Uncertainty Assessment in Deep Learning-based Plant Trait Retrievals from Hyperspectral data

Eya Cherif[1,2], Teja Kattenborn[5,3], Luke A. Brown[6], Michael Ewald[7], Katja Berger[8], Phuong D. Dao[9,10,11], Tobias B. Hank[12], Etienne Laliberté[13], Bing Lu[14], Hannes Feilhauer[1,2,3,4]

[1] Institute for Earth system Science and Remote Sensing, Leipzig University, Leipzig, 04103, Germany
[2] Center for scalable data analytics and artificial intelligence (ScaDS.AI), Leipzig University, 04105, Leipzig, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany
[4] Helmholtz-Centre for Environmental Research (UFZ), 04318, Leipzig, Germany
[5] Sensor-based Geoinformatics (geosense), University of Freiburg, 79116, Freiburg, Germany
[6] School of Science, Engineering & Environment, University of Salford, Manchester, M5 4WT, UK
[7] Institute of Geography and Geoecology, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
[8] GFZ Helmholtz Centre for Geosciences, Potsdam, 14473, Germany
[9] Department of Agricultural Biology, Colorado State University, Fort Collins, CO 80523, USA
[10] Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO 80523, USA
[11] School of Global Environmental Sustainability, Colorado State University, Fort Collins, CO 80523, USA
[12] Department of Geography, Faculty of Geosciences, Ludwig-Maximilians-Universit¨at München (LMU), 80333, Munich, Germany
[13] Département de Sciences Biologiques et Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, H1X 2B2, Canada
[14] Department of Geography, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Correspondence to: Eya Cherif (eya.cherif@informatik.uni-leipzig.de)

Abstract

Large-scale mapping of plant biophysical and biochemical traits is essential for ecological and environmental applications. Given their finer spectral resolution and unprecedented data availability, hyperspectral data, in concert with machine and particularly deep learning models, have emerged as a promising, non-destructive tool for accurately retrieving these traits. However, when deploying these methods on a large scale, reliably quantifying the associated uncertainty remains a critical challenge, especially when models encounter out-of-domain (OOD) data, i.e. samples that differ substantially from those of the training data, such as unseen geographical regions, species, biomes, data acquisition modalities, or scene components (e.g., clouds and water bodies). Traditional uncertainty quantification methods for deep learning models, including deep ensembles (~~Ens_det_UN~~ deterministic and ~~Ens_prob_UN~~probabilistic) and Monte Carlo dropout ~~(MCdrop_UN),~~ rely on the variance of predictions but often fail to capture uncertainty in OOD scenarios, leading to overly optimistic and possibly misleading uncertainty estimates. To address this limitation, we propose a distance-based uncertainty estimation method (Dis_UN) that quantifies prediction uncertainty by measuring dissimilarity in the predictor space (spectral inputs) and embedding space (features learned by the deep model) between the training and test data. Dis_UN leverages residuals as a

proxy for uncertainty and employs dissimilarity indices in data manifolds to estimate worst-case errors via 95-quantile regression. We evaluate Dis_UN using a pretrained deep learning model to predict multiple plant traits from hyperspectral images, analyzing its performance across OOD data, such as pixels containing spectral variations from urban surfaces, bare ground, water, clouds, or open surface waters. In this study, we target six leaf and canopy traits: leaf mass per area (LMA), chlorophylls ~~(Chl)~~, carotenoids ~~(Car)~~, nitrogen ~~(N)~~ content, equivalent water thickness ~~leaf area index (LAI)~~, and leaf area index ~~equivalent water thickness (EWT)~~. ~~The results indicate that Dis_UN effectively differentiates between OOD components and provides more reliable uncertainty estimates than traditional methods, which tend to underestimate the range of uncertainty (on average over traits 44.8% for Ens_prob_UN, 26.7% for Ens_det_UN and 6.5% for MCdropout_UN).~~ Compared to scaled variance-based methods, Dis_UN provides (1) a superior estimation of uncertainty in OOD scenarios, achieving 36% higher contrast (KS distances: 0.648 vs 0.475) between non-vegetation pixels, particularly under mixed-pixel conditions at medium resolution (30m); (2) uncertainty quantification without requiring normality or symmetry assumptions, accommodating asymmetric error patterns; (3) enhanced interpretability of uncertainty sources, as uncertainty is directly linked to sample dissimilarity from the training data; and (4) computational efficiency at inference (2.6-7.7× faster), requiring only a single forward pass compared to multiple passes for ensemble-based methods. ~~However, e~~Challenges remain for traits that are affected by spectral saturation. These findings highlight the advantages of distance-aware uncertainty quantification methods and underscore the necessity of diverse training datasets to minimize sampling biases and enhance model robustness. The proposed framework improves the reliability of uncertainty estimation in vegetation monitoring and offers a promising approach for broader applications.

1 Introduction

Plant functional traits, including structural, biochemical, physiological, and phenological properties, are key to understanding ecosystem structure, function, and resilience (Lavorel and Garnier, 2002; Reich, 2014; Funk et al., 2017). These traits regulate fundamental ecological processes, such as photosynthesis, nutrient cycling, stress response, and productivity (Serbin and Townsend, 2020). The large-scale mapping of plant traits, such as leaf chlorophyll, nitrogen, and water contents, is essential for a range of ecological and environmental applications. These include biodiversity monitoring, Earth system modeling, and vegetation health assessment (Briottet et al., 2022; Cavender-Bares et al., 2020; Houborg et al., 2015; Kissling et al., 2018; Sakschewski et al., 2015; Van Bodegom et al., 2014). However, traditional methods of measuring plant traits via field sampling are resource-intensive, spatially limited, and insufficient to capture global variability. In this respect, hyperspectral data from Earth observation (EO) satellites and airborne sensors have emerged as valuable data sources for predicting functional plant traits (Cavender-Bares et al., 2017; Jetz et al., 2016). These sensors enable the measurement of reflectance across hundreds of narrow and contiguous wavelength bands that are sensitive to subtle biophysical, biochemical, and structural variations within plant canopies (Jacquemoud and Ustin, 2019). With the recent launch of hyperspectral satellite missions, such as Gaofen-5 (GF-5, Ge et al., 2022), Hyperspectral Imaging Satellite (HySIS, Garg et al., 2024), PRecursore IperSpettrale della Missione Applicativa (PRISMA, Cogliati et al., 2021), Environmental Mapping and Analysis

Program (EnMAP, Chabrillat et al., 2024), and upcoming Surface Biology and Geology (SBG, Cawse-Nicholson et al., 2021) and Copernicus Hyperspectral Imaging Mission for the Environment (CHIME, Nieke et al., 2023), the volume of hyperspectral data will provide unprecedented opportunities to map plant traits on a global scale, thus advancing ecosystem monitoring (Asner and Martin, 2016; Briottet et al., 2022; Hank et al., 2019).

Machine learning models, particularly deep learning, have been highly successful in predicting plant traits from hyperspectral data (Cherif et al., 2023; Pullanagari et al., 2021; Serbin et al., 2019; Singh et al., 2015; Wang et al., 2019, 2020). However, when these models are applied to unseen data, for example, from different geographical regions, biomes, with unknown scene components (e.g., clouds or shadows), or new sensors, it becomes crucial to assess the uncertainty of the predicted values. This is particularly important when the predictor space of unseen data deviates from that of the training dataset, resulting in out-of-domain (OOD) observations. In addition, the relationship between predictors and the response variable may vary across different geographical regions or biomes due to spatial structures or non-stationarity. While some efforts have been made to quantify uncertainty in the context of hyperspectral plant trait retrieval (García-Soria et al., 2024; Singh et al., 2015; Wang et al., 2019), the results are often not comparable as the definition and interpretability of the uncertainty estimates vary depending on the methods used. Uncertainty quantification is particularly prevalent in EO for vegetation monitoring, where training data are typically sparse, and models are often applied to new, unseen regions; hence, data that are OOD (Kattenborn et al., 2022; Ploton et al., 2020, Meyer and Pebesma 2021).

In addition to providing crucial information on the quality of OOD predictions, quantitative estimates of uncertainty are increasingly utilized in a range of downstream ecological and environmental applications and are often required by data assimilation schemes in order to appropriately weigh all available observations (Chernetskiy et al., 2017; Lewis et al., 2012; Mathieu and O'Niell, 2008). Furthermore, incorporating trait-level uncertainty in ecological models allows for realistic error propagation, thereby increasing the robustness of simulations related to vegetation dynamics, biodiversity assessments, and Earth system forecasts. For example, in the assessment of land surface phenology, recent studies have explored the propagation of plant trait prediction uncertainties to derived phenological metrics (Graf et al., 2023), enabling a more robust detection of changes in phenophases (e.g., due to the effects of climate change).

Uncertainty estimates are also valuable for identifying underrepresented conditions in the training set. By highlighting regions with high uncertainty, they can inform active learning strategies and guide targeted data acquisition campaigns, ultimately improving model generalization and data representativeness. The increasing importance of uncertainty estimates is reflected in the recent efforts of space agencies and data providers (Brown et al., 2021b; Gorroño et al., 2018, 2017; Goryl et al., 2023), and uncertainties are now a goal of Analysis Ready Data (ARD) standards (Committee on Earth Observation Satellites, 2024, https://ceos.org/ard/). In addition, recent reports from the European Commission (Camia et al., 2024) highlight uncertainty estimations as a specific quantitative requirement for various European Union land-related environmental and agricultural policies (Berger et al. 2025).

Uncertainty in model predictions arises from two primary sources: aleatoric uncertainty, which stems from inherent data variability and measurement noise, and epistemic uncertainty, which reflects the model's lack of knowledge or representation

related to the modeling choice (García-Soria et al. 2024; Lang et al. 2022; Martínez-Ferrer et al.2022). While aleatoric uncertainty is irreducible as it originates from stochastic measurement errors, epistemic uncertainty can be mitigated by incorporating additional data, refining model complexity, or improving feature representation. Popular methods for estimating epistemic uncertainty, particularly for deep learning models, include bootstrapping (Efron and Tibshirani, 1993), Monte Carlo dropout (Gal and Ghahramani, 2016), and deep ensembles (Lakshminarayanan et al., 2017). These methods rely on variations in model predictions to estimate uncertainty. For example, with deep ensembles or bootstrapping, the uncertainty is estimated from the variance of the predictions obtained from multiple models trained with different subsets of the training data. As they are inherently based on model training data, the capacity of such approaches to estimate uncertainty in OOD data is limited. For example, multiple predictions of OOD data obtained from an ensemble approach might all be very biased, even if the variance is low, which would translate into an underestimation of the uncertainty (Gal et al., 2016). Instead of building on the variance in the predictions, uncertainty estimation for EO should focus on the dissimilarity between the training and new data. In other words, if an observation is very different from what the model has learned, it is likely to be very uncertain (Meyer and Pebesma, 2021, Linnenbrink et al., 2024). Therefore, there is a need for an uncertainty estimation approach that accounts for dissimilarities between training and unseen data (Silvan-Cardenas et al., 2008; Khatami et al., 2017; Feilhauer et al., 2021).

Distance-based methods have emerged as promising solutions to address the challenges of uncertainty quantification, particularly in OOD scenarios. Earlier studies have applied similarity-based metrics in the context of classification (Silvan-Cardenas et al., 2008; Khatami et al., 2017; Feilhauer et al., 2021). These approaches remain tied to discrete, categorical problems and shallow empirical models. More recently, distance-based methods have been extended to regression and spatial prediction tasks. For instance, Janet et al. (2019) proposed a low-cost uncertainty metric for predictions of chemical properties of unknown substances/materials based on the distance of new inputs from the training data in latent space, outperforming traditional uncertainty metrics such as Monte Carlo dropout and ensembles, particularly for data points far from the training set. Meyer and Pebesma (2021) discussed the importance of defining an "area of applicability" for spatial models, emphasizing the use of dissimilarity metrics to assess model confidence when dealing with new data. Building on this, Papacharalampous et al. (2024) and Linnenbrink et al. (2024) illustrated the effectiveness of distance-based metrics in enhancing uncertainty quantification and improving the reliability of spatial predictions.

In the context of plant trait predictions from hyperspectral data, distance-based uncertainty metrics could offer similar advantages. Hence, this study aims to develop and evaluate a distance-based method for quantifying the uncertainty of deep learning models used for plant trait retrievals. Specifically, we propose a distance-based method, hereafter referred to as Dis_UN, to evaluate a previously established multi-trait model when used for inference on OOD data (Cherif et al., 2023). Our approach uses residuals as proxies for total uncertainty and employs dissimilarity indices in data manifolds to predict uncertainty. By performing 95-quantile regression, we estimate the upper-bound limit of residuals. Specifically, with an established multi-trait model for trait predictions, we aim to:

a) evaluate the efficacy of this method in quantifying uncertainty for OOD vegetation samples at the local scale for six leaf and canopy traits: leaf mass per area (LMA), chlorophyll (Chl), carotenoids (Car), nitrogen (N) content, equivalent water thickness (EWT)~~leaf area index (LAI)~~, and ~~leaf area index (LAI)equivalent water thickness (EWT)~~. To achieve this, we iterate over 50 datasets, each time training a multi-trait model that serves as the basis for generating and evaluating our distance-based uncertainty estimates.

b) demonstrate the method's potential at the landscape scale with hyperspectral scenes, investigating how the model performs when confronted with OOD observations, such as pixels containing spectral variation from urban surfaces, bare ground, water, clouds, or open surface waters.

2 Methods

2.1 Distance-aware U~~u~~ncertainty Q~~q~~uantification (Dis_UN)

We introduce a distance-based approach to estimate the absolute residuals of a deep learning model (referred to as uncertainty) based on the degree of dissimilarity between the training data and new, unseen samples (Fig. 1). We acknowledge that error is not exactly the same as uncertainty in the probabilistic sense but can be a significant contributing component (JCGM, 2008; Widlowski, 2015). However, for the purposes of this analysis, we adopt this definition as we approximate the upper bound of the residuals. To accomplish this, we adopt~~ed~~ the 'dissimilarity index' (DI) originally presented by Meyer and Pebesma (2021). Our proposed methodology builds on data splits of a cross-validation obtained from evaluating the deep learning model. Specifically, we distinguished between two subsets: the training sets, which included the data used to train the deep learning model, and the test sets, which comprised unseen data (ODD) used for evaluation (Fig. 1). These splits enable the systematic quantification of uncertainty by comparing unseen samples with training data through several manifolds. Dis_UN was subsequently modelled using 95-quantile regression with the calculated DIs serving as predictors. We applied this method to hyperspectral imagery to evaluate the estimated uncertainties for the OOD data. To assess the performance of Dis_UN at the local-scale on OOD vegetation data, we used a one hold-out set. For landscape-scale OOD data, we assess~~ed~~ if the uncertainty is elevated on unseen or unrelated scene components compared to the uncertainty of vegetated areas, such as clouds, shadows, urban areas, or waterbodies. In this setting, we benchmark~~ed~~ our method against two state-of-the-art approaches: Monte Carlo dropout (MCdrop_UN) and deep ensemble methods.

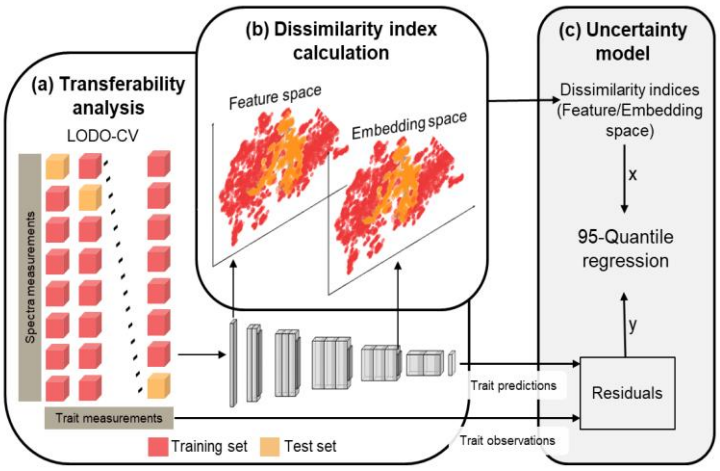


Figure 1: Overview~~Workflow~~ of the distance-based uncertainty method (Dis_UN) for assessing the uncertainty of a deep learning model~~. The method consists of three phases~~, including: (a) Leave-one-dataset-out cross-validation (LODO-CV) on the deep learning model, (b) Training data generation for uncertainty estimation using the LODO-CV, and (c) uncertainty modeling. The uncertainty modelling~~, 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.

2.1.1 The Multi-trait Model

The deep learning model evaluated in our study was built upon a Convolutional Neural Network (CNN) originally proposed by Cherif et al. (2023). This model is based on the EfficientNet-B0 architecture with a customization that optimizes the model for one-dimensional spectral data, making it well-suited for predicting multiple plant traits from hyperspectral reflectance. The structure allows the model to capture both localized spectral features and broader spectral patterns, enabling it to learn the relationships between spectral signals and plant traits.

The dataset used for this model is a curation of multiple datasets, incorporating spectra and trait observations from diverse ecosystems, including forests, grasslands, shrublands, and agricultural regions (Tables S1 and S2). Reflectance data spanning wavelengths from 400 to 2500 nm were collected using various 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 S1). This curation resulted in a sparse dataset with limited trait observations and an imbalanced number of samples across the original datasets (Table S3). In line with Cherif et al. (2023) (S1), all datasets were resampled to a common 1 nm ~~resolution~~ spectral step across the 400–2500 nm range to harmonize diverse measurements. We chose to upsample rather than downsample as most datasets were originally acquired at a 1 nm spectral sampling interval~~resolution~~, thereby minimizing data manipulation. ~~–~~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 consistently applied across all datasets to ensure comparability within the curated collection. A total of 1522 interpolated spectral bands were retained for analysis. The corresponding trait observations encompassed both biochemical (such as N and pigment contents) and structural traits (e.g., LMA and LAI), chosen to represent a diverse range of plant functions. For this analysis, we focused on six leaf and canopy traits: LMA, Chl, Car, N content, EWT~~LAI~~, and ~~–~~LAI ~~EWT~~. This heterogeneous training set was intended to ensure broader ecological and environmental representativeness, thereby enhancing the model's generalizability. Yet, the collected data do not provide a fully global representation due to the labor-intensive nature of data collection and the inherent bias in available data measurements (Table S3). To reduce the over-representation of large datasets during the training of the multi-trait CNN, we (i) performed random upsampling with replacement so that each source dataset contributed approximately equally per training epoch, and (ii) applied per-sample loss weights inversely proportional to the number of labeled samples in the corresponding source dataset, which down-weighted over-represented datasets and up-weighted under-represented ones (Cherif et al. 2023).

The feature space of the CNN model is constructed from the full spectral range of the reflectance data. This feature space allows the model to distinguish nuanced spectral differences, effectively mapping them to plant traits. Additionally, an embedding space is generated from the final convolutional layers of the model, producing a high-dimensional, condensed representation of the spectral data (Fig. S1). This embedding effectively distills essential spectral patterns while enhancing the model's ability to identify trait relationships.

The model was evaluated using a leave-one-dataset-out cross-validation (LODO-CV), ensuring that each dataset (n = 50) was held out once as a test set. This cross-validation facilitated the assessment of the model performance across different vegetation types and spectral configurations. Using the LODO-CV and the obtained models' residuals of the trait predictions, training samples were created to estimate the uncertainty in the multi-trait model. This dataset included 1) the feature space, that is the hyperspectral data, 2) the embedding space of the multi-trait model, 3) the trait predictions obtained from the deep learning models and 4) the true trait observations (Fig.1).

2.1.2 Dissimilarity ~~I~~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 the spectral shape while minimizing the influence of amplitude variations that occur uniformly across the spectrum. This helps to mitigate the brightness changes caused by heterogeneous illumination and internal shading (Feilhauer et al. 2010).

Formally, the cosine distance between a test spectrum x_i and a training spectrum z_i is defined as:

$$CosineDist(x_i, z_j) = 1 - \frac{x_i . z_j}{||x_i|| . ||z_j||} \quad (1)$$

This DI was calculated in both the feature and the embedding spaces of the models (Fig. S3). As a first step, we calculated the cosine distances between each sample of the test dataset x_i and the samples of the training dataset z_i. 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:

$$DI_i = median\{CosineDist(x_i, z_j)\}_{j=1}^{50} \quad (2)$$

We chose 50 neighbors as a compromise between preserving local similarity and avoiding excessive signal dilution, given the relatively small size of the training dataset (~7000 samples). Smaller values would lead to very fine-grained distance distributions that are overly sensitive to individual training outliers, while larger values progressively dilute the local dissimilarity signal by incorporating increasingly dissimilar samples.

To ensure comparability across samples, the indices were normalized against the mean DI value of the entire training set (Meyer and Pebesma, 2021):

$$DI_i^{norm} = \frac{DI_i}{\mu_{train}}, \text{ with } \mu_{train} = \frac{1}{n}\sum_{j=1}^{n} DI_i \text{ where n is the number of training samples} \quad (3)$$

As reference data for the uncertainty models (Dis_UN), the residuals were used as a proxy for the uncertainty for each trait. This corresponds to the difference between the reference trait data and the trait predictions from the multi-trait models.

The two indices, representing dissimilarities in the feature and embedding spaces, and the absolute residuals were then used as input data for the supervised uncertainty estimation (Dis_UN).

2.1.3 Dis_UN Model Training

We developed a distinct Dis_UN model for each plant trait, utilizing the DIs calculated for each sample of the test datasets within the LODO-CV as predictors. We partitioned the dataset into training and validation subsets using an 80/20 hold-out split for evaluating the model performance. The Dis_UN models were trained using 95-quantile regression, a statistical technique that allows for the estimation of the 95th percentile of the target distribution (Koenker and Hallock, 2001). This approach is particularly advantageous for uncertainty quantification as it focuses on the upper tail of the error distribution,

providing insight into how large the errors can be in the worst-case scenarios (up to 5% of the cases, JCGM 2008). Quantile regression better captures variability because it directly models specific points in the distribution of the target variable, allowing it to represent tail behaviors that mean-based approaches inherently overlook or smooth out. To further support the choice of the 95th quantile, we conducted a sensitivity analysis across a range of quantiles (τ between 75 and 99) for all traits (Fig. S2). The results showed that $\tau = 0.95$ provides a good balance between capturing a high proportion of large errors and avoiding overly wide and unstable uncertainty bounds. Lower quantiles ($\tau \leq 0.93$) tended to underestimate the extent of potential errors, missing a fraction of extreme cases, while higher quantiles ($\tau \geq 0.97$) led to unnecessarily conservative bounds that can fluctuate sharply. This balance makes the 95th quantile a robust choice for representing worst-case uncertainty across variables, avoiding both underestimation and over-conservatism. Additionally, the empirical coverage analysis (Fig. S2, top panels) provides a direct validation of Dis_UN's calibration, demonstrating that the 95th percentile predictions achieve their target coverage of approximately 95% across all traits. The fit criterion for quantile regression is the pinball loss function, which is specifically designed to penalize deviations from the predicted quantile (Koenker and Hallock, 2001).

We addressed potential imbalances in the distribution of the target variable using a histogram-based weighting scheme. We divided the target variable (residuals for each trait) into five bins according to its distribution. Samples in less populated bins, which typically correspond to more extreme or uncommon uncertainty values, were assigned higher weights. These weights were calculated as the inverse of the proportion of samples within each bin, effectively emphasizing the importance of accurately predicting higher levels of uncertainty. This weighting scheme was integral to the training process, as it enhanced the model's sensitivity to the tails of the distribution, where uncertainty is usually highest. Additionally, we applied data transformations on the training data (calculated DIs) as the dissimilarity indices exhibit long-tailed distributions and varying value ranges (Fig. S3). Specifically, a Box-Cox transformation was applied to the predictor variables followed by a standardization to normalize their distribution and reduce skewness.

2.2 State-of-the-Art Uncertainty Estimation Methods

Monte Carlo Dropout and deep ensembles are considered the gold standard methods for quantifying the epistemic uncertainty (Abdar et al. 2021, Liu et al. 2023). Below, we outline these approaches.

2.2.1 Monte Carlo Dropout for Uncertainty Estimation (MCdrop_UN)

Monte Carlo dropout is a state-of-the-art method that approximates the posterior distribution in Bayesian deep learning (Gal and Ghahramani, 2016). The core concept of MCdrop_UN is to use dropout not only during the training phase but also during inference, thus enabling the estimation of model uncertainty in a simple, probabilistic manner without requiring significant architectural modifications. Originally, dropout was introduced as a regularization method and was implemented by randomly deactivating a fraction of neurons during training (Srivastava et al. 2014). In MCdrop_UN, this randomness is extended to inference, allowing the creation of an ensemble of predictions without having multiple models.

To quantify the 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 forward pass. This rate is widely adopted in practice because it provides a good balance between preserving sufficient network capacity and introducing stochasticity for both regularization and uncertainty quantification (Kendall and Gal, 2017; Gal and Ghahramani, 2016).

2.2.2 Deep Ensembles for Uncertainty Estimation (Ens_UN)

Deep ensembles represent another approach for uncertainty estimation in deep learning (Deng et al., 2022). Two main variations of deep ensembles exist in the literature (Ashukha et al. 2020; Lakshminarayanan et al., 2017). We here refer to them as deterministic deep ensembles Ens_det_UN and probabilistic deep ensembles Ens_prob_UN. In our experiments, we employed both Ens_det_UN and Ens_prob_UN to provide a comprehensive comparison.

Deterministic Deep Ensembles for Uncertainty Estimation (Ens_det_UN)

Deterministic ensemble leverages the power of multiple, independently trained models to achieve more reliable predictions and uncertainty quantification. The key idea behind this method is to train several models from scratch, each with random initialization and potentially using different subsets of data, often achieved through bootstrapping. By using different random seeds for weight initialization, each model takes a unique path through the parameter space, ensuring a diverse collection of models. During inference, all the models in the ensemble make predictions for the same input data. The final prediction is calculated by averaging the outputs of all ensemble members, while the variance among these predictions provides an estimate of the epistemic uncertainty. For our analysis, we utilized the mean and standard deviation of the predictions from 50 independently trained models, which were established during the LODO-CV process of the multi-trait model, as described in Section 2.1.1.

Probabilistic Deep Ensembles for Uncertainty Estimation (Ens_prob_UN)

Probabilistic ensembles extend this framework by having each model predict not only a mean value but also an associated variance by minimizing the negative log-likelihood loss (NLL) (Lakshminarayanan et al., 2017; Lang et al., 2022). In this setting, each model outputs the parameters of a probability distribution, where $\mu_k(x)$ and $\sigma_k^2(x)$ are the predicted mean and variance of the k^{th} ensemble member, respectively. For an ensemble of K models, the total predictive uncertainty is expressed as follows:

$$Var_{tot} = \frac{1}{M}\sum_{k=1}^{M}\sigma_k^2 + \frac{1}{M}\sum_{k=1}^{M}\mu_k^2 - \left(\frac{1}{K}\sum_{k=1}^{M}\mu_k\right)^2 \quad (4)$$

Formatted: Font: (Default) +Headings (Times New Roman), Subscript

Formatted: Font: (Default) +Headings (Times New Roman), Subscript

Formatted: Font: (Default) +Headings (Times New Roman), Superscript

For our analysis, we also utilized the mean and standard deviation of the predictions from 50 independently trained, probabilistic models.

2.3 Evaluation

2.3.1 Evaluation Using Hyperspectral Imagery (HSI): EnMAP and NEON

To evaluate the different uncertainty estimation methods under extreme OOD conditions, we selected two distinct hyperspectral datasets, each sourced from a different sensor platform: the EnMAP satellite and the National Ecological Observatory Network (NEON) airborne observation platform (AOP), each differing in spatial resolution and environmental context. The first dataset is sourced from the Environmental Mapping and Analysis Program (EnMAP, Chabrillat et al., 2024). The EnMAP mission carries a hyperspectral instrument on a satellite platform and has been designed as a scientific precursor for future more operational spectroscopy missions. EnMAP captures spectral data across the visible, near-infrared (VNIR), and short-wave infrared (SWIR), covering wavelengths from 420 to 2450 nm and comprising 224 spectral bands in total. This dataset offers a spatial resolution of 30 m and a swath width of 30 km, enabling analysis of medium-scale land cover features. The dataset includes a clip of a scene of the city of Leipzig, captured on June 27, 2022 (Fig. 2). This product was downloaded in L2A format (orthorectified and atmospherically corrected, De Los Reyes et al., 2023) from the EnMAP portal (https://planning.enmap.org/). The second dataset comes from NEON's AOP, which provides high spatial and spectral resolution hyperspectral imagery collected annually over various ecological monitoring sites across the United States (Kampe et al., 2010). The NEON AOP spectrometer covers a spectral range from 380 to 2510 nm, resulting in approximately 426 spectral bands with 6 nm spectral sampling (Kampe et al., 2010; Wang et al., 2020). The NEON dataset features a spatial resolution of about 1 m, which allows for the detailed analysis of fine-scale land cover features. This image is a clipped section of a scene at Little Rock Lake (LIRO) from NEON's Great Lakes Domain (D05) captured on August 16th 2020, and processed as a Level 3 (L3): Mosaicked and orthorectified Surface Directional Reflectance (Fig. 2, see details in data.neonscience.org/api/v0/documents/NEON.DOC.001288vB and data.neonscience.org/api/v0/documents/NEON.DOC.004365vB). This product was downloaded from the NEON portal (https://data.neonscience.org/data-products/DP3.30006.001). Both datasets include a mix of scene components, such as vegetation, urban areas, water bodies, clouds, and shadows, leading to pixels with mixed signals from different elements. This diversity was intentionally selected to evaluate the robustness of uncertainty estimation methods in detecting OOD samples.

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

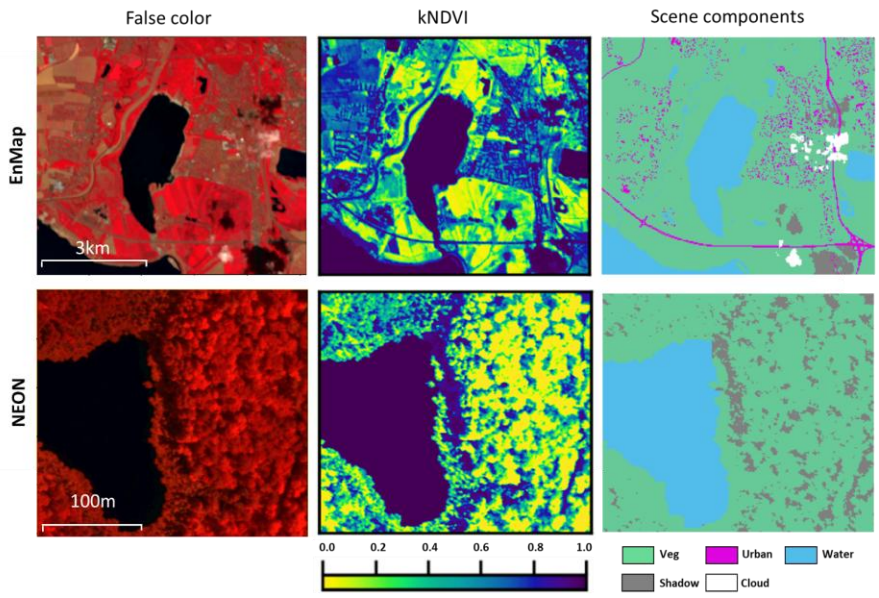


Figure 2: Two study cases for uncertainty inference. On the ~~right~~left, a false color representations of a spaceborne scene from EnMAP with 30 m spatial resolution (South of Leipzig, Germany) and an airborne scene from the NEON observatory with 1 m spatial resolution (Little Rock Lake (LIRO) from NEON's Great Lakes Domain (D05), US). In the middle is the kernel normalized difference vegetation index (kNDVI, Camps-Valls et al.; 2021) and on the right side a map of the scene components.

2.3.2 Scene Components' Identification

To identify the different components within each scene, we utilized a combination of data sources: OpenStreetMap (OSM: https://download.geofabrik.de/) and manual labeling for certain features, such as clouds, cloud shadows, and tree shadows. The land cover data for urban areas and water bodies were primarily derived from OSM. The maps were resampled to match the spatial resolution of the hyperspectral images. The urban areas included only buildings and highway features. Manual labeling was employed for features like clouds, cloud shadows, and tree shadows. Particularly for tree shadows of the NEON scene, we delineated cloud and tree shadows with a carefully selected threshold of the NIR band (~824 nm). For cloud detection and delineation, we used an aggregation band of RED (~650 nm) and BLUE (~444 nm). The final scene component maps were generated by integrating the OSM data with the manually labeled elements (Fig. 2).

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

2.3.3 Uncertainty Evaluation Metrics

To evaluate the performance of our uncertainty estimation methods, we employed different metrics for OOD on local and landscape scale analyses. These metrics quantify how well each model's predicted uncertainties align with actual residuals, as well as the model's ability to identify OOD samples based on elevated uncertainty predictions when compared to vegetated areas. For the local scale OOD analysis, we used the Expected Normalized Calibration Error (ENCE, eq1, Levi et al., 2022; Scalia et al., 2020) as a primary metric to assess how well predicted uncertainties aligned with the actual residuals of the deep learning model predictions. Lower ENCE values signify better model calibration, whereas higher values indicate poorer alignment between predictions and actual residuals, meaning that the model's uncertainty predictions do not accurately reflect the observed errors. The ENCE was computed as Eq. (5):

$$\text{ENCE} = \frac{1}{K}\sum_{i=1}^{K}\frac{|RMSE(b_i) - UE(b_i)|}{UE(b_i)} \quad (5)$$

where K is the total number of bins, $RMSE(b_i)$ is the root mean squared in bin i and $UE(b_i)$ is the uncertainty estimation in bin i.

To evaluate the extent to which the predicted uncertainty range matches the observed range for each trait, we also computed a quantile-based ratio (QuRatio, Eq. (2)). This metric quantifies the proportion of the actual uncertainty range that the model's predictions cover and is calculated as Eq. (5):

$$\text{QuRatio}(\%) = \frac{\text{qu95(pred\_UN)} - \text{qu5(pred\_UN)}}{\text{qu95(obs\_UN)} - \text{qu5(obs\_UN)}}, \quad (6)$$

where *qu*95 and *qu*5 represent the 95th and 5th quantiles, respectively, of the predicted uncertainty values (*pred_UN*) and observed uncertainty values (*obs_UN*).

For the OOD analysis, we aimed to evaluate the model's ability to detect OOD samples by predicting elevated uncertainty values. To quantify this, we used two statistical metrics: the Kolmogorov–Smirnov (K-S) statistical test and the related KS~~Jeffries-Matusita (JM)~~ distance. These metrics measure the degree of separation between the distributions of predicted uncertainties for different scene components, specifically between vegetated and non-vegetated pixels (~~Richards et al., 2022;~~ Wacker and Landgrebe, 1972). For each OOD component, we balanced the sample count by sampling to match the least represented component within each category (vegetated and non-vegetated). This approach ensures a fair comparison, as different categories are not equally represented in a scene. We then calculated and compared the distribution of predicted uncertainties across these samples. The KS distance, ranging from 0 to 1, quantifies the maximum difference between the empirical cumulative distribution functions of the two distributions, with 0 indicating identical distributions and larger values indicating stronger separation.~~The JM distance, ranging from 0 to 2, quantifies the separation between distributions, with 0 indicating perfect overlap and 2 representing perfect separation~~. Higher ~~JM~~ KS values thus reflect a stronger ability of the model to distinguish OOD samples through uncertainty predictions.

Formatted: Font: (Default) +Headings (Times New Roman), Subscript

Formatted: Font: (Default) +Headings (Times New Roman), Subscript

3 Results

3.1 Uncertainty for OOD Vegetation data

To ensure a consistent comparison between Dis_UN and the variance-based approaches, we scaled the ensemble- and dropout-based uncertainties to the 95% confidence interval, i.e 1.96 × σ (Fig. 3). Results with the default scale, i.e 1 × σ as commonly used, are presented in Fig. S6 in the Supplementary. The MCdrop_UN method showed the strongest underestimation of residuals, independent if the uncertainty estimate was scaled or not and only covered 3.1% to 7.9% (QuRatio) of the observed residual variability (Fig. 3, Table. S5). This reflects a very narrow predicted interval that does not correspond well to the actual errors. The ENCE values, used to quantify the uncertainty prediction calibration, were the highest among all methods (13.1–20.8), indicating a weak relation between actual residuals and predicted uncertainty (Table S4).

The two ensemble-based methods exhibited better calibration only after scaling the predicted uncertainty (1.96 × σ). Ens_det_UN achieved the closest alignment with observed residuals among the variance-based methods, with ENCE values between 0.12 and 0.65 and ~~coverage (~~QuRatio~~)~~ between 43.2% to 60.1%. Note that without scaling, which is the common approach in literature, the ensemble approaches greatly underestimate the range of the observed residuals (Fig. S6).

~~Both the Ens_det_UN and MCdrop_UN methods demonstrated substantial underestimation of predicted residuals (5th to 95th quantile range) across all traits (Fig. 3). For the Ens_det_UN method, the predicted uncertainty for LMA covered only 29% of the observed range in residuals. This underestimation was even more pronounced for N, where only 26% was captured, and for Chl and Car, where just 22–25% of the observed range was captured. This pattern highlights a limited ability of Ens_det_UN to reflect true variability in uncertainty across traits. The ENCE values for Ens_det_UN varied between 1.4 and 2.2, indicating inconsistent calibration across traits (Table S4).~~

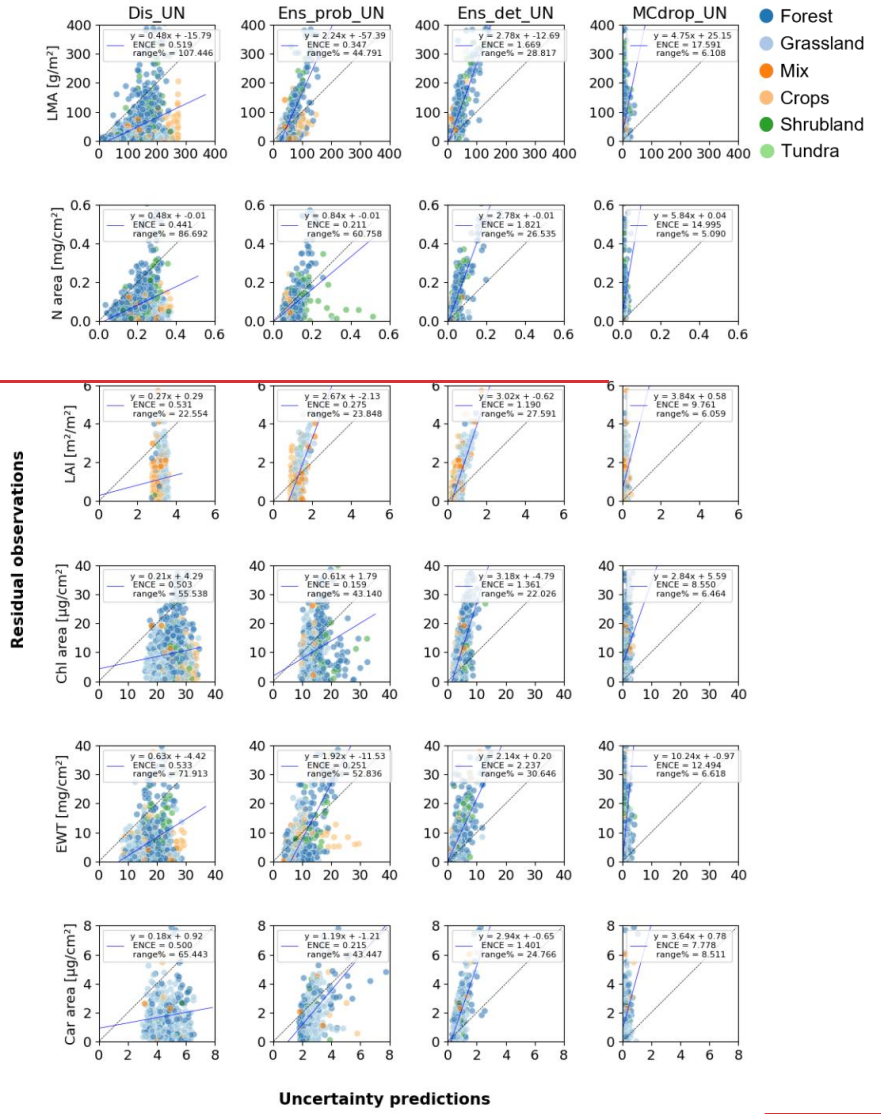
~~The MCdrop_UN method exhibited an even greater underestimation of residuals, with predicted uncertainties covering only 5% to 6% of the observed range (Fig. 3), suggesting a particularly narrow predicted interval that does not align well with actual residuals. Correspondingly, MCdrop_UN's ENCE values ranged from 7.8 to 17.6, indicating a substantial misalignment between predicted uncertainties and observed residuals (Table S4).~~

~~The Ens_prob_UN showed a better alignment with the residuals than the other two state-of-the-art methods with ENCE values between 0.16 and 0.35. However, it still shows the same tendency of underestimation in the trait coverage range (between 24 and 60%).~~

~~In contrast, t~~The distance-based method, Dis_UN, provided a broader ~~and more balanced~~ range of uncertainty estimates. For traits like LMA and N, Dis_UN produced uncertainties that exceeded the observed range of residuals (QuRatio 107.4%, 86.7%; Table S5), indicating a more robust approach of the upper-bound error estimation. For other traits, such as Chl, LAI, EWT, and Car, Dis_UN captured 22.6% to 77.9% of the observed range (Table S5), suggesting that it achieves a more reliable and well-calibrated uncertainty estimate. Dis_UN's ENCE values ranged from 0.44 to 0.53 (Table S4), with data points clustering below the 1:1 line across all traits. This pattern indicates a tendency to overestimate uncertainties, ensuring a conservative uncertainty estimation in which the predicted uncertainties encompass the residuals.

Across the different vegetation types represented in the training samples, grassland pixels consistently displayed lower uncertainty values, particularly with Dis_UN, while more heterogeneous vegetation types, such as shrubland and forest, exhibited a broader spread in uncertainty (Fig. 3).

Across different plant traits, we observed different relationships between modelled uncertainty and the spectral and embedding distances, respectively, as indicated by the regression coefficients of the 95-quantile regression. The standardized spectral distance consistently shows higher coefficients compared to embedding space distance across all traits (Table S6). For instance, the spectral distance coefficient for LMA is 53.74, significantly higher than the embedding distance coefficient of 7.07. Similarly, for EWT, the spectral distance coefficient is 4.36, while the embedding distance contributes only 0.86. This trend is also observed across traits like Carotenoid content, where spectral distance (0.81) outweighs embedding distance (0.10). These differences suggest that spectral distance provides a stronger signal in the regression model than the distances in the embedding space.


Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN
Forest
Grassland
Mix
Crops
Shrubland
Tundra
LMA [g/m²]
N area [mg/cm²]
LAI [m²/m²]
Chl area [µg/cm²]
EWT [mg/cm²]
Car area [µg/cm²]
Residual observations
Uncertainty predictions

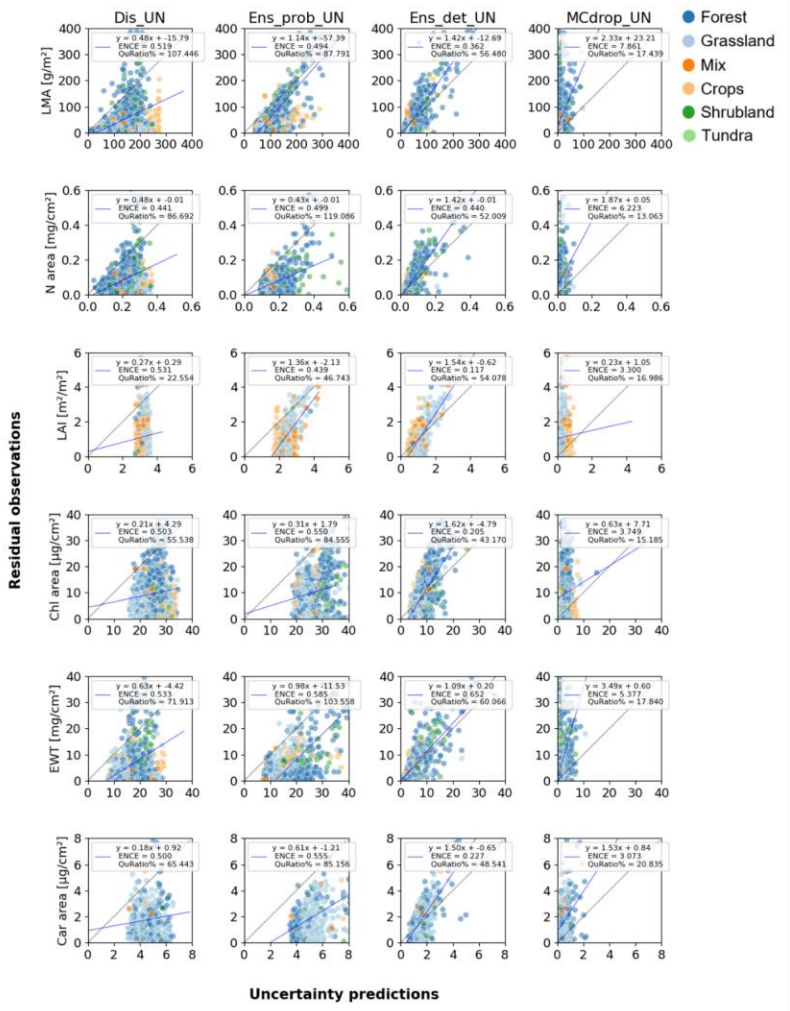


Figure 3: ~~Scatter plots comparing the predicted uncertainties (x-axis) from four methods—distance-based (Dis_UN), probabilistic ensemble (Ens_prob_UN), deterministic ensemble (Ens_det_UN), and Monte Carlo dropout (MCdrop_UN)—against observed residuals (y-axis) of the multi-trait models across six traits: Leaf Mass per Area (LMA), Nitrogen content, Chlorophyll content~~

~~(Chl), Equivalent Water Thickness (EWT), Leaf Area Index (LAI), and Carotenoid content (Car). Each point represents a sample, colored by vegetation type. For each trait–method combination, the regression line (blue) is compared to the 1:1 line (black) to visualize alignment between predicted and observed errors. Model calibration is quantified by the Expected Normalized Calibration Error (ENCE), while the ratio of predicted to observed uncertainty ranges indicates the coverage of residual variability (see also Table S4 and S5).~~ Scatter plots comparing the predicted uncertainties (x-axis) from four methods— Distance-based (Dis_UN), probabilistic ensemble (Ens_prob_UN), deterministic ensemble (Ens_det_UN) and Monte Carlo dropout (MCdrop_UN) calibrated by a factor of 1.96 × standard deviation—against observed residuals (y-axis) of the multi-trait models across six traits: Leaf Mass per Area (LMA), Nitrogen content, Chlorophyll content (Chl), Equivalent Water Thickness (EWT), Leaf Area Index (LAI), and Carotenoid content (Car). Each point represents a sample, colored by vegetation type. For each trait–method combination, the regression line (blue) is compared to the 1:1 line (black) to visualize alignment between predicted and observed errors. Model calibration is quantified by the Expected Normalized Calibration Error (ENCE), while the ratio of predicted to observed uncertainty ranges (QuRatio) indicates the coverage of residual variability (see also Table S4 and S5). Note that the default definition of uncertainty with the Ensemble and Drop out approaches did systematically underestimate the observed residuals (Fig. S6).

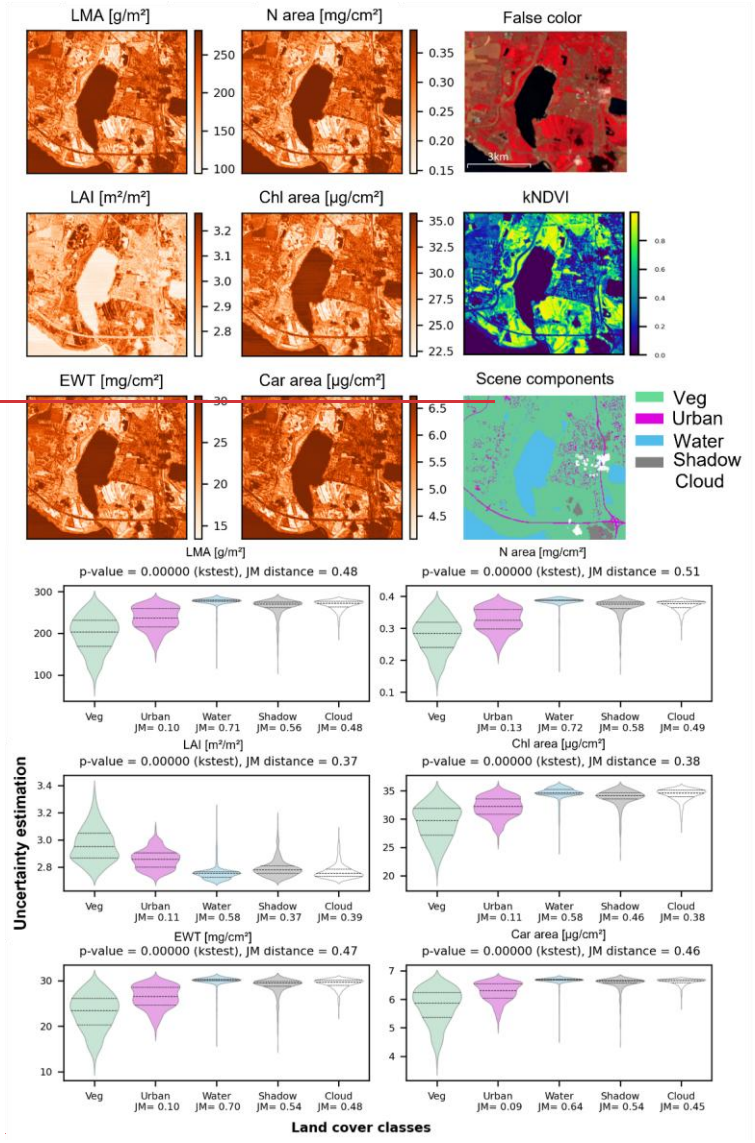
3.2 Uncertainty for OOD data

Predicted uncertainty values from the Dis_UN model consistenly differed~~were consistent~~ across different scene components, with vegetation components showing lower uncertainty and OOD components, such as clouds or water bodies, exhibiting higher uncertainty (Figs. 4 and 5).

In the EnMAP scene (Fig. 4), notable spatial variations were observed in the predicted uncertainties from the Dis_UN method across the different traits, particularly for non-vegetative pixels, including shadows, clouds, and waterbodies. These components showed the largest uncertainty, as reflected by their respective KS ~~JM~~ distances. Uncertainty predictions for traits such as LMA, N and EWT as well as pigments followed similar patterns, with clear differentiation between vegetation and non-vegetation pixels, achieving a higher KS ~~JM~~ distance of 0.~~51~~66.

Among the traits, LAI showed distinct uncertainty patterns. The coefficient of variations (CV) of the predicted uncertainty values varied between 11.34 and 23.27% for all the traits except for LAI showing a uniform predicted values of CV 4.79%. Despite the low variation, LAI exhibited relatively lower uncertainty for water, shadow and cloud pixels and elevated uncertainties in vegetated areas. Across all traits, uncertainty value distributions from OOD samples (cloud, urban and shadow) were significantly different when compared to vegetated samples (KS test, $p < 0.005$ for all traits).

Similar results were obtained for the NEON scene (Fig. 5), with predicted uncertainties consistently lower for vegetation components and higher for OOD scene components like tree shadows and water. The KS ~~JM~~ distances were higher than those of the EnMAP scene ranging between 0.~~49~~ 77 to 0.~~62~~ 83 among traits. Tree-covered areas (forest) exhibited greater variance with bimodal distribution, while water pixels had the highest uncertainty levels.


LMA [g/m²]
N area [mg/cm²]
False color
3km
LAI [m²/m²]
Chl area [µg/cm²]
kNDVI
EWT [mg/cm²]
Car area [µg/cm²]
Scene components
Veg
Urban
Water
Shadow
Cloud
LMA [g/m²]
p-value = 0.00000 (kstest), JM distance = 0.48
N area [mg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.51
LAI [m²/m²]
p-value = 0.00000 (kstest), JM distance = 0.37
Chl area [µg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.38
EWT [mg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.47
Car area [µg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.46
Uncertainty estimation
Land cover classes

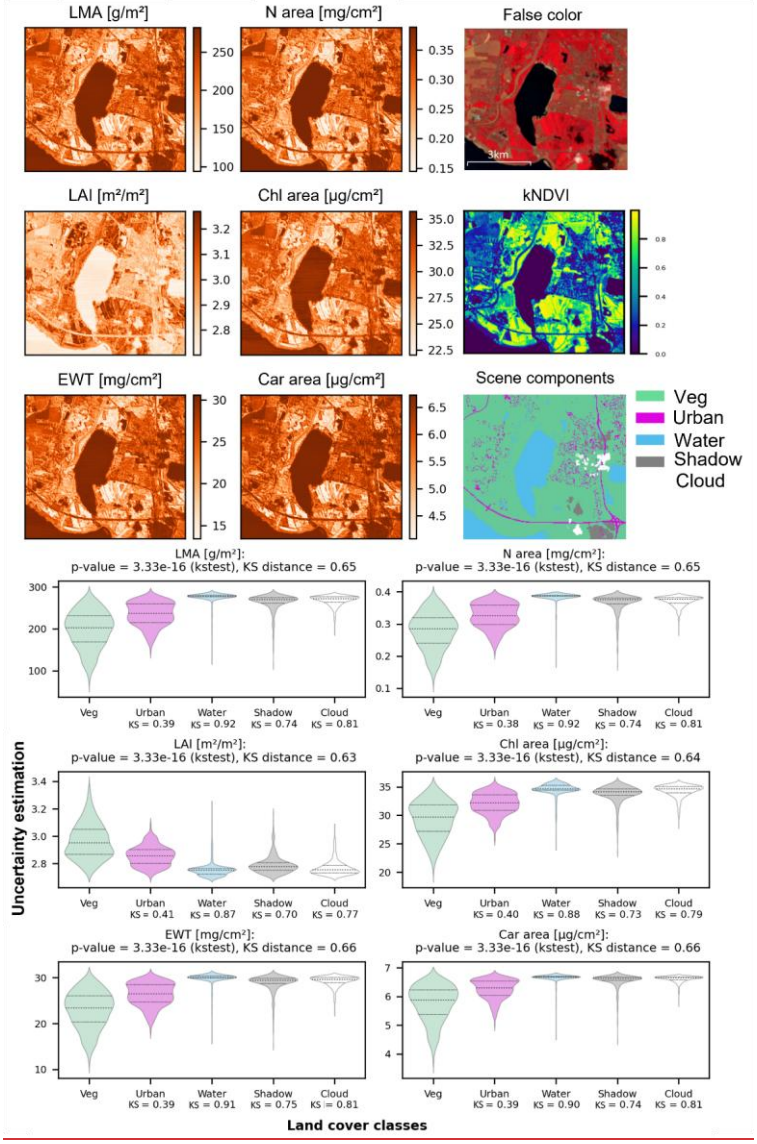
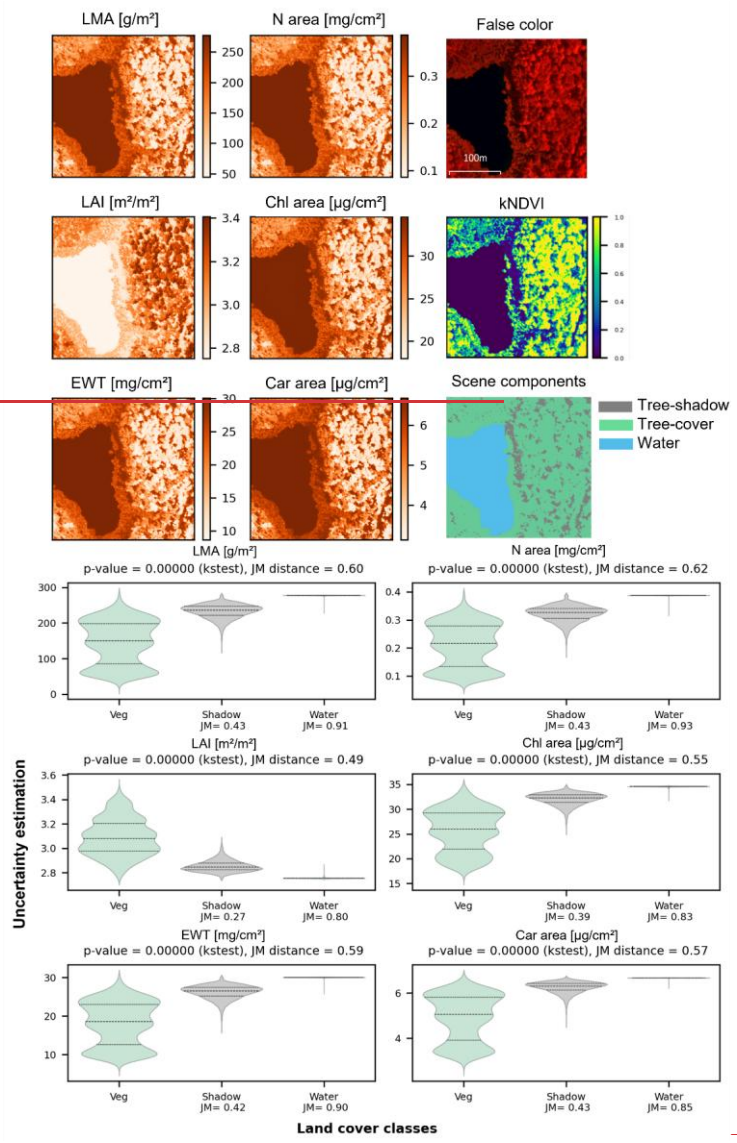

LMA [g/m²]
N area [mg/cm²]
False color
3km
LAI [m²/m²]
Chl area [µg/cm²]
kNDVI
EWT [mg/cm²]
Car area [µg/cm²]
Scene components
Veg
Urban
Water
Shadow
Cloud
LMA [g/m²]:
p-value = 3.33e-16 (kstest), KS distance = 0.65
N area [mg/cm²]:
p-value = 3.33e-16 (kstest), KS distance = 0.65
LAI [m²/m²]:
p-value = 3.33e-16 (kstest), KS distance = 0.63
Chl area [µg/cm²]:
p-value = 3.33e-16 (kstest), KS distance = 0.64
EWT [mg/cm²]:
p-value = 3.33e-16 (kstest), KS distance = 0.66
Car area [µg/cm²]:
p-value = 3.33e-16 (kstest), KS distance = 0.66
Uncertainty estimation
Land cover classes

Figure 4: EnMAP scene (right panel) Description of the scene including false-color composite highlighting vegetation in red, kernel normalized difference vegetation index (kNDVI) map showing higher values for denser vegetation area and component scene map. (Left panels) Spatial distribution of uncertainty estimation for six plant traits: Leaf Mass per Area (LMA), Nitrogen content (N), Leaf Area Index (LAI), Chlorophyll content (Chl), Equivalent Water Thickness (EWT), and Carotenoid content (Car), showing how uncertainty varies across the scene. (Bottom panel) Trait-wise violin plots of predicted uncertainty distributions from randomly sampled pixels across the scene components. To evaluate the plausibility of uncertainty estimates, Kolmogorov–Smirnov (KS) ~~Jeffries–Matusita (JM)~~ distances measure the separability between vegetation and non-vegetation distributions, while KS ~~Kolmogorov–Smirnov (KS)~~ p-values indicate whether these differences are statistically significant.


LMA [g/m²]
N area [mg/cm²]
False color
100m
LAI [m²/m²]
Chl area [µg/cm²]
kNDVI
EWT [mg/cm²]
Car area [µg/cm²]
Scene components
Tree-shadow
Tree-cover
Water
LMA [g/m²]
p-value = 0.00000 (kstest), JM distance = 0.60
N area [mg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.62
LAI [m²/m²]
p-value = 0.00000 (kstest), JM distance = 0.49
Chl area [µg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.55
EWT [mg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.59
Car area [µg/cm²]
p-value = 0.00000 (kstest), JM distance = 0.57
Veg
Shadow
JM= 0.43
Water
JM= 0.91
JM= 0.93
JM= 0.27
JM= 0.80
JM= 0.39
JM= 0.83
JM= 0.42
JM= 0.90
JM= 0.85
Uncertainty estimation
Land cover classes

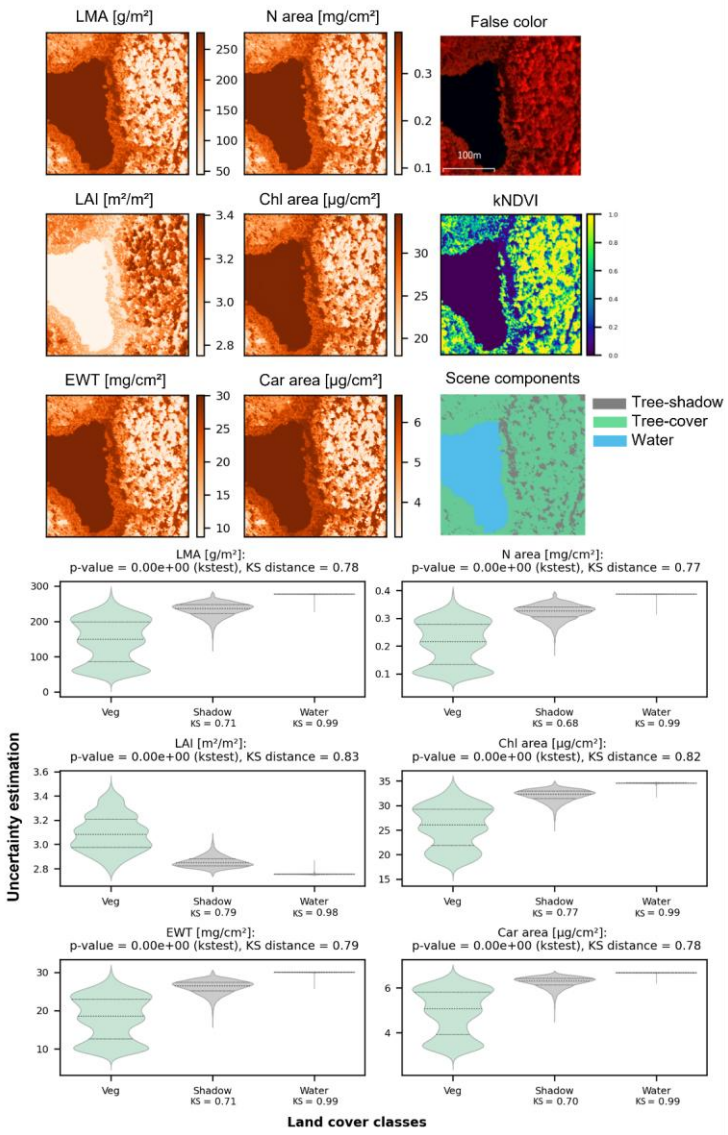
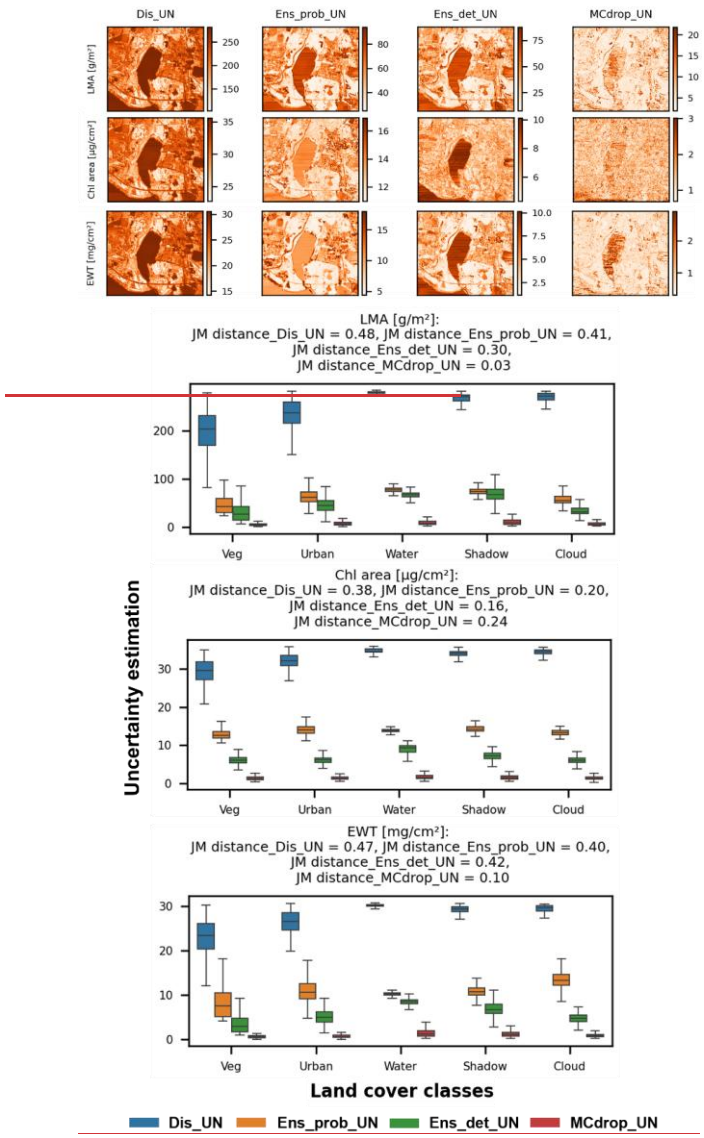

LMA [g/m²]
N area [mg/cm²]
False color
LAI [m²/m²]
Chl area [µg/cm²]
kNDVI
EWT [mg/cm²]
Car area [µg/cm²]
Scene components
Tree-shadow
Tree-cover
Water
100m
LMA [g/m²]:
p-value = 0.00e+00 (kstest), KS distance = 0.78
N area [mg/cm²]:
p-value = 0.00e+00 (kstest), KS distance = 0.77
LAI [m²/m²]:
p-value = 0.00e+00 (kstest), KS distance = 0.83
Chl area [µg/cm²]:
p-value = 0.00e+00 (kstest), KS distance = 0.82
EWT [mg/cm²]:
p-value = 0.00e+00 (kstest), KS distance = 0.79
Car area [µg/cm²]:
p-value = 0.00e+00 (kstest), KS distance = 0.78
Veg
Shadow
KS = 0.71
KS = 0.68
KS = 0.79
KS = 0.77
KS = 0.71
KS = 0.70
Water
KS = 0.99
KS = 0.98
Uncertainty estimation
Land cover classes

Figure 5: NEON scene (Right panels) Description of the scene including false-color composite highlighting vegetation in red, kernel normalized difference vegetation index (kNDVI) map showing higher values for denser vegetation area and component scene map. (Left panels) Spatial distribution of uncertainty estimation for six plant traits: Leaf Mass per Area (LMA), Nitrogen content (N), Leaf Area Index (LAI), Chlorophyll content (Chl), Equivalent Water Thickness (EWT), and Carotenoid content (Car), showing how uncertainty varies across the scene. (Bottom panels) Trait-wise violin plots of predicted uncertainty distributions from randomly sampled pixels across the scene components. To evaluate the plausibility of uncertainty estimates, Kolmogorov–Smirnov (KS) ~~Jeffries–Matusita (JM)~~ distances measure the separability between vegetation and non-vegetation distributions, while KS ~~Kolmogorov–Smirnov (KS)~~ p-values indicate whether these differences are statistically significant.

Consistent with the findings from the OOD vegetation analysis (Section 3.1), MCdrop_UN displayed the lowest differences ~~performance in distinguishing~~ between scene components, with the smallest ~~JM~~ KS distances relative to the vegetation related uncertainty distribution (Fig. 6 and 7, Fig. S7 and S9). In contrast, the spatial pattern of estimated uncertainties was generally similar for Ens_prob_UN, Ens_det_UN and Dis_UN, ~~with both methods~~ showing lower uncertainties for vegetation and higher uncertainties for water and cloud-shadow pixels. However, Dis_UN provided a clearer contrast between non-vegetated and vegetated pixels. Yet, Ens_prob_UN predictions exhibited lower sensitivity to non-vegetated OOD areas.

More specifically, for the EnMAP scene (30m spatial resolution, where mixed pixels are more prevalent), Dis_UN achieved substantially higher contrast than all variance-based methods. Across the six traits, Dis_UN attained KS distances ranging from 0.63 to 0.66 (mean: 0.648) when comparing vegetation versus non-vegetation (water, clouds, shadows, urban) uncertainty distributions. In comparison, scaled Ens_prob_UN achieved KS distances of 0.43-0.51 (mean: 0.475), scaled Ens_det_UN achieved 0.05-0.50 (mean: 0.337), and scaled MCdrop_UN achieved 0.12-0.32 (mean: 0.212). This represents a 36% higher contrast for Dis_UN compared to the best-performing ensemble method (Ens_prob_UN), and a 206% improvement over MCdrop_UN. Critically, Dis_UN maintained consistently high KS distances across all traits (CV KS: 1.8%), while ensemble methods showed higher variability (CV KS 7%-48%).

For the NEON scene (1 m spatial resolution, where less sub-pixel variation is present), we observed similar patterns, while overall the contrast in uncertainty estimates was higher among scene components with the Dis_UN. Dis_UN achieved KS distances of 0.78–0.83 (mean = 0.795), while scaled Ens_det_UN and scaled Ens_prob_UN reached 0.14–0.64 (mean = 0.465) and 0.43–0.51 (mean = 0.470), respectively. MCdrop_UN again performed weakest, with KS distances of 0.25–0.46 (mean = 0.328).


Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN
LMA [g/m²]
Chl area [µg/cm²]
EWT [mg/cm²]
LMA [g/m²]:
JM distance_Dis_UN = 0.48, JM distance_Ens_prob_UN = 0.41,
JM distance_Ens_det_UN = 0.30,
JM distance_MCdrop_UN = 0.03
Chl area [µg/cm²]:
JM distance_Dis_UN = 0.38, JM distance_Ens_prob_UN = 0.20,
JM distance_Ens_det_UN = 0.16,
JM distance_MCdrop_UN = 0.24
EWT [mg/cm²]:
JM distance_Dis_UN = 0.47, JM distance_Ens_prob_UN = 0.40,
JM distance_Ens_det_UN = 0.42,
JM distance_MCdrop_UN = 0.10
Uncertainty estimation
Veg
Urban
Water
Shadow
Cloud
Land cover classes
Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN

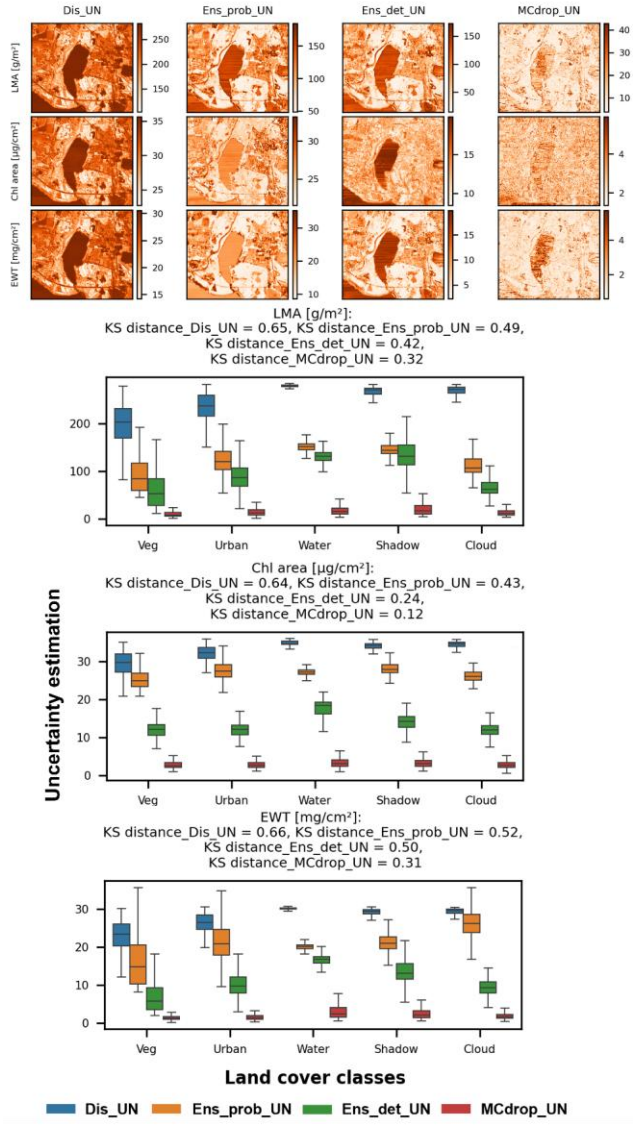
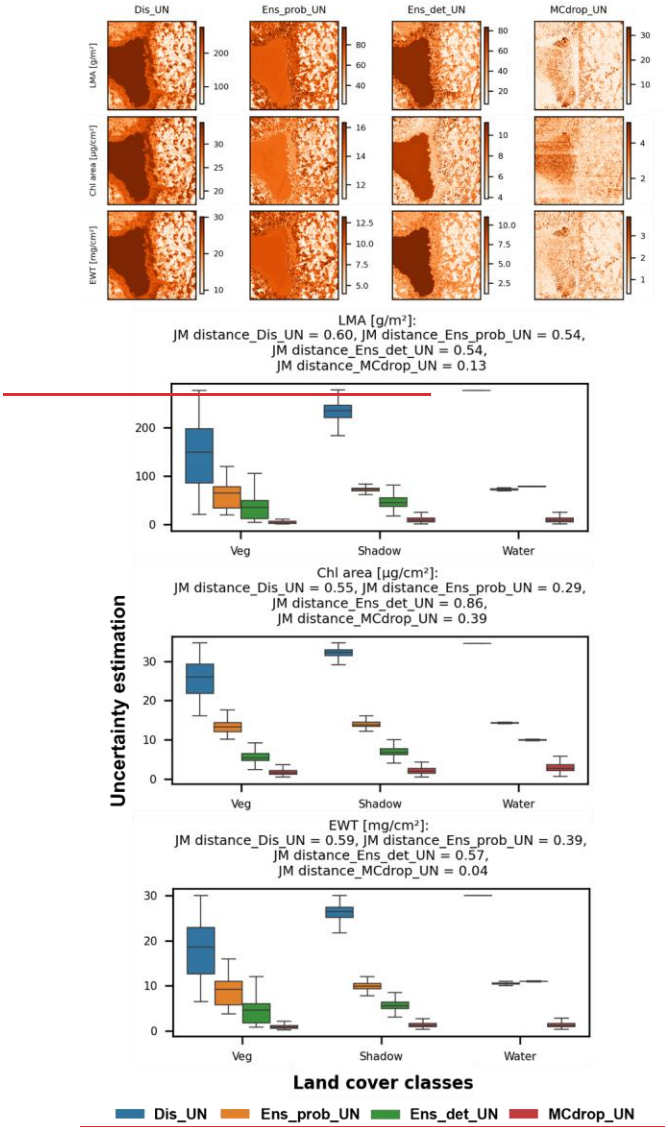

Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN
LMA [g/m²]
Chl area [µg/cm²]
EWT [mg/cm²]
LMA [g/m²]:
KS distance_Dis_UN = 0.65, KS distance_Ens_prob_UN = 0.49,
KS distance_Ens_det_UN = 0.42,
KS distance_MCdrop_UN = 0.32
Chl area [µg/cm²]:
KS distance_Dis_UN = 0.64, KS distance_Ens_prob_UN = 0.43,
KS distance_Ens_det_UN = 0.24,
KS distance_MCdrop_UN = 0.12
EWT [mg/cm²]:
KS distance_Dis_UN = 0.66, KS distance_Ens_prob_UN = 0.52,
KS distance_Ens_det_UN = 0.50,
KS distance_MCdrop_UN = 0.31
Uncertainty estimation
Veg
Urban
Water
Shadow
Cloud
Land cover classes
Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN

Figure 6: EnMAP scene: Comparison of uncertainty estimations for three traits (Leaf Mass per Area (LMA), Chlorophyll content (Chl) and Equivalent Water Thickness (EWT)) using four methods: Distance-based (Dis_UN), probabilistic ensemble (Ens_prob_UN), deterministic ensemble (Ens_det_UN), and Monte Carlo dropout (MCdrop_UN). All state-of-the-art methods were scaled by 1.96 × standard deviation: (Upper panesl): Spatial maps of the predicted uncertainty distribution across the scene, allowing comparison of how each method assigns uncertainty to different scene components. (Bottom panels): Trait-wise box plots of uncertainty distributions from the different methods, based on randomly sampled pixels across scene components (vegetation, cloud, shadow, urban, water). Associated Kolmogorov–Smirnov (KS)~~Jeffries–Matusita (JM)~~ distances quantify the separability between vegetation and non-vegetation uncertainty distributions.


Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN
LMA [g/m²]
Chl area [µg/cm²]
EWT [mg/cm²]
LMA [g/m²]:
JM distance_Dis_UN = 0.60, JM distance_Ens_prob_UN = 0.54,
JM distance_Ens_det_UN = 0.54,
JM distance_MCdrop_UN = 0.13
Chl area [µg/cm²]:
JM distance_Dis_UN = 0.55, JM distance_Ens_prob_UN = 0.29,
JM distance_Ens_det_UN = 0.86,
JM distance_MCdrop_UN = 0.39
EWT [mg/cm²]:
JM distance_Dis_UN = 0.59, JM distance_Ens_prob_UN = 0.39,
JM distance_Ens_det_UN = 0.57,
JM distance_MCdrop_UN = 0.04
Uncertainty estimation
Veg
Shadow
Water
Land cover classes
Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN

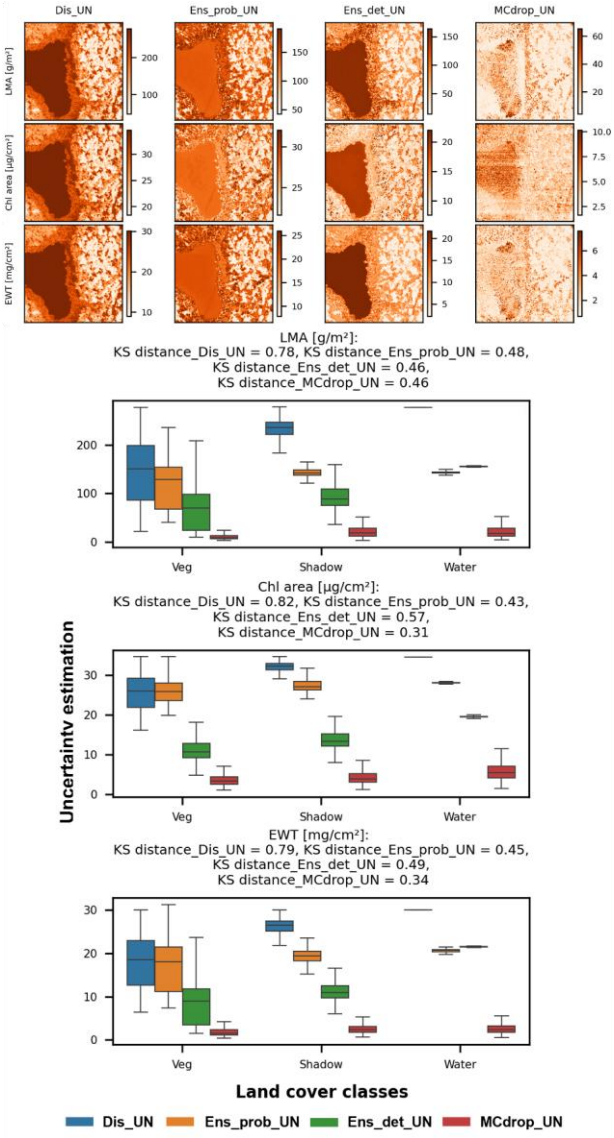

Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN
LMA [g/m²]
Chl area [µg/cm²]
EWT [mg/cm²]
LMA [g/m²]:
KS distance_Dis_UN = 0.78, KS distance_Ens_prob_UN = 0.48,
KS distance_Ens_det_UN = 0.46,
KS distance_MCdrop_UN = 0.46
Chl area [µg/cm²]:
KS distance_Dis_UN = 0.82, KS distance_Ens_prob_UN = 0.43,
KS distance_Ens_det_UN = 0.57,
KS distance_MCdrop_UN = 0.31
EWT [mg/cm²]:
KS distance_Dis_UN = 0.79, KS distance_Ens_prob_UN = 0.45,
KS distance_Ens_det_UN = 0.49,
KS distance_MCdrop_UN = 0.34
Uncertaintv estimation
Veg
Shadow
Water
Land cover classes
Dis_UN
Ens_prob_UN
Ens_det_UN
MCdrop_UN

Figure 7: NEON scene: Comparison of uncertainty estimations for three traits (Leaf Mass per Area (LMA), Chlorophyll content (Chl) and Equivalent Water Thickness (EWT)) using four methods: Distance-based (Dis_UN), probabilistic ensemble (Ens_prob_UN), deterministic ensemble (Ens_det_UN), and Monte Carlo dropout (MCdrop_UN). All state-of-the-art methods were scaled by 1.96 × standard deviation: (Upper panels): Spatial maps of the predicted uncertainty distribution across the scene, allowing comparison of how each method assigns uncertainty to different scene components. (Bottom panels): Trait-wise box plots of uncertainty distributions from the different methods, based on randomly sampled pixels across scene components (vegetation, shadow, water). Associated Kolmogorov–Smirnov (KS) ~~Jeffries-Matusita (JM)~~ distances quantify the separability between vegetation and non-vegetation uncertainty distributions.

4 Discussion

Understanding and accurately quantifying uncertainty is essential for assessing the reliability of model predictions, particularly in OOD scenarios where the model encounters unseen data. In this section, we evaluate the performance of different uncertainty estimation methods at both the local and landscape scales, highlighting their strengths and limitations in capturing uncertainty across various vegetation types and scene components.

4.1 Local-Scale Uncertainty in OOD Vegetation data

~~In this study, both the deterministic ensemble Ens_det_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 S5 and Fig. S5). 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). This underestimation, especially for higher predicted values, is a known limitation of ensemble and Monte Carlo-based approaches (Hu et al., 2022; Klotz et al., 2022; Liu at al. 2021; Lakshminarayanan et al., 2017), which tend to be optimistic in their uncertainty estimates. A similar trend of predicted uncertainty was observed across all traits with both methods (Fig. 3). The Monte Carlo approach, in particular, performed poorly (high ENCE values), indicating its failure to capture variability within samples from different vegetation types.~~

~~These findings emphasize the importance of carefully selecting and interpreting uncertainty estimation methods, especially under distribution shift (Liu et al. 2021; Ovadia et al., 2019). 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_det_UN improved when scaling uncertainty by a factor of 1.96 (Fig. S6), although this adjustment did not improve for 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. Among the tested ensemble-based methods, the probabilistic ensemble (Ens_prob_UN) achieved lower ENCE values,~~

~~suggesting good average-case calibration for OOD vegetation samples. The relatively good calibration of Ens_prob_UN observed for OOD vegetation samples is consistent with Gustafsson et al. (2020) showing that probabilistic deep ensembles can remain well-calibrated under natural distribution shifts, that is, when new data differ in source, or different conditions but still represent the same underlying object class. However, the predictive ranges were comparatively narrow, which limited their ability to capture extreme residuals.~~

Applying the ensemble and Monte Carlo methods with their default settings resulted in a critical underestimation of observed residuals (Fig. S6). After scaling variance-based uncertainties to approximate 95% confidence intervals (1.96 × σ), the ensemble approaches showed substantially improved alignment with observed residuals (Fig. 3), in contrast to their unscaled performance shown in Fig. S6. The improved performance of both ensemble methods for OOD vegetation samples indicates the average alignment under natural distribution shifts (Gustafsson et al., 2020), that is, when new data differ in source or environmental conditions but still represent the same underlying object class (vegetation). However, these methods do not automatically adapt to distributional changes if not validated with an independent set.

This limitation highlights the need for post-hoc scaling and introduces practical challenges in operational settings: practitioners must determine appropriate scaling factors for each trait and application context, requiring either assumptions of normality or empirical calibration on held-out data, which may not be available for novel sensors or ecosystems. Recent studies have increasingly emphasized the need for robust recalibration strategies for variance-based uncertainty methods, particularly under distributional shift (Liu et al., 2021; Ovadia et al., 2019; Palmer et al., 2022). While a comprehensive evaluation of such calibration strategies lies beyond the scope of this study, our results underscore that naïve application of variance-based methods without careful consideration of error distribution characteristics can lead to systematically biased uncertainty estimates.

However, the Monte Carlo approach continued to perform poorly (high ENCE values) even after scaling, indicating its failure to capture variability within samples from different vegetation types. 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). This underestimation, especially for higher predicted values, is a known limitation of Monte Carlo-based approaches (Hu et al., 2022; Klotz et al., 2022; Liu at al. 2021), which tend to be optimistic in their uncertainty estimates.

In contrast, the distance-based quantile regression method (Dis_UN) provides a complementary perspective. It demonstrated a stronger alignment with predicted uncertainty values and the residuals of the multi-trait model but with exaggerated values (Fig. 3, worst-case uncertainty). This indicates that the conservative estimates provided by Dis_UN effectively contain the residuals, a valuable characteristic for uncertainty modeling as highlighted by Brown et al. (2021b). A key advantage of the Dis_UN method is the dissociation of predicted uncertainty values of different vegetation types (Fig. 3), which enhances interpretability. This can be attributed to the incorporation of spectral distance as a predictor (feature space). By leveraging spectral distance, the Dis_UN method quantifies how far a given sample is from typical spectral signatures in the training dataset, allowing for more accurate adjustment of the predicted uncertainty intervals. The regression analysis supports this observation, highlighting the higher importance of the scaled spectral distance compared to embedding space distance in

predicting residuals across all traits (Table S6). For instance, the spectral distance coefficients for traits such as LMA, EWT, and Car are significantly higher than the corresponding coefficients for embedding distance. This indicates that spectral distance contributes more strongly to the regression model's ability to predict uncertainty. The prominence of spectral distance ~~suggests that it is more related to the diversity in trait values, as it directly reflects the physical and chemical properties captured in the hyperspectral reflectance data. In contrast, the embedding distance, while useful, abstracts the data into a latent space, which may lose some trait-specific spectral information~~can be explained by its sensitivity to spectral variability that is not exclusively driven by trait variation, which becomes particularly important under OOD conditions. While the embedding space is optimized to capture trait-relevant features learned during training, it may abstract away spectral characteristics that are not directly informative for trait prediction.

The predicted uncertainty was directly related to the vegetation characteristics. For example, grassland samples tended to have lower uncertainty across most traits compared to forest and shrubland (Fig. 3). 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 S2) and, from a radiative transfer point of view, it is considered structurally simpler and more homogeneous compared to more complex vegetation types like 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. The behavior of uncertainty also varied across different traits, influenced by both the inherent properties of each trait, the trait variability and representativeness of data samples from various vegetation types (e.g., forest, grassland, crops). For example, the uncertainty modeling for LMA and N showed a better fit to the training data compared to LAI, Chl, and Car content (Fig. 3 and Fig. S4). This can be attributed to differences in how these traits influence spectral reflectance. LAI is highly prone to spectral saturation, where the spectral signal becomes less sensitive to changes in the trait, reducing sensitivity to variation (Brantley et al., 2011; Gamon et al., 1995; Sellers, 1985; Wang et al., 2005). In contrast, traits that primarily affect specific spectral regions, such as Chl and Car, which mainly influence the visible spectrum, may not be fully captured by the distance-based approach, potentially leading to an underestimation of uncertainty.

4.2 Landscape-Scale Uncertainty in OOD data

We further tested the performance of these methods on the landscape scale with two scenes from distinct sensors, each containing a variety of scene components that are not typically in the training data. This served as a proof of concept to assess models' behavior under extreme OOD conditions, as it is hard to visually validate the predictions with no reference data. In such cases, a higher range of uncertainty should indicate regions where the model is not confident.

4.2.1 Comparison with other methods

Despite being trained exclusively on vegetation samples, the Dis_UN method demonstrated robust performance in detecting shifts in the distribution of different scene components - even if they are not vegetation. The Dis_UN performance was comparable to that of the ensemble methods and superior to the MCdrop_UN approach, as evidenced by the ~~JM~~ KS distance metric (Fig 6 and 7). The Dis_UN method and both ensemble methods showed similar trends in uncertainty estimation: regions with high and low uncertainty values were consistently identified by both methods. This comparability with Ens_det_UN is expected, given that the same trained models were used to develop both methods. However, while Dis_UN produced markedly higher contrast in predicted uncertainty between vegetated and non-vegetated areas, the Ens_prob_UN method appeared less sensitive to these strongly OOD pixels, assigning lower uncertainty values to them. The Dis_UN method exhibited a broader range of uncertainty, allowing for finer differentiation across different regions. However, when examining the spatial uncertainty maps, the Dis_UN method produced more homogeneous boundaries within the scene components, highlighting its ability to provide more consistent spatial representations of uncertainty (Figs 4, 5, 6, 7). Results from the Kolmogorov-Smirnov (KS) test further demonstrated that the Dis_UN method produced distinct uncertainty distributions for vegetated and non-vegetated pixels, with non-vegetated areas consistently showing higher residuals across all traits (Figs. 4, 5).

In contrast, the MCdrop_UN method consistently demonstrated the weakest performance, raising concerns about its reliability in these contexts, particularly when assessing model uncertainty. One key issue is that low standard deviation values may create a false sense of confidence in the model's predictions, suggesting better trait model performance than what is actually the case. This can be misleading, as the narrow range of uncertainty estimates does not necessarily reflect the true error in the model's outputs. Such behavior has been observed in previous studies (e.g., Pullanagari et al. 2021, Padarian et al. 2022, García-Soria et al. 2024), where dropout-based uncertainty appeared overly optimistic in comparison to other uncertainty assessment methods. To avoid potential misinterpretations, it is crucial to consider alternative metrics and methods when evaluating uncertainty estimation approaches. For instance, relying solely on standard deviation values or similar variance-based metrics may overlook important aspects of model performance, such as the method's ability to capture OOD uncertainty or account for heterogeneity in complex datasets.

Furthermore, the variation in spatial resolution reveals a critical operational advantage of Dis_UN. The distance-based method maintains robust OOD detection during sub-pixel variation. At 30m resolution (EnMAP), individual pixels frequently contain mixtures of vegetation and non-vegetation components, for example, urban pixels containing street trees, or forest edges mixing canopy and bare ground. Such mixed pixels exhibit intermediate spectral signatures that fall between pure scene components. For variance-based methods, particularly ensemble approaches, these mixed-signature pixels can produce moderate prediction variance that fails to clearly flag them as problematic, since ensemble members may converge on intermediate predictions with modest disagreement (Figs. 6, 7 and S8). This is evident in the low minimum KS values for ensemble methods at 30m (as low as 0.05), indicating poor separation for certain traits where mixed pixels dominate the scene. In contrast, Dis_UN's distance-based predictors explicitly measure dissimilarity relative to the training set, which consists predominantly of pure vegetation samples. Mixed pixels, even if spectrally intermediate, are recognized as

dissimilar from the pure vegetation training manifold, resulting in elevated uncertainty predictions. This mechanism remains effective regardless of pixel purity, explaining Dis_UN's consistent performance (mean KS: 0.648 at 30 m) and its particularly strong advantage over ensemble methods at coarser resolution (36% higher than Ens_prob_UN at 30 m vs. 69% higher at 1m).

At higher spatial resolution (NEON, 1 m), where less sub-pixel variation is present than in the EnMAP scene and component boundaries are sharper, the performance of variance-based methods improves relative to the medium-resolution case (EnMAP) (mean KS ≈ 0.33-0.47). However, despite this improvement, Dis_UN still achieved the highest separability (mean KS = 0.795), indicating a stronger and more consistent contrast between vegetated and non-vegetated components. Importantly, the improved ensemble performance at 1 m resolution remains resolution-dependent and cannot be assumed in operational medium-resolution satellite applications, which dominate current global monitoring systems (e.g., EnMAP, PRISMA at 30 m).

Beyond enhancing uncertainty prediction performances, the proposed distance-based uncertainty estimation method provides substantial computational advantages over variance-based approaches. Unlike variance-based methods that require multiple forward passes to compute prediction variance, the distance-based approach allows for straightforward application once the uncertainty model is trained. This eliminates the need for repeated inference runs, making it significantly more computationally efficient. Such efficiency is particularly valuable for large-scale remote sensing applications, where fast and scalable uncertainty estimation is crucial. Though, it is important to distinguish between the training and inference costs of the proposed method (Tables S7 and S8). In our experimental setup, the training time of the distance-based uncertainty model was of a similar order to that of deep ensembles, as we adopted a leave-one-dataset-out (LODO) transferability analysis to explicitly evaluate out-of-distribution conditions, requiring the training of 50 models. However, this design reflects a specific validation strategy rather than an intrinsic requirement of the approach. In practice, distance-based uncertainty estimation can be integrated into more conventional validation schemes, such as k-fold cross-validation, thereby substantially reducing the training overhead.

4.2.2 Uncertainty Patterns of Dis_UN Across scene components and Spatial Resolutions

For the EnMAP scene, with 30 m spatial resolution, the uncertainty maps from the Dis_UN approach showed a variation in uncertainty among traits. This is expected because the different traits depend on distinct spectral regions, which are affected differently by various scene components (e.g., water, urban areas, or vegetation). Higher uncertainty was particularly evident in non-vegetative land cover types such as water, cloud and cloud-shadow regions, which are clearly OOD relative to the model's primary focus on vegetation. These areas lack spectral similarity to vegetation and therefore result in increased uncertainty in the model's predictions. In contrast, urban areas exhibited lower uncertainty, a pattern that can be explained by the mixed component in such areas. Urban areas often contain green spaces, like trees and meadow patches. As a result, these mixed pixels are closer to vegetation data of the training set and do not show as strong uncertainties as pure non-vegetative classes like water. These results highlight that the method works as expected independently from the land cover

types and scene components, so that with gradual dissimilarity from vegetation spectra, the uncertainty increases. As shown here at the example of the 30m EnMap data, such an uncertainty quantification is particularly important to evaluate the robustness of a prediction in complex scenes, where more than one land cover type can be present per pixel and the resolution of the sensor and existing land cover products are not detailed enough to explicitly resolve the scene component.

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. S10 and S13), 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. S17). 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. S12 and S15). 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 densely vegetated areas, while OOD pixels such as water, shadow, and urban regions showed lower and less variable uncertainty.

For the EnMAP scene, we identified pure vegetated pixels representing different vegetation types, including crops, tree cover, and grassland. As discussed in Section 4.1, grassland pixels exhibited the lowest uncertainty with minimal variation, likely due to their simpler spectral properties. Crops showed the highest uncertainty values but with lower variation, while tree cover and shrubland displayed high variability in predicted uncertainty, reflecting their greater spectral and structural complexity (Fig. S8). We observed a similar pattern for the 1m resolution NEON data. Uncertainty in the vegetated areas (forest) exhibited greater variability (bimodal distribution, Fig. 5) in the NEON scene compared to vegetated areas in the EnMAP scene. This suggests that uncertainty in forested regions may be more sensitive to fine-scale shadows and canopy structural complexity, resulting in a more complex spectral response. These small-scale confounding factors, such as those from canopy gaps and canopy structures, substantially affect the spectral response (Nagendra and Rocchini, 2008).

4.3 Challenges in training uncertainty models

Training uncertainty models for trait prediction in remote sensing is subject to substantial challenges due to the sparse and non-uniform distribution of available trait data. This lack of comprehensive data coverage is particularly problematic at the low and high ends of the trait distribution. Very low trait values often are associated with weak signals while high trait

values can lead to saturation effects, both resulting in high uncertainty estimation, when using Dis_UN. This data sparsity not only limits the model's ability to make robust trait predictions but also prevents the training of models to robustly estimate uncertainty across the full trait range. Moving forward, it is crucial to collect more data, particularly from the tails of the trait distributions (e.g., very high or low values). This means not only gathering data from plants at the peak of their growing season, when they exhibit high vitality, but also from senescent or stressed plants, which are typically underrepresented in current datasets (Schiefer et al. 2021, Brown et al., 2020). Such efforts underscore the importance of data sharing within the scientific community to improve the robustness and reliability of uncertainty models (Cherif et al. 2023).

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 presents challenges, particularly due to the underlying assumptions of each approach (Ovadia et al., 2019). Traditional methods, such as ensemble 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. While their focus was on hydrological modeling, similar asymmetries are present in plant trait values, making their insights applicable to our context. 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 modeling extreme deviations. This approach focuses on extreme residuals, allowing for a more conservative and distribution-agnostic uncertainty estimate.

Moreover, Janet et al. (2019) reveal that traditional uncertainty estimation methods such as MCdrop_UN and ensemble techniques often produce overly confident predictions in OOD regions. In our study, MCdrop_UN and ensemble methods exhibited similar tendencies, particularly when applied to OOD samples with unfamiliar spectral characteristics due to different sensor properties and scene elements like water and clouds. This overconfidence in OOD regions is problematic, as it can lead to a false sense of reliability in model predictions where the model is actually less certain.

Variance-based approaches have notable limitations in uncertainty estimation. While they provide a measure of dispersion, their uncertainties do not always scale appropriately with the actual errors, making direct interpretation and practical application challenging (Fig. 3). The ensemble methods, in particular, has been observed to underestimate uncertainty, providing an optimistic assessment of model performance (Janet et al., 2019; Meyer and Pebesma, 2021). This can be misleading, especially in cases where the predictor is inherently biased, but the variations within the ensemble are small,

giving a false and often optimistic impression of reliability. Additionally, uncertainty is trait-specific, influenced by the intrinsic nature of the trait being predicted. Thus, interpreting uncertainty in these models requires careful consideration of both the method used and the specific trait characteristics, to avoid misinterpretation and ensure the reliability of predictions. Unlike conventional methods, which largely reflect the uncertainty inherent in the training data, the distance-based approach adapts to new data by comparing it with the training set, offering a more comprehensive and flexible assessment of uncertainty.

4.5 Outlook: Uncertainty in the Context of Global Trait Mapping

Estimating uncertainty from remote sensing-based machine learning algorithms remains challenging, as current methods struggle to capture the full complexity and diversity of real-world data and datasets are often limited in representativeness and quantity. Despite these issues, the distance-based method shows promise due to its increased interpretability compared to more complex probabilistic approaches. The distance-based method offers a clearer view of where and why high uncertainty arises, particularly in areas substantially different from the training data. By measuring the distance from known data, it intuitively identifies OOD regions, providing insights into the model's reliability under shifting conditions. At the same time, it is important to recognize that distance-based uncertainty estimation cannot by itself overcome data-intrinsic limitations. Structural traits such as leaf area index are affected by the long-recognized problem of spectral saturation, where top-of-canopy reflectance becomes insensitive to additional foliage at high canopy densities (e.g., Sellers, 1985; Myneni et al., 1995; Gitelson, 2004, Steltzer and Welker. 2006, Zheng et al. 2009, Xu et al. 2020). In such cases, saturation arises from the inherent distribution of the data and constrains both training and inference. Distance-based uncertainty is therefore best understood as a diagnostic tool that reveals these information gaps, rather than as a mechanism to eliminate them. Progress will require more sophisticated sensing strategies, where recent work has shown promising directions (Mutanga et al., 2023; Wan et al., 2022). However, no purely optical method fully overcomes saturation, as this limitation is rooted in the physics of canopy reflectance. This limitation, in turn, motivates the continued development of distance-aware uncertainty methods that more explicitly link the training samples to unseen data. Importantly, such methods are not limited to vegetation trait retrieval but can be applied and extended to a wide range of applications and datasets where robustness under distribution shift and reliable uncertainty quantification are crucial.

Looking forward, further refinement of the distance-based method could involve testing on more diverse datasets and exploring hybrid approaches that combine it with complementary probabilistic techniques.

In addition, while directional errors were not explicitly modeled in this study, analyzing signed residuals could help reveal trait- or vegetation-specific biases. The systematic underestimation of high LAI values due to spectral saturation (Fig. S17, Section 4.2.2) exemplifies such directional errors. 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.

5 Conclusion

Accurate methods for predicting uncertainty from machine learning models are lacking for vegetation monitoring. As an example of trait prediction from hyperspectral data, this study demonstrates that traditional uncertainty quantification methods, such as ensemble and dropout Bayesian approaches, often struggle to adapt to unseen (OOD) data, resulting in overconfident or misleading uncertainty estimates. The presented distance-based method, however, offers improved adaptability across land cover types. Compared to scaled variance-based methods, Dis_UN demonstrates four key operational advantages: (1) superior estimation of uncertainty in OOD scenarios, particularly under mixed-pixel conditions at medium resolution (30m) common in operational satellite monitoring; (2) uncertainty quantification without requiring normality or symmetry assumptions, accommodating asymmetric error patterns; (3) enhanced interpretability of uncertainty sources, as uncertainty is directly linked to sample dissimilarity from the training data; and (4) computational efficiency at inference (2.6-7.7× faster), critical for processing large-scale hyperspectral data. Our results highlight the importance of incorporating diverse datasets to mitigate distributional biases. Furthermore, the proposed uncertainty method was successfully tested across two scenes acquired by different sensors, with varying resolutions, across different biomes and land cover types, demonstrating its robustness across heterogeneous conditions. Accordingly, such an approach could be used for prediction of uncertainty in large-scale assessments, where OOD data is prevalent and reliable uncertainty estimation is crucial.

Formatted: Font: (Default) +Headings (Times New Roman), Not Italic, English (United States)

Data and code availability

The code and data for this study are available at: https://github.com/echerif18/Multi_trait_Uncertainty/

The code and data of the multi-trait model is available at: https://github.com/echerif18/multiTraitPredictions

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships.

Author Contributions

EC contributed to the conceptualization, methodology, formal analysis, data curation, visualization, and writing of the original draft. HF and TK contributed to conceptualization, data curation, supervision, and writing of the original draft. LK, KB, PD, ME, TH, EL and BL provided data and contributed to writing, review and editing.

Acknowledgements

We thank all data owners for sharing the data either by request or through the public Ecological Spectral Information System (EcoSIS), Data Publisher for Earth & Environmental Science (PANGEA) and DRYAD platforms. EC and HF acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft Kultur und Tourismus in the program Center of Excellence for AI-research "Center for

Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI. HF and TK acknowledge support for this work from the Federal Ministry for Economic Affairs and Climate Action (BMWK) and the German Aerospace Center (DLR) through the project AIResVeg (grant 50EE2203A). LK acknowledges that this activity was partly carried out under the Living Planet Fellowship, a program funded by the European Space Agency. The view expressed in this publication can in no way be taken to reflect the official opinion of the European Space Agency. This study has been undertaken using data from "Fiducial Reference Measurements for Vegetation" (FRM4VEG), which was funded by the European Commission and managed by the European Space Agency under the Copernicus program, and "Fiducial Reference Measurements for Vegetation—Phase 2" (FRM4VEG—Phase 2), which was funded by the European Space Agency. This work was supported in part by the European Space Agency and European Commission through the Sentinel-3 Mission Performance Centre. This work was supported by MetEOC-4. The project 19ENV07 MetEOC-4 has received funding from the EMPIR program co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation program. Support for PD is from his faculty start-up package provided by Colorado State University.

References:

- Abdar, M., and others: A review of uncertainty quantification in deep learning: Techniques, applications, and challenges, Inf. Fusion, 76, 243–297, 2021.
- Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, arXiv preprint arXiv:2002.06470, 2020.
- Asner, G. P.: Biophysical and biochemical sources of variability in canopy reflectance, Remote Sens. Environ., 64, 234–253, 1998.
- Asner, G. P., Scurlock, J. M. O., and Hicke, J. A.: Global synthesis of leaf area index observations: implications for ecological and remote sensing studies, Glob. Ecol. Biogeogr., 12, 191–205, 2003.
- Asner, G. P. and Martin, R. E.: Spectranomics: Emerging science and conservation opportunities at the interface of biodiversity and remote sensing, Glob. Ecol. Conserv., 8, 212–219, 2016.
- Bethell, D., Gerasimou, S., and Calinescu, R.: Robust uncertainty quantification using conformalised Monte Carlo prediction, Proc. AAAI Conf. Artif. Intell., 38, 20939–20948, 2024.
- Berger, K., Foerster, S., Szantoi, Z., Hostert, P., Foerster, M., Van De Kerchove, R., and Herold, M.: Evolving Earth observation capabilities for recent land-related EU policies, *Land Use Policy*, **158**, 107749, https://doi.org/10.1016/j.landusepol.2025.107749, 2025.
- Briottet, X., Bajjouk, T., Chami, M., Delacourt, C., Feret, J.-B., Jacquemoud, S., Minghelli, A., Sheeren, D., Weber, C., Fabre, S., and others: BIODIVERSITY–A new space mission to monitor Earth ecosystems at fine scale, Rev. Fr. Photogramm. Télédétection, 224, 33–58, 2022.

Formatted: Font: (Default) +Headings (Times New Roman), English (United States)

- Brantley, S. T., Zinnert, J. C., and Young, D. R.: Application of hyperspectral vegetation indices to detect variations in high leaf area index temperate shrub thicket canopies, Remote Sens. Environ., 115, 514–523, 2011.
- Brodrick, P. G., Pavlick, R., Bernas, M., Chapman, J. W., Eckert, R., Helmlinger, M., Hess-Flores, M., Rios, L. M., Schneider, F. D., Smyth, M. M., Eastwood, M., Green, R. O., Thompson, D. R., Chadwick, K. D., and Schimel, D. S.: SHIFT: AVIRIS-NG L2A Unrectified Surface Reflectance Version 1, ORNL DAAC, Oak Ridge, Tennessee, USA, 2023.
- Brown, L. A., Dash, J., Lidón, A. L., Lopez-Baeza, E., and Dransfeld, S.: Synergetic exploitation of the Sentinel-2 missions for validating the Sentinel-3 Ocean and Land Color Instrument terrestrial chlorophyll index over a vineyard dominated Mediterranean environment, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 12, 2244–2251, 2019.
- Brown, L. A., Meier, C., Morris, H., Pastor-Guzman, J., Bai, G., Lerebourg, C., Gobron, N., Lanconelli, C., Clerici, M., and Dash, J.: Evaluation of global leaf area index and fraction of absorbed photosynthetically active radiation products over North America using Copernicus Ground Based Observations for Validation data, Remote Sens. Environ., 247, 111935, 2020.
- Brown, L. A., Camacho, F., García-Santos, V., Origo, N., Fuster, B., Morris, H., Pastor-Guzman, J., Sánchez-Zapero, J., Morrone, R., Ryder, J., and others: Fiducial Reference Measurements for Vegetation Bio-Geophysical Variables: An End-to-End Uncertainty Evaluation Framework, Remote Sens. (Basel), 13, 3194, 2021a.
- Brown, L. A., Fernandes, R., Djamai, N., Meier, C., Gobron, N., Morris, H., Canisius, F., Bai, G., Lerebourg, C., Lanconelli, C., and others: Validation of baseline and modified Sentinel-2 Level 2 Prototype Processor leaf area index retrievals over the United States, ISPRS J. Photogramm. Remote Sens., 175, 71–87, 2021b.
- Brown, L. A., Morris, H., MacLachlan, A., D'Adamo, F., Adams, J., Lopez-Baeza, E., Albero, E., Martínez, B., Sánchez-Ruiz, S., Campos-Taberner, M., and others: Hyperspectral leaf area index and chlorophyll retrieval over forest and row-structured vineyard canopies, Remote Sens. (Basel), 16, 2066, 2024.
- Burnett, A. C., Serbin, S. P., and Rogers, A.: Source-sink imbalance detected with leaf- and canopy-level spectroscopy in a field-grown crop, Plant Cell Environ., 44, 2466–2479, 2021.
- Chadwick, K. D., Queally, N., Zheng, T., Cryer, J., Vanden Heuvel, C., Villanueva-Weeks, C., Ade, C., Anderegg, L., Angel, Y., Baker, B., and others: SHIFT: Photosynthetic and Leaf Traits, Santa Barbara County, 2022, ORNL DAAC, Oak Ridge, Tennessee, USA, 2023.
- Camia, A., Gliottone, I., Dowell, M., Vancutsem, C., Bertoglio, C.: Earth Observation for Biodiversity, European Commission, Ispra (Italy), 2024.
- Camps-Valls, G., Campos-Taberner, M., Moreno-Martínez, Á., Walther, S., Duveiller, G., Cescatti, A., Mahecha, M. D., Muñoz-Marí, J., García-Haro, F. J., Guanter, L., and others: A unified vegetation index for quantifying the terrestrial biosphere, Sci. Adv., 7, eabc7447, 2021.

- Cavender-Bares, J., Gamon, J. A., Hobbie, S. E., Madritch, M. D., Meireles, J. E., Schweiger, A. K., and Townsend, P. A.: Harnessing plant spectra to integrate the biodiversity sciences across biological and spatial scales, Am. J. Bot., 104, 966–969, 2017.
- Cavender-Bares, J., Gamon, J. A., and Townsend, P. A.: Remote sensing of plant biodiversity, Springer Nature, 2020.
- Cawse-Nicholson, K., Townsend, P. A., Schimel, D., Assiri, A. M., Blake, P. L., Buongiorno, M. F., Campbell, P., Carmon, N., Casey, K. A., Correa-Pabón, R. E., and others: NASA's surface biology and geology designated observable: A perspective on surface imaging algorithms, Remote Sens. Environ., 257, 112349, 2021.
- Cerasoli, S., Campagnolo, M., Faria, J., Nogueira, C., and Caldeira, M. C.: On estimating the gross primary productivity of Mediterranean grasslands under different fertilization regimes using vegetation indices and hyperspectral reflectance, Biogeosciences, 15, 5455–5471, 2018.
- Chabrillat, S., Foerster, S., Segl, K., Beamish, A., Brell, M., Asadzadeh, S., Milewski, R., Ward, K. J., Brosinsky, A., Koch, K., and others: The EnMAP spaceborne imaging spectroscopy mission: Initial scientific results two years after launch, Remote Sens. Environ., 114379, 2024.
- Cherif, E., Feilhauer, H., Berger, K., Dao, P. D., Ewald, M., Hank, T. B., He, Y., Kovach, K. R., Lu, B., Townsend, P. A., and others: From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data, Remote Sens. Environ., 292, 113580, 2023.
- Chernetskiy, M., Gómez-Dans, J., Gobron, N., Morgan, O., Lewis, P., Truckenbrodt, S., and Schmullius, C.: Estimation of FAPAR over croplands using MISR data and the earth observation land data assimilation system (EO-LDAS), Remote Sens. (Basel), 9, 656, 2017.
- Chlus, A., Kruger, E. L., and Townsend, P. A.: Mapping three-dimensional variation in leaf mass per area with imaging spectroscopy and LiDAR in a temperate broadleaf forest, Remote Sens. Environ., 250, 112043, 2020.
- Cogliati, S., Sarti, F., Chiarantini, L., Cosi, M., Lorusso, R., Lopinto, E., Miglietta, F., Genesio, L., Guanter, L., Damm, A., and others: The PRISMA imaging spectroscopy mission: Overview and first performance analysis, Remote Sens. Environ., 262, 112499, 2021.
- Dao, P. D., Axiotis, A., and He, Y.: Mapping native and invasive grassland species and characterizing topography-driven species dynamics using high spatial resolution hyperspectral imagery, Int. J. Appl. Earth Obs. Geoinf., 104, 102542, 2021.
- De Los Reyes, R., Langheinrich, M., and Bachmann, M.: EnMAP ground segment-level 2A Processor (atmospheric correction over land) ATBD, EN-PCV-TN-6007, 2, 2023.
- Douze, M., and others: The FAISS library, arXiv preprint, arXiv:2401.08281, 2024.
- Deng, Z., Zhou, F., Chen, J., Wu, G., and Zhu, J.: Deep ensemble as a Gaussian process approximate posterior, arXiv preprint arXiv:2205.00163, 2022.

- Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, Vol. 57, CRC Press, 1993.
- Egele, R., Maulik, R., Raghavan, K., Lusch, B., Guyon, I., and Balaprakash, P.: Autodeuq: Automated deep ensemble with uncertainty quantification, Proc. Int. Conf. Pattern Recognit. (ICPR), 1908–1914, 2022.
- Ewald, M., Skowronek, S., Aerts, R., Dolos, K., Lenoir, J., and others: Analyzing remotely sensed structural and chemical canopy traits of a forest invaded by Prunus serotina over multiple spatial scales, Biol. Invasions, 20, 2257–2271, 2018.
- Ewald, M., Skowronek, S., Aerts, R., Lenoir, J., Feilhauer, H., and others: Assessing the impact of an invasive bryophyte on plant species richness using high-resolution imaging spectroscopy, Ecol. Indic., 110, 105882, 2020.
- Feilhauer, H., Zlinszky, A., Kania, A., Foody, G. M., Doktor, D., Lausch, A., and Schmidtlein, S.: Let your maps be fuzzy!—Class probabilities and floristic gradients as alternatives to crisp mapping for remote sensing of vegetation, Remote Sens. Ecol. Conserv., 7, 292–305, 2021.
- Feilhauer, H., and others: Brightness-normalized partial least squares regression for hyperspectral data, J. Quant. Spectrosc. Radiat. Transf., 111, 1947–1957, 2010.
- Funk, J. L., Larson, J. E., Ames, G. M., Butterfield, B. J., Cavender-Bares, J., Firn, J., Laughlin, D. C., Sutton-Grier, A. E., Williams, L., and Wright, J.: Revisiting the H oly G rail: using plant functional traits to understand ecological processes, Biological Reviews, 92, 1156–1173, 2017.
- Gal, Y. and others: Uncertainty in deep learning, 2016.
- Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 1050–1059, 2016.
- García-Soria, J. L., Morata, M., Berger, K., Pascual-Venteo, A. B., Rivera-Caicedo, J. P., and Verrelst, J.: Evaluating epistemic uncertainty estimation strategies in vegetation trait retrieval using hybrid models and imaging spectroscopy data, Remote Sens. Environ., 310, 114228, 2024.
- Gamon, J. A., Field, C. B., Goulden, M. L., Griffin, K. L., Hartley, A. E., Joel, G., Penuelas, J., and Valentini, R.: Relationships between NDVI, canopy structure, and photosynthesis in three Californian vegetation types, Ecol. Appl., 5, 28–41, 1995.
- Garg, A., Patil, A., Sarkar, M., Moorthi, S. M., and Dhar, D.: Advancements in data processing and calibration for the Hyperspectral Imaging Satellite (HySIS), arXiv preprint arXiv:2411.08917, 2024.
- Ge, X., Ding, J., Teng, D., Xie, B., Zhang, X., Wang, J., Han, L., Bao, Q., and Wang, J.: Exploring the capability of Gaofen-5 hyperspectral data for assessing soil salinity risks, Int. J. Appl. Earth Obs. Geoinf., 112, 102969, 2022.

Gitelson, A. A.: Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation, J. Plant Physiol., 161, 165–173, https://doi.org/10.1078/0176-1617-01176, 2004.

- Gorroño, J., Fomferra, N., Peters, M., Gascon, F., Underwood, C. I., Fox, N. P., Kirches, G., and Brockmann, C.: A radiometric uncertainty tool for the Sentinel-2 mission, Remote Sens. (Basel), 9, 178, 2017.

- Gorroño, J., Hunt, S., Scanlon, T., Banks, A., Fox, N., Woolliams, E., Underwood, C., Gascon, F., Peters, M., Fomferra, N., and others: Providing uncertainty estimates of the Sentinel-2 top-of-atmosphere measurements for radiometric validation activities, Eur. J. Remote Sens., 51, 650–666, 2018.
- Goryl, P., Fox, N., Donlon, C., and Castracane, P.: Fiducial reference measurements (FRMs): What are they?, Remote Sens. (Basel), 15, 5017, 2023.
- Graf, L. V., Gorroño, J., Hueni, A., Walter, A., and Aasen, H.: Propagating Sentinel-2 top-of-atmosphere radiometric uncertainty into land surface phenology metrics using a Monte Carlo framework, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 16, 8632–8654, 2023.
- Gravel, A., Laliberté, E., and Kalacska, M.: Foliar functional trait mapping of a mixed temperate forest using imaging spectroscopy, Federated Research Data Repository, 2024.
- Gustafsson, F. K., Danelljan, M., and Schön, T. B.: How reliable is your regression model's uncertainty under real-world distribution shifts?, arXiv preprint arXiv:2302.03679, 2023.
- Hank, T. B., Berger, K., Bach, H., Clevers, J. G. P. W., Gitelson, A., Zarco-Tejada, P., and Mauser, W.: Spaceborne imaging spectroscopy for sustainable agriculture: Contributions and challenges, Surv. Geophys., 40, 515–551, 2019.
- Hank, T., Locherer, M., Richter, K., Mauser, W., Consortium, E., and others: Neusling (Landau ad Isar) 2012—a multitemporal and multisensoral agricultural EnMAP Preparatory Flight Campaign, 2016.
- Hank, T., Richter, K., Mauser, W., Consortium, E., and others: Neusling (Landau ad Isar) 2009—an agricultural EnMAP Preparatory Flight Campaign using the HyMap instrument, 2015.
- Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., and Bonfil, D. J.: LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands, Remote Sens. Environ., 115, 2141–2151, 2011.
- Houborg, R., Fisher, J. B., and Skidmore, A. K.: Advances in remote sensing of vegetation function and traits, Int. J. Appl. Earth Obs. Geoinf., 2015.
- Hu, Y., Musielewicz, J., Ulissi, Z. W., and Medford, A. J.: Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials, Mach. Learn. Sci. Technol., 3, 45028, 2022.
- Jacquemoud, S. and Ustin, S.: Leaf optical properties, Cambridge University Press, 2019.
- Janet, J. P., Duan, C., Yang, T., Nandy, A., and Kulik, H. J.: A quantitative uncertainty metric controls error in neural network-driven chemical discovery, Chem. Sci., 10, 7913–7922, 2019.
- Joint Committee for Guides in Metrology (JCGM): Evaluation of measurement data—Guide to the expression of uncertainty in measurement, Int. Organ. Stand. Geneva ISBN, 50, 134, 2008.
- Jetz, W., Cavender-Bares, J., Pavlick, R., Schimel, D., Davis, F. W., Asner, G. P., Guralnick, R., Kattge, J., Latimer, A. M., and Moorcroft, P.: Monitoring plant functional diversity from space, Nat. Plants, 2, 1–5, 2016.

- Kampe, T. U., Johnson, B. R., Kuester, M. A., and Keller, M.: NEON: The first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure, J. Appl. Remote Sens., 4, 43510, 2010.
- Kattenborn, T., Fassnacht, F. E., and Schmidtlein, S.: Differentiating plant functional types using reflectance: Which traits make the difference?, Remote Sens. Ecol. Conserv., 5, 5–19, 2019.
- Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., and Dormann, C. F.: Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks, ISPRS Open J. Photogramm. Remote Sens., 5, 100018, 2022.
- 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.
- Khatami, R., Mountrakis, G., and Stehman, S. V.: Mapping per-pixel predicted accuracy of classified remote sensing images, Remote Sens. Environ., 191, 156–167, 2017.
- Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., Amengual, J., Basset, A., Van Bodegom, P. M., Cornelissen, J. H. C., and others: Towards global data products of Essential Biodiversity Variables on species traits, Nat. Ecol. Evol., 2, 1531–1540, 2018.
- Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, 2022.
- Koenker, R., and Hallock, K. F.: Quantile regression, J. Econ. Perspect., 15, 143–156, 2001.
- Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., and Goetz, A. F. H.: The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data, Remote Sens. Environ., 44, 145–163, 1993.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst., 30, 2017.
- Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., and Wegner, J. D.: Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles, Remote Sens. Environ., 268, 112760, 2022.
- Lavorel, S. and Garnier, É.: Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail, Funct Ecol, 16, 545–556, 2002.
- Levi, D., Gispan, L., Giladi, N., and Fetaya, E.: Evaluating and calibrating uncertainty prediction in regression tasks, Sensors, 22, 2022.
- Lewis, P., Gómez-Dans, J., Kaminski, T., Settle, J., Quaife, T., Gobron, N., Styles, J., and Berger, M.: An earth observation land data assimilation system (EO-LDAS), Remote Sens. Environ., 120, 219–235, 2012.

- Linnenbrink, J., Milà, C., Ludwig, M., and Meyer, H.: kNNDM CV: k-fold nearest-neighbour distance matching cross-validation for map accuracy estimation, Geosci. Model Dev., 17, 5897–5912, 2024.
- Liu, J. Z., and others: A simple approach to improve single-model deep uncertainty via distance-awareness, J. Mach. Learn. Res., 24, 42, 1–63, 2023.
- Liu, Y., Pagliardini, M., Chavdarova, T., and Stich, S. U.: The peril of popular deep learning uncertainty estimation methods, arXiv preprint arXiv:2112.05000, 2021.
- Martínez-Ferrer, L., Moreno-Martínez, Á., Campos-Taberner, M., García-Haro, F. J., Muñoz-Marí, J., Running, S. W., Kimball, J., Clinton, N., and Camps-Valls, G.: Quantifying uncertainty in high-resolution biophysical variable retrieval with machine learning, Remote Sens. Environ., 280, 113199, 2022.
- Mathieu, P.-P. and O'Neill, A.: Data assimilation: From photon counts to Earth System forecasts, Remote Sens. Environ., 112, 1258–1267, 2008.
- Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods Ecol. Evol., 12, 1620–1633, 2021.
- Mutanga, O., Masenyama, A., and Sibanda, M.: Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects, ISPRS J. Photogramm. Remote Sens., 198, 297–309, 2023.
- Myneni, R. B., Hall, F. G., Sellers, P. J., and Marshak, A. L.: The interpretation of spectral vegetation indexes, IEEE Trans. Geosci. Remote Sens., 33, 481–486, https://doi.org/10.1109/36.377948, 1995.
- Nagendra, H. and Rocchini, D.: High-resolution satellite imagery for tropical biodiversity studies: the devil is in the detail, Biodivers. Conserv., 17, 3431–3442, 2008.
- Nieke, J., Despoisse, L., Gabriele, A., Weber, H., Strese, H., Ghasemi, N., Gascon, F., Alonso, K., Boccia, V., Tsonevska, B., Choukroun, P., Ottavianelli, G., and Celesti, M.: The Copernicus Hyperspectral Imaging Mission for the Environment (CHIME): an overview of its mission, system, and planning status, in: Sensors, Systems, and Next-Generation Satellites XXVII, 1272909, 2023.
- Ollinger, S. V.: Sources of variability in canopy reflectance and the convergent properties of plants, New Phytol., 189, 375–394, 2011.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., et al.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst., 32, 2019.
- Palmer, G., Du, S., Politowicz, A., Emory, J. P., Yang, X., Gautam, A., et al.: Calibration after bootstrap for accurate uncertainty quantification in regression models, npj Comput. Mater., 8, 115, 2022.
- ~~Padarian, J., Minasny, B., & McBratney, A. B. (2022). Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout. *Geoderma, 425*, 116063.~~Padarian, J., Minasny, B., and McBratney, A. B.:

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout, Geoderma, 425, 116063, https://doi.org/10.1016/j.geoderma.2022.116063, 2022.

- Papacharalampous, G., and others: Uncertainty estimation of machine learning spatial precipitation predictions from satellite data, Mach. Learn. Sci. Technol., 5, 035044, 2024.
- Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., and others: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 4540, 2020.
- Pottier, J., Malenovský, Z., Psomas, A., Homolová, L., Schaepman, M. E., and others: Modelling plant species distribution in alpine grasslands using airborne imaging spectroscopy, Biol. Lett., 10, 20140347, 2014.
- Pullanagari, R. R., Dehghan-Shoar, M., Yule, I. J., and Bhatia, N.: Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network, Remote Sens. Environ., 257, 112353, 2021.
- Rahaman, R.: Uncertainty quantification and deep ensembles, Adv. Neural Inf. Process. Syst., 34, 20063–20075, 2021.
- ~~Reich PB (2014) The world-wide 'fast–slow' plant economics spectrum: a traits manifesto. J Ecol 102:275–301~~Reich, P. B.: The world-wide "fast–slow" plant economics spectrum: a traits manifesto, J. Ecol., 102, 275–301, https://doi.org/10.1111/1365-2745.12211, 2014.

~~Richards, J. A., Richards, J. A., and others: Remote sensing digital image analysis, Springer, 2022.~~

- Rogers, A., Serbin, S., and Ely, K.: Leaf mass area, leaf carbon and nitrogen content, Kougarok Road and Teller Road, Seward Peninsula, Alaska, 2016, 2021.
- Sakschewski, B., von Bloh, W., Boit, A., Rammig, A., Kattge, J., Poorter, L., Peñuelas, J., and Thonicke, K.: Leaf and stem economics spectra drive diversity of functional plant traits in a dynamic global vegetation model, Glob. Chang. Biol., 21, 2711–2725, 2015.
- Savitzky, A. and Golay, M. J. E.: Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem., 36, 1627–1639, 1964.
- Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., and Green, W. H.: Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction, J. Chem. Inf. Model., 60, 2697–2717, 2020.
- Schiefer, F., Schmidtlein, S., and Kattenborn, T.: The retrieval of plant functional traits from canopy spectra through RTM-inversions and statistical models are both critically affected by plant phenology, Ecol. Indic., 121, 107062, 2021.
- Sellers, P. J.: Canopy reflectance, photosynthesis, and transpiration, Int. J. Remote Sens., 6, 1335–1372, 1985.

Formatted: Font: (Default) +Headings (Times New Roman),

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Formatted: Font: (Default) +Headings (Times New Roman),

- Serbin, S. P., Wu, J., Ely, K. S., Kruger, E. L., Townsend, P. A., Meng, R., Wolfe, B. T., Chlus, A., Wang, Z., and Rogers, A.: From the Arctic to the tropics: multibiome prediction of leaf mass per area using leaf reflectance, New Phytol., 224, 1557–1568, 2019.
- Serbin, S. P., and Townsend, P. A.: Scaling functional traits from leaves to canopies, in: Remote Sensing of Plant Biodiversity, pp. 43–82, 2020
- Silvan-Cardenas, J. L. and Wang, L.: Sub-pixel confusion–uncertainty matrix for assessing soft classifications, Remote Sens. Environ., 112, 1081–1195, 2008.
- Singh, A., Serbin, S. P., McNeil, B. E., Kingdon, C. C., and Townsend, P. A.: Imaging spectroscopy algorithms for mapping canopy foliar chemical and morphological traits and their uncertainties, Ecol. Appl., 25, 2180–2197, 2015.

~~Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.~~ Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

- Steltzer, H. and Welker, J. M.: Modeling the effect of photosynthetic vegetation properties on the NDVI–LAI relationship, *Ecology*, 87, 2765–2772, https://doi.org/10.1890/0012-9658(2006)87[2765:MTEOPV]2.0.CO;2, 2006.
- ~~Ting, Z., Ye, Z., Singh, A., Desai, A. R., Krishnayya, N. S. R., Dave, M. G., and Townsend, P. A.: Plot-level traits and their corresponding image spectra from the Western Ghats of India, Ecological Spectral Information System (EcoSIS), 2023.~~
- Van Cleemput, E., Roberts, D. A., Honnay, O., and Somers, B.: A novel procedure for measuring functional traits of herbaceous species through field spectroscopy, Methods Ecol. Evol., 10, 1332–1338, 2019.
- Van Bodegom, P. M., Douma, J. C., and Verheijen, L. M.: A fully traits-based approach to modeling global vegetation distribution, Proc. Natl. Acad. Sci. USA, 111, 13733–13738, 2014.
- Wacker, A. G. and Landgrebe, D. A.: Minimum distance classification in remote sensing, LARS Tech. Rep., 25, 1972.

~~Wang, Q., Putri, N. A., Gan, Y., and Song, G.: Combining both spectral and textural indices for alleviating saturation problem in forest LAI estimation using Sentinel-2 data, *Geocarto Int.*, 37, 10511–10531, 2022.~~

- Wang, Z., Chlus, A., Geygan, R., Ye, Z., Zheng, T., Singh, A., and others: Foliar functional traits from imaging spectroscopy across biomes in eastern North America, New Phytol., 228, 494–511, 2020.
- Wang, Z., Townsend, P. A., Schweiger, A. K., Couture, J. J., Singh, A., Hobbie, S. E., and Cavender-Bares, J.: Mapping foliar functional traits and their uncertainties across three years in a grassland experiment, Remote Sens. Environ., 221, 405–416, 2019.
- Wang, D. R., Wolfrum, E. J., Virk, P., Ismail, A., Greenberg, A. J., and McCouch, S. R.: Robust phenotyping strategies for evaluation of stem non-structural carbohydrates (NSC) in rice, J. Exp. Bot., 67, 6125–6138, 2016.

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight

- Wang, Q., Adiku, S., Tenhunen, J., and Granier, A.: On the relationship of NDVI with leaf area index in a deciduous forest site, Remote Sens. Environ., 94, 244–255, 2005.
- Wocher, M., Berger, K., Danner, M., Mauser, W., and Hank, T.: Physically-based retrieval of canopy equivalent water thickness using hyperspectral data, Remote Sens. (Basel), 10, 1924, 2018.
- Widlowski, J.-L.: Conformity testing of satellite-derived quantitative surface variables, Environ. Sci. Policy, 51, 149–169, 2015.

Xu, D., An, D., and Guo, X.: The impact of non-photosynthetic vegetation on LAI estimation by NDVI in mixed grassland, *Remote Sens.*, 12, 1979, https://doi.org/10.3390/rs12121979, 2020.

~~Yang, S., and Yee, K.: Towards reliable uncertainty quantification via deep ensemble in multi-output regression task, Eng. Appl. Artif. Intell., 132, 107871, 2024.~~

~~Zeevi, T., Venkataraman, R., Staib, L. H., and Onofrey, J. A.: Monte Carlo frequency dropout for predictive uncertainty estimation in deep learning, Proc. IEEE Int. Symp. Biomed. Imaging (ISBI), 1–5, 2024.~~

- Zheng, T., Ye, Z., Singh, A., Desai, A. R., Krishnayya, N. S. R., Dave, M. G., and Townsend, P. A.: Plot-level traits and their corresponding image spectra from the Western Ghats of India, Ecological Spectral Information System (EcoSIS), 2023.
- Zheng, G., and Moskal, L. M.: Retrieving leaf area index (LAI) using remote sensing: theories, methods and sensors, Sensors, 9, 2719–2745, 2009.

Formatted: Font: (Default) +Headings (Times New Roman), Not Highlight