IRJIEST 11, 2025 XX-XX

Available online at https://ojs.batstate-u.edu.ph

International Research Journal on Innovations in Engineering, Science and Technology

Preprocessing cloud-contaminated satellite data: Balancing samples across cloud levels for super-resolution

Alvin Alexander Idin, Nor Erne Nazira Bazin*

Universiti Teknologi Malaysia, Johor Bahru, Malaysia

ABSTRACT

Satellite imagery is used in data-driven applications such as environmental monitoring, agriculture, and disaster management. However, while multispectral imagery suffers from cloud contamination that reduces quality and consistency, radar data remain unaffected but have differing signal characteristics. This study introduces a reproducible preprocessing pipeline for super-resolution that quantifies cloud levels and applies sampling strategies to balance representation across low, medium, and high cloud contamination. Because super-resolution models are typically trained on clean multispectral imagery yet must operate on cloud-contaminated inputs, cloud-level imbalance can bias learning and destabilise optimisation. The pipeline was tested on the SEN12MS-CR dataset, which includes multispectral Sentinel-2 imagery and radar Sentinel-1 data, with image degradation applied to simulate low-resolution inputs for super-resolution tasks. Weighted sampling balanced cloud categories, preserved spectral diversity, and improved repeat-level stability, while random sampling produced pronounced imbalance and variability. This study extends existing preprocessing research by quantifying how cloud-level imbalance influences entropy, variance, and repeat-level stability, which are factors that are rarely examined together in a super-resolution workflow. By ensuring controlled cloud-level representation and preserving diversity, the proposed approach improves dataset representativeness and strengthens the robustness of super-resolution applications.

Keywords: balanced sampling, preprocessing, multi-sensor, data diversity, cloud-aware

Received: November 6, 2025 **Revised:** December 9, 2025 **Accepted:** December 22, 2025 (Article in Press)

1. Introduction

Multi-sensor satellite imagery is widely applied in Earth observation domains such as environmental monitoring [1], precision agriculture [2], and disaster management [3]. Sentinel-1 provides Synthetic Aperture Radar (SAR) data with dual polarisation bands (VV and VH), and Sentinel-2 provides thirteen multispectral bands covering the visible, near-infrared, and shortwave-infrared regions [4]. These sensors are complementary and valuable for machine-learning tasks that require both spectral richness and structural detail, including super-resolution.

A persistent challenge in multi-sensor imagery is cloud contamination. Sentinel-2 observations are obscured by clouds, which remove large fractions of usable pixels [5]. Sentinel-1 SAR data are unaffected by atmospheric conditions, yet combining SAR with cloud-contaminated optical imagery introduces heterogeneity in feature space and sampling distributions. This imbalance biases model training, limits generalisation, and reduces predictive accuracy of downstream performance [6].

Most preprocessing pipelines for optical imagery address clouds by detecting and masking contaminated pixels to provide clean inputs. The main operational processors for Sentinel-2, including MAJA, Sen2Cor, and Fmask, have been validated and compared by Baetens et al. [7]. These processors remove cloud artefacts effectively, but the statistical representativeness of the resulting datasets has received limited quantitative analysis. However, existing preprocessing studies rarely examine how cloud-level distributions influence downstream learning behaviour.

Super-resolution models are usually trained on clean, cloud-free imagery, yet operational Sentinel-2 data often contain varying cloud levels. When cloudy low-resolution inputs are used, the model must infer high-resolution structure even where thin clouds or haze obscure the surface, making cloud effects part of the learning process. In this context, the distribution of cloud levels matters: imbalance can bias the model toward common regimes and weaken gradient signals for rarer conditions, reflected in shifts in spectral entropy and increased variance across repeats. Figure 1 contrasts conventional clean-image super-resolution pipelines with cloud-aware scenarios where cloud distortions propagate from the inputs. These considerations motivate a preprocessing approach that explicitly accounts for cloud-level imbalance.

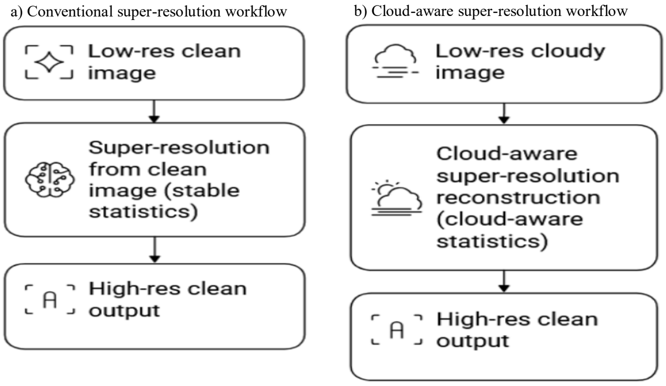

a) Conventional super-resolution workflow
Low-res clean image
Super-resolution from clean image (stable statistics)
High-res clean output
b) Cloud-aware super-resolution workflow
Low-res cloudy image
Cloud-aware super-resolution reconstruction (cloud-aware statistics)
High-res clean output

Figure 1. Illustrative comparison of conventional and cloud-aware super-resolution workflows. The cloud-aware setting requires reconstruction from cloudy inputs, introducing different statistical properties compared with clean-image training.

*Corresponding author

Email address: erne@utm.my

This study develops a cloud-aware preprocessing and sampling framework for the SEN12MS-CR dataset [4]. Cloud levels are quantified for each image patch, and inverse-probability weighting is applied to balance low, mid, and high cloud categories. The analysis evaluates how balancing affects entropy-based diversity, variance, and repeat stability. This work addresses a gap in understanding how cloud-level imbalance influences dataset variance, entropy, and repeat-level consistency. These factors directly affect optimisation stability in super-resolution models but remain unquantified in existing preprocessing studies. By characterising these statistical effects, the study establishes both a cloud-aware preprocessing workflow and a statistical evaluation procedure for assessing dataset representativeness and stability before model training. This combined preprocessing and evaluation approach provides a reproducible foundation for preparing more reliable datasets under heterogeneous cloud conditions.

2. Materials and methods

This section describes the dataset, preprocessing steps, and analytical procedures used to evaluate the effects of cloud-aware sampling. The workflow includes cloud quantification, entropy-based characterisation, statistical testing, and image degradation for super-resolution evaluation. Each step was designed to ensure reproducibility and consistent comparison between random and weighted sampling strategies.

2.1. Dataset

This study used the SEN12MS-CR dataset [4], which provides paired Sentinel-1 SAR (VV, VH) and Sentinel-2 multispectral (13 bands) imagery across seasons and cloud conditions. Each patch is provided on a common 10 m spatial resolution grid and delivered as 256 × 256 pixel tiles, corresponding to an area of approximately 2.56 × 2.56 km. The dataset contains globally sampled regions of interest: each composed of cloudy and cloud-free Sentinel-2 images together with co-registered Sentinel-1 acquisitions. In total, SEN12MS-CR includes 122,218 patches across spring (29,117), summer (33,827), fall (40,941), and winter (18,333). After excluding patches with zero cloud score, 59,906 patches remained across spring (16,206), summer (15,999), fall (20,122), and winter (7,579). Within this filtered subset, the cloud-level composition consists of 43,642 low-cloud, 10,377 mid-cloud, and 5,887 high-cloud patches. These clouded samples form the basis for all subsequent preprocessing, sampling, and statistical analyses. The overall preprocessing workflow, including data inputs, cloud-level categorisation, entropy computation, and sampling strategy, is summarised in Figure 2.

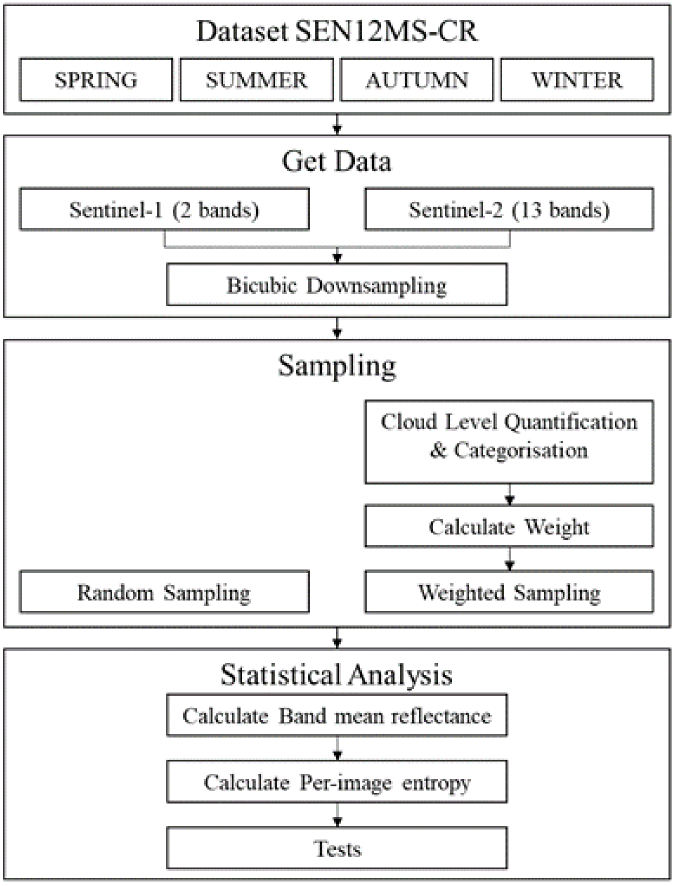


Figure 2. Workflow of the preprocessing pipeline showing inputs (Sentinel-1 SAR and Sentinel-2 multispectral imagery), cloud-level quantification and categorisation, entropy and statistical calculations, and generation of random versus weighted samples used in the analysis.

SEN12MS-CR was designed for cloud-removal research, providing paired cloudy and cloud-free imagery to train and validate restoration models. In this study, the dataset was adapted for super-resolution experiments by degrading the high-resolution Sentinel-2 images to simulate low-resolution inputs using bicubic downsampling while retaining the original clouded images. This modification preserves the dataset's cloud-aware characteristics but adapts it to analyse preprocessing effects on super-resolution tasks. Processing and analysis were performed in Python 3.12 / PyTorch 2.4 (CUDA 11.8, Ubuntu 22.04 LTS) on an NVIDIA RTX 4080 SUPER GPU.

2.2. Cloud level quantification

Cloud contamination was quantified using a method adapted from Meraner et al. [4]. Reflectance values were normalised to [0, 1], and cloud scores were derived from the blue, aerosol, cirrus, and visible bands. The Normalised Difference Snow Index (NDSI) was used to exclude snow-covered pixels, and the cloud-probability maps were smoothed using morphological closing. Morphological closing employed 5×5 square kernels for both dilation and erosion, followed by additional 7×7 average smoothing. Images were categorised into low (≤ 0.3), mid (0.3–0.6), and high (≥ 0.6) cloud levels. The thresholds of 0.30 and 0.60 were selected as simple heuristic cut-points, providing an interpretable categorisation (Figure 3) that enables the subsequent sampling strategy to address the inherent imbalance in cloud conditions explicitly. Cloud-free images were excluded from analysis. Figure 4 shows example patches illustrating the defined cloud levels, including Sentinel-1 SAR, cloud-free Sentinel-2, downsampled input, and Sentinel-2 scenes with low, mid, and high cloud conditions.

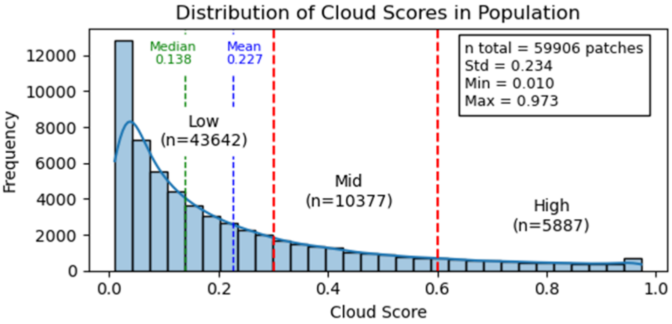


Figure 3. A histogram with KDE illustrates the population-level frequency of raw cloud scores. Vertical dashed lines at 0.30 and 0.60 mark the thresholds used to define low, mid, and high cloud levels. Mean and median values are shown for reference, along with summary statistics indicating the strongly right-skewed distribution that motivates the categorical separation used in subsequent sampling analyses.

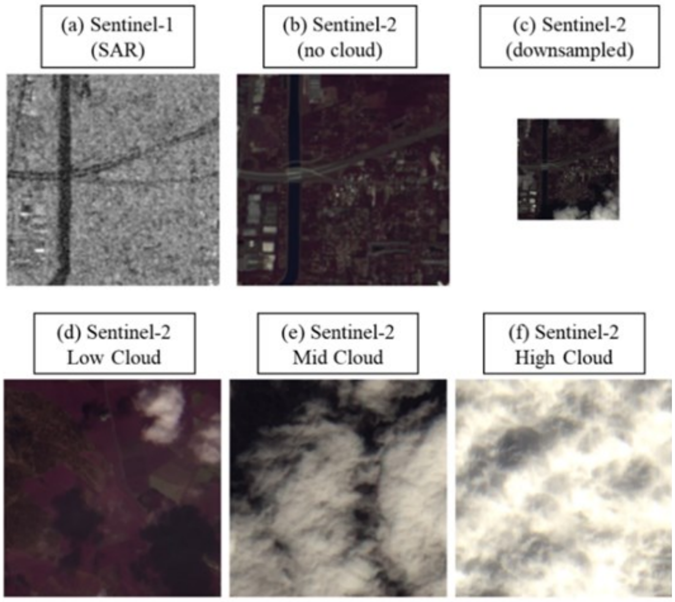

(a) Sentinel-1 (SAR)
(b) Sentinel-2 (no cloud)
(c) Sentinel-2 (downsampled)
(d) Sentinel-2 Low Cloud
(e) Sentinel-2 Mid Cloud
(f) Sentinel-2 High Cloud

Figure 4. Example SEN12MS-CR samples showing (a) Sentinel-1 SAR, (b) Sentinel-2 cloud-free reference, (c) downsampled input, and (d–f) Sentinel-2 scenes with low, mid, and high cloud levels.

2.3. Spectral statistics and entropy

Per-image entropy quantified spectral diversity, interpreted as image information content of multispectral imagery [8]. For every image, the mean reflectance of each band was computed, producing a vector of fifteen per-band means. The 15-band entropy vector consists of the 13 multispectral Sentinel-2 bands together with the two Sentinel-1 SAR polarisation channels (VV and VH), as SAR backscatter penetrates cloud and contributes structural information absent from the optical bands, yielding an entropy measure that reflects complete multisensor scene variability. These means were normalised and converted into a probability distribution representing the relative contribution of each band to the image's overall spectral composition. Entropy (H) was calculated as

$$H =- \sum_{i=1}^{n} p_i log^2\left(p_i\right) \quad (1)$$

where p_i is the normalised probability of intensity level i across all spectral bands, and N is the number of discrete intensity levels. Higher H values indicate higher spectral variability, while lower values reflect homogeneous spectral responses.

This formulation follows recent studies that use entropy to measure information content in optical and multisensor imagery. It was shown that entropy-based metrics capture multi-feature image diversity by combining spectral, spatial, and textural attributes [8]. Consistent with that framework, the present study focuses on the spectral component. Entropy values were standardised with z-normalisation to ensure comparability across bands and cloud categories, computed as

$$H' = \frac{H-\mu_H}{\sigma_H} \quad (2)$$

where μ_H and σ_H are the mean and standard deviation of entropy across all images in the dataset. This study analyses image-level spectral summaries only. Pixel-level spatial structure is not included, as all computations use aggregated per-image statistics.

2.4. Sampling strategy

Two sampling strategies were used. In random sampling, images were drawn uniformly without considering the cloud level. In weighted sampling, selection probabilities were calculated as the inverse of each category's relative frequency, normalised to sum to one:

$$w_k = \frac{1/f_k}{\sum_{j=1}^{3}\left(1/f_j\right)} \quad (2)$$

where f_k is the relative frequency of category k (low, mid, high). This weighting increases the contribution of under-represented categories and decreases that of common categories, producing balanced subsets. For each strategy, 150 images per category per season were selected, based on evidence that most performance gains in balanced image-based deep-learning tasks occur before roughly 150 samples per class, with limited improvement beyond that point [9]. With three categories and four seasons, this produced 1,800 images per replicate sample. The sampling process was repeated 100 times to generate replicate subsets.

2.5. Statistical tests

The effects of random and weighted sampling were evaluated using non-parametric tests, following standard statistical procedures [10]. Class balance was assessed with the Chi-square test. At the same time, differences in central tendency were examined using the Mann–Whitney U test for pairwise comparisons and the Kruskal-Wallis H test for multi-group comparisons. Variance was analysed with the Brown–Forsythe test, a median-based variant of Levene's test. Repeat stability was further assessed by applying the Brown–Forsythe test to per-repeat medians across the 100 replicate samples. These procedures provided a statistical comparison of class distributions, entropy medians, and variance patterns between sampling strategies, both overall and within individual cloud-level categories.

2.6. Degradation process for super-resolution

High-resolution Sentinel-2 images were degraded to create corresponding low-resolution inputs for supervised super-resolution training and evaluation. Each image was downsampled by a factor of two (degrade factor = 0.5) using bicubic interpolation, a standard baseline in super-resolution research [11]. This process produced paired low-resolution and original high-resolution images for each patch, ensuring spatial alignment across all spectral bands.

3. Results and discussion

This section presents the statistical outcomes of the analyses and interprets their implications for cloud-aware preprocessing. Results are organised by key aspects of data quality and stability: class balance, central tendency, variance, and repeat consistency. The discussion links these results to the goal of improving dataset representativeness for training super-resolution models under varying cloud conditions. Table 1 summarises all hypotheses tested and the corresponding statistical decisions, with Tables 2–12 presenting the detailed results for each analysis.

Table 1. Summary of hypotheses tested in this study

Table No.	Hypotheses	Decision
2	Does weighted sampling produce balanced class proportions across cloud levels? H_0: Cloud-level proportions are equal under random and weighted sampling. H_1: Cloud-level proportions differ between random and weighted sampling.	Reject H_0
3	Does weighting affect the overall median entropy (spectral information content)? H_0: Median entropy is equal between random and weighted samples. H_1: Median entropy differs between random and weighted samples.	Reject H_0
4	Does weighting affect the median entropy within each cloud-level category? H_0: Median entropy within each cloud level is equal between random and weighted sampling. H_1: Median entropy within each cloud level differs between random and weighted sampling.	Fail to reject H_0
5	Does cloud level influence entropy regardless of sampling method? H_0: Median entropy does not differ across cloud levels. H_1: Median entropy differs among low, mid, and high cloud levels.	Reject H_0
6	Does weighting change the overall variance of entropy values? H_0: Variance of entropy is equal between random and weighted sampling. H_1: Variance of entropy differs between random and weighted sampling.	Fail to reject H_0
7	Does weighting affect variance within individual cloud levels? H_0: Variance within each cloud level is equal between random and weighted sampling. H_1: Variance within each cloud level differs between random and weighted sampling.	Reject H_0 (marginal)
8	Does cloud level influence variance regardless of sampling method? H_0: Variance does not differ across cloud levels. H_1: Variance differs among low, mid, and high cloud levels.	Reject H_0
9	Does weighting improve repeat stability across sampling iterations? H_0: Repeat-level variance is equal between random and weighted methods. H_1: Repeat-level variance differs between random and weighted methods.	Reject H_0
10	Does weighting affect repeat-level variance within each cloud category? H_0: Repeat-level variance within each cloud level is equal between random and weighted sampling. H_1: Repeat-level variance within each cloud level differs between random and weighted sampling.	Reject H_0 (partial)
11	Does cloud level influence repeat-level variance patterns? H_0: Repeat-level variance does not differ across cloud levels. H_1: Repeat-level variance differs among low, mid, and high cloud levels.	Reject H_0
12	Does weighting interact with cloud level to influence repeat-level variance? H_0: Method × cloud-level interaction has no effect on repeat-level variance. H_1: Interaction affects repeat-level variance (patterns differ by method and level).	Reject H_0 (partial)

3.1. Class balance

The first analysis tested whether random and weighted sampling produced equal representation across low, mid, and high cloud categories. A Chi-square test was applied, and the results are summarised in Table 2.

Table 2. Class balance check across random vs. weighted sampling (χ^2 test)

Method	Samples Reject H₀	Samples Not Reject H₀	% Reject H₀	X² Statistic (mean ± SD)	p-value (mean ± SD)
Random	100	0	100.0%	1225.52 ± 67.29	p < 0.001
Weighted	7	93	7.0%	2.09 ± 2.05	0.482 ± 0.279

Random sampling produced highly unbalanced class distributions. Across one hundred repeated draws, all samples rejected the null hypothesis of equal proportions, with large Chi-square values (mean = 1225.52 ± 67.29) and near-zero p-values. Weighted sampling maintained uniformity; only seven per cent of replicates rejected the null, with much smaller statistics (mean = 2.09 ± 2.05) and higher p-values (0.482 ± 0.279). Figure 5 confirms that random sampling yields skewed proportions, while weighted sampling produces nearly uniform distributions.

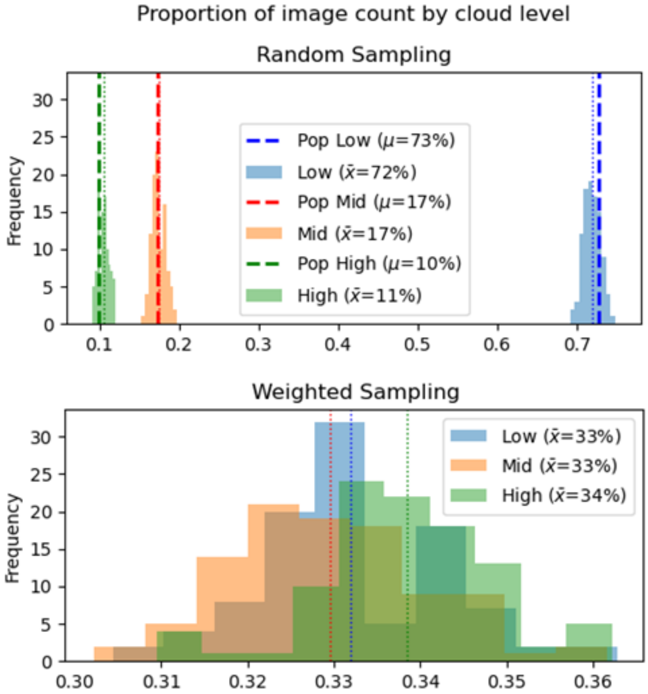


Figure 5. Proportion of samples in low, mid, and high cloud categories under random vs. weighted sampling. Weighted shows near-uniform distribution; random is skewed.

For super-resolution models, balanced sampling ensures that the network does not overfit to the most frequent cloud type and that performance remains reliable across rare conditions. Weighted sampling, therefore, supports better generalisation across cloud levels.

3.2. Central tendency

Central tendency was compared between random and weighted subsets using the Mann–Whitney U test. Both overall entropy distributions and within-level comparisons were evaluated. The outcomes are presented in Tables 3–5.

Table 3. Effect of balancing on overall entropy median (Mann–Whitney U test)

Comparison	Statistic (U)	p-value	Median Random	Median Weighted	Ratio W/R	Ratio R/W
Random vs Weighted	0.0	p < 0.001	3.544	3.669	1.035	0.966

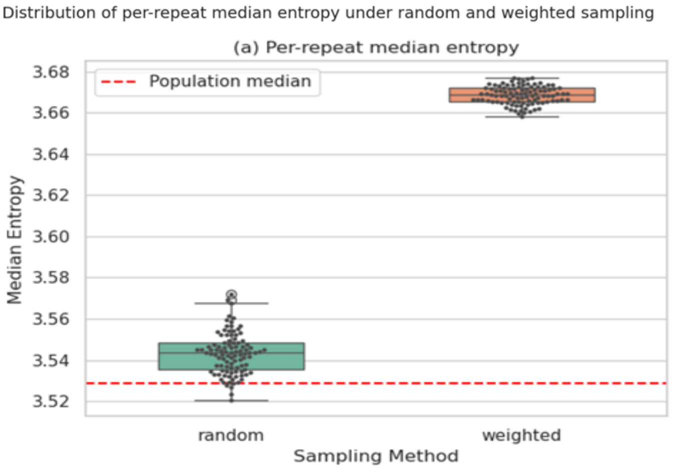
Table 4. Effect of balancing on entropy median within cloud levels (Mann–Whitney U test)

Cloud Level	Statistic (U)	p-value	Median Random	Median Weighted	Ratio W/R	Ratio R/W
Low	5656.0	0.109	3.440	3.437	0.999	1.001
Mid	5725.5	0.076	3.694	3.693	1.000	1.000
High	5891.5	0.029	3.711	3.709	0.999	1.001

Table 5. Cloud-level effect on median entropy (Kruskal–Wallis H test)

Method	Statistic (H)	p-value	Median Low	Median Mid	Median High	Ratio Low/ Mid	Ratio Mid/ High	Ratio Low/ High
Random	261.88	p < 0.001	3.440	3.694	3.711	0.931	0.995	0.927
Weighted	265.78	p < 0.001	3.437	3.693	3.709	0.931	0.996	0.927

Entropy distributions differed significantly between random and weighted sampling (U = 0, $p < 0.001$). The weighted method produced slightly higher entropy (median = 3.67 vs 3.54), representing an increase of about 3.5 per cent. Figure 6 shows that weighted medians cluster above random, and the distribution of paired differences is positive. Balancing raised overall entropy, indicating greater spectral variability.



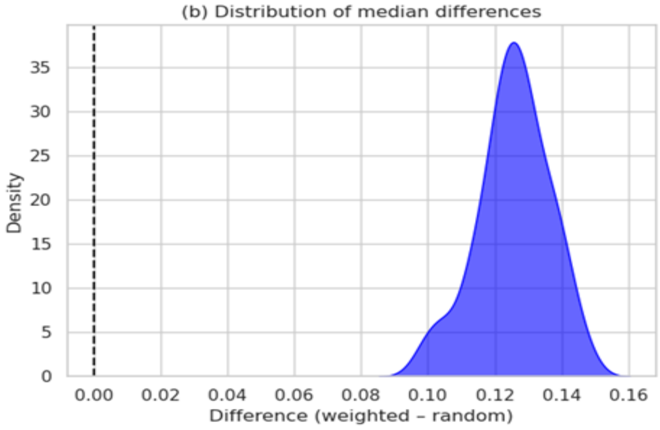


Figure 6. Distribution of per-repeat median entropy under random and weighted sampling, where panel (a) shows boxplots with swarmplots comparing repeat-level medians against the population reference, and panel (b) presents the density of differences (weighted–random), highlighting that weighted medians are consistently higher.

Within individual cloud levels, differences were negligible. Thus, weighting did not significantly alter within-level entropy. Low and mid clouds showed no significant difference ($p = 0.109$ and 0.076). High clouds showed a small but detectable change ($p = 0.029$), though the effect size was minor. Figure 7 illustrates these results with violin plots of median entropy across cloud levels for both sampling methods.

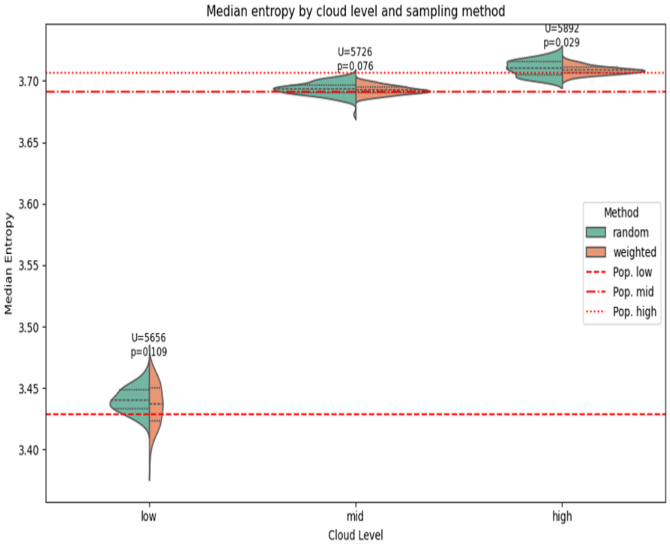


Figure 7. Comparison of median entropy by method and cloud level, shown with split violin plots for random and weighted sampling across low, mid, and high categories. Population medians are marked with horizontal reference lines, and statistical test annotations confirm that only the high-cloud category shows a small but statistically detectable difference, while low and mid clouds remain comparable across methods.

The cloud level itself had the dominant effect on entropy. Median entropy increased from low to mid by about seven per cent, but remained stable from mid to high. Kruskal–Wallis tests confirmed significant differences ($p < 0.001$) for both methods. These results indicate that entropy variation is driven by cloud level rather than sampling strategy.

Entropy reflects spectral richness and textural diversity. Weighted sampling preserved or increased entropy, increasing the diversity of spectral inputs and improving their capacity to reconstruct fine-scale details instead of learning biased representations from over-represented conditions.

3.3. Variance

Variance in entropy values was analysed using the Brown–Forsythe test. Comparisons included overall variance, variance within cloud levels, and cross-level effects. Results are provided in Tables 6–8.

Comparison	**Statistic (F)**	**p-value**	**Variance Random**	**Variance Weighted**	**Ratio W/R**	**Ratio R/W**
Random vs Weighted	0.029	0.866	7.95×10^{-2}	5.71×10^{-2}	0.718	1.393

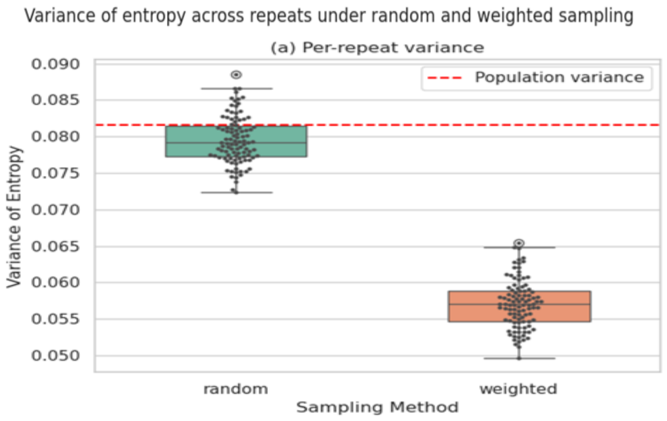
Table 7. Effect of balancing on variance within cloud levels (Brown–Forsythe test)

Cloud Level	**Statistic (F)**	**p-value**	**Var Random**	**Var Weighted**	**Ratio W/R**	**Ratio R/W**
Low	17.77	$p < 0.001$	8.30×10^{-2}	8.24×10^{-2}	0.993	1.008
Mid	19.70	$p < 0.001$	2.98×10^{-2}	3.09×10^{-2}	1.040	0.962
High	17.79	$p < 0.001$	1.08×10^{-2}	1.08×10^{-2}	1.003	0.997

Table 8. Cloud-level effect on variance (Brown–Forsythe test)

Method	**Statistic (H)**	**p-value**	**Var Low**	**Var Mid**	**Var High**	**Ratio Low/ Mid**	**Ratio Mid/ High**	**Ratio Low/ High**
Random	24.13	$p < 0.001$	1.04×10^{-4}	2.79×10^{-5}	3.95×10^{-5}	3.746	0.707	2.648
Weighted	124.58	$p < 0.001$	2.91×10^{-4}	1.27×10^{-5}	1.16×10^{-5}	22.930	1.091	25.023

Overall variance did not differ significantly between random and weighted sampling ($p = 0.866$). Weighted variance appeared lower by about twenty-eight per cent, but this difference was not significant. Balancing, therefore, does not alter total variance systematically. Figure 8 visualises these findings: both sampling methods follow similar variance distributions.



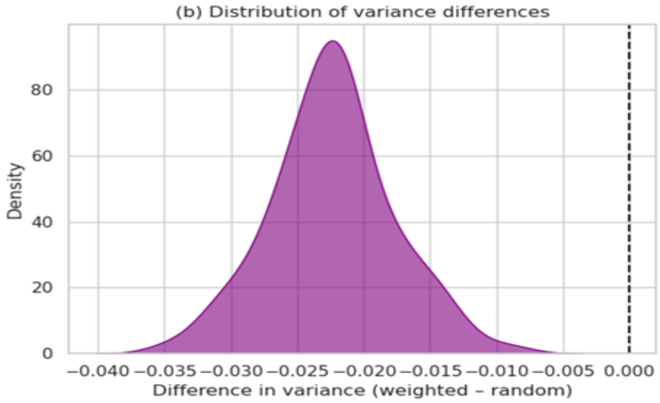


Figure 8. Overall variance of entropy under random and weighted sampling, where panel (a) shows per-sample variances compared with the population reference, and panel (b) presents the density of variance differences (weighted – random). Both visualisations confirm the Brown–Forsythe test results in Table 6, showing that weighted sampling yields slightly lower variance on average, but the difference is not statistically significant.

Within cloud levels, variance differences were statistically significant but small in magnitude. Weighted variance was 0.8 per cent lower for low clouds, about 4 per cent higher for mid clouds, and 0.3 per cent higher for high clouds. The Brown–Forsythe results confirm that the differences are detectable but not practically meaningful. Figure 9 shows that the variance patterns under both methods are nearly identical.

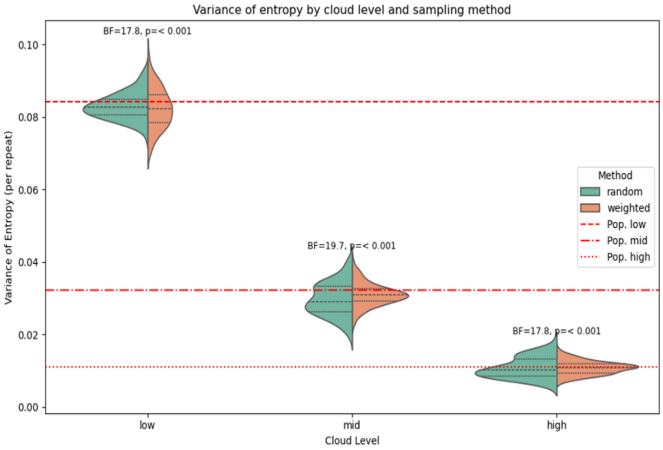


Figure 9. Variance of entropy by cloud level under random and weighted sampling. Split violin plots with quartile markers show the spread of per-sample variances for low, mid, and high clouds. Population reference lines and Brown–Forsythe test annotations indicate that although differences are statistically significant, the effect sizes are negligible, with weighted variances differing from random by less than 5%.

Variance differed markedly across cloud levels. In the random method, low-cloud variance was three to four times higher than mid-cloud and roughly twice that of high clouds. Under weighting, low-cloud variance increased relative to mid and high by more than twentyfold, while mid and high were nearly equal. These patterns show that cloud level is the primary driver of variance, and weighting amplifies this contrast for low-cloud cases.

Stable variance across categories indicates consistent training signals. Random sampling produces uneven variance that can cause noisy or biased learning. Weighted sampling improves proportional balance but increases low-cloud variability, suggesting that models trained on these subsets may still show instability unless further balanced or filtered.

3.4. Repeat stability

Reproducibility was assessed by analysing variance across one hundred sampling iterations. Brown–Forsythe tests compared overall and per-level repeat variance. The findings are shown in Tables 9–12.

Table 9. Repeat-level variance comparison across methods (Brown–Forsythe test)

Comparison	Statistic (F)	p-value	Var Random	Var Weighted	Ratio W/R	Ratio R/W
Random vs Weighted	36.50	$p < 0.001$	9.76×10^{-5}	1.90×10^{-5}	0.195	5.128

Weighted sampling reduced repeat variance by roughly fivefold (1.90×10^{-5} vs 9.76×10^{-5}; $p < 0.001$). The result indicates more consistent resampling behaviour.

Table 10. Repeat-level variance within cloud levels (Brown–Forsythe test)

Cloud Level	Statistic (F)	p-value	Var Random	Var Weighted	Ratio W/R
Low	25.63	$p < 0.001$	1.04×10^{-4}	2.91×10^{-4}	2.787
Mid	10.23	0.002	2.79×10^{-5}	1.27×10^{-5}	0.455
High	44.34	$p < 0.001$	3.95×10^{-5}	1.16×10^{-5}	0.295

Variance patterns varied by cloud level. Hence, weighting improved mid/high-cloud stability but increased low-cloud variability. For low clouds, weighted variance was higher (2.91×10^{-4} vs 1.04×10^{-4}). For mid and high clouds, weighted variance was lower, roughly half to one-third of random. Weighting stabilised mid and high clouds but inflated variability at low clouds.

Table 11. Cloud-level effect on repeat-level variance (Brown –Forsythe test)

Method	Statistic (H)	p-value	Median Low	Median Mid	Median High	Ratio Low/ Mid	Ratio Mid/ High	Ratio Low/ High
Random	24.13	$p < 0.001$	1.04×10^{-4}	2.79×10^{-5}	3.95×10^{-5}	3.746	0.707	2.648
Weighted	124.58	$p < 0.001$	2.91×10^{-4}	1.27×10^{-5}	1.16×10^{-5}	22.930	1.091	25.023

Both methods showed strong cloud-level variance effects ($p < 0.001$). In random sampling, low-cloud variance was about three to four times higher than mid or high. In weighted sampling, the difference rose to roughly twentyfold. The dominance of low-cloud variability persists regardless of method.

Table 12. Pairwise comparisons of repeat-level variance across cloud levels (Brown–Forsythe test)

Method	Pair	Statistic (F)	p-value	Var A	Var B	Ratio A/B	Ratio B/A
Random	Low vs Mid	37.58	$p < 0.001$	1.04×10^{-4}	2.79×10^{-5}	3.746	0.267
Random	Low vs High	19.24	$p < 0.001$	1.04×10^{-4}	3.95×10^{-5}	2.648	0.378
Random	Mid vs High	6.23	0.013	2.79×10^{-5}	3.95×10^{-5}	0.707	1.415
Weighted	Low vs Mid	127.73	$p < 0.001$	2.91×10^{-4}	1.27×10^{-5}	22.930	0.044
Weighted	Low vs High	132.85	$p < 0.001$	2.91×10^{-4}	1.16×10^{-5}	25.023	0.040
Weighted	Mid vs High	0.61	0.435	1.27×10^{-5}	1.16×10^{-5}	1.091	0.916

These results indicate that while weighting improves reproducibility overall, an interaction between sampling method and cloud level remains evident. Stability under mid- and high-cloud conditions improves markedly, but low-cloud subsets continue to drive residual variance, confirming that weighting stabilises some categories while amplifying variability in others. This interaction effect suggests that cloud-level heterogeneity, rather than sampling method alone, governs repeat-level variance patterns. Careful control of low-cloud representation is therefore necessary to maintain consistent performance in repeated sampling. For super-resolution models, this stability ensures predictable convergence and reliable evaluation across training runs. In contrast, persistent variability in low-cloud cases highlights the need for further refinement of preprocessing steps to minimise noise and enhance model robustness.

3.5. Key insights

The analyses show that cloud-aware sampling improves both dataset balance and training stability without altering the essential statistical structure of the imagery. Weighted sampling equalises the representation of low, mid, and high cloud levels, ensuring that models are trained on data that reflect a wider range of atmospheric conditions. Entropy results confirm that balancing maintains or slightly increases information content, indicating that diversity in spectral and spatial content is preserved.

Variance tests demonstrate that differences across cloud levels remain the primary source of instability. Weighted sampling reduces overall variability between repeated samples but amplifies variance within low-cloud subsets, revealing that these cases are the most unpredictable. Repeat-level analysis confirms that balanced sampling improves reproducibility and provides consistent behaviour across resampling iterations.

Together, these results indicate that the proposed preprocessing and sampling framework produces datasets that are statistically representative and stable across repeated experiments. The findings establish a foundation for reliable model training and evaluation in cloud-contaminated satellite imagery, supporting the development of robust super-resolution methods that generalise across diverse atmospheric conditions.

Across all analyses, low-cloud scenes showed the highest variance and the largest fluctuations across repeated samples. This reflects the mixed composition of partially obscured scenes, where surface reflectance, thin cloud layers, haze, and mixed surface–atmosphere contributions jointly influence the spectral signal. Such mixed conditions produce more heterogeneous responses than fully unobscured or fully clouded scenes, leading to wider variability in entropy and variance statistics. This pattern explains why low-cloud categories remain the most variable component of the dataset, even when overall sampling balance is improved.

These stability gains also carry practical value for applications that rely on consistent satellite-derived products, such as crop monitoring, flood assessment, and hazard mapping, where unstable inputs may distort yield assessments, underestimate flood spread, or shift risk zones. By stabilising the inputs to super-resolution workflows, the framework supports operational decisions such as crop management, response coordination, or hazard warnings, in contexts where inconsistent outputs can incur economic or societal costs.

3.6. Implications for super-resolution model training

The improvements in class balance, entropy distribution, and repeat-level variance indicate that the balanced subsets produced by the preprocessing framework offer more stable and representative inputs for super-resolution training. Consistent data statistics are important for super-resolution because variability in spectral composition or cloud conditions can influence optimisation behaviour and lead to inconsistent reconstruction quality.

Balanced subsets that maintain comparable representation of low-, mid-, and high-cloud scenes while preserving spectral diversity provide a more reliable basis for learning across atmospheric conditions. The reduction in repeat-level variance further suggests that training on these subsets would yield more stable gradient signals and more consistent optimisation behaviour.

Taken together, these statistical gains strengthen the reliability of datasets prepared for super-resolution workflows and provide a sound foundation for downstream model development under heterogeneous atmospheric conditions. Although full model training is beyond the scope of this study, the stabilised entropy and variance distributions documented here represent prerequisites for predictable super-resolution optimisation behaviour, especially under heterogeneous cloud conditions.

4. Conclusions

This study introduced a reproducible preprocessing framework for cloud-contaminated satellite imagery that balances data representation across low, mid, and high cloud levels. Using the SEN12MS-CR dataset, the analysis compared random and weighted sampling across multiple statistical dimensions, including class balance, entropy, variance, and repeat stability. Weighted sampling achieved balanced class proportions and preserved spectral diversity, producing datasets that more accurately represent the range of cloud conditions observed in optical imagery. Entropy differences confirmed that weighting preserved the information content of the imagery, while variance and repeat-level tests showed that weighting improves overall stability. The results also revealed that low-cloud subsets remain the primary source of variability, indicating the need for refined control in these cases.

The improvements in balance, entropy distribution, and repeat-level stability show that the proposed preprocessing framework provides more representative and reliable inputs for subsequent learning. For super-resolution tasks, balanced cloud-level representation helps prevent bias toward over-represented atmospheric conditions and supports more stable optimisation across cloud regimes. Low-cloud scenes, however, remained the most variable, indicating that models trained on such conditions may experience less stable optimisation unless their representation is carefully managed.

The limitations of this study include its focus on spectral entropy without integrating spatial or texture-based analysis that could capture structural variability, as well as the SEN12MS-CR dataset-specific nature of the findings, which may have limited broader applicability across different sensors or regional contexts. Future work may build on the balanced and variance-controlled datasets developed here as training inputs for super-resolution models, enabling direct evaluation of how stabilised cloud-level representation affects convergence behaviour and reconstruction quality across varying atmospheric conditions. This can be complemented with spatial-based analysis and cloud-free patches.

References

[1] Alfieri L, Avanzi F, Delogu F, Gabellani S, Bruno G, Campo L, Libertino A, Massari C, Tarpanelli A, Rains D, Miralles DG, Quast R, Vreugdenhil M, Wu H, Brocca L. High-resolution satellite products improve hydrological modeling in northern Italy. Hydrol Earth Syst Sci [Internet]. 2022 Jul 29;26(14):3921-39. Available from: https://doi.org/10.5194/hess-26-3921-2022

[2] Vallentin C, Harfenmeister K, Itzerott S, Kleinschmit B, Conrad C, Spengler D. Suitability of satellite remote sensing data for yield estimation in northeast Germany. Precis Agric [Internet]. 2021 Jun 17. Available from: https://doi.org/10.1007/s11119-021-09827-6

[3] Zheng E, Zhang Y, Zhang J, Zhu J, Yan J, Liu G. Deep learning-based study on assessment and enhancement strategy for geological disaster emergency evacuation capacity in Changbai Mountain North Scenic Area. Sci Rep [Internet]. 2024 Dec 28;14(1). Available from: https://doi.org/10.1038/s41598-024-81583-9

[4] Meraner A, Ebel P, Zhu XX, Schmitt M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J Photogramm Remote Sens [Internet]. 2020 Aug;166:333-46. Available from: https://doi.org/10.1016/j.isprsjprs.2020.05.013

[5] Stubenrauch CJ, Rossow WB, Kinne S, Ackerman S, Cesana G, Chepfer H, Di Girolamo L, Getzewich B, Guignard A, Heidinger A, Maddux BC, Menzel WP, Minnis P, Pearl C, Platnick S, Poulsen C, Riedi J, Sun-Mack S, Walther A, Winker D, Zeng S, Zhao G. Assessment of global cloud datasets from satellites: project and database initiated by the GEWEX radiation panel. Bull Am Meteorol Soc [Internet]. 2013 Jul 1;94 (7):1031-49. Available from: https://doi.org/10.1175/bams-d-12-00117.1

[6] Enomoto K, Sakurada K, Wang W, Fukui H, Matsuoka M, Nakamura R, Kawaguchi N. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW) [Internet]; 2017 Jul 21-26; Honolulu, HI, USA. New York, NY, USA: IEEE; 2017. Available from: https://doi.org/10.1109/cvprw.2017.197

[7] Baetens L, Desjardins C, Hagolle O. Validation of copernicus sentinel-2 cloud masks obtained from MAJA, sen2cor, and fmask processors using reference cloud masks generated with a supervised active learning procedure. Remote Sens [Internet]. 2019 Feb 20;11 (4):433. Available from: https://doi.org/10.3390/rs11040433

[8] Silong L, Xiaoguang Z, Dongyang H, Ali N, Qiankun K, Sijia W. A multi-feature framework for quantifying information content of optical remote sensing imagery. Remote Sens [Internet]. 2022 Aug 20;14(16):4068. Available from: https://doi.org/10.3390/rs14164068

[9] Shahinfar S, Meek P, Falzon G. "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecol Inform [Internet]. 2020 May;57:101085. Available from: https://doi.org/10.1016/j.ecoinf.2020.101085

[10] Zar J. Biostatistical analysis [Internet]. 5th ed. Harlow: Pearson Education Limited; 2013. 760 p. Available from: https://elibrary.pearson.de book/99.150005/9781292037110?utm_source=copilot.com

[11] Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Loy CC, Qiao Y, Tang X. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [Internet]. arXiv preprint arXiv:1809.00219. 2018. Available from: https://arxiv.org/abs/1809.00219