Eya Cherif
Remote Sensing Centre for Earth
System Research (RSC4Earth)

Leipzig University
cherif@informatik.uni-leipzig.de

To Biogeosciences

27.08.2025

Ref. No.: egusphere-2025-1284- "Uncertainty Assessment in Deep Learning-based Plant Trait Retrievals from Hyperspectral data"

Dear Reviewer, dear Editor,

Thank you for your time reviewing our manuscript and for your constructive and helpful comments. In response to the comments and suggestions provided, the main changes to the manuscript will include:

1. **Methodology presentation:** We will improve the methodological section by adding equations describing the dissimilarity indices and by including a clearer and more comprehensive workflow diagram.
2. **Spectral data description and preprocessing:** We will clarify the description of datasets and provide an explanation of the spectral resampling and smoothing strategy, including its rationale.
3. **Uncertainty modeling and fairness of comparison:** We will clarify in the discussion section our rationale for comparing different uncertainty estimation methods as they are typically applied in literature, while also providing recalibrated results in the appendix for additional context.
4. **Vegetation type interpretation:** We will revise the discussion of grasslands to avoid ambiguity around the term "simple," now framing them in terms of structural homogeneity. We also included additional references.
5. **Technical corrections:** We will address all technical comments, including supplementary material citations, terminology consistency, unit reporting in figures and tables, and clarification of dropout rate usage.

In addition to addressing the reviewer's comments, we have enriched the comparison to state-of-the-art uncertainty estimation methods. In the literature, both probabilistic and deterministic deep ensemble approaches are used; we now explicitly include both methods.

A detailed, point-by-point response, including the proposed changes in the manuscript, are attached.

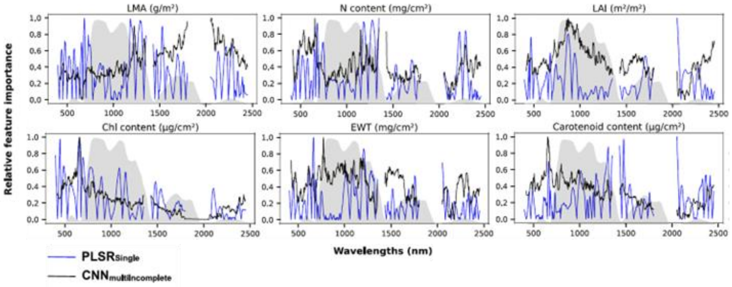
Kind regards,

Eya Cherif (on behalf of all co-authors).

Responses to individual comments from RC1

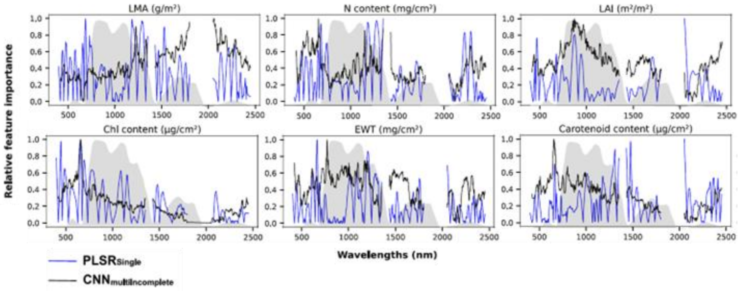
ID	Comment	Response
1	This is an interesting work that deals with the capability of inferring the uncertainty of machine (i.e., Deep) learning models' predictions based on the dissimilarity between the seen and input data of the model. The proposed methodology is tested using a complete dataset of observations, and the results suggest that, particularly for unseen data, the uncertainty estimates are more accurate or at least more conservative than those provided by other methods, which appear to underestimate uncertainty systematically. The results depend on the biophysical variable predicted by the model, with large differences in performance or behaviour in some cases. The manuscript and the results are well presented, and the discussion is consistent with them; still, some relevant questions remain to be clarified. Overall, the work is relevant to the domain of remote sensing and machine learning, and the proposed method appears to improve upon the state-of-the-art alternatives.	We would like to thank the reviewer for these very constructive comments that were very helpful to further improve the manuscript. Please find below a detailed answer to the comments.
1a	The methodology could be more clearly presented (e.g., including equations and a flowchart). Additionally, some results, particularly those regarding LAI, require a deeper inspection that justifies the hypothesis presented by the authors to justify their findings. Some points made in the discussion should be reviewed or	While a flowchart was already included in the original manuscript (Fig. 1), we recognize that it may not have clearly conveyed the methodology. We have therefore revised and improved the flowchart to provide a clearer representation of the workflow (see response to comment 4). In addition, we have elaborated on the specific case of LAI (see response to comment 10), with further clarification in the discussion to better justify the interpretation of saturation effects. We will also revise the discussion in response to comments 5, 6, and 9 to improve clarity of the manuscript.

	linked to the methodological choices made by the authors.	
	Specific comments:	
1	Lines 162-166: Clarify to the reader that spectral data correspond to proximal sensing and airborne canopy reflectance factors so that it is not necessary to access Table S3 to know this detail. Also, specify whether these datasets were gathered from open-access repositories or were privately lent by the producers for this study.	We thank the reviewer for pointing out the ambiguity in the description of the spectral datasets. We will revise the paragraph to more clearly state at the beginning that the data include proximal and airborne reflectance sources and clarify the origin of the datasets (open-access and private contributions). Below a copy of the updated paragraph: *“**2.1.1 The Multi-trait Model*** *…* *The dataset used for this study is a curation of multiple datasets, incorporating spectra and trait observations from diverse ecosystems, including forests, grasslands, shrublands, and agricultural regions. Reflectance data, spanning wavelengths from 400 to 2500 nm, were collected using a variety of hyperspectral sensors, including proximal field spectrometers and airborne imaging instruments. These datasets were gathered from both open-access repositories and privately shared contributions. In total, 50 datasets were integrated into this study (Herrmann et al., 2011; Pottier et al., 2014; Singh et al., 2015; Hank et al., 2015, 2016; Wang et al., 2016; Wocher et al., 2018; Ewald et al., 2018; Cerasoli et al., 2018; Ewald et al., 2020; Kattenborn et al., 2019; van Cleemput et al., 2019; Brown, 2019; Chlus et al., 2020; Wang et al., 2020; Burnett et al., 2021; Dao et al., 2021; Rogers et al., 2021; Brown et al., 2021a; Brodrick et al., 2023; Chadwick et al., 2023; Zheng et al., 2023; Gravel et al., 2024; Table S3). “*
2	Lines 170-172: While the approach is generally accepted (e.g., the ASD field spectroradiometers interpolate to one nm step in their output), I wonder how the authors pondered this choice. Overall, interpolation will not be able to improve the information of the datasets with the coarsest spectral resolution.	We thank the reviewer for raising this important point. To clarify, the interpolation step was not intended to improve the spectral resolution or enhance the information content of lower-resolution sensors. Instead, it was a necessary step to harmonize reflectance data from multiple sources with varying spectral sampling and band positions. This harmonization strategy was developed and described in detail in our earlier work (Cherif et al. 2023), and the current study builds directly upon that unified dataset. Maintaining the same resolution across all spectra is required for the deep learning model. We acknowledge that interpolating coarse-resolution spectra cannot recover fine spectral features not originally measured. However, downsampling all data to the lowest resolution would result in a net loss of useful information for sensors that do capture fine-grained features and may reduce model performance overall. Additionally, even with such downsampling, interpolation would still be necessary to harmonize the central band positions, since the sensors differ not only in resolution but also in band centers.

		In the revised manuscript we will provide an elaborated explanation and rationale on how the data was processed: *"**2.1.1 The Multi-trait Model*** *…* *In line with Cherif et al. (2023), all datasets were resampled to a common 1 nm resolution across the 400–2500 nm range to harmonize the diverse measurements. We chose to upsample rather than downsample, as most of the samples were originally acquired at 1 nm resolution, thereby minimizing manipulation of the data. To address known challenges associated with atmospheric water absorption in open-sky canopy reflectance spectra, we excluded the water absorption regions (1251–1529 nm, 1801–2050 nm, and 2451–2501 nm). The remaining three spectral segments were independently smoothed using a Savitzky-Golay filter (Savitzky and Golay, 1964) with a 65 nm window size. As no sensor-specific noise information was available, the same preprocessing procedure was applied consistently across all datasets to ensure comparability within the curated collection."*
2a	What impact do the authors expect from this imbalance in the information rendered in each dataset? Do they expect that the information contained in the narrow spectral features, which remain only in the datasets with the highest spectral performance, will be learned by the model, even if not present in all the datasets, or would this mixture just be a source of confusion and uncertainty for the training? In the second case, wouldn't it be more robust to downgrade the spectral resolution to the lowest among the datasets?	We thank the reviewer for raising this important point. We acknowledge the potential concern regarding variable spectral resolution across datasets. However, we would like to clarify that vegetation reflectance spectra are typically smooth and continuous reflectance curves without narrow absorption features. Therefore, the fact that some datasets have coarser spectral resolutions than others is unlikely to pose a significant problem for trait retrieval. Fine-scale spectral detail is generally less critical in vegetation studies (see figure below extracted from Cherif et al. 2023, RSE) than in, for example, mineralogy applications.  Consequently, we do not expect that the interpolation to a common 1 nm grid introduces substantial bias or confusion into the model training. We have to choose one common spectral resolution (see comment above) and 1 nm resolution is the most common resolution across all datasets and ensures that we introduce minimal alterations across datasets. While we agree that exploring the impact of spectral resolution and downsampling could be valuable, we consider this beyond the scope

		of the current study, which builds directly on a previously established modeling pipeline (Cherif et al. 2023). Future work could certainly examine how harmonization strategies affect model performance across sensors in both training and inference contexts. Cherif, E., Feilhauer, H., Berger, K., Dao, P. D., Ewald, M., Hank, T. B., ... & Kattenborn, T. (2023). From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data. Remote Sensing of Environment, 292, 113580. In the revised manuscript, we will provide an elaborated rationale on spectral resampling (see comment 2 above).
3	Lines 172-174: Spectra with different noise levels are smoothed with the same window width. Likely, the field spectroradiometers are less noisy than airborne imagers; thus, there's a risk of over-smoothing already smooth data. Overall, the aim should be to achieve a comparable noise level for each dataset. Could processing data according to their noise levels improve the learning process? Considering this further, I understand that the levels of uncertainty are unknown, but perhaps different degrees of reliability could be provided for field spectroscopy and airborne imagery. Would giving different weights to each type of data improve the results?	We thank the reviewer for this comment. While this preprocessing step is not the central focus of the current manuscript, we note that the spectral data used here were pre-processed following the protocol established in a previous study (Cherif et al. 2023). However, to address the reviewer's concern, we would like to clarify the rationale behind using the same smoothing setup across all datasets. We intentionally applied identical smoothing parameters to all spectra to ensure consistent data treatment. Using different smoothing parameters for each dataset, even if motivated by differences in sensor noise, would introduce non-stationarity in the data (e.g. dampening of some spectral features for some sensor types). Furthermore, noise levels are influenced by multiple factors, such as illumination conditions, calibration protocols, reference target quality, and atmospheric correction procedures, that vary across datasets and are often undocumented. This makes it infeasible to develop a simple sensor-specific or noise-aware smoothing strategy. Moreover, the optimal setup for smoothing is not determined solely by the noise characteristics of the sensor, but also by the nature of the spectral features that must be preserved. In our case, we are dealing exclusively with vegetation spectra, which exhibit consistent and broad biophysical features (e.g., the red-edge, green reflectance peak, and chlorophyll absorption features). The smoothing method (Savitzky-Golay filter with the selected parameters) is similarly applied in a range of related studies and were carefully chosen to suppress sensor noise while preserving these critical spectral features. Vegetation spectra are not characterized by narrow absorption features (as is common in mineral spectroscopy), so moderate smoothing does not risk removing meaningful information. This is also supported by the feature importance analysis in Cherif et al. (2023), where SHAP values (Figure below) showed that trait

predictions largely depend on broader spectral patterns rather than isolated narrow bands. This further confirms that minor differences in spectral resolution or smoothing have limited impact on model learning.



In the revised manuscript, we will provide an elaborate rationale on spectral smoothing (see comment 2 above).

We will also add a supplementary section, adding detailed information for our data processing. The section will read as follows:

"S1. Preprocessing pipeline

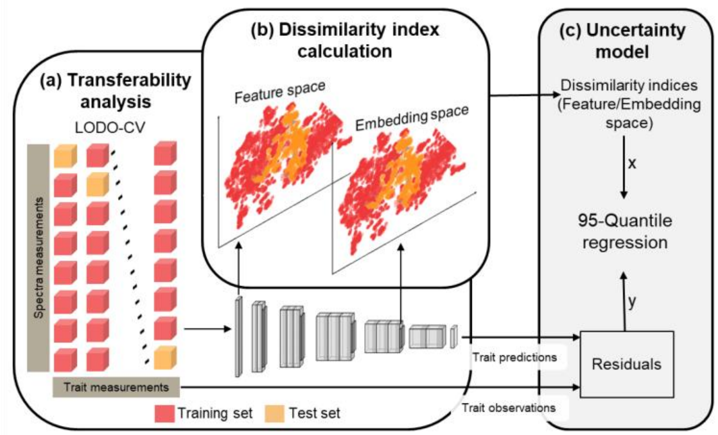
All 50 compiled datasets were pre-processed using the same standardized pipeline, without any dataset-specific deviations. The procedure followed here is based on the analysis of Cherif et al. (2023) and summarized as follows:

First, reflectance spectra were quality-checked. Reflectance values outside the physical range were masked: values below zero were set to missing, and values greater than one were treated as spurious spikes. These missing values were then replaced by the mean of the nearest valid neighbors.

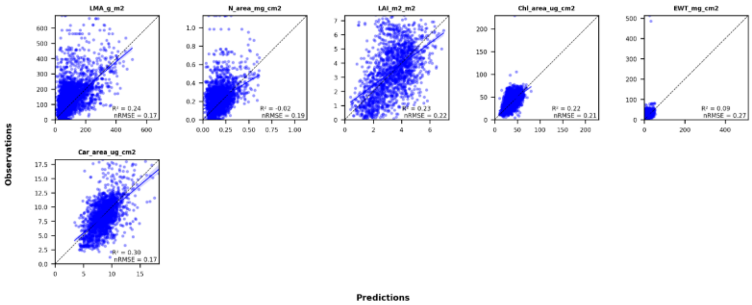
Second, all datasets were resampled to a common 1 nm resolution to harmonize the diverse measurements from different sensors (from proximal and airborne), which varied in spectral sampling and band centers. This resampling was not intended to enhance spectral resolution or recover fine-scale features absent in coarser sensors, but to provide a uniform input representation required for the deep learning model. Most datasets were already acquired at 1 nm resolution, so upsampling was preferred over downsampling to minimize manipulation of the data and avoid loss of information from higher-resolution sensors.

Third, spectral intervals strongly affected by atmospheric water absorption were excluded uniformly across all datasets. In the implementation, the following wavelength ranges were removed: 1351–1430 nm, 1801–2050 nm, and 2451–2500 nm.

		Finally, the remaining reflectance data were smoothed using a Savitzky–Golay filter applied independently to three contiguous segments of the spectrum: 400–1350 nm, 1431–1800 nm, and 2051–2451 nm. Each segment was filtered with a window size of 65 nm and a polynomial order of one.”
4	Section 2.1.2: This section’s clarity could benefit from some equations. Additionally, a flowchart summarizing the methodology would provide a clearer view for the reader at an early stage of the manuscript.	We thank the reviewer for this constructive suggestion, which helped improve the clarity and readability of our methodology. In response, we will revise Section 2.1.2 by including a mathematical formulation of the dissimilarity index (DI), which was used as a core predictor in our uncertainty estimation approach. The added equations describe the cosine distance calculation between test and training spectra, the procedure for summarizing distances via the median of the 50 nearest neighbors, and the normalization against the training set mean (Equations 1–3). We will add the corresponding clarification under the 2.1.2 Dissimilarity indices (predictors) section: *“**2.1.2 Dissimilarity indices (predictors)*** *The DI, used as a predictor in this study, was calculated using the cosine distance, a well-suited metric for analyzing reflectance data. The cosine distance effectively captures the angular relationship between two spectra (Kruse et al., 1993), emphasizing spectral shape while minimizing the influence of amplitude variations that occur uniformly across the spectrum. This helps mitigate brightness changes caused by heterogeneous illumination and internal shading (Feilhauer et al. 2010).* *Formally, the cosine distance between a test spectrum xi and a training spectrum zi is defined as:* $CosineDist(x_i, z_j) = 1 - \frac{x_i.z_j}{\Vert x_i \Vert . \Vert z_j \Vert}$ *(1)* *This DI was applied in both the feature space and the embedding space of the models (Fig. S2). As a first step, we calculated cosine distances between each sample of the test dataset xi and the samples of the training data set zi. These calculations were performed using the Python package FAISS (Douze et al., 2024), which is optimized for fast similarity search and clustering of large datasets. As a next step, each DI was calculated as the median of the distance distribution between a test sample and its 50 nearest neighbors in the training set:* $DI_i = median\{CosineDist(x_i, z_j)\}_{j=1}^{50}$ *(2)*

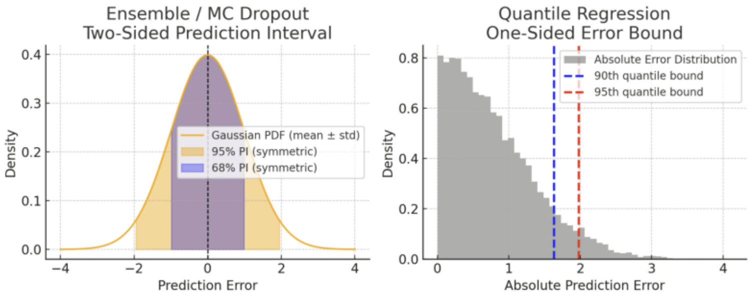
		To ensure comparability across samples, the indices were normalized against the mean DI value of the entire training set (Meyer and Pebesma, 2021): $DI_i^{norm} = \frac{DI_i}{\mu_{train}}$, *with* $\mu_{train} = \frac{1}{n}\sum_{j=1}^{n} DI_i$ *where n is the number of training samples* (3) " Furthermore, we have updated **Fig. 1** to include a clearer and more comprehensive workflow diagram of our distance-based uncertainty method (Dis_UN). The revised figure provides an overview of the entire pipeline and will be presented earlier in the manuscript to guide the reader through the subsequent sections.  (a) Transferability analysis LODO-CV Spectra measurements Trait measurements (b) Dissimilarity index calculation Feature space Embedding space (c) Uncertainty model Dissimilarity indices (Feature/Embedding space) x 95-Quantile regression y Residuals Trait predictions Trait observations Training set Test set *Figure 1: Workflow of the distance-based uncertainty method (Dis_UN) to assess uncertainty of a deep learning model. The method consists of three phases: (a) Leave-one-dataset-out cross-validation (LODO-CV) on the deep learning model, (b) Training data generation for uncertainty estimation using the LODO-C), and (c) uncertainty modeling, which incorporates the following inputs: dissimilarity indices between the training and the test samples in feature and embedding space of the multi-trait model, the trait predictions obtained from the deep learning models and the true trait observations.*
5	Lines 205-207, Section 4.4: The absolute value of the errors is taken. Do the authors expect the errors to be symmetric and centered on zero, or was this checked? Would any knowledge be gained if the error and the 2.5- and 97.5-quantile regressions were applied instead? (e.g., biases in	We thank the reviewer for this comment. We would like to clarify two important aspects in this comment. First, we used absolute errors to focus on the magnitude of prediction errors, independent of direction. This choice aligns with standard practice for evaluating and comparing uncertainty quantification methods such as Ens_UN and MCdrop_UN, which estimate uncertainty through predictive variance (a measure inherently tied to the dispersion of predictions, not their sign). While this approach does not capture directional bias, we recognize the

	specific directions for specific vegetation types). In the discussion (Section 4.4), they precisely raise the issue of assuming or forcing symmetric error distributions, but their analyses start from absolute error values. Could they comment on the potential impact of their choice in the context of symmetric distributions, and maybe foresee future lines of research at least?	importance of analyzing signed residuals, and we will highlight this as a potential extension in future work: ***"4.5 Outlook: Uncertainty in the Context of Global Trait Mapping*** *In addition, while directional errors were not explicitly modeled in this study, analyzing signed residuals could help reveal trait- or vegetation-specific biases. We recognize this as a valuable avenue for future research and recommend that future developments in uncertainty modeling explore the use of signed residuals and the estimation of both lower and upper quantiles."* Second, we emphasize that our approach does not make any assumptions about the symmetry or shape of the residual distribution. In fact, a key motivation behind our method is to provide a conservative and distribution-agnostic estimate of uncertainty. Specifically, we model only the upper bound (95th quantile) of the absolute residuals, focusing on extreme errors rather than assuming any distributional form. This is particularly valuable in contexts where errors are skewed or heavy-tailed, as it avoids the limitations of traditional variance-based methods. In contrast, approaches such as Ens_UN and MCdrop_UN typically estimate uncertainty via the mean ± std, implicitly assuming that the prediction errors follow a symmetric (often Gaussian) distribution. This assumption can lead to underestimation of uncertainty in the presence of asymmetric error distributions, especially in ecological applications where such asymmetries are common. We clarify this point further in the discussion section: ***"4.4 Challenges in comparing and interpreting the uncertainty of state-of-the-art methods*** *Comparing and interpreting the uncertainty estimates produced by different state-of-the-art methods is challenging due to the underlying assumptions of each approach. Traditional methods, such as Ens_UN techniques and MCdrop_UN, often assume Gaussian uncertainty, implying that prediction errors are symmetrically distributed around the mean (i.e., mean ± std) (Hu et al., 2022; Klotz et al., 2022). However, this assumption does not hold true for many plant trait distributions, which are inherently skewed and variable due to ecological and physiological factors across diverse vegetation types—including forests, grasslands, and crops. Models that cannot account for this asymmetry will produce biased or inaccurate uncertainty estimates, as they assume that the data's spread around the mean is similar on both sides. For instance, Klotz et al. (2022) emphasize the importance of accounting for asymmetric distributions in natural data, noting how uncertainty estimates can be improved by modeling heavy tails and skewed data. When plant trait distributions are skewed, their corresponding uncertainty*

		estimates should reflect this asymmetry. Our approach addresses this by not assuming any specific distributional form. Instead, we estimate the upper bound of residuals directly using the 95th quantile of absolute errors, which allows for a distribution-agnostic uncertainty estimate."
6	Dis_UN model and training, and lines 440-445: The performance of this model's training (and test) is not presented; therefore, the reader can't know whether the predictions were expected to be accurate or precise when applied.	We thank the reviewer for this comment. To clarify this point, we will add the evaluation results of the model's predictive performance under leave-one-dataset-out cross-validation (LODO-CV) in the Appendix. The results are presented as scatter plots comparing predicted versus observed trait values (Figure below). 
6a	Despite being more "conservative" than the other methods, Dis_UN predictions are also uncertain. Fig. 3 compares the absolute value of the error with the expected 95 % of their distribution predicted by Dis_UN, but the 68 % (one standard deviation) for the others, which may not be the most appropriate comparison. The statement in lines 440-445 raises the question of whether this comparison is then fair, or whether, comparing the same uncertainty coverages, the difference between the uncertainty predictions would become lower. Perhaps the 95% tail should be calculated and compared for all methods (e.g., multiplying the Ens_UN and MCdrop_UN estimates by 1.96), or the 68-quantile regression (one standard deviation) used for Dis_UN.	We thank the reviewer for this important observation. We agree that comparing different uncertainty estimation methods under varying coverage levels may raise concerns regarding fairness and interpretability. Nonetheless, we acknowledge that the uncertainty estimation methods differ fundamentally in their rationale, quantile regression (Dis_UN) estimates an upper bound on prediction error, while variance-based methods like Ens_UN and MCdrop_UN estimate the spread of model predictions. As a result, their uncertainty values are not directly comparable in magnitude, as they convey different interpretations. However, the intent of Figure 3 is not to compare these absolute values, but to assess how well each method's predicted uncertainty aligns with the actual model residuals, that is, how well-calibrated the uncertainties are in practice. Our focus is on evaluating this alignment rather than calibrating all methods to a common nominal confidence level. While calibration of variance-based methods is an active area of research (e.g., Rahaman et al., 2021; Egele et al., 2022; Bethell et al., 2024), its evaluation is beyond the scope of this study. Nonetheless, we acknowledge its practical relevance: for example, scaling Ens_UN by ~1.96 under Gaussian assumptions improved alignment with residuals, whereas MCdrop_UN did not benefit similarly. These results are included in the Appendix and mentioned in the main text.

Importantly, Fig. 3 presents all methods as they are typically applied in the literature (e.g., Pullanagari et al., 2021; Lang et al., 2022; Palmer et al., 2022; García-Soria et al., 2024), without post hoc adjustments, to reflect current real-world practice.

Moreover, the concept of “coverage” differs between the approaches (Figure below). In quantile regression, coverage is the proportion of residuals below the predicted quantile bound; for example, a 95% quantile model aims for 95% of residuals to fall beneath this bound. In variance-based methods, coverage is the proportion of residuals within a symmetric prediction interval [μ ± zσ] under an assumed parametric distribution (often Gaussian). These are not statistically equivalent: prediction interval coverage reflects calibration of a symmetric probabilistic model, whereas quantile regression coverage measures the accuracy of an asymmetric, one-tailed bound. For this reason, we evaluate each method in its native uncertainty representation rather than forcing them to a common nominal level.



To reduce potential misunderstanding, we have revised the corresponding paragraph in the discussion to clarify our rationale. The revised text (Discussion section) now reads:

“4.1 Local-Scale Uncertainty in OOD Vegetation data

These findings emphasize the importance of carefully selecting and interpreting uncertainty estimation methods. Recalibration of variance-based approaches has been increasingly recommended (JCGM, 2008), and several recent efforts have proposed post hoc methods to better align their uncertainty estimates with observed residuals. For example, we observed that the underestimation in Ens_UN improved when scaling uncertainty by a factor of 1.96 (Fig. S5), although this adjustment did not improve MCdrop_UN. In Fig.3 , however, we present all methods as they are typically applied in the literature (e.g., Pullanagari et al., 2021; Lang et al., 2022; Palmer et al., 2022; García-Soria et al., 2024), without additional adjustments. More broadly, calibration of predictive uncertainty remains an active research area (Rahaman et al., 2021; Egele et al., 2022; Palmer et al., *2022; Bethell et al., 2024; Yang et al., 2024; Zeevi et al., 2024), but the evaluation of such strategies lies beyond the scope of the present study."*

References:

Bethell, D., Gerasimou, S., & Calinescu, R. (2024, March). Robust uncertainty quantification using conformalised Monte Carlo prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 19, pp. 20939-20948).

Zeevi, T., Venkataraman, R., Staib, L. H., & Onofrey, J. A. (2024, May). Monte-carlo frequency dropout for predictive uncertainty estimation in deep learning. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.

Yang, S., & Yee, K. (2024). Towards reliable uncertainty quantification via deep ensemble in multi-output regression task. *Engineering Applications of Artificial Intelligence*, *132*, 107871.

Egele, R., Maulik, R., Raghavan, K., Lusch, B., Guyon, I., & Balaprakash, P. (2022, August). Autodeuq: Automated deep ensemble with uncertainty quantification. In *2022 26th International conference on pattern recognition (ICPR)* (pp. 1908-1914). IEEE.

Rahaman, R. (2021). Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, *34*, 20063-20075.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

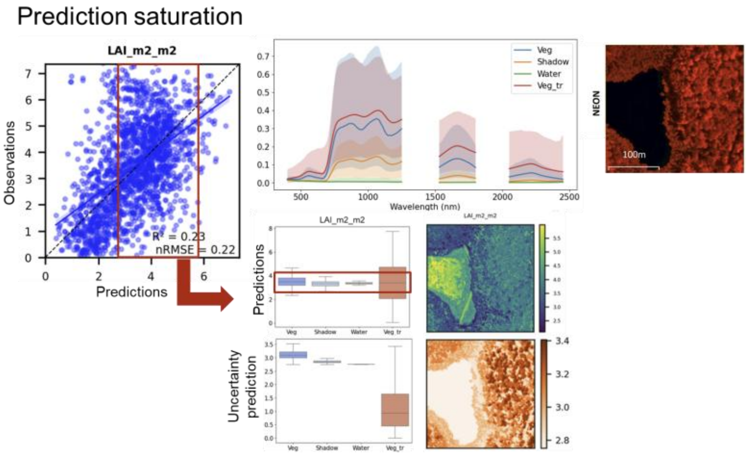
Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., & Wegner, J. D. (2022). Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote sensing of environment*, *268*, 112760

García-Soria, J. L., Morata, M., Berger, K., Pascual-Venteo, A. B., Rivera-Caicedo, J. P., & Verrelst, J. (2024). Evaluating epistemic uncertainty estimation strategies in vegetation trait retrieval using hybrid models and imaging spectroscopy data. *Remote Sensing of Environment*, *310*, 114228.

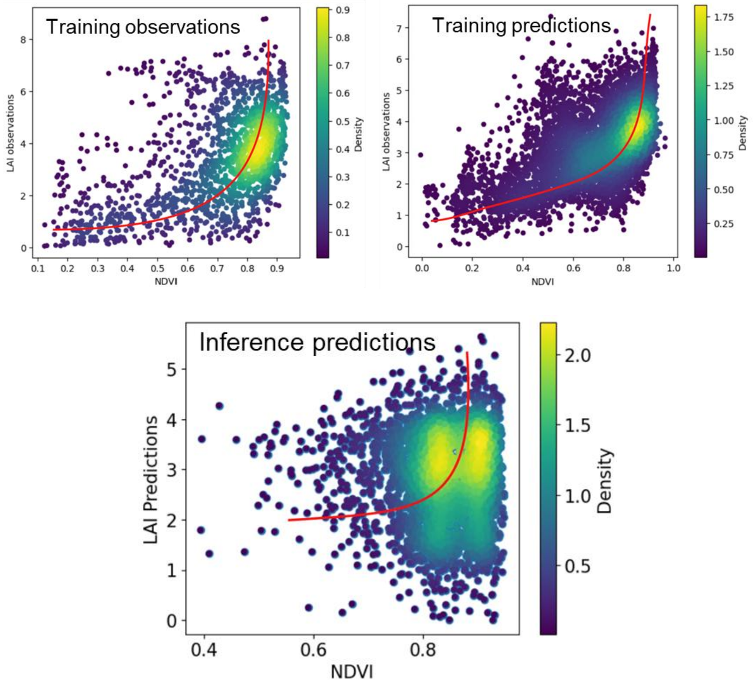
Pullanagari, R. R., Dehghan-Shoar, M., Yule, I. J., & Bhatia, N. (2021). Field spectroscopy of canopy nitrogen concentration in temperate

		grasslands using a convolutional neural network. *Remote Sensing of Environment*, *257*, 112353. Palmer, G., Du, S., Politowicz, A., Emory, J. P., Yang, X., Gautam, A., ... & Morgan, D. (2022). Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Computational Materials*, *8*(1), 115.
7	Lines 431-433: Perhaps "correlation" is not the most representative term for the problem; for example, ens_UN might feature higher Pearson coefficient correlations than the other methods (Fig. 3). The coefficient of determination might neither represent the achievement the authors report, thus, another term should be used instead.	We thank the reviewer for this important observation. We will revise the paragraph as follows: *"In this study, both the Ens_UN and MCdrop_UN methods tended to largely underestimate residuals when applied to OOD vegetation data (on average 26.7% and 6.5% respectively, Table S4 and Fig. S3). The observed low alignment between predicted uncertainty and residuals suggests that the uncertainty estimates produced by these models do not fully represent the model's errors (Fig. 3)."*
8	Lines 456-457: If the embedded space misses some spectral information that might be different between vegetated and non-vegetated surfaces, do these differences matter when the traits are predicted?	We thank the reviewer for this comment. To clarify, the current trait model was optimized exclusively for vegetation trait prediction. Consequently, when confronted with non-vegetated spectra (e.g., water, urban, clouds), the model assigns them the closest values represented in the embedding, however, these outputs are not intended to be meaningful trait estimates. To address this, we could explicitly include non-vegetated spectra in the training process by assigning trait values of zero to spectra from water, rocks, urban materials, or clouds would allow the embedding space to capture these differences more explicitly and yield more reliable predictions in mixed or OOD conditions. While such an approach was beyond the scope of this study, we see it as a promising extension for our trait model.
9	Lines 460-466: Grasslands are not so simple; they can include non-green elements, such as standing senescent material (e.g., Pacheco-Labrador et al, 2021), flowers (Perrone et al, 2024), and pixels usually mix numerous species (Darvishzadeh et al, 2008), which hamper the relationships between spectral and biophysical properties. Phenology during sampling is not	We thank the reviewer for their instructive comment. We agree that the term "simple" in reference to grasslands may have been misunderstood, and we appreciate the helpful suggestions and references provided. In response, we will revise the corresponding paragraph and now explicitly refer to structural homogeneity rather from a radiative transfer modeling perspective than broad "simplicity" that could be indeed confused with ecological or spectral uniformity per se. We also acknowledged the lower BRDF effects in grasslands and added that most in-situ measurements were collected during the green-peak period, which likely minimized background contributions and contributed to lower observed uncertainty.

	reported in the manuscript, but if all the datasets correspond to the green-peak period, grasses will cover most of the soil, and unlike in the shrublands, the background contribution will be minimized. Lower uncertainties might result from bias in the sampling time of the grasslands towards the green peak, if this is indeed the case (which could be confirmed). Unlike forests and shrublands, grasslands exhibit a significantly lower geometrical BRDF component, which may explain the differences between cover types. Forests, in addition to shrublands, will present a more complex vertical profile with a distinct understory of vegetation. There are arguments to justify the findings, but grasslands should not be regarded as "simple". The issue of the phenology bias is, in fact, commented on in the discussion (Section 4.3). References: Darvishzadeh, R., Skidmore, A., Schlerf, M., and Atzberger, C.: Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland, Remote Sensing of Environment, 112, 2592-2604, https://doi.org/10.1016/j.rse.2007.12.003, 2008. Pacheco-Labrador, J., El-Madany, T. S., van der Tol, C., Martin, M. P., Gonzalez-Cascon, R., Perez-Priego, O., Guan, J., Moreno, G., Carrara, A., Reichstein, M., and Migliavacca, M.: senSCOPE: Modeling mixed canopies combining green and brown senesced leaves. Evaluation in a Mediterranean Grassland, Remote Sensing of Environment, 257, 112352,	The revised text (Discussion section) now reads: *"**4.1 Local-Scale Uncertainty in OOD Vegetation data*** *…* *This can be explained by the fact that grassland is one of the more highly represented land cover types in the dataset (1403 of 5573 samples, Table S1) and, from a radiative transfer point of view, it is considered structurally more homogeneous compared to forests and shrubland (Asner, 1998; Ollinger, 2011; Brown et al., 2024). Grasslands typically exhibit lower 3D canopy complexity, and reduced geometric BRDF components, which may reduce spectral variability and residual errors (Jacquemoud et al. 2009). Forests and shrublands are structurally more complex, often containing many scene components beyond green leaves, such as bare ground in canopy gaps, stems, bark, canopy shadow, and other non-photosynthetic components, that contribute to the spectral measurements but are not directly related to the plant traits being measured ."* *references* *Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P. J., Asner, G. P., … & Ustin, S. L. (2009). PROSPECT+ SAIL models: A review of use for vegetation characterization. Remote sensing of environment, 113, S56-S66.*

	https://doi.org/10.1016/j.rse.2021.112352, 2021. Perrone, M., Conti, L., Galland, T., Komárek, J., Lagner, O., Torresani, M., Rossi, C., Carmona, C. P., de Bello, F., Rocchini, D., Moudrý, V., Šímová, P., Bagella, S., and Malavasi, M.: “Flower power”: How flowering affects spectral diversity metrics and their relationship with plant diversity, Ecological Informatics, 81, 102589, https://doi.org/10.1016/j.ecoinf.2024.102589, 2024.	
10	LAI and saturation: The authors argue that the different performance of Dis_UN for LAI is due to saturation; however, in the training datasets, maximum LAI barely reaches 6. I think this hypothesis should be more robustly explored, which may have been done but not presented to the reader. I am not sure saturation can justify the negative coefficients presented in Table S6. Usually, LAI and canopy-scale vegetation variables are easier to retrieve from remote sensing than leaf-level measurements, which should raise some flags. The authors could start by checking how well the DL model performs in predicting LAI compared to the other variables; i.e., the training and test statistics of the model could be presented in the supplementary material. I would expect LAI to be more predictable than foliar traits. If there is saturation, where does it happen? The authors could plot, for example, NDVI (or other index, e.g., NIRv) vs. LAI from their training	We thank the reviewer for this constructive comment. We agree that the behavior of the LAI uncertainty model deserves further clarification, especially in relation to saturation and spectral distance. When examining LAI prediction maps for inference, we observe that the deep learning model tends to predict LAI values close to the median of the training distribution (LAI of ~4), across the land cover components including unseen vegetation, shadow, and water (Figure Figure S11b and figure below).  As shown in a range of studies, LAI estimation is, in comparison to foliar traits, not trivial and particularly for large LAI gradients expected to have severe uncertainty due to saturation effects (Schiefer et al. 2021, Cherif et al. 2023, Mederer et al. 2024) LAI estimation can suffer from saturation, where the model struggles to differentiate between high and very high LAI due to limited spectral sensitivity at those levels, which is a common and unresolved physical problem for remote sensing of LAI (Zheng et al. 2009,

	datasets and explore above which LAI level their dataset saturates. Then, could they check whether the problems for Dis_UN to predict uncertainty occur above the threshold? In the case of non-vegetated surfaces, the DL model might have learnt to predict low LAI for clouds, buildings, or soils even without having seen them. Under this hypothesis, it would also be worth checking whether the estimation uncertainty is low because, indeed, the LAI prediction is accurate. Therefore, the authors may also want to explore the maps of predicted variables to confirm whether the predicted values are reasonable for any of the variables (i.e., LAI being close to 0) for the OOD pixels.	Camps-Valls at al. 2021). Increasing LAI beyond a certain threshold (e.g., LAI > 4–5) does not result in a change of the spectral signal. As a result, the model's predictive distribution becomes compressed at the upper range, which may also lead to residual skewness. This is consistent with the distribution of LAI residuals (median: 0.94, max: 5.88), which shows a right-skewed pattern driven by the model's underestimation of high-LAI cases due to spectral and predictive saturation. In terms of spectral distance, we observe that high-LAI vegetation often appears spectrally similar, even across unseen ecosystems, because dense green canopies produce similar reflectance patterns. In contrast, low-LAI samples exhibit greater spectral variability, due to background effects (e.g., soil, understory, senescent material). This creates a counterintuitive scenario: high spectral distances may correspond to low LAI, and low distances may occur in high-LAI cases, despite the latter being harder to predict. This mismatch explains the inverse relationship between spectral distance (and embedding distance) and LAI uncertainty, as seen in our regression analysis (Table S6). And this is indirectly linked to the reflectance saturation effect with LAI. That said, we will clarify our conclusion related to these challenges and emphasize that our distance-based uncertainty method provides clear advantages over state-of-the-art variance-based approaches like MC Dropout or Deep Ensembles. We find that integrating spectral or latent distance improves the alignment between predicted uncertainty and actual residuals, especially in out-of-distribution settings. We will clarify this point further in the revised discussion section and highlight that the performance of distance-based uncertainty is trait-dependent, with structural traits like LAI posing more challenges due to spectral saturation and narrow value ranges. The corresponding discussion section now reads: "***4.2.2 Uncertainty Patterns Across scene components and Spatial Resolutions*** *…While most of the traits showed a similar spatial pattern in the predicted uncertainties (Fig. 4 and 5), also when compared to the range of uncertainty values of training data samples (Fig. S7 and S10), LAI was distinguishingly different. Traits, such as LMA, EWT and N, exhibit lower uncertainty in areas with dense canopies, where a strong leaf signal is present. This contrasts with LAI, which shows greater uncertainty in dense vegetation, likely due to saturation effects—where increases in leaf area are no longer detectable by the*

		sensor. This saturation issue is common for LAI that have limited sensitivity in dense vegetation conditions (Asner et al., 2003; Mutanga et al., 2023) and is reflected in our training data (Fig. below to be added). Specifically, scatter plots of observed and predicted LAI against NDVI show that while LAI observations continue to increase with NDVI up to ~6, the predicted values plateau around LAI ≈ 4–5 once NDVI exceeds ~0.8. This indicates that the model systematically underestimates high-LAI cases, producing a compressed predictive distribution and a right-skewed residual pattern. This behavior diverges from that of other traits, where uncertainties were typically higher in OOD regions due to substantial deviations between predicted trait values and the training data distributions of the multi-trait model (Fig. S8 and S11). In the case of LAI, high values produce spectrally similar signals across ecosystems, reducing distances in both feature and embedding spaces, while low-LAI samples are more spectrally variable due to background effects (e.g., soil, litter, understory). This explains the negative regression coefficients observed in Table S6 and the unique behavior of LAI uncertainty predictions: higher uncertainties were detected in dense vegetation areas, while OOD pixels such as water, shadow, and urban regions showed lower and less variable uncertainty.”  References: Mederer, D., Feilhauer, H., Cherif, E., Berger, K., Hank, T. B., Kovach, K. R., ... & Kattenborn, T. (2025). Plant trait retrieval from hyperspectral data: Collective efforts in scientific data curation

		outperform simulated data derived from the PROSAIL model. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *15*, 100080. Cherif, E., Feilhauer, H., Berger, K., Dao, P. D., Ewald, M., Hank, T. B., ... & Kattenborn, T. (2023). From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data. *Remote Sensing of Environment*, *292*, 113580. Schiefer, F., Schmidtlein, S., & Kattenborn, T. (2021). The retrieval of plant functional traits from canopy spectra through RTM-inversions and statistical models are both critically affected by plant phenology. Ecological Indicators, 121, 107062 Camps-Valls, G., Campos-Taberner, M., Moreno-Martínez, Á., Walther, S., Duveiller, G., Cescatti, A., ... & Running, S. W. (2021). A unified vegetation index for quantifying the terrestrial biosphere. Science Advances, 7(9), eabc7447. Zheng, G., & Moskal, L. M. (2009). Retrieving leaf area index (LAI) using remote sensing: theories, methods and sensors. Sensors, 9(4), 2719-2745.
11	Uncertainty modeling: The variable-dependence on the capability of Dis_UN to predict the uncertainty might be alleviated by computing the dissimilarity in different spectral regions (E.g., Visible, Red-Edge, NIR, and SWIR). While I do not ask the authors to apply this approach, they might want to consider it in the Outlook section (4.5)	Here, we follow a data-driven approach to predict the uncertainty that incorporates all spectral regions. The uncertainty of each trait is modelled separately. Thus, the data-driven approach will automatically identify the relevant spectral information (represented by the feature space and embedding space). Thus, it is unlikely reducing the spectral information to sub-regions will enhance the performance.
	TECHNICAL CORRECTIONS	
1	Supplementary materials' citations**:** I am unsure whether the journal requires them to be presented in the order of appearance in the main text, but it might make more sense or facilitate the reader's search.	We will change the supplementary citations accordingly.
2	Line 245: For the least familiarized readers, define what a "dropout rate of 0.5" means and maybe justify why this rate is chosen.	*We will elaborate the description as follows:* *" **2.2.1 Monte Carlo Dropout for Uncertainty Estimation (MCdrop_UN)*** *….*

		To quantify uncertainty, multiple forward passes are performed on the input data while keeping dropout active. Each pass generates a different set of neuron activations, effectively simulating different sub-networks. By aggregating these predictions, the mean serves as the final output, while the variability among the predictions (i.e., the standard deviation) reflects the epistemic uncertainty. In our analysis, we calculated the standard deviation of 50 repeated forward passes of the multi-trait model on unseen data with a dropout rate of 0.5 enabled during inference. A dropout rate of 0.5 means that each neuron has a 50% probability of being turned off during a given forward pass. This rate is widely adopted in practice, as it provides a good balance between preserving sufficient network capacity and introducing stochasticity for both regularization and uncertainty quantification (Gal & Ghahramani, 2016; Kendall & Gal, 2017).” *Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 1050-1059.* *Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584.*
3	Line 296: Maybe better: “For clouds delineation we used…”	We will change the text accordingly
4	Table S2, and overall, figures and tables: Provide the units of the variables presented.	We will add the units in the revised version.
5	Table S4: Indicate that the ratio is expressed in percentage.	We will correct the unit in the revised version.
6	Supplementary material: Enhance the presentation and ensure that units and symbols are properly introduced.	We will carefully revise the supplementary material in the revised version.
7	Terminology: Review and homogenize terminology in the main text, tables, and figures. For example, in the paper and figures, the terms “Ens_UN” and “ens_UN” can be found.	We will carefully revise the terminology in the revised version.