Reviewer 1 Responses

We thank the reviewer for the detailed and constructive evaluation of our manuscript. The comments provided have been instrumental in improving the clarity, methodological rigor, and interpretation of the study.

In response, we plan to revise the manuscript to better align the claims with the supporting evidence, strengthen the statistical evaluation of the modelling framework, and refine the discussion of key findings. We believe these revisions would result in a more balanced and robust presentation of the work.

Below we provide the reviewers full comments in blue and with the responses following each comment.

1. The manuscript repeatedly claims to be the first to integrate LiDAR CH, multi-layer GF, LiDAR intensity, MS, and TIR for biomass estimation. This is not accurate. Multi-sensor fusion combining LiDAR and MS has been done, including in wheat. LiDAR + TIR exists in related cereal studies. The unique contribution here is the specific configuration and systematic comparison, not the existence of the fusion itself. The novelty claim must be rewritten. As written, it will not pass expert scrutiny.

We would like to clarify that it was not our intention to claim that multi-sensor fusion of LiDAR, multispectral, and thermal data has not been previously explored. We agree that such approaches have been applied in crop monitoring, including in wheat.

Our intention was to emphasize the specific combination and systematic evaluation of multiple LiDAR-derived features, particularly multi-layer gap fraction and normalized intensity, together with multispectral and thermal data within a unified framework. We acknowledge that the original wording may have given the impression of a broader novelty claim than intended.

In response, we plan to revise the manuscript to remove or rephrase statements that could be interpreted as claiming novelty in the fusion concept itself. The revised text now more clearly positions this study as an extension of existing multi-sensor approaches, focusing on the integration and comparative assessment of underused LiDAR features rather than the fusion concept alone. In addition, we will revise the manuscript more broadly to reduce overly strong wording and ensure a more neutral and scientifically precise tone throughout.

2. The paper makes strong claims about LiDAR intensity encoding physiological and structural canopy traits. These claims are not sufficiently supported.

- Intensity is highly sensitive to range, angle, moisture, and instrument drift.
- Only ground-return normalization is applied, which is insufficient for cross-date physiological interpretation.
- No radiometric calibration or angular correction is performed.
- No independent evidence is presented linking intensity to nitrogen status, pigment changes, or biochemical traits.

The conclusions around LiDAR intensity are overstated and require substantial tempering. At present, the manuscript treats INT as if it were a calibrated spectral variable, which it is not.

We fully agree that LiDAR intensity is influenced by multiple factors, including range, scan angle, surface moisture, and that it should not be interpreted as a radiometrically calibrated spectral measurement. Our intention was not to treat intensity as calibrated reflectance, but rather as an empirical signal that, when appropriately normalized, can provide useful information related to canopy structure and, indirectly, vegetation condition. We acknowledge that the current manuscript may convey a stronger interpretation than intended and will revise the text to ensure a more cautious and scientifically appropriate framing.

This interpretation is supported by previous studies demonstrating relationships between LiDAR-derived signals and vegetation condition, including green canopy dynamics (e.g., GAI/GLAI; Liu et al., 2017), plant water status (Junttila et al., 2021), responses to nitrogen treatment (Hütt et al., 2023), and broader bio-/geophysical parameters (Scaioni et al., 2018).

In this study, we apply ground-return normalization to reduce temporal variability in the intensity signal, following approaches used in previous work (e.g., You et al., 2017), which demonstrated improved LAI estimation using normalized LiDAR signals. While this does not constitute full radiometric calibration, it represents a practical baseline method to improve temporal consistency under typical UAV acquisition conditions. We will clarify this distinction more explicitly in the revised manuscript.

We will also expand the Discussion (currently briefly addressed in Lines 443–444) to acknowledge that additional corrections (e.g., scan angle, range, and atmospheric effects) are often required in LiDAR intensity analysis. Although these effects may be reduced in our case due to low flight altitude (~50 m), relatively small scan angles (~25°), double-grid flight patterns, and flat terrain, they are not fully eliminated. Previous studies have shown that incidence angle effects become more pronounced at higher angles, while remaining limited at lower angles (e.g., Hütt et al., 2024), and that range effects may be less influential under UAV acquisition conditions compared to airborne systems (e.g., Bakuła et al., 2020). We will emphasize that incorporating such corrections could further improve intensity-based analyses, particularly in more complex scenarios.

To further support the interpretation of LiDAR intensity, we will include additional analysis comparing normalized intensity with independently measured plant area index (PAI), leaf area index (LAI), green leaf area index (GLAI), and brown leaf area index (BLAI) derived from destructive sampling. This will allow us to assess whether intensity is more closely related to canopy structure or to green foliage components.

We will present these results to demonstrate that LiDAR intensity may reflect aspects of both canopy structure and greenness, which are themselves linked to vegetation condition. Accordingly, we will position it as an empirical structural–spectral proxy that complements purely structural metrics such as crop height and gap fraction. Overall, we will revise the manuscript to temper its interpretation, explicitly state its limitations, and strengthen the supporting evidence to ensure a more accurate and balanced presentation.

3. **The ANN modelling framework lacks statistical rigor**

The modelling approach is a weak point of the paper.

- **Only 86 destructive samples** are available, which is extremely limited for training multi-feature ANN models.
- A simple 70/30 split is insufficient for robust validation; no cross-validation, no repeated sampling, and no uncertainty estimates are provided.
- There is no demonstration that spatial autocorrelation was controlled. The model may be learning subplot-level patterns, not generalizable relationships.
- Feature dimensionality is high relative to sample size, yet model selection is described loosely and without formal procedure.

As it stands, the modelling is not statistically robust enough to support the strong performance claims made throughout the manuscript.

We agree that the original evaluation of the ANN models did not sufficiently demonstrate statistical robustness, and we appreciate the emphasis on strengthening the validation framework. In the revised manuscript, we will substantially improve the modelling evaluation by implementing repeated k-fold cross-validation in addition to the original 70/30 holdout approach. We will report mean performance metrics along with measures of variability (e.g., standard deviation) across folds to provide uncertainty estimates. We will also clarify the model selection procedure and constraints applied to limit overfitting given the relatively small sample size.

To provide a clearer and more structured evaluation, we will adopt a three-step framework: (1) initial comparison of sensor feature combinations using the holdout dataset; (2) robustness assessment of top-performing and commonly used feature sets using cross-validation; and (3) where possible, independent validation using an external dataset.

In this context, we will include an additional analysis step by applying cross-validation to the best-performing and representative sensor configurations identified in the initial comparison. This will provide stronger statistical support for the observed performance differences between sensor feature sets.

If feasible within the revision timeline, we will further evaluate model generalizability by training models on the full 2021 dataset and testing them on an independent dataset collected in 2023 over a separate winter wheat field with identical sensor configurations and destructive measurements. This additional validation will help assess transferability and provide insight into potential spatial dependence effects beyond subplot-level sampling.

We will also expand the Discussion to explicitly acknowledge limitations related to sample size, feature dimensionality, and potential spatial autocorrelation, noting that the primary objective of this study is to compare sensor feature contributions rather than to develop a fully generalizable predictive model. These additions will provide a more robust, transparent, and statistically supported evaluation of model performance while maintaining the focus on comparative analysis of sensor-derived features.

4. **Multi-layer GF method needs deeper justification**

The multi-layer GF analysis is a valuable idea, but the implementation and interpretation need more discipline.

- The segmentation thresholds (10/20/30 cm) appear arbitrary.
- No empirical or theoretical justification is given for why these scales represent meaningful canopy stratification.
- The comparison to 3DPI and classical GF methods is underdeveloped.
- The model’s sensitivity to point density and scan geometry is not addressed.

The method shows potential, but the manuscript overstates its generality and does not provide enough evidence for the proposed optimal configuration beyond this specific dataset.

We agree that the justification and interpretation of the multi-layer gap fraction (GF) approach require further clarification and more careful framing.

In the revised manuscript, we will clarify that the selected vertical layer depths (10–30 cm) and horizontal resolutions (10–30 cm) were chosen as practical discretization levels based on LiDAR range precision (~4 cm), UAV point density, wheat canopy height (~1–1.2 m), and plot dimensions, rather than representing theoretically optimal canopy stratification. Layers thinner than 10 cm approach the sensor precision and lead to sparse voxel representation at the given flight altitude, while layers thicker than 30 cm would result in too few strata to meaningfully describe canopy structure. Similarly, smaller grid cells would contain insufficient returns for stable gap fraction estimation, whereas larger cells would not fully exploit the spatial resolution of the data.

We will revise the manuscript to emphasize that the identified configuration (e.g., 5 layers at 20 cm with 30 cm GSD) is specific to this dataset and sensor setup, and we will avoid presenting it as a universally optimal solution. Instead, we will frame the analysis as an exploration of parameter sensitivity and practical implementation constraints.

We will also clarify that the 3D Plant Index (3DPI) is cited as conceptual background for multi-layer canopy representation rather than as a direct benchmark. Although this is briefly addressed in the current manuscript, we will further expand the discussion to more clearly distinguish our approach from classical gap fraction methods and previous 3DPI-based implementations. In addition, we note that Line 385 already highlights a key distinction, where we state that our approach directly incorporates multi-layer GF inputs into the ANN, allowing the model to learn their relative importance throughout the season without additional parameterization, in contrast to 3DPI-based methods. We will revise this section to clarify that this is intended to explain methodological differences rather than serve as a direct comparison. If deemed necessary, and depending on the scope of the revision, we would also consider including a direct 3DPI-based analysis to further contextualize our approach within existing methods.

Regarding scan geometry and point density, we acknowledge that these factors were not sufficiently addressed. In the revised manuscript, we will expand this section to describe the acquisition conditions, including relatively low scan angles, double-grid flight patterns with high overlap, and flat terrain, which reduce variability in scan geometry and point density. We will also include additional analysis to evaluate the influence of scan angle, including its relationship with AGB and its effect on model performance, noting that its inclusion in preliminary modelling did not improve predictive accuracy. Finally, we will explicitly acknowledge that the sensitivity of the GF approach to point density and scan geometry may become more significant under different acquisition conditions, and that further investigation across crop types, sensor configurations, and flight parameters is needed.

5. **Sensor dominance over time is overstated**

The temporal analysis suggests shifts in sensor utility across growth stages. While the general trends are plausible, the manuscript repeatedly makes categorical statements (e.g., "MS dominates during senescence," "CH dominates early") without rigorous statistical backing.

These conclusions need to be presented as observations from this dataset, not generalizable statements about sensor behavior

We agree that the description of temporal "sensor dominance" was phrased too broadly in the original manuscript. In the revised manuscript, we will reframe these results as observations specific to this dataset rather than generalizable statements about sensor behavior. We will also moderate the language to avoid categorical terms such as "dominates," and instead describe relative performance trends across growth stages. Where appropriate, we will support these observations with references to prior studies that report similar patterns. Finally, we will expand the Discussion to emphasize that these trends are context-dependent and may vary with crop type, site conditions, and season, thereby avoiding overgeneralization.

6. **Nitrogen treatment analysis draws conclusions not supported by data**

   Figure 12 shows expected spatial smoothing from UAV-based predictions compared to subplot-level destructive samples. This does not prove that UAV "captures management effects more effectively." It only shows that UAV sampling is spatially denser.

   The manuscript conflates spatial resolution advantages with biological sensitivity. This needs correction.

We agree that the original wording may have conflated spatial sampling advantages with biological sensitivity. Our intention was not to imply that UAV-based estimates possess greater biological sensitivity than destructive sampling, which is not possible. In the revised manuscript, we will clarify that UAV-derived biomass maps provide spatially continuous coverage, allowing management-related patterns (e.g., nitrogen gradients) to be visualized more comprehensively than with sparse destructive measurements.

We will revise the text to explicitly distinguish between spatial completeness and biological sensitivity, emphasizing that the observed differences reflect sampling density rather than an inherent improvement in physiological or biological detection capability. These revisions will ensure a more accurate and appropriately constrained interpretation of the results

Other Critical Points

7. The abstract uses promotional language ("breakthrough," "promising tool") that is not supported by the analysis.

We will revise the Abstract to remove promotional language and ensure that all statements are aligned with the strength of the presented analysis and supported by the results.

8. Figures are overly dense, especially Figures 7–9; interpretation is difficult.

We agree that the original figures were overly dense and could hinder interpretation. We will improve figure clarity by reorganizing the visual content. Specifically, the scatter plots will be moved to the Appendix, where they will be referenced in the main text. These plots are useful for illustrating bias and feature behavior across the growing season. However, their density makes them less suitable for the main body of the manuscript.

In the main text, we will retain and emphasize bar plot comparisons, which more clearly summarize model performance across sensor features and combinations. Following the initial 70/30 training/testing comparison, we will also include bar plots of cross-validation results for the best-performing and commonly used feature sets, providing a clearer and more structured presentation of model performance.

9. No error bars or confidence intervals are provided anywhere, this weakens all conclusions.

We agree that the original manuscript lacked appropriate uncertainty representation. As part of the revised modelling framework, we will include variability metrics derived from cross-validation (e.g., standard deviation across folds), providing a clearer and more robust assessment of model uncertainty.

10. The discussion section extrapolates beyond the evidence, particularly regarding the physiological meaning of intensity and the claimed operational advantages.

We will revise the Discussion to reduce over interpretation and ensure that all conclusions are directly supported by the presented results. In particular, we will temper statements related to LiDAR intensity and operational implications to reflect a more cautious and evidence-based interpretation. We will also emphasize that the observed benefits are specific to the conditions and use case investigated in this study, and should be interpreted as demonstrations within this experimental context rather than generalizable conclusions.

References

Bakuła, K., Pilarska, M., Ostrowski, W., Nowicki, A., & Kurczyński, Z. (2020). UAV LIDAR DATA PROCESSING: INFLUENCE OF FLIGHT HEIGHT ON GEOMETRIC ACCURACY, RADIOMETRIC INFORMATION AND PARAMETER SETTING IN DTM PRODUCTION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B1-2020, 21–26. https://doi.org/10.5194/isprs-archives-XLIII-B1-2020-21-2020

Hütt, C., Bolten, A., Hüging, H., & Bareth, G. (2023). UAV LiDAR Metrics for Monitoring Crop Height, Biomass and Nitrogen Uptake: A Case Study on a Winter Wheat Field Trial. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 91(2), 65–76. https://doi.org/10.1007/s41064-022-00228-6

Hütt, C., Bolten, A., Firl, H., Hüging, H., Jenal, A., & Reddig, F. (2024). LiDAR intensity variability in UAV-based agricultural monitoring: Insights from a winter wheat field trial. DGPF-Jahrestagung 2024 - Stadt, Land, Fluss - Daten Vernetzen, 334–341. https://doi.org/10.24407/KXP:1885708025

Junttila, S., Hölttä, T., Puttonen, E., Katoh, M., Vastaranta, M., Kaartinen, H., Holopainen, M., & Hyyppä, H. (2021). Terrestrial laser scanning intensity captures diurnal variation in leaf water potential. Remote Sensing of Environment, 255, 112274. https://doi.org/10.1016/j.rse.2020.112274

Liu, S., Baret, F., Abichou, M., Boudon, F., Thomas, S., Zhao, K., Fournier, C., Andrieu, B., Irfan, K., Hemmerlé, M., & Solan, B. de. (2017). Estimating wheat green area index from ground-based LiDAR measurement using a 3D canopy structure model. Agricultural and Forest Meteorology, 247, 12–20. https://doi.org/10.1016/j.agrformet.2017.07.007

Scaioni, M., Höfle, B., Baungarten Kersting, A. P., Barazzetti, L., Previtali, M., & Wujanz, D. (2018). METHODS FROM INFORMATION EXTRACTION FROM LIDAR INTENSITY DATA AND MULTISPECTRAL LIDAR TECHNOLOGY. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII–3, 1503–1510. https://doi.org/10.5194/isprs-archives-XLII-3-1503-2018

You, H., Wang, T., Skidmore, A. K., & Xing, Y. (2017). Quantifying the Effects of Normalisation of Airborne LiDAR Intensity on Coniferous Forest Leaf Area Index Estimations. Remote Sensing, 9(2), Article 2. https://doi.org/10.3390/rs9020163