changes in distribution. The reason may be as follows: specific humidity has large difference in magnitude between the minimum value and the maximum value for every selected pressure level, while most of the specific humidity profiles in the original dataset S are included in the first few bins. Consequently, the increase of sum entropy in Eq. 3 mainly depends on another two variables (i.e., atmospheric temperature and 2 meter temperature) during sampling, which results in a better uniform distribution in atmospheric temperature and 2 meter temperature, a relatively skewed distribution in specific humidity.

![Figure 1](image1.png)

**Figure 1.** Proportion of situations belonging to the 20 bins in February for category-3. (a, c) the initial randomly sampled dataset; (b, d) the new entropy-based dataset. (a, b) atmospheric temperature; (c, d) specific humidity; (e) 2 meter temperature, blue line represent the initial randomly sampled dataset and green line represent the new entropy-based dataset.

Figure 2 shows the statistical differences between the initial randomly sampled dataset O and the new dataset B of category-3 in February. For atmospheric temperature (figure 2 (a)), the min value, max value and mean value have not changed too much. For specific humidity (figure 2(b)), the mean and max value of the most pressure levels have increased. The standard deviation (std) of atmospheric temperature and specific humidity in the new dataset is significantly larger than that of the initial dataset, which indicates that the dissimilarities of two variables between samples in the new dataset has increased when compared with the initial dataset.

Three variables’ seasonal variation is taken into account in sampling process by independently sampling m samples for every month in every category (see Section 2). For category-4, atmospheric temperature profiles have wider variation ranges and larger dissimilarities in February (figure 3(a))
and November (figure 3(d)) compared to May (figure 3(b)) and August (figure 3(c)). In August, humidity (not shown here) is more abundant and has a larger std than other three months for every selected levels. 2 meter temperature (not shown here) has the similar tendency with atmospheric temperature.

Figure 2. Comparison of statistical results between the initial randomly sampled dataset (O) and new entropy based dataset (B) in February for category-3. (a) atmospheric temperature; (b) specific humidity. The outer curves show the minimum and the maximum values (yellow for O, green for B). The horizontal bars represent the standard deviation (black for O, pink for B) and the dots represent the mean (red for O and blue for B).

Figure 3. Comparison of atmospheric temperature in the new dataset between four months of category-4. (a) February; (b) May; (c) August; (d) November. The outer curves show the minimum and the maximum value values (green). The horizontal bars represent the standard deviation (grey) and the dots represent the mean (blue).

Usually, the sampling method have two objectives: (1) the statistics of the sampled dataset are required to reproduce the statistics of the original dataset (i.e. the mean value, std); (2) a uniformly distributed sampling results of different variables. The entropy-based sampling method are mainly focused on the second objective. But here, variables’ statistical results between the new dataset B and the original ERA5 dataset S in February are also compared in figure 4. The first row represents the statistical comparison results of atmospheric temperature and specific humidity of category-1 in February and the second row represent that of category-2. It is clearly shown that atmospheric temperature of the new dataset (figure 4(a, c)) has a particularly similar statistical distribution to that of the original dataset, while specific humidity (figure 4(b, d)) in B has slightly smaller mean value and variation ranges than in S for most of the pressure levels.
Figure 4. Comparisons of statistical results of atmospheric temperature and specific humidity between the original dataset S and the new dataset B of two categories in February. (a, c) atmospheric temperature; (b, d) specific humidity; (a, b) category-1; (c, d) category-2. The outer curves show the minimum and the maximum values (yellow for S, green for B). The horizontal bars represent the standard deviation (black for S, pink for B) and the dots represent the mean (red for S and blue for B).

Figure 5 shows the spatial distribution of the new atmospheric dataset of four months. Spatial information is not incorporated in the sampling process, but the shuffling operation before the iterative process may initially contribute a lot to a scattered distribution. Then in the sampling process, the distribution is further changed. It is clearly seen from figure 5 that samples are more concentrated in the middle and high latitudes in August and in the low latitudes in February.

Figure 5. Spatial distribution of the new atmospheric dataset: (a) February; (b) May; (c) August; (d) November.

4. Discussion and Conclusions

In this paper we build a new regional atmospheric dataset for the first geostationary hyperspectral infrared sounder GIIRS on board Fengyun-4A. The new dataset is sampled from ERA5 2019 dataset based on Shannon entropy method. Considering that the climate difference and varied vertical pressure structures of regions within the studied geographical area, the new dataset is classified into six categories according to the topographic height. Besides, seasonal variation is also considered in the sampling process. The new dataset’s main advantages are summarized as follows: 1) The new dataset presents a uniform distribution in both the spatial coverage and variation ranges of multiple variables; 2) Atmospheric variables’ seasonal variation are clearly shown in the new dataset; 3) Six datasets of different topographic height category are independently sampled. Such classification provides
convenience to the researchers who are interested in the targeted areas. And it will also help build more precise transmittance parameterization coefficients of GIIRS in the fast radiative transfer model.

Compared with the existing atmospheric dataset, the following two aspects are mainly improved in our new dataset: 1) By comparison with ECMWF atmospheric datasets, the sampled datasets B in this paper based on Shannon entropy preserve higher uniformity of samples distributed within the variation ranges of multiple variables at the same time; 2) The new dataset addresses the deficiencies in TIGR series datasets that lack representative atmospheric samples in China.

There exists a potential application of the new dataset in the study of the utility of new infrared and microwave instruments in numerical weather forecast. The new atmospheric dataset can be used as a training dataset for statistical retrieval of temperature and water-vapor profiles from the GIIRS observations. Besides, the new dataset can also be used to train the transmittance parameterization scheme in the fast radiative transfer model.

References
[1] Menzel W P, Schmit T J, Zhang P, Li J 2017 Satellite based atmospheric infrared sounder development and applications Bull. Am. Meteorol. Soc. 99(3): 583-603.
[2] Saunders R, Matricardi, M, Brune P 1999 An improved fast radiative transfer model for assimilation of satellite radianc observations Q. J. R. Meteorol. Soc. 125(556): 1407-1425.
[3] Matricardi M, Chevallier F, Kelly G, Thepaut J 2004 An improved general fast radiative transfer model for the assimilation of radiance observations Q. J. R. Meteorol. Soc. 130(596): 153-173.
[4] Zhou S, Cheng J 2019 A framework for estimating clear-sky atmospheric total precipitable water (TPW) from VIIRS/S-NPP Remote Sens. 11(8): 916-.
[5] Pinker R T, Chen W, Ma Y T, Islam T, Borbas E, Hain C, Hulley G, Hook S 2017 MEaSUREsLand Surface Temperature from GOES satellites Proceedings of the European Geosciences Union General Assembly, Vienna, Austria.
[6] Matricardi M, Saunders R 1999 Fast radiative transfer model for simulation of infrared atmospheric sounding interferometer radiances Appl. Optics. 38(27): 5679-91.
[7] Di D, Li J, Han W, Bai W G, Wu C Q, Menze W P 2018 Enhancing the fast radiative transfer model for fengyun-4 giirs by using local training profiles J. Geophys. Res-Atmos. 123(22): 12583–12596.
[8] Yang J, Zhang Z, Wei C, Lu F, Guo Q 2016 Introducing the new generation of Chinese geostationary weather satellites, Fengyun-4(FY-4) Bull. Am. Meteorol. Soc. 98(8): 1637–1658.
[9] Cai X, Bao Y S, Petropoulos GP, Lu f, Lu Q F, Zhu L H, Wu Y 2020 Temperature and humidity profile retrieval from FY4-GIIRS hyperspectral data using artificial neural networks Remote Sens. 12(11): 1872.
[10] Eresmaa R, McNally A 2014 Diverse Profile Datasets from the ECMWF 137-level Short-Range Forecasts European Centre for Medium Range Weather Forecasts: Reading, UK.
[11] Borbas E E, Suzanne W S, Hung L H, Li J, Menzel W P 2005 Global profile training database for satellite regression retrievals with estimates of skin temperature and emissivity Proceedings of the 14th International ATOVS Study Conference, Beijing, China.
[12] Yang S, Sun G J, He W Y, Zhao H, Li F M, Xiong Y C 2011 Correlation of topographic factors with precipitation and surface temperature in arid and cold region of northwest china: a case study in gansu province Acta Entomol. Sin. 31(9): 2414-2420.
[13] Deng D, Gao S, Hu L, Du X, Wang C 2016 The impact of guizhou topography on the distribution of freezing rain in early january 2011 Quart. J. Roy. Meteor. Soc. 141(693): 3252-3267.