

Revision note

Manuscript number: essd-2026-64

Manuscript title: HyBEAR: A Hyperspectral Benchmark for Bare Soil Detection

We would like to thank the Editor and Reviewers for considering the paper and for very insightful suggestions.

We have refined the paper carefully following all the remarks. Please find below our point-to-point responses to the specific comments.

Reviewer #1

This paper provides a curated benchmark about hyperspectral images for bare soil recognition along with the cross-validation procedures, as well as a comparative of the different ML models (and their code) used in this task.

The paper is very well-written, it is clear that the authors have been working on this topic for a long time and revised the document carefully. The contribution is notable, as this kind of annotated benchmark is hard to find in open data repositories. The results shown in the paper are sound and the validation procedure is methodologically correct.

Response: We would like to thank the Reviewer for the very positive assessment of our work and the encouraging feedback. We are pleased that the quality of the HyBEAR benchmark, its documentation, and the soundness of our experimental validation were recognized. We appreciate the Reviewer’s confirmation regarding the importance of providing such annotated datasets in open data repositories to foster research in bare soil detection for precision agriculture.

Still, there are room for some improvement and clarification as listed below:

1. Introduction section. Perhaps it could be a better idea to split this section in two, being the new one named “Background” or “Related Work” or similar. In this manner, it would be clearer the context and goal of the paper (in the introduction) and the state of the arte and the contribution (in the new section).

Response: We thank the Reviewer for this constructive suggestion regarding the structure of our manuscript. We agree that the current Introduction is quite extensive and that separating the background information from the primary objectives will enhance the paper’s clarity. In the revised version of the manuscript, we will restructure this section by splitting it as follows:

- Section 1: Introduction – This section will be focused strictly on the motivation, the general context of precision agriculture, and the core goals of the HyBEAR benchmark.
- Section 2: Background / Related Work – We will move the detailed state-of-the-art review and the discussion of existing remote sensing techniques here. This section will conclude with a clear description of our Contribution and the Structure of the paper, providing a more logical flow and better positioning of our work.

2. For the NO-DATA in the images, it should be clarified why there is the case that no data is captured. Is due to limitation on the device used? Or due to absence/noisy data?

Response: We thank the Reviewer for this question and the opportunity to clarify this technical aspect. The presence of NO-DATA pixels in certain patches is not due to sensor noise or device failure. Instead, it is a consequence of the irregular geographical boundaries of the flight strips and the subsequent orthorectification process.

When the large source orthophotomaps are merged from multiple flight strips and then divided into a regular grid of square patches (250×250 pixels), the areas that fall outside the actual coverage of the acquired data are left empty. These pixels are encoded with the specific value of -9999 in the hyperspectral images and are marked as background in the ground-truth images to support their automated filtration during any data analysis or model training. We will include this clarification in Section 2.3 of the revised manuscript.

3. Can be the cost of the hyperspectral data collection measured? It would be convenient to show this information in section 2.

Response: We agree with the Reviewer that the “cost” of such a large-scale hyperspectral campaign is an important factor. However, providing a precise monetary value is difficult due to the commercial nature of the flight operations and varying market rates for specialized equipment. Instead, to address this point, we will update Section 2.1 to more clearly describe the scale of the investment and the resources involved, which serve as a proxy for the data’s high value:

- Specialized equipment—the data was acquired using a manned Piper PA-31 Navajo aircraft equipped with a state-of-the-art dual-sensor HySpex system (VNIR and SWIR).
- Scale of operation—the campaign covered over 43,000 hectares (approx. 108 million pixels) in two distinct geographical regions.
- Institutional support—we will explicitly mention that this extensive data collection was made possible through major funding from the SmartSoil (POIR.01.01.01-00-0287/21) and DELTA (no. 314/25) projects.

By highlighting these technical and logistical details, we demonstrate the significant effort and investment required to produce a benchmark of this magnitude without disclosing confidential financial data.

4. In section 2.2, when describing the limitations of identifying different elements in Figure 2 (eg., dirt road, tree shadows), it would be very useful to circle or signaling these elements in the figure itself, to quickly spot the problem.

Response: We thank the Reviewer for this practical suggestion to improve the visual clarity of our examples. We agree that adding explicit markers will help the reader quickly identify the specific features that make bare soil detection challenging.

In the revised version of the manuscript, we will update Figure 2 by adding visual indicators (such as circles or arrows) to patches (c) and (d). These markers will specifically point out the dirt roads and tree shadows mentioned in Section 2.2, clearly illustrating why these elements can be misleading when relying solely on RGB or CIR imagery during the ground truth preparation process.

5. Regarding the results, I miss an explanation on the low performance of DT, Random Forest and AB models with Folder 0. Could it be due to the fact that this folder is the one with the minimum percentage of soil pixel ratio?

Response: We thank the Reviewer for this insightful observation regarding the performance drop in Fold 0. The Reviewer is correct that Fold 0 has the lowest soil pixel ratio in the dataset (28.3% for the FULL version and 25.6% for MINI), which may indeed contribute to the difficulty.

However, we believe the primary reason for the lower performance of non-linear models like Decision Trees, Random Forest, and AdaBoost on this specific fold is that Fold 0 represents the most challenging test of spatial generalization within our benchmark. While Folds 1–4 are all derived from the same source orthophotomap (P2), Fold 0 is the only fold extracted from a completely different geographic location (P1). These two areas are located more than 60 km apart and, although acquired on the same day, they were captured under different lighting and atmospheric conditions due to the dynamic position of the sun and clouds.

Our results suggest that while complex models like Random Forests can highly accurately fit the characteristics of the P2 scene, they are more prone to performance degradation when faced with the domain shift present in the P1 scene (Fold 0). In contrast, simpler linear models (LR, SVM) proved to be more robust and universal in this cross-location scenario. We will expand the discussion in Section 3.2 to include this explanation, clarifying the role of both the class imbalance and the geographic disjointness of Fold 0.

6. An explicit paragraph indicating the limitations of this work is needed, as well as a more clear description of specific application that can be benefited from this work.

Response: We thank the Reviewer for this suggestion to provide a more balanced and practical perspective on our work. In the revised manuscript, we will address this by adding a dedicated paragraph regarding the limitations and expanding on the practical applications.

We will explicitly state in the Conclusions section of Limitations that while HyBEAR is a large-scale collection, it is currently limited by its regional focus (Southern Poland) and temporal snapshot (data acquired on a single day in early spring). Consequently, it may not capture the full spectral variability of all global soil types or the influence of diverse seasonal and climatic conditions. We will also acknowledge that the current baselines are primarily pixel-level, which provides a starting point but does not yet fully exploit the field-level nature of the annotations for parcel-based decision-making.

We will expand Section 1.3 to provide a clearer description of specific applications that directly benefit from this benchmark:

- Precision agriculture—using the provided bare soil detections as masks for targeted fertilization and soil property mapping, ensuring that analytical models (e.g., for moisture or organic matter) are applied only to relevant pixels.
- Environmental monitoring—facilitating the large-scale tracking of agricultural practices such as tillage and fallow periods, as well as identifying areas at risk of soil erosion
- On-board data compression—supporting the development of algorithms for satellite edge devices, where pruning non-soil areas can drastically reduce data transmission volume and accelerate on-orbit analysis.

We believe these additions will significantly help the community understand both the immediate utility and the boundaries of the current HyBEAR benchmark.

Reviewer #2

General comments

The manuscript presents HyBEAR, a hyperspectral benchmark dataset for bare soil detection at the agricultural field scale. The topic is relevant for Earth system science data, precision agriculture, hyperspectral remote sensing, and machine learning applications. The paper clearly motivates the need for bare soil detection as a preprocessing step for soil property estimation and other precision agriculture workflows.

The dataset is potentially valuable for the community. It provides high-resolution airborne hyperspectral imagery, ground sampling distance of 2 m, 430 spectral bands, manually prepared and expert-verified bare soil annotations, predefined cross-validation folds, baseline machine learning results, and code/data availability through Zenodo. The size of the dataset is also substantial for this specific task.

I appreciate that the authors provide both FULL and MINI versions of the dataset. This is useful because the FULL dataset is relatively large, while the MINI version allows users to test workflows and reproduce some baseline results without immediately downloading and processing the full 96 GB dataset.

I also checked the data availability and basic usability of the supplementary materials. The DOI is valid, the data can be downloaded without an additional request, and the files can be opened both in QGIS and using the provided code. The MINI dataset and the provided trained models allow reproduction of the reported results for Logistic Regression and SVM, and the reproduced values are consistent with the reported tables. This is an important positive aspect of the manuscript.

However, I think the manuscript requires revision before publication. The main issues concern metadata completeness, reproducibility of the full baseline workflow, the limited spatial diversity of the dataset, the interpretation of the cross-validation protocol, and the mismatch between the stated “whole-field bare soil” objective and the primarily pixel-level baseline evaluation.

Response: We would like to thank Reviewer for the very thorough and constructive assessment of our manuscript. We specifically appreciate the Reviewer’s time and effort invested in verifying the data availability and the usability of our supplementary materials.

We are pleased that the Reviewer successfully independently reproduced our baseline results for Logistic Regression and SVM, confirming the transparency of our work. We also value the recognition of the HyBEAR benchmark’s potential for the Earth system science community and the practical utility of providing both FULL and MINI versions of the dataset.

We have carefully considered all the points raised regarding metadata, reproducibility, and spatial diversity. We believe that the resulting revisions, which we discuss in detail in the following point-by-point responses, have significantly strengthened the quality and reliability of this benchmark.

Specific comments

1. Data availability, metadata, and usability

The data are available through Zenodo and can be downloaded without restriction. The file structure is generally understandable, and the TIFF/GT files can be opened in standard geospatial software and through Python. This is a strong point of the paper.

However, the metadata description should be improved. The hyperspectral image files contain 430 bands, but the bands are not clearly named or documented in a user-friendly way. For a hyperspectral benchmark, the mapping between band index and wavelength is essential. At the moment, users need to inspect the code to understand which bands correspond to RGB visualization, for example. This creates an unnecessary barrier for reuse, especially for users who are not already familiar with this sensor configuration.

I recommend adding a separate metadata file, for example `bands.csv` or `wavelengths.csv`, with band index, central wavelength, spectral range, and sensor source if applicable. The manuscript should also clearly state whether this wavelength information is encoded inside the TIFF metadata or only provided externally.

Response: We thank the Reviewer for this practical recommendation to improve the accessibility and usability of the HyBEAR benchmark. We agree that a clear mapping between band indices and wavelengths is essential for the effective reuse of hyperspectral data, particularly for those not familiar with the specific sensor configuration. To address this, we have added a dedicated metadata file named `bands.csv` to the Zenodo repository. This file provides a comprehensive reference for all 430 spectral bands included in the dataset. Specifically, it includes:

- Band Index (0 to 429).
- Central Wavelength [nm] (covering the range from 414.1 nm to 2357.4 nm).
- Spectral Resolution (3.26 nm for the VNIR range and 5.45 nm for the SWIR range).
- Sensor Source (identifying which bands originate from the VNIR-1800 and SWIR-384 sensors).

Furthermore, the `bands.csv` file explicitly documents the band indices used for RGB and CIR visualizations (as specified in Section 2.2), ensuring that users can easily create visual representations without inspecting the baseline code. In the revised manuscript, we will clarify that while the TIFF files contain basic geospatial and encoding metadata, the detailed wavelength and sensor mapping is provided primarily through this external metadata file to ensure maximum user-friendliness and ease of access.

2. Ground truth quality and annotation protocol

The description of the ground truth preparation is generally clear. The authors explain that the annotation process used RGB, CIR, and NDVI representations, followed by expert verification. The distinction between SOIL and MAYBE-SOIL during the annotation process is useful, and the conversion of ambiguous cases into final SOIL or NON-SOIL classes is reasonable.

Nevertheless, the annotation protocol should be described in more detail. Since this is a benchmark dataset, confidence in the ground truth is central. The authors should clarify whether several annotators independently labeled the same areas, whether inter-annotator agreement was assessed, and how disagreements or uncertain cases were resolved. It would also be useful to provide more explicit rules for sparse vegetation, crop residues, shadows, soil moisture effects, roads, field margins, and other ambiguous cases.

At present, the trust in the GT relies mainly on the statement that the annotations were manually checked and verified by domain experts. This is acceptable as a starting point, but for a benchmark dataset the quality control procedure should be more transparent.

Response: We thank the Reviewer for this suggestion to increase the transparency of our ground-truth (GT) preparation. We agree that for a benchmark dataset, a detailed and rigorous quality control protocol is a fundamental pillar of trust. In the revised manuscript, we will expand Section 2.2.1 to provide the following procedural details:

- Annotator Team and Resolution of Disagreements—the annotation team consisted of three experts with 1, 4, and 10 years of experience in remote sensing and agricultural data

analysis. The workflow was designed to be iterative: initial delineations were performed, and any area deemed ambiguous was marked with the MAYBE-SOIL label. All such cases were then subjected to a consensus-based group discussion involving all three annotators. Final decisions for the most difficult cases were moderated by the most senior expert (10 years of experience) to ensure long-term consistency across both geographic locations (P1 and P2).

- Explicit Rules for Ambiguous Cases:
 - Sparse vegetation—a parcel was labeled as SOIL only if the visible bare soil constituted at least approximately 85% of the surface, with no distinct indications of mature vegetation
 - Crop residues and stubble—these were distinguished from active vegetation by cross-referencing CIR and NDVI compositions. If the infrared signal showed no signs of active photosynthesis despite the presence of surface matter, the area was retained as bare soil.
 - Roads and field margins—dirt roads and technical infrastructure were manually excluded from the SOIL polygons by inspecting both the hyperspectral data and high-resolution RGB imagery to ensure only the interior of agricultural parcels remained.
 - Shadows and moisture—shaded areas (e.g., near tree lines) and high-moisture zones were carefully inspected; if the shadow was deep enough to obscure the spectral characteristics of the soil, the area was excluded from the SOIL class to ensure the benchmark remains focused on high-quality, interpretable pixels.
- Quality control transparency—the labeling was performed using a dedicated Python application based on the LabelMe library, which allowed for precise, parcel-level polygon management. We believe these additional details regarding the expert-led consensus and the “85% visibility rule” clarify the rigor of our quality control procedure and strengthen the reliability of the HyBEAR benchmark.

3. Whole-field bare soil concept versus pixel-level evaluation

One of the most interesting aspects of the paper is the claim that HyBEAR supports bare soil detection at the level of entire agricultural fields rather than only isolated bare-soil pixels. This is important and relevant for precision agriculture workflows, where the decision is often made at the field or parcel level.

However, the baseline experiments are still mainly formulated as pixel-wise classification. The models operate on 430-band spectral vectors for individual pixels, and the reported metrics are also pixel-level metrics. This creates a methodological tension between the stated “whole-field bare soil” objective and the actual evaluation protocol.

I recommend that the authors clarify this point. If the dataset is intended to support whole-field bare soil detection, then field-level or parcel-level evaluation metrics should be added. For example, the authors could aggregate pixel predictions within each field and evaluate whether the field is correctly classified as bare soil or non-bare soil. This would make the benchmark more consistent with the stated objective.

At minimum, the manuscript should explicitly state that the current baselines are pixel-wise baselines and do not fully exploit the field-level nature of the annotations. One of the most interesting aspects of the paper is the claim that HyBEAR supports bare soil detection at the level of entire agricultural fields rather than only isolated bare-soil pixels. This is important and relevant for precision agriculture workflows, where the decision is often made at the field or parcel level.

Response: We appreciate the Reviewer’s emphasis on the field-level potential of the HyBEAR benchmark. We agree that the transition from pixel-level detection to whole-field decision-making is a key objective for precision agriculture. To better align the manuscript with this vision while maintaining the integrity of our established baseline, we have refined the text in Sections 4.1 and 5.1 to more clearly distinguish between our current pixel-wise performance reference and the future potential for parcel-based evaluation enabled by our unique annotations. We have also extended the discussion in Section 5.2 (Future Work) to highlight how the community can now leverage these parcel-level ground-truth delineations to develop and validate more advanced field-level strategies, such as majority voting.

4. Spatial diversity and cross-validation protocol

The five-fold cross-validation protocol is useful and clearly described. The use of spatially disjoint folds is appropriate and helps reduce the risk of overly optimistic results due to spatial leakage.

However, the interpretation of this protocol should be more cautious. The dataset is large in terms of patches and pixels, but it comes from only two geographic locations in southern Poland, acquired on the same day and within a short time interval. Therefore, the dataset has limited spatial, seasonal, and agroecological diversity.

This limitation is visible in the results. Fold 0 behaves quite differently from Folds 1–4 for several models. In many cases, performance on Fold 0 is substantially lower, especially in terms of sensitivity and F1-score. This suggests that Fold 0 is a more challenging cross-location test case, whereas Folds 1–4 may mainly reflect variability within the second scene rather than fully independent geographic generalization.

I recommend that the authors tone down claims about robustness and generalization. The dataset is valuable, but it should be presented as a benchmark based on two spatial areas rather than as a broadly representative benchmark for different soil types, regions, seasons, and agricultural systems. A dedicated limitations paragraph would be very helpful.

Response: We thank the Reviewer for this insightful critique of our spatial diversity and the interpretation of our cross-validation results. We agree that while the HyBEAR benchmark is large in terms of pixel count and coverage, it represents a specific regional and temporal snapshot (Southern Poland on a single day in early spring), which inherently limits its seasonal and agroecological diversity. In the revised manuscript, we will implement the following changes to address these concerns:

- Toning down generalization claims—we will revise the Introduction and Conclusions to tone down claims regarding global robustness. Instead, we will more accurately present HyBEAR as a benchmark focused on regional cross-location generalization.
- Interpretation of fold 0—we agree with the Reviewer’s observation that Fold 0 represents a more challenging cross-location test case (moving from scene P2 to P1), while Folds 1–4 primarily reflect spatial variability within the larger P2 scene. This explains the significantly lower sensitivity and F1-scores observed on Fold 0 for models like Random Forest and AdaBoost. We will add a detailed discussion in Section 3.2 to clarify this distinction for the reader.
- Dedicated limitations paragraph—as requested by both reviewers, we will add a dedicated paragraph in the Conclusions section explicitly detailing the limitations of the work. This

will specify that the dataset is currently bounded by its regional focus and the specific climatic/seasonal conditions of the acquisition day, serving as a tool for testing cross-location robustness rather than a globally representative soil index.

We believe these revisions provide a more cautious and scientifically sound interpretation of our results while maintaining the value of HyBEAR as a rigorous regional benchmark.

5. No-data pixels and patch composition

During inspection of the dataset, I noticed that some patches contain a large proportion of no-data pixels. The manuscript states that no-data pixels are encoded as -9999 and can be filtered automatically. This is technically acceptable, but the effect of no-data pixels on the dataset statistics and evaluation should be described more clearly.

The authors should clarify whether no-data pixels are excluded from all reported metrics, including ACC, F1, IoU, MCC, and AUC. It would also be useful to provide statistics on the distribution of valid pixels per patch, or at least to indicate whether patches with a very high no-data fraction were retained without additional filtering.

Response: We thank the Reviewer for this technical query, as it allows us to clarify a fundamental aspect of our evaluation methodology. We confirm that no-data pixels (encoded as -9999 in the HSIs and marked as background in the GT) were strictly excluded from the calculation of all reported performance metrics, including Accuracy (ACC), F-score (F1), IoU, MCC, and AUC. The performance was evaluated only on the “useful” pixels belonging to either the SOIL or NON-SOIL classes. To address your request for more clarity regarding the dataset’s composition:

- Retention of patches—we chose to retain all patches generated by the grid-based extraction process, even those with a high no-data fraction. This was done to preserve the spatial continuity of the regular grid (250×250 pixels) as overlaid on the source orthophotomaps.
- Distribution statistics—in the current version of the manuscript, Table 1 provides the “Background Pixel Ratio [%]” for each fold, which ranges from approximately 6.2% to 15.1% for the FULL dataset. This ratio represents the percentage of no-data pixels within each fold.

In the revised manuscript, we will update Section 2.3 and Section 3.1 to explicitly state the exclusion of background pixels from metric computations. We will also add a more detailed statistical breakdown of the valid (non-background) pixel distribution per fold to Section 2.3 to provide users with a clearer understanding of the effective data density in different parts of the HyBEAR benchmark.

6. Reproducibility of baseline results

The provided code and trained models allow reproduction of some reported results, at least for the MINI dataset and for Logistic Regression and SVM. This is a positive point.

However, it appears that the complete training code for the baseline models is not provided. If only model evaluation code and trained models are available, then the benchmark is only partially reproducible. For an ESSD data paper, and especially for a benchmark dataset, it would be preferable to provide the full training workflow: preprocessing, normalization, sampling strategy, handling of class imbalance, random seeds, model hyperparameters, fold handling, and model training scripts.

If the authors decide not to provide full training code, this limitation should be clearly stated. Otherwise, I recommend adding the missing training scripts to the Zenodo package or to a linked repository.

Response: We thank the Reviewer for highlighting this crucial point regarding the reproducibility of our benchmark. We fully agree that for a data paper in ESSD, providing clear scripts for independent verification is essential. To ensure transparency, we have updated the Zenodo repository (<https://doi.org/10.5281/zenodo.17607897>) to include the full testing procedure. The updated code package now provides:

- Complete evaluation scripts for all six machine learning models reported in this study (Logistic Regression, Linear with L2 regularization, AdaBoost, SVM, Decision Trees, and Random Forest).
- Detailed documentation of preprocessing and normalization procedures for the all spectral bands.
- The exact hyperparameters, random seeds, and fold-handling logic used during the 5-fold cross-validation process to ensure the results reported in Table 2 and Table A1 are fully reproducible.

We have also updated Section 5 Conclusion of the manuscript to explicitly mention that the complete testing workflow is now publicly available, allowing researchers to independently verify the presented benchmark results.

7. Baseline results and interpretation

The baseline results are useful. It is interesting that relatively simple linear models, especially Logistic Regression and SVM, perform very well and sometimes outperform more complex tree-based models. This suggests that spectral information alone is already highly informative for this task.

At the same time, the qualitative examples show that some patches remain difficult, especially in Fold 0. The manuscript would benefit from a deeper discussion of failure cases. It would be helpful to explain whether these errors are likely caused by spectral differences between P1 and P2, illumination changes, soil moisture, crop residues, shadows, boundary ambiguity, or annotation uncertainty.

Response: We thank the Reviewer for this thoughtful analysis of our baseline results and are pleased that the effectiveness of relatively simple linear models, such as Logistic Regression and SVM, was recognized, confirming that the high-dimensional spectral information in our 430-band data is highly discriminative for the bare soil detection task. Regarding the specific difficulties observed in Fold 0, we agree that this fold serves as the primary benchmark for cross-location generalization. In the revised manuscript, we will include a more comprehensive discussion of the failure cases, noting that Fold 0 is extracted from scene P1, while the other folds originate from scene P2, located more than 60 km away. Even though the data were acquired on the same day, the dynamic position of the sun and clouds resulted in varying lighting conditions and reflectance levels between these geographic areas. This domain shift is the primary reason why non-linear models, which may overfit the specific characteristics of P2, show a performance drop when applied to Fold 0. We will also address physical challenges such as tree shadows and dirt roads that can spectrally mimic bare soil, alongside the presence of crop residues and early-stage vegetation that introduces spectral mixing. Furthermore, we acknowledge that Fold 0 contains the lowest soil pixel ratio in the dataset (28.3%), which can exacerbate identification

difficulties in a new geographical context. While we mitigated annotation uncertainty through expert consensus involving specialists with up to 10 years of experience, we recognize that boundary ambiguity remains a challenge that pixel-level models struggle to resolve without exploiting contextual information. We will update Section 3.2 and add detailed comments to Figure 5 to explicitly discuss these failure modes and provide a clearer understanding of the hurdles in cross-location bare soil detection.

8. Length and focus of the manuscript

The manuscript is generally well structured, but some parts of the introduction and literature review could be shortened. Since this is a data description article, the paper would benefit from a stronger focus on dataset construction, annotation quality, metadata, data access, reproducibility, limitations, and reuse scenarios.

Response: We thank the Reviewer for this constructive feedback regarding the manuscript’s focus. We agree that as a data description article, the technical details of the benchmark itself should be the primary emphasis. In the revised version, we will condense the Introduction and restructure the literature review into a separate, more concise Related Work section, which effectively shortens the opening of the paper. This restructuring allows us to place a much stronger focus on the core technical aspects of the HyBEAR benchmark. Specifically, we have expanded the sections covering dataset construction and annotation quality by providing a transparent description of our expert-led quality control protocol and the specific rules used for ambiguous cases. To improve metadata and reproducibility, we have released a comprehensive ‘bands.csv’ wavelength mapping file and the complete training workflow scripts in our Zenodo repository. Furthermore, we have added a dedicated paragraph addressing the limitations of our work and expanded on practical reuse scenarios, such as on-board data compression and parcel-based fertilization planning. We believe these revisions significantly improve the balance of the manuscript, ensuring it serves as a practical and well-documented resource for the scientific community.

9. Technical corrections

- (a) All formulas in Section 2.5 should be checked and reformatted. The current rendering contains several notation errors: missing equality signs after metric names, missing or unclear fraction bars, missing arithmetic operators, and incomplete brackets/parentheses. Please revise the entire section to ensure that all metric formulas are mathematically correct and properly typeset.
- (b) Please use consistent terminology for ROC, AUC, and ROC-AUC. In the tables, the column is labeled “ROC”, but it appears to refer to AUC or ROC-AUC.
- (c) There is a typo: “othophotops”.
- (d) Please add a clear table or metadata file with band indices and wavelengths.
- (e) Please clarify whether the TIFF files contain wavelength metadata internally.
- (f) Please provide the random seed and the exact stratification procedure used to select the MINI subset.
- (g) Please explicitly state whether no-data pixels are excluded from all reported metrics.
- (h) Please add the training scripts for the baseline models, or clearly state that only evaluation scripts and trained models are provided.
- (i) In Figure 5, it would be useful to add a short explanation of why IMG_0022_F0 is a difficult case and what this example shows about cross-location generalization.

Response: We thank the Reviewer for this detailed list of technical corrections. We have carefully implemented all of them in the revised manuscript. Specifically, we have corrected the formulas in Section 2.5, standardized the terminology (using AUC/ROC-AUC), fixed the identified typos, and provided the requested metadata (wavelengths mapping) and training scripts in the Zenodo repository. We have also clarified the exclusion of no-data pixels from metrics and added the requested explanation for Figure 5. We believe these corrections have significantly improved the formal quality and transparency of our work.

10. Recommendation

Overall, I consider HyBEAR a valuable and relevant dataset for ESSD. The data are accessible, the file structure is mostly clear, and part of the baseline results can be reproduced. However, before publication, the manuscript should be substantially revised to improve metadata completeness, clarify the annotation quality control procedure, provide the full training workflow or clearly state the limits of reproducibility, better address the limited spatial diversity of the dataset, and align the evaluation protocol more closely with the stated whole-field bare soil detection objective.

Response: We thank the Reviewer for the positive recommendation and the constructive feedback. We have carefully implemented all suggested revisions—including improvements to metadata, reproducibility, while more clearly distinguishing our pixel-wise baseline results from the field-level potential of the HyBEAR dataset—which have significantly enhanced the quality and transparency of the HyBEAR benchmark. We are grateful for the effort invested in improving our work.

Once again, we would like to thank the Reviewers and Editor for the constructive comments and suggestions—they helped us improve the paper substantially.

Best Regards,

Authors