Supplement of Earth Syst. Sci. Data, 18, 5895–5913, 2026
https://doi.org/10.5194/essd-18-5895-2026-supplement

Open Access Earth System Science Data

Supplement of

A global dataset of forest disturbance regimes derived from satellite biomass observations

Siyuan Wang et al.

Correspondence to: Siyuan Wang (siyuan.wang@bgc-jena.mpg.de) and Nuno Carvalhais (nuno.carvalhais@bgc-jena.mpg.de)

Supplement

S1. The Simulation Framework: Linking Disturbance and Biomass

Building on our previous work (Wang et al., 2024), we use a forward-modeling framework to quantify the link between disturbance regime and the resulting biomass statistics. The core idea is to first simulate how a wide array of known, parameterized disturbance regimes manifest as synthetic, spatially explicit patterns of aboveground biomass. By systematically generating these cause-and effect pairings, we can establish a statistical link that later allows us to retrieve the problem: inferring the characteristics of a known disturbance regime from observable biomass patterns.

The workflow proceeds in four main stages: 1) we parameterize a disturbance regime using a set of key attributes that control the disturbance rate, gap-size distribution, disturbance severity and background state of forest mortality; 2) we stochastically generate multi-year sequences of disturbance events on a simulated landscape, where the size, location, and shape of each event are governed by the defined regime parameters; 3) we apply these disturbance sequences to a simple carbon cycle model to simulate the dynamic change in AGB across the landscape over centuries. 4) once the simulated landscape reaches a dynamically stable state, that is, a shifting mosaic of different successional stages, a comprehensive suite of statistical metrics from the AGB map is extracted, which quantitatively describes the emergent biomass pattern. This forward-modeling approach creates a large synthetic dataset where known disturbance regime parameters are linked to unique biomass pattern signatures, forming the basis for training a predictive model.

S1.1 Parameterization of disturbance regime

We characterize the disturbance regime acting upon a landscape using four key parameters, three of them—disturbance rate (μ), gap-size distribution (α), and disturbance severity (β)—are adapted from our previous framework (Wang et al., 2024), but with a significant modification to the parameter range for α. Specifically, we now include lower values for the clustering degree, which allows for the simulation of fewer, larger, and more contiguous events. This adjustment is crucial for representing large-scale, intense disturbances such as stand-replacing fires. Another major advancement in our parameterization is the explicit inclusion of the background mortality rate (K_b) as a fourth independent parameter to be predicted, allowing us to disentangle the effects of baseline mortality from episodic disturbances and address a potential source of equifinality in the previous framework. The four parameters are detailed in Table 1.

Table S1. Parameters Defining the Simulated Disturbance Regimes

Parameter	Symbol	Definition	Proposed Range	Intervals	Count	Role in Simulation
Disturbance Rate	μ	The mean fraction of the total domain area affected by disturbance events annually.	0.01 - 0.05	0.005	9	Controls the overall frequency and extent of disturbances.
Gap-size Distribution	α	The scaling exponent in the power-law function ($n \propto z^{-\alpha}$) relating the number of events (n) to their size (z).	0.5 - 1.8	0.05	27	Governs the spatial pattern of disturbance, where increasing values shift the pattern from large, contiguous events to many small, fragmented disturbances.
Disturbance Severity	β	The slope of the relationship defining the fraction of AGB	0.03 - 0.5	0.01/0.05 /0.1	14	Determines the severity of disturbance events.

		lost as a function of disturbance event size.				
Background Mortality	K_b	The constant, first-order mortality rate representing baseline carbon loss process like litterfall, root exudates, and herbivory.	0.025 - 0.2	0.025	8	Controls the background carbon turnover time ($\tau = \frac{1}{K_h}$) and influences the maximum potential biomass.

S1.2 Simulating Spatially-Explicit Disturbance Events

To translate our parameterized disturbance regime into concrete spatial patterns, we developed an advanced disturbance event generator. This tool is a key advancement from our previous framework, which was limited to computationally efficient but unrealistic rectangular shapes. The new generator is capable of producing disturbance events with a wide range of morphological complexities, allowing us to enhance the realism of our simulations and test the influence of event geometry on emergent biomass patterns.

For each combination of disturbance regime parameters, the generator creates a set of 200 annual disturbance maps on a 1000 × 1000 pixel grid, with event shapes corresponding to one of the six settings detailed in Table 2. These 200 maps are temporally independent, representing a stochastic sequence of events over time. This design choice—the absence of temporal autocorrelation—is crucial as it allows the set of maps to be shuffled, enabling multiple, unique simulation runs for the same underlying disturbance regime.

Table S2. Disturbance Event Shape Settings

Category	Shape Setting	Description
Simple Regular	Rectangle	The baseline setting from (Wang et at., 2024); all events are four-sided rectangles.
	Triangle	All events are simple three-sides polygons.
	Circle	All events are uniform, non-directional circular patches.
Complex Convex	Gradient	Event complexity (number of sides) is proportional to event size; larger events have more sides.
	Complex	All events are generated with maximum complexity (49 sides), regardless of their size.
	Random	The complexity of each event is a random integer of sides between 3 and 49.

The complete spatiotemporal output of the generator for a single disturbance regime is stored in a disturbance reference cube. This three-dimensional array (1000 × 1000 pixels × 200 years) serves as the precise blueprint of all disturbance locations, sizes, and shapes over the entire simulation period. A unique disturbance reference cube is generated for every combination of the six shape settings and the disturbance regime parameters (μ, α and β). This systematic approach allows us to isolate the impact of event morphology on emergent biomass patterns. This methodology tests the robustness of our framework, with the hypothesis that landscape-level biomass statistics are more sensitive to the fundamental regime parameters than to the specific geometry of individual disturbance patches.

S1.3 Generating Synthetic Biomass Statistics with a Carbon Cycle Model

The carbon modeling workflow is conceptually identical to our previous framework. Each disturbance reference cube is used to drive a simple, dynamic carbon cycle model at the pixel level. The annual change in aboveground biomass is governed by the balance between carbon gains and losses:

$$\frac{dAGB}{dt} = NPP_{AGB} - L_{total}$$

where the total loss, L_{total}, is partitioned into episodic disturbance loss (L_d) and continuous background mortality ($L_b = AGB \times K_b$).

The simulation experiment was designed to be comprehensive. Our full factorial design included every combination of the four disturbance regime parameters (9 μ values × 27 α values × 14 β values × 8 Kb values), five different levels of primary productivity (Photosynthetic capacity), and all six disturbance shape settings. To account for stochasticity, each of these scenarios was replicated 10 times using a different random shuffle of the annual disturbance maps, resulting in a total of 8,164,800 individual simulation runs.

Simulations are initialized from bare ground (zero AGB); because the landscape is allowed 200 years to reach a dynamically stable equilibrium, these initial conditions do not influence the final statistical extraction. To characterize this stable state, we extracted the Gross Primary Production (GPP) from the final simulation year and the average aboveground biomass (AGB) map over the last 10 years. It is from this 10-year average AGB map that we derived a comprehensive vector of spatial statistics to serve as the quantitative signature of the biomass pattern. The computational methodology used to extract this suite of metrics, encompassing first-order distribution statistics and second-order texture metrics from a Gray-Level Co-occurrence Matrix (GLCM), is identical to the algorithm applied to the observed biomass data described in the following section, ensuring a consistent basis for comparison.

The final output is a massive dataset linking each unique set of input parameters to a corresponding vector of output features (Table S3, 16 biomass pattern statistics plus the average GPP). This dataset provides the foundation for training a machine learning algorithm to infer disturbance regimes from biomass patterns.

Table S3. Features of Biomass and GPP used to train the Random Forest model

Type	Feature	Meaning
Histogram Features	Mean	Mean biomass value of domain
	Median	Median biomass value of domain
	Variance	Statistical Variance of biomass values in the domain
	Standard Deviation	Statistical deviation of biomass values in the domain
	Coefficient of Variation	Ratio of the standard deviation to the mean of biomass values
	Skewness	Measure of the asymmetry of the biomass

		value distribution
	Kurtosis	Measure of the weight of the tails or the sharpness of the central peak of the biomass value distribution
	Percentile 25%	The 25th percentile biomass value of the domain (P25)
	Percentile 75%	The 75th percentile biomass value of the domain (P75)
	Range	Distance between 90th percentile biomass and 20th percentile biomass (P90 - P20)
	Trimean	The Tukey's trimean of the biomass values, calculated as (P25 + 2*Median + P75)/4
Informative Feature	Shannon Entropy	A measure of the diversity or uncertainty in the distribution of biomass values
Texture Features	Contrast	Measures how "sharp" or "different" neighboring biomass values are. (High contrast = big jumps between neighbors)
	Correlation	Measures how related neighboring biomass values are. (High correlation = neighbors are usually very similar)
	Energy	Measures how "uniform" or "orderly" the biomass pattern is. (High energy = a very simple, repetitive pattern)
	Homogeneity	Measures the "smoothness" of the biomass pattern. (High homogeneity = most neighbors have very similar values)
Photosynthesis Feature	GPP	Mean Gross Primary Production at the end of the simulation

S2. Determining the Optimal Aggregation Scale

S2.1 Mismatch between Model Simulation and EO

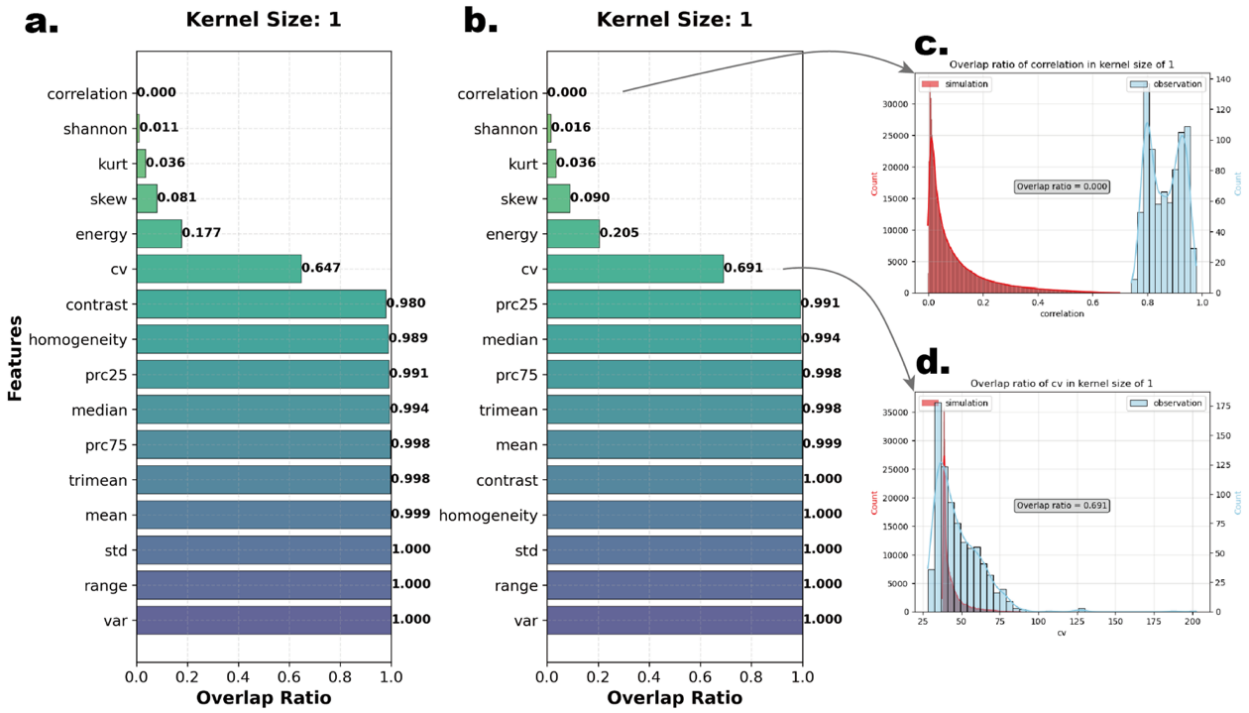


Figure S1. Statistical overlap between simulated and observed biomass feature distributions. The panels compare the percentage of overlap for 17 statistical features derived from (a) the original forward modeling framework and (b) the improved framework, which incorporates an expanded parameter space and non-rectangular disturbance shapes (see Supplementary S1 for details). The overlap percentage is calculated based on the intersecting range between the frequency distributions of simulated and observed values. Panels (c) and (d) show the specific distributions for the two most important predictive features identified in Wang et al. (2024): GLCM Correlation and Coefficient of Variation, respectively.

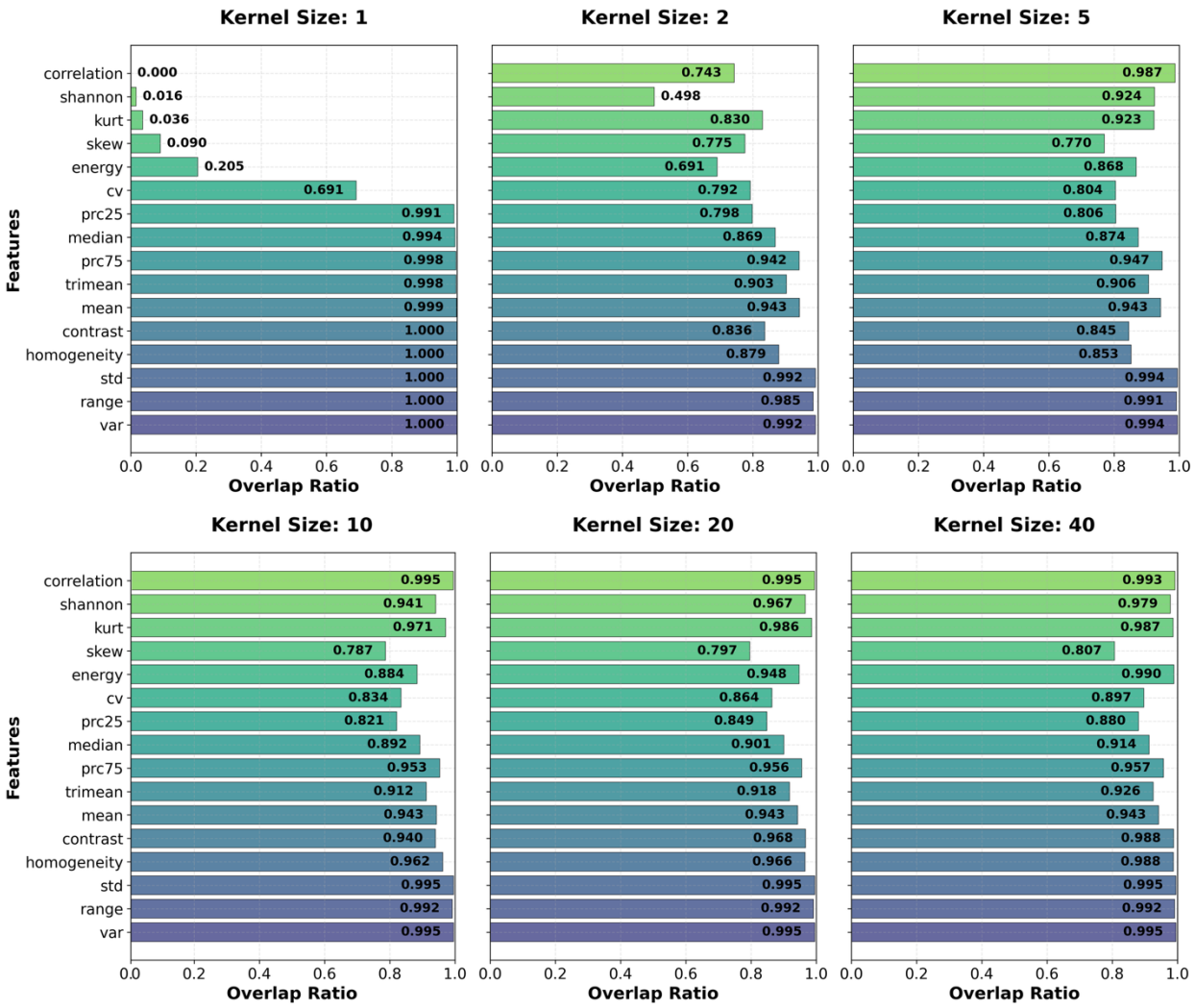


Figure S2. The effect of spatial aggregation on the statistical overlap between simulated and observed biomass features. The figure illustrates how statistical overlap changes as a function of aggregation scale, which is represented by the kernel size used for averaging (larger kernels correspond to coarser spatial resolutions). The results demonstrate a clear positive trend: as the level of aggregation increases, the statistical overlap between the simulated and observed datasets consistently improves. This confirms that spatial aggregation is an effective strategy for reducing the discrepancy between the two domains, an effect that is particularly pronounced for key predictive features such as GLCM Correlation and the Coefficient of Variation (CV).

S2.2 Weighted Overlap Ratio

To objectively identify the optimal aggregation scale, we developed the Feature Importance Weighted Overlap Ratio (WOR), a robust metric that quantifies the similarity between the multi-dimensional feature spaces of the simulated and observed datasets. The WOR calculation prioritizes the most influential features (Correlation, Kurtosis, Skewness, and CV) by assessing their pairwise overlap and weighting the result by their combined, scale-dependent feature importance ($FI'_{i,j}$). The pairwise overlap ratio ($OR_{i,j}$) for any two feature distributions, $\mathrm{p}(\mathrm{x},\mathrm{y})$and $\mathrm{q}(\mathrm{x},\mathrm{y})$, is calculated as the intersecting volume of their probability density function:

$$OR_{i,j} = \iint min(p(x,y), q(x,y))dxdy$$

The final WOR score is a single value between 0 and 1 that measures statistical alignment, calculated as:

$$\mathrm{WOR} = \sum_{\mathrm{i,j}} \mathrm{FI}'_{\mathrm{i,j}} \times \mathrm{OR}_{\mathrm{i,j}}$$

A higher WOR values signifies greater similarity and thus higher confidence in the model's predictive capability at that scale.

S2.3 Optimal Aggregation Scale

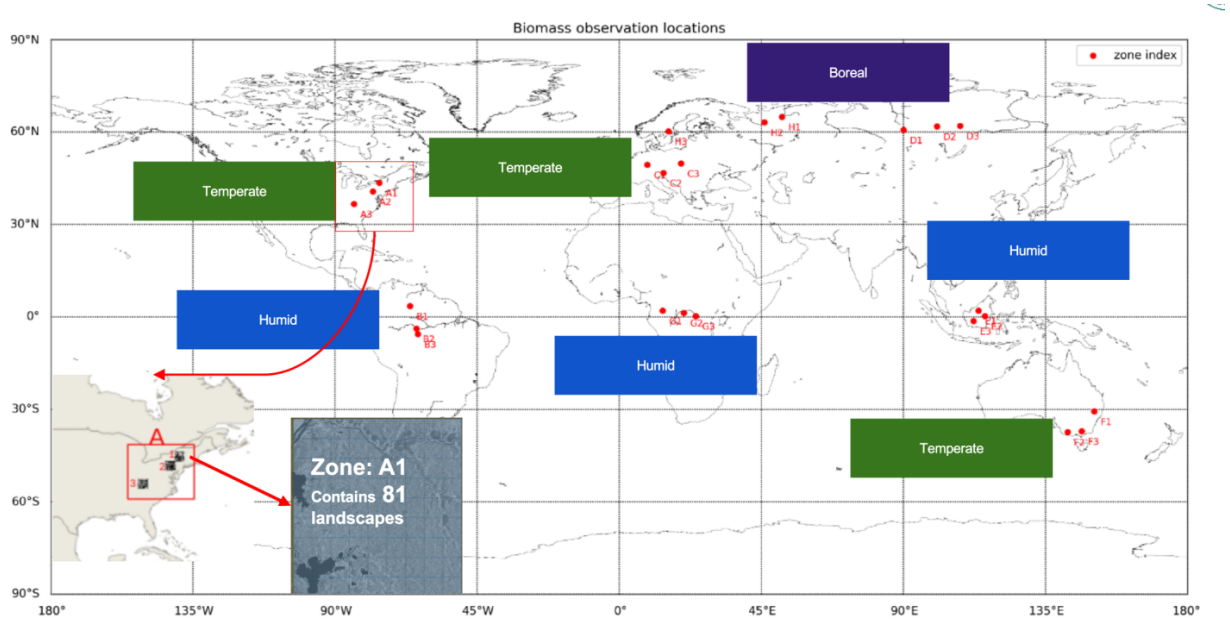


Figure S3. Global distribution of biomass observation locations for optimal aggregation analysis. The map displays the locations of the 24 globally distributed zones randomly selected for the multi-level overlap analysis, categorized by biome. Each zone (e.g., A1, inset) is composed of 81 individual landscapes, where each landscape consists of a 1000×1000 pixel grid corresponding to approximately 25×25 km at the equator. This design forms the hierarchical structure used for evaluating the statistical discrepancy between simulated and observed biomass features at the zone, region, biome, and global levels.

The optimal aggregation scale for generating global product was determined through a comprehensive, multi-level sensitivity analysis using the Weighted Overlap Ratio (WOR). This was conducted across a nested hierarchy of observational levels, based on a global distribution of landscape zones sampled from Boreal, Temperate, and Humid Biomes **(Figure S3)**. The WOR was calculated at each level of the hierarchy by progressively expanding the observational pool: from individual zones to regions (e.g., Region A= pool of A1, A2, A3), to biomes, and finally to the entire global dataset.

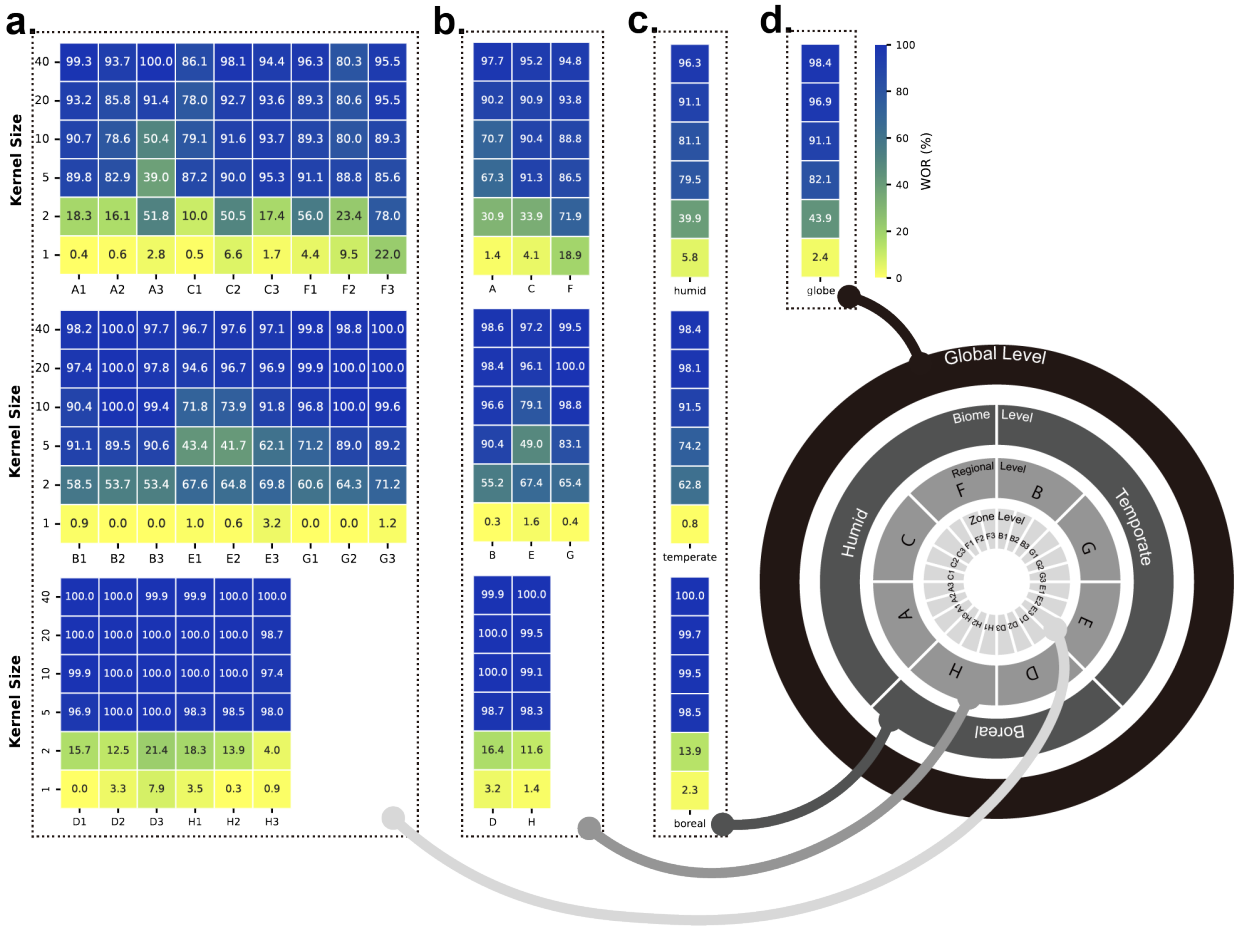


Figure S4. Multi-scale Weighted Overlap Ratio (WOR) analysis across hierarchical spatial domains. The figure displays WOR heatmaps for four distinct observational levels: (a) individual landscape zones (A1-H3), (b) regional groups (e.g., A, B, C), (c) biome groups (Humid, Temperate, Boreal), and (d) the global scale. For each heatmap, the rows represent different aggregation kernel sizes, and the columns represent the different spatial domains (e.g., individual zones, regions). Color intensity represents the WOR percentage (0-100%), indicating the statistical similarity between observed and simulated datasets when both are processed with the same aggregation scale. The conceptual diagram (bottom right) illustrates the hierarchical spatial organization from landscape zones through regional and biome aggregations to global scale.

The result of the hierarchical WOR analysis **(Figure S4)** reveals several key findings. First, a consistent trend was observed across all landscape zones: as the degree of aggregation increases, the WOR value improves, highlighting the general effectiveness of this strategy in narrowing the simulation-observation gap. For most zones, an aggregation with a kernel size of 10 is sufficient to achieve a WOR greater than 90%, indicating good statistical consistency. Second, this trend holds at higher spatial levels (region, biome, and global), confirming the value of aggregation. Specifically, boreal ecosystems consistently exhibit a relatively higher WOR compared to temperate and humid biomes at equivalent aggregation scales. Third, by setting a target threshold of 90% for the WOR and considering that prediction accuracy is highest at smaller kernel sizes, a kernel size of 10 emerges as the optimal choice for global-scale prediction. While this represents the global optimum, the framework allows for this scale to be adjusted for specific regional or biome-level analyses.

S3. Uncertainty Assessment based on Meyer and Pebesma's framework

S3.1 Theoretical Formulation

The DI metric quantifies the relative dissimilarity of a prediction point to the training data space:

$$DI_k = \frac{D_I}{D_{AVG}}$$

where DI_k is the minimum weighted Euclidean distance from the prediction point to any training sample, D_{AVG} is the average weighted distance between training samples. The weighting is based on feature importance derived from aggregated Random Forest models, ensuring that more informative features contribute more to the dissimilarity calculation.

For a prediction point x_p and a set of training sample vectors $\{x_i\}_{i=1}^{n}$, the weighted distance is computed as:

$$D_I = \min_i \sqrt{\sum_{j=1}^{d} w_j \left(x_{p,j} - x_{i,j}\right)^2}$$

where w_j represents the importance weight for feature j,

The D_{AVG} (Distance Average) computes the average pairwise distances between training samples to establish a baseline for the DI metric:

$$D_{AVG} = \frac{1}{n(n-1)/2} \sum_{i<k} \sqrt{\sum_{j=1}^{d} w_j \left(x_{i,j} - x_{k,j}\right)^2}$$

S3.2 Scalable Implementation Workflow

Directly applying the DI formulation described in Section S3.1 to large-scale datasets is computationally prohibitive due to the quadratic complexity of calculating the pairwise distance matrix for the full training dataset (T). To address this challenge, we develop a scalable workflow that decouples the calculation into an efficient pre-calculation for each prediction point. This process involves three key steps: data standardization, baseline dissimilarity pre-calculation, and the final DI computation.

To ensure that all features contribute proportionally to the distance metric and to mitigate the influence of extreme outliers, each feature j is independently standardized, we use a robust percentile-based normalization where each value $x_{i,j}$ of a sample i for a feature j is scaled to the range defined by the 1st and 99th percentiles of that feature in the training data:

$$x'_{i,j} = \frac{x_{i,j} - P_1(j)}{P_{99}(j) - P_1(j)}$$

where $x'_{i,j}$ is the normalized feature value, and $P_{99}(j)$ and $P_1(j)$ are the 99th and 1st percentiles of feature j across the training set T, respectively. All subsequent calculations are performed on these normalized features.

The average pairwise distance between all samples in the training data, D_{avg}, serves as a stable baseline for the DI metric. To compute this efficiently, we pre-calculate it on a representative and computationally tractable subset of the training data. First, a small subset $T_s \subset T$ is sampled (typically 0.1% of the full dataset) to maintain statistical representativeness while ensuring computational feasibility. For each feature j , the unweighted average distance, $D_{avg}(j)$, is computed across all pairs of points within this subset:

$$D_{avg}(j) = \frac{1}{\binom{n_s}{2}} \sum_{i=1}^{n_s-1} \sum_{k=i+1}^{n_s} \left| x'_{i,j} - x'_{k,j} \right|$$

where n_s is the number of samples in the subset T_s. These single-feature average distances are pre-calculated and stored for each cross-validation fold, forming the building blocks for the final weighted baseline dissimilarity, D_{avg}. With the component values pre-calculated, the final DI for a new prediction point (observed statistical features from a realistic landscape), x_{pred}, is computed efficiently.

First, the weighted average distance $\bar{d}$ is assembled by combining the pre-calculated single-feature distances with their corresponding feature importance weights, w_j, derived from the trained Random Forest model:

$$\bar{d} = \sum_{j=1}^{d} w_j \cdot D_{avg}(j)$$

where d is the total number of features, 17 in this study.

Next, the minimum weighted Euclidean distance, d_k, from the new prediction point x_{pred} to any sample in the full training set T is calculated:

$$d_k(x_{pred}, \mathrm{T}) = \min_{i \in T} \left(\sqrt{\sum_{j=1}^{d} w_j (x'_{pred,j} - x'_{i,j})^2} \right)$$

Finally, the Dissimilarity Index for the prediction point is computed as the ratio of these two values:

$$DI(x_{pred}) = \frac{d_k(x_{pred})}{\bar{d}}$$

This scalable workflow enables the practical application of the DI framework to extensive datasets by strategically avoiding the most computationally expensive operation while preserving the theoretical integrity of the uncertainty metric.

S4. Predictions

S4.1 Sub-grid Sampling Density and Grid Mapping

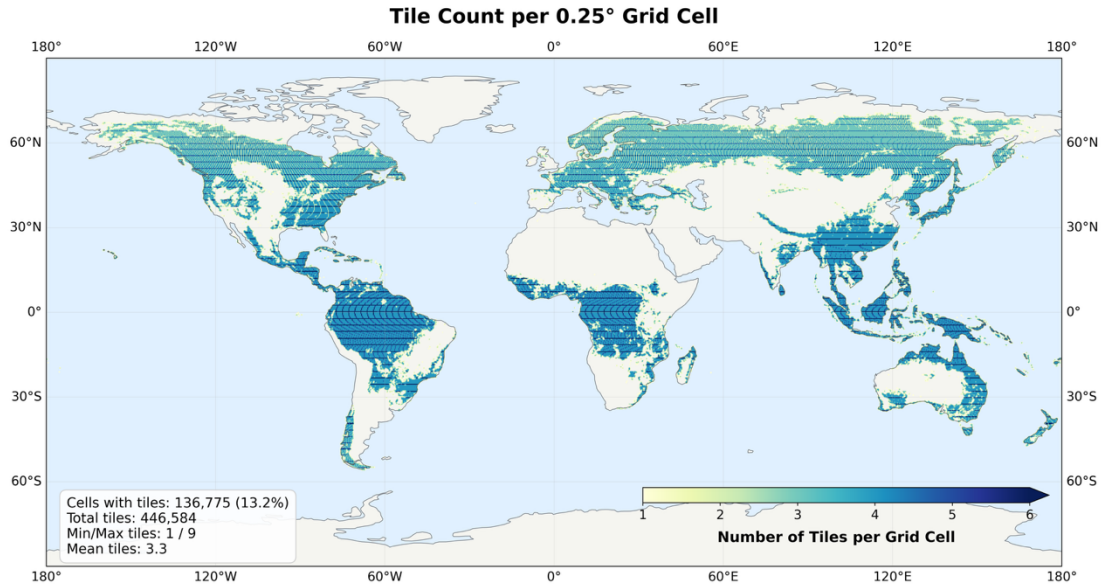


Figure S5. Global map of the number of 25 km x 25 km landscape tiles aggregated into each 0.25° grid cell. This tile_count layer serves as an indicator of sampling density, showing the number of underlying tile-level predictions used to calculate each grid cell value. Higher values indicate that the gridded cell value is derived from a more robust spatial sample. The regular, wave-like patterns of high overlap naturally occur when fixed equal-area landscapes (25 km × 25 km) are projected onto a 0.25° geographic grid, causing periodic boundary overlaps as the grid cells physically shrink toward the poles.

S4.2 Consistency Across Baseline Biomass Products

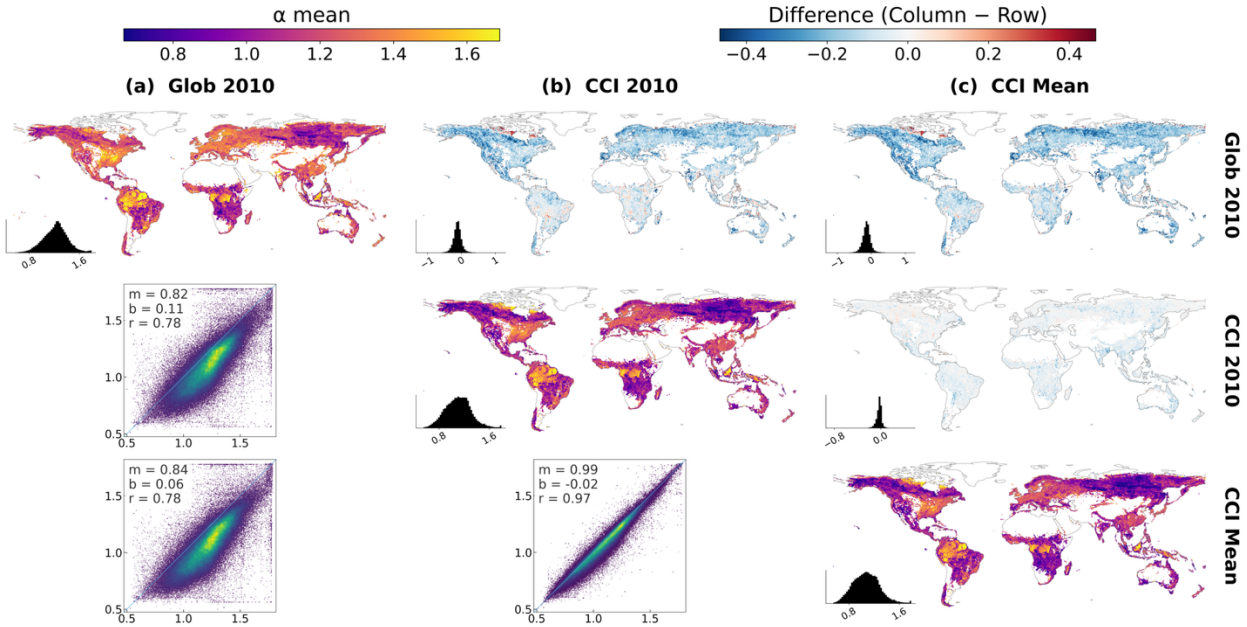


Figure S6. Spatial cross comparison of retrieved α derived from different baseline biomass products. The diagonal panels display the global spatial distributions of α predictions gridded at 0.25° x 0.25° resolution derived from (a) GlobBiomass (Glob 2010), (b) ESA CCI Biomass (CCI 2010), (c) the multi-year mean of ESA CCI Biomass (CCI Mean, 2005-2011), with a unified color bar. The upper off-diagonal panels display the spatial difference maps between each corresponding pair of datasets (column dataset minus row dataset) to illustrate localized discrepancies, using a separate symmetrical color bar centered at 0. The bottom off-diagonal panels present the pixel-to-pixel scatter plots between each pair of products, and annotated statistics including the regression slope (m), intercept (b), and correlation coefficient (r). Hereafter, all figures comparing different spatial maps include the information in a similar manner.

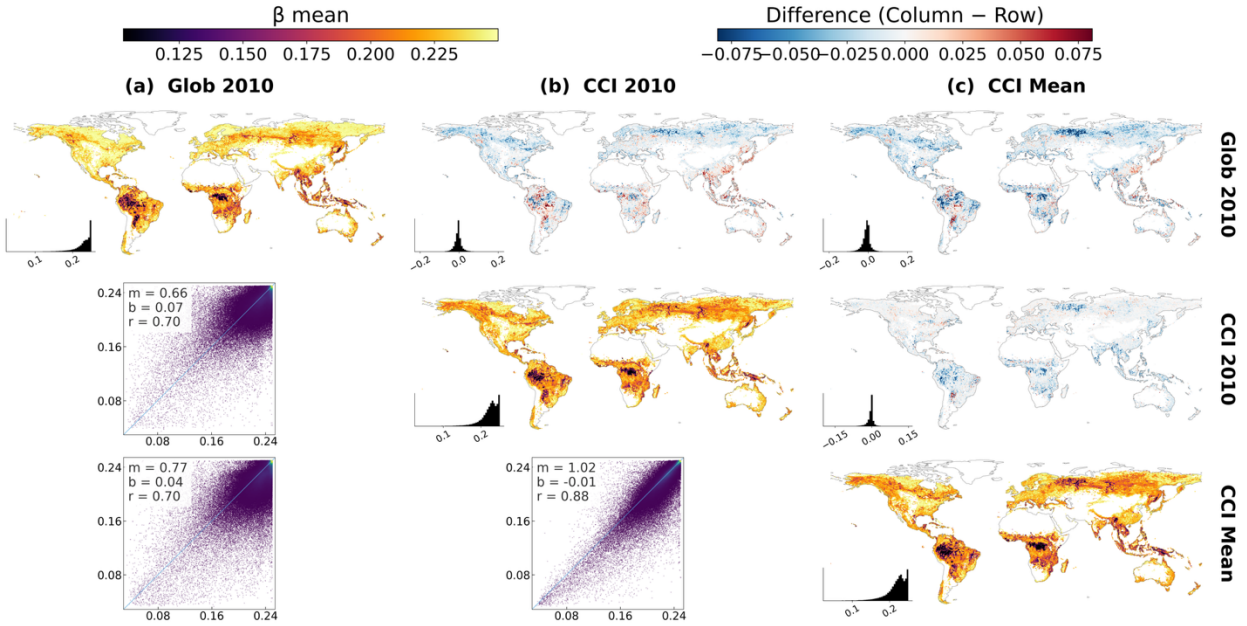


Figure S7. Spatial cross comparison of retrieved β derived from different baseline biomass products.

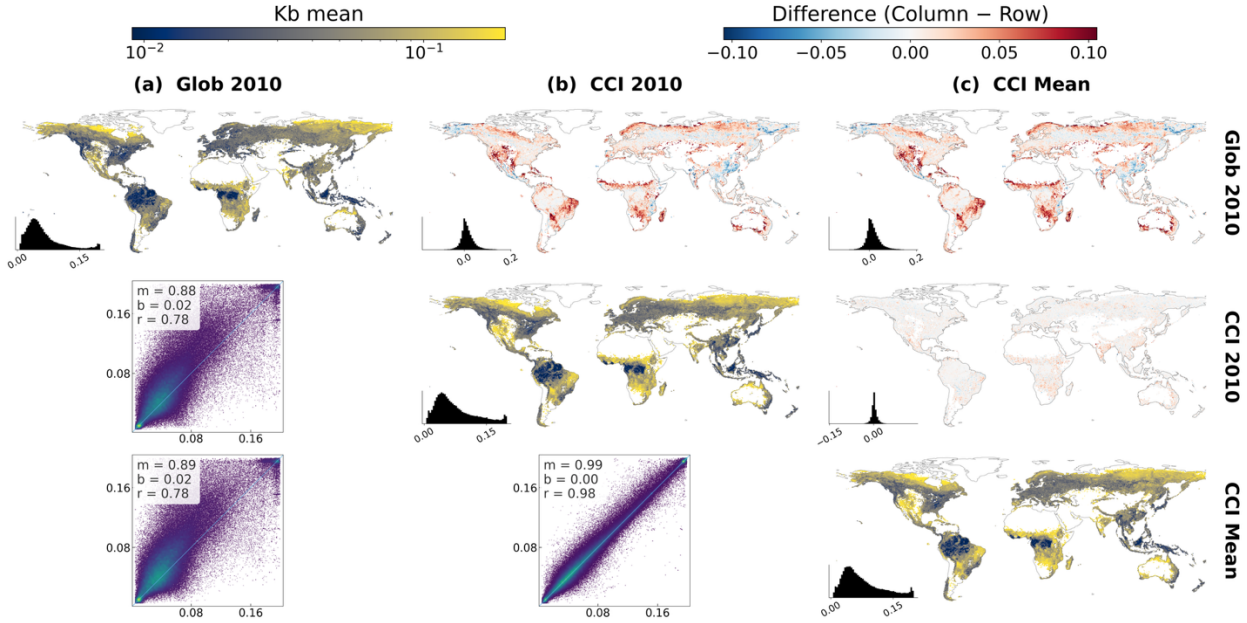


Figure S8. Spatial cross comparison of retrieved K_b derived from different baseline biomass products.

S4.3 Single-Year versus Multi-Year GPP

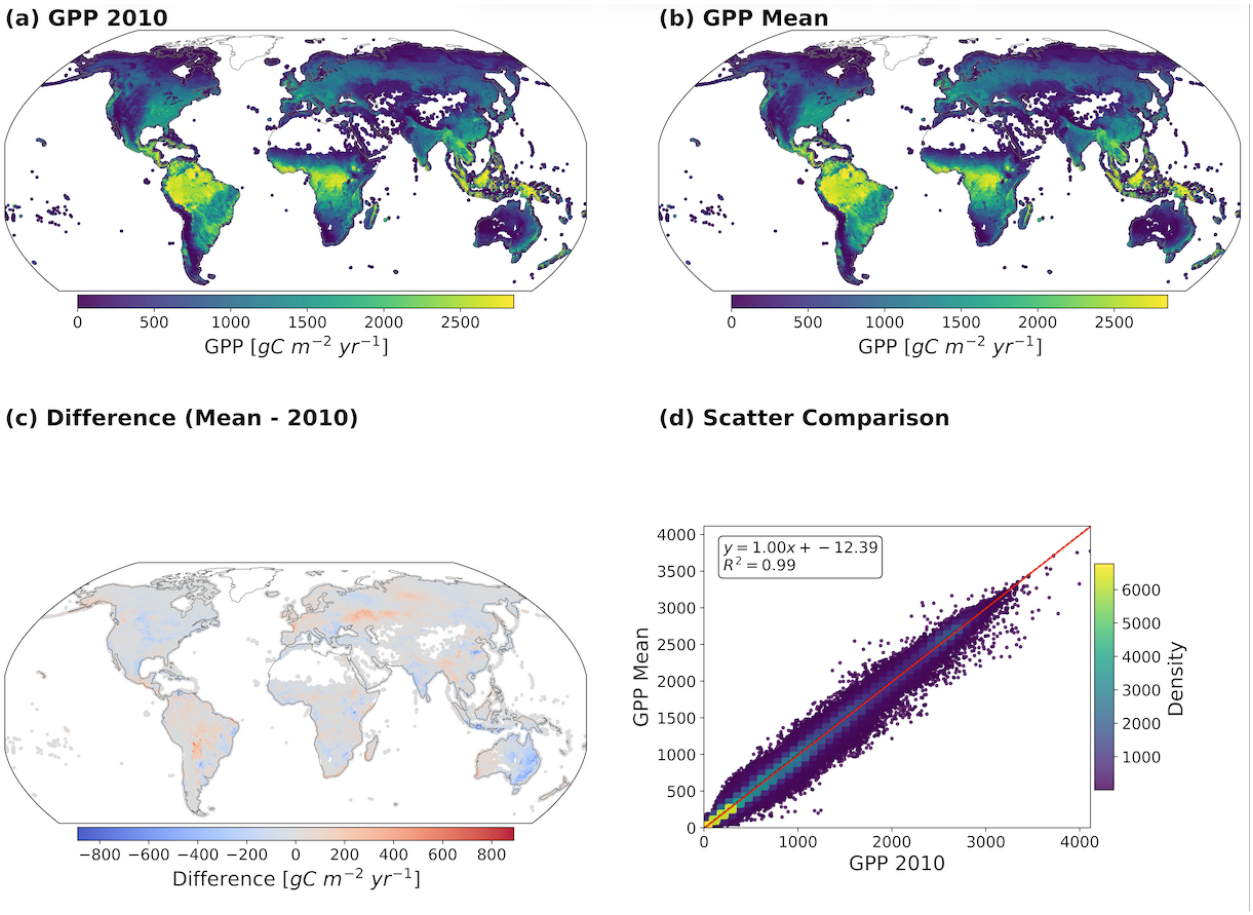


Figure S9. Spatial comparison of FLUXCOM-X GPP between the single year 2010 and the 2001–2010 multi-year mean. The figure displays (a) the spatial distribution of GPP for the year 2010, (b) the multi-year mean GPP averaged from 2001 to 2010, (c) the spatial difference map between the two datasets, and (d) the corresponding pixel-to-pixel scatter density plot. The minimal spatial deviations (c) and strong correlation (d) across the analyzed forest landscapes demonstrate that the macro-scale spatial gradients of GPP remain highly stable over time. While the single-year baseline and the multi-year mean are highly consistent, the 10-year average is explicitly utilized in our framework to mitigate transient interannual variability and short-term climatic noise.

S4.4 Preliminary Regional Validation over European Forests

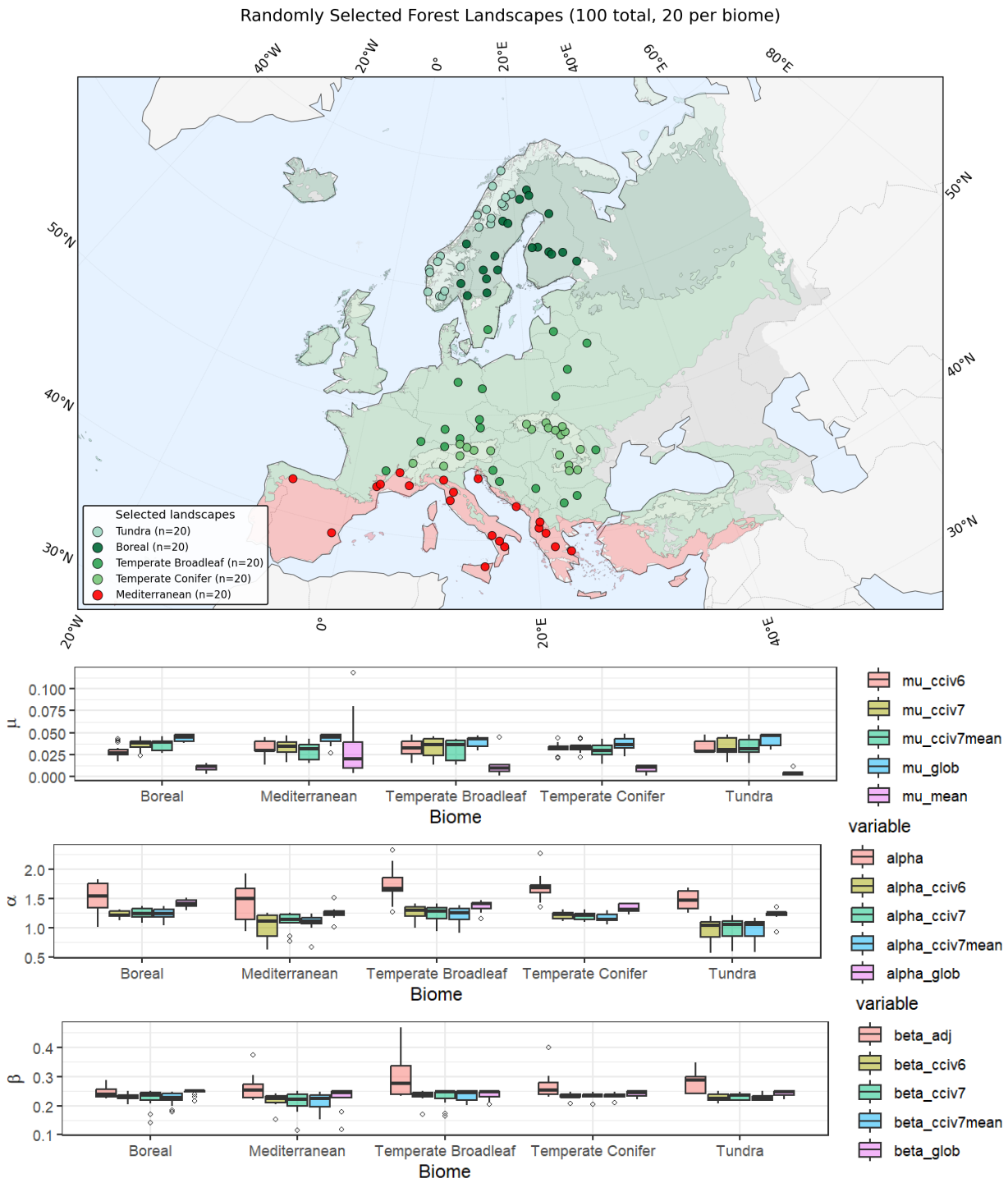


Figure S10. Regional independent validation of predicted DRPs across European biomes. The top map illustrates the geographic distribution of the 100 randomly selected validation landscapes (20 tiles per biome) across Europe. The bottom boxplots show the comparison between Landsat-observed DRPs and our model predictions derived from multiple alternative biomass input layers (cci v6, cci v7, cci v7 mean, and glob).

To evaluate the empirical reliability of the retrieved DRPs, we conducted random cross-biome validations across European regions using independent forest disturbance data. These reference data were derived from Landsat time-series images by detecting stand-replacing disturbance events that occurred between 1985 and 2010. To derive spatially coherent disturbance units from the satellite records, events were defined as contiguous groups of disturbed pixels (using an 8-neighbor connectivity rule) that shared the exact same year of disturbance.

The reference DRPs from the Landsat data were calculated at the landscape level as follows:

- Disturbance Rate (μ): Calculated as the mean yearly fraction of the forest area affected by disturbances within a particular landscape tile.
- Gap-size Distribution (α): Calculated as the fitted power-law exponent between disturbance size classes and their spatial frequency.
- Disturbance Severity (β): Derived from the exponential relationship between the fraction of carbon loss and the spatial size of disturbances. The fraction of carbon loss was proxied using the disturbance magnitude, defined as the difference in spectral values between the stable state of the forest (pre-disturbance) and its disturbed state.

Although these two frameworks exhibit fundamental operational differences, the empirical comparison across a macro-ecological gradient revealed that the retrieved parameter values are broadly comparable in their numerical ranges to the Landsat-based estimates across biomes (**Figure S10**).

S5. Synthetic Variance Reconstruction Experiment

S5.1 Mathematical formulation of Synthetic Un-saturation Recovery Methods

In high-biomass regions like the Amazon and Congo basins, L-band Synthetic Aperture Radar (SAR) signals, alongside algorithmic constraints and spatial filtering, can lead to a compressed dynamic range. Aboveground biomass (AGB) observations exceeding high-biomass thresholds (e.g., $T = 150$ Mg ha^{-1}) often exhibit dampened spatial variance. This limits the observable biomass spatial variance in mature canopies, potentially leading to an underestimation of both the mean biomass and the relative spatial contrast within these dense forest ecosystems.

To address this, we implement a sensitivity testing framework designed to systematically inject variance into these high-biomass regions. By expanding the compressed high-biomass distribution tail, we can evaluate how modifying the spatial structure affects our spatial statistical metrics and the resulting DRP predictions.

This reconstruction is performed directly on the biomass data at its native spatial resolution prior to grid aggregation, using a continuous blending method to prevent artificial spatial discontinuities. We define two comparable physical and mathematical formulations for the target un-saturated biomass state (y_{target}):

S5.1.1 Linear Recovery

The linear recovery model applies a proportional expansion to scale the biomass distribution tail above the threshold:

$$y_{\text{linear}}(x) = T + (x - T) \cdot (1 + k)$$

where:

x is the original native-resolution biomass pixel value (Mg ha^{-1}).

T is the evaluation threshold (tested across $T \in [150, 200, 250, 300]$ Mg ha^{-1})

k is the scaling parameter controlling the intensity of the linear recovery, evaluated across $k \in \{0.1,0.3,0.5,1.0,1.5\}$.

S5.1.2 Inverse Recovery

The inverse logarithmic recovery model simulates a non-linear expansion curve to restore the compressed tail using an asymptotic inverse function:

$$y_{\log}(x) = -S \cdot \ln\left(1 - \min\left[\frac{x}{S+10}, 0.95\right]\right)$$

where: - S is the asymptotic saturation parameter dictating the theoretical maximum capacity of the backscatter-biomass relationship, evaluated at $S \in \{200,250,300,500,800,1200,2000\}$ Mg ha^{-1}. The clipping term $\min[\cdot, 0.95]$ prevents mathematical undefined states in the logarithmic domain.

S5.1.3 Sigmoid Weighting Methods

To ensure spatial continuity across the saturation boundary, we implement a soft thresholding framework governed by a continuous Sigmoid weight function, $W(x)$:

$$W(x) = \frac{1}{1 + e^{-c(x-T)}}$$

where: $c = 0.05$ is the transition parameter controlling the steepness of the blending curve around the threshold T.

The final reconstructed biomass field, $y(x)$, is calculated as a continuous weighted blend of the native saturated observations and the target un-saturated models (y_{target}):

$$y(x) = \big(1 - W(x)\big) \cdot x + W(x) \cdot y_{\text{target}}(x)$$

where $y_{\text{target}}(x)$ represents either $y_{\text{linear}}(x)$ or $y_{\log}(x)$.

This mathematical integration guarantees that Low-biomass observations remain unmodified ($W(x) \approx 0 \Longrightarrow y(x) \approx x$); saturated pixels above the threshold T are fully stretched to their recovered states ($W(x) \approx 1 \Longrightarrow y(x) \approx y_{\text{target}}(x)$); and the first-order spatial derivative across the boundary T remains perfectly smooth, preventing the introduction of artificial boundary edge artifacts that could bias spatial texture inference.

S5.2 Aggregation methods Protocol Audit and Selection

To conduct a rigorous pixel-to-pixel comparison, we mapped the varying native resolution biomass arrays (100 m for ESA CCI and 25 m for GlobBiomass) onto a strictly standardized 100×100 target grid (~ 250 m target resolution for a 25 km $\times$ 25 km landscape tile).

The choice of spatial resampling protocol fundamentally alters the captured biomass distribution, spatial texture (GLCM), and regional carbon totals. Table S4 provides a comparative audit of the evaluated aggregation methods.

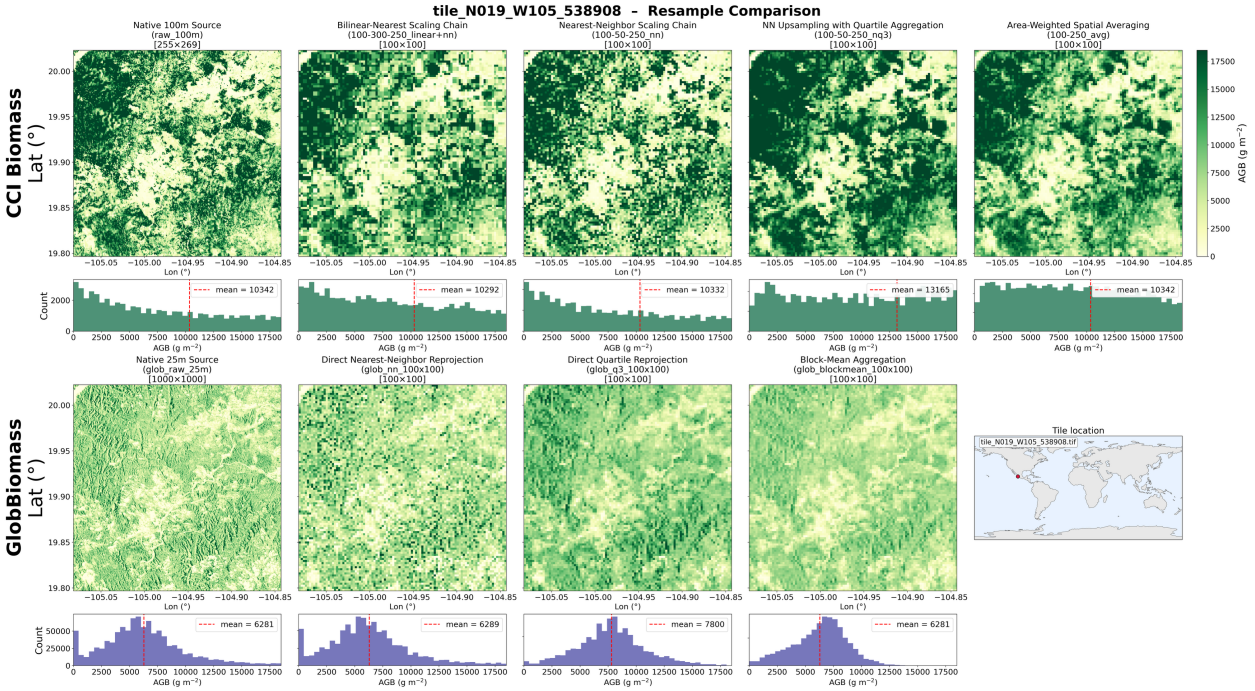


Figure S11. Spatial and histogram comparison of grid resampling and aggregation protocols across a representative 25 km × 25 km forest landscape. Leftmost panels show the original, unresampled native-resolution datasets (100m ESA CCI and 25m GlobBiomass), while the right panels display different resampling results with their aligned pixel-frequency histograms directly beneath (red dashed lines indicate domain-wide means). The bottom-right inset shows the active tile location. Spatial integration methods (Area-Weighted Spatial Averaging and Block-Mean Aggregation, selected) strictly conserve baseline mass, whereas quartile-based methods (nq3 and q3) systematically skew distribution shapes to preserve upper canopy limits.

Table S4: Evaluation of Grid Resampling Protocols

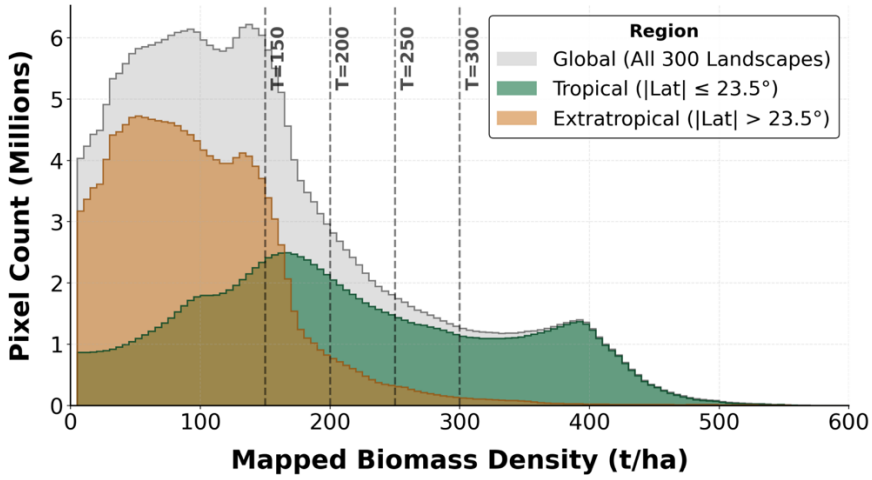
Protocol	Method	Mean AGB ($g\ m^{-2}$) (Baseline: 10342 / 6281)	Statistical & Spatial Evaluation	Selection Status
ESA CCI METHODS				
100-300-250_linear+nn	Performs bilinear downsampling by 1/3 (to 300m), followed by nearest-neighbor upscaling by 1.2 (to 250m) via scipy.ndimage.zoom, and final nearest-neighbor reprojection to the target 100×100 grid.	10292 (-0.5% bias)	Minor negative bias; generates artificial grid-alignment boundaries and smoothing signatures.	Rejected
100-50-250_nn	Performs nearest-neighbor upscaling by 2.0 (to 50m), followed by nearest-neighbor downsampling by 1/5 (to 250m) via scipy.ndimage.zoom, and final nearest-neighbor reprojection to the target 100×100 grid.	10332 (-0.1% bias)	Negligible mean bias; preserves range boundaries but causes severe spatial aliasing and pixelation noise.	Rejected
100-50-250_nq3	Performs nearest-neighbor upscaling by 2.0 (to 50m) via scipy.ndimage.zoom, followed by third-quartile (75th percentile) aggregation reprojection onto the target 100×100 grid.	13165 (+27.3% bias)	Severe positive mass bias; distorts spatial GLCM textures; heavily overestimates canopy density.	Rejected
100-250_avg	Performs direct area-weighted average reprojection of the native 100m grid onto the target 100×100 grid using a spatial integration low-pass filter.	10342 (0.0% bias)	Perfect mass conservation; smooths local high-frequency noise; preserves macro-scale spatial correlation.	Selected (CCI)
GLOBBIOMASS METHODS				
glob_nn_100x100	Performs direct point-sampling of the native 25m grid onto the target 100×100 grid by	6289 (+0.1% bias)	Negligible mean bias; discards 15/16 of source pixels;	Rejected

Protocol	Method	Mean AGB ($g\ m^{-2}$) (Baseline: 10342 / 6281)	Statistical & Spatial Evaluation	Selection Status
	extracting the nearest-neighbor grid-cell centroids.		introduces severe centroid point-sampling aliasing.	
glob_q3_100x100	Performs direct spatial aggregation reprojection of the native 25m grid onto the target 100×100 grid by calculating the 75th percentile of overlapping sub-pixels.	7800 (+24.2% bias)	Severe positive mass bias; violates mass conservation; skews GLCM spatial texture patterns.	Rejected
glob_blockmean_100x100	Performs direct 2D arithmetic block-averaging of non-overlapping 10×10 pixels from the native 25m grid using a nanmean filter to construct the standardized 100×100 target grid.	6281 (0.0% bias)	Perfect mass conservation; reduces micro-pixel noise; stabilizes and aligns local spatial covariance.	Selected (GlobBiomass)

Based on empirical spatial diagnostics (illustrated in Figure S11), we selected Area-Weighted Spatial Averaging (for ESA CCI) and Block-Mean Aggregation (for GlobBiomass) as our standard spatial aggregation protocols. This selection is mathematically justified by two critical criteria: the strict preservation of the ecological mass balance and the mitigation of spatial aliasing. While the primary objective of the quartile-based aggregation methods (nq3 and q3) was to maximize the preservation of high biomass extremes, the diagnostic histograms confirm that this intentional stretching systematically distorts and shifts the underlying biomass distributions. These methods introduce substantial systematic positive biases ($+27.3\%$ for nq3 and $+24.2\%$ for q3). Conversely, while discrete nearest-neighbor reprojections maintain the nominal range of native histograms, they introduce centroid sampling bias, manifesting as artificial high-frequency grid-line boundaries and checkerboard artifacts over fragmented landscapes. By utilizing spatial averaging and block aggregation, our selected protocols preserve the exact mean AGB of the raw source datasets (10342 g m^{-2} for ESA CCI and 6281 g m^{-2} for GlobBiomass without bias) and suppress sub-pixel high-frequency noise. This maintains the continuous physical spatial gradients necessary for robust biomass statistics extraction, thereby stabilizing the feature space for subsequent multi-output Random Forest inversion.

S5.3 Reconstructed scenarios for high-biomass regions

(a)



(b)

Region	> 150 t/ha	> 200 t/ha	> 250 t/ha	> 300 t/ha
Global	37.56%	23.91%	16.47%	11.49%
Tropical	63.57%	46.12%	33.48%	24.03%
Extratropical	15.98%	5.49%	2.35%	1.09%

Figure S12. Original GlobBiomass distributions and high-biomass pixel fractions across 300 sampled landscapes. (a) histograms showing the distribution of mapped biomass density (t/ha, equals to Mg ha^{-1}) for all global landscapes (grey), subset into tropical (green) and extratropical (orange) regions. Vertical dashed lines denote the four high-biomass thresholds ($T \in [150, 200, 250, 300]$ Mg ha^{-1}) used in the synthetic reconstructed scenarios. (b) summary table detailing the proportion of valid pixels exceeding each respective threshold within regional categories.

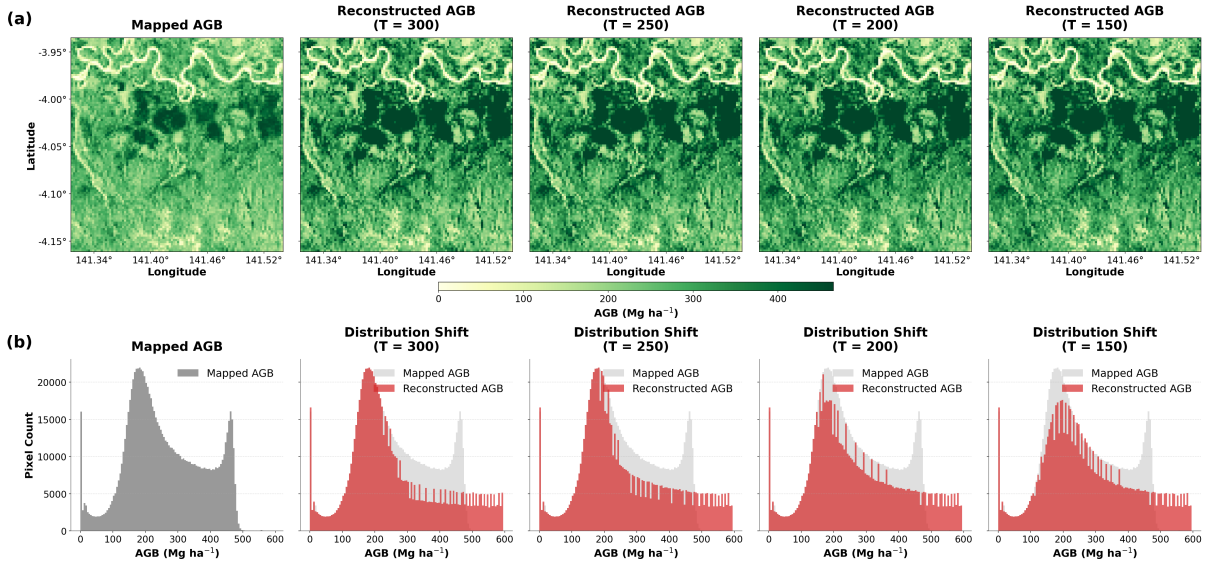


Figure S13. Impact of varying reconstructed high-biomass thresholds on spatial and statistical AGB distribution. Comparison between the original baseline (mapped AGB) and four reconstructed scenarios ($T \in [150, 200, 250, 300]$ Mg ha^{-1}) for a representative landscape. (a) AGB maps show that as T decreases, the reconstruction encompassed a larger fraction of the dense forest. (b) Corresponding pixel-level histograms illustrating the distribution shifts. The baseline mapped AGB (gray) is overlaid with the reconstructed distribution (red).

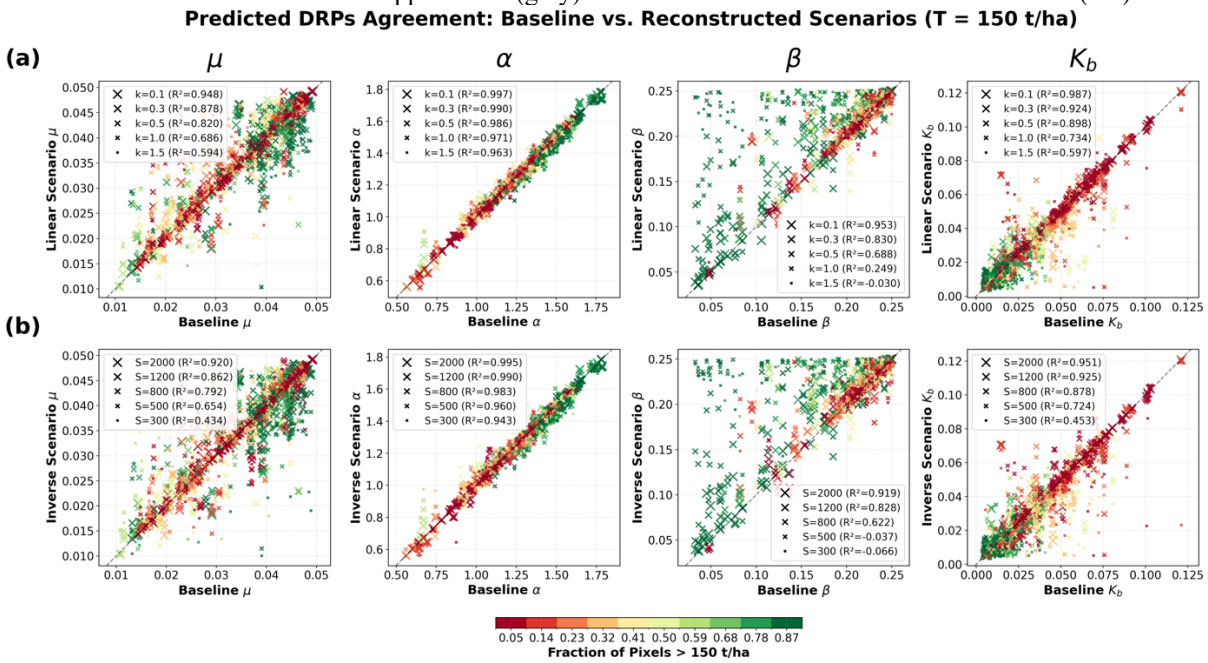


Figure S14. Robustness analysis of predicted DRPs via 1:1 scatter comparisons between the original baseline (x-axis) and reconstructed recovery scenarios (y-axis) . The multi-panel scatter plots compare the predicted parameter values under (a) five linear recovery scenarios ($k \in [0.1,1.5]$) and (b) five inverse recovery scenarios ($S \in [300,2000]$) against the original baseline across 300 randomly selected global landscapes. Scatters are colored by the fraction of pixels within the landscape exceeding threshold (150 Mg ha^{-1} in this case). Across all configurations, α maintains strict alignment along the 1:1 line with consistently high coefficients of determination ($R^2 > 0.94$), confirming its resilience to the dampened range of high biomass regions. In contrast, as the intensity of the reconstructed recovery increases (i.e., k increasing to 1.5, or S decreasing to 300), the prediction variance for μ, β, and K_b expands systematically, deviating from 1:1 line and leading to a progressive drop in R^2 scores. This trend is strongly linked to the high biomass fraction within the landscapes; area with a greater proportion of high biomass

(indicated by greener hues) exhibit more pronounced discrepancies from the baseline. Hereafter, all figures comparing the baseline with reconstructed scenarios follow this same convention.

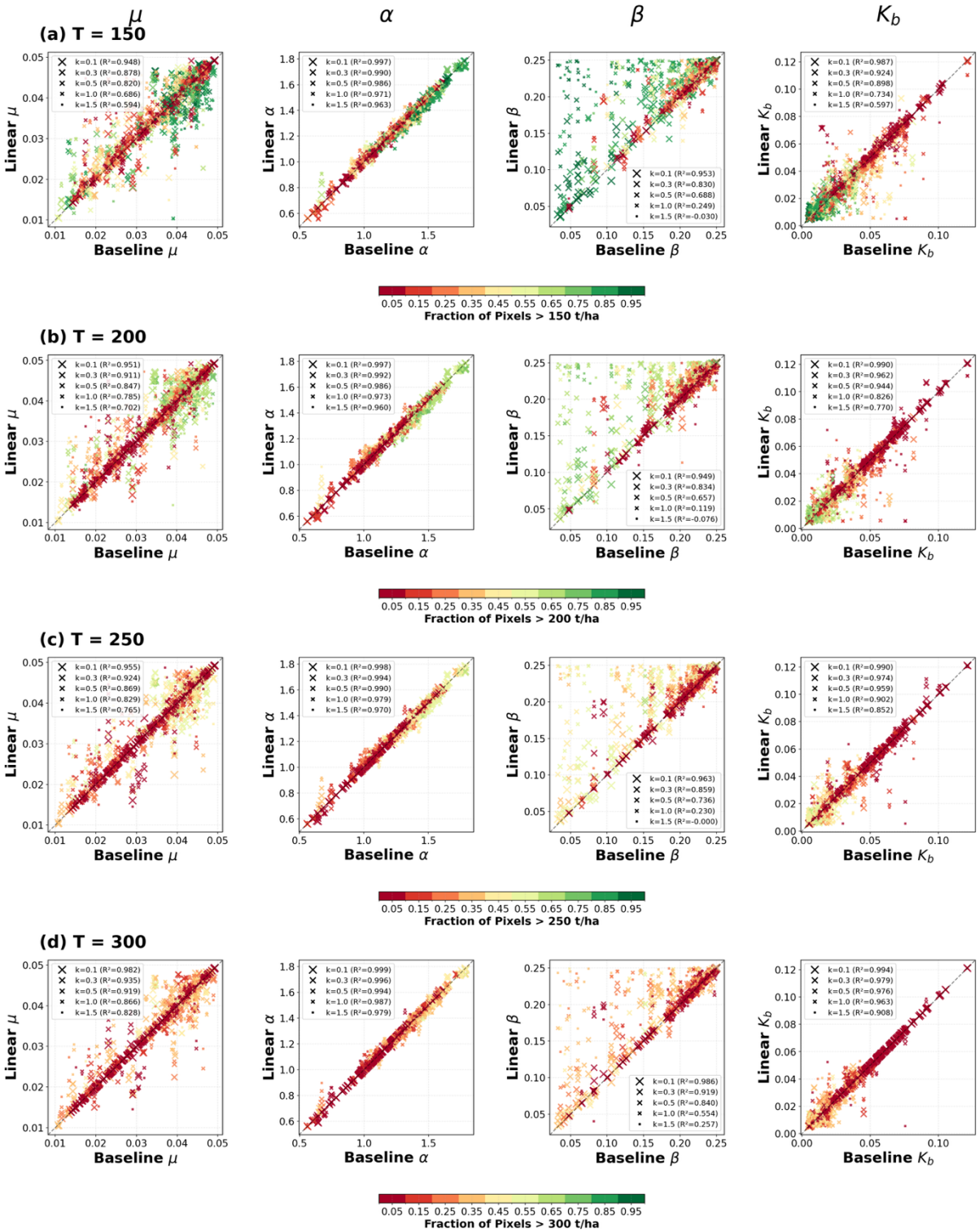
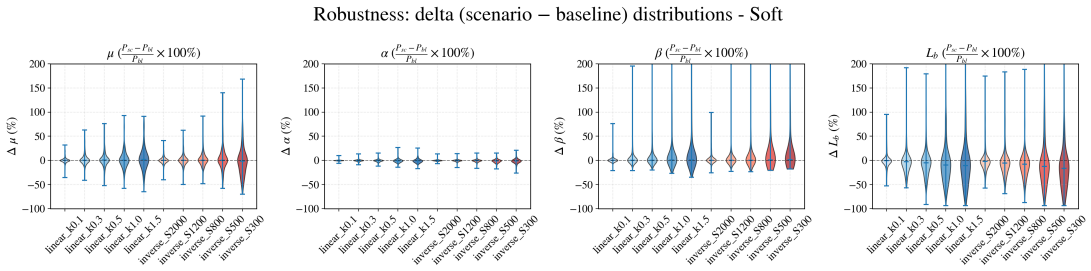
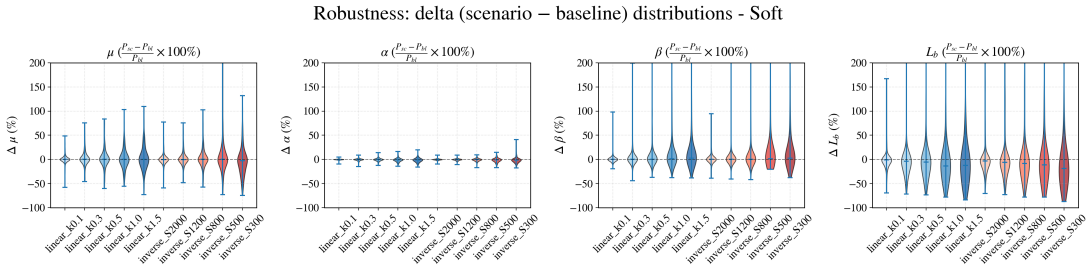


Figure S15. Impact of varying high-biomass reconstruction thresholds. 1:1 scatter comparison between baseline predictions (x-axis) and linear reconstructed scenarios. Columns represent different predicted DRPs, rows represent different high-biomass thresholds: (a) T = 150, (b) = 200, (c) = 250, (d) = 300 t/ha.

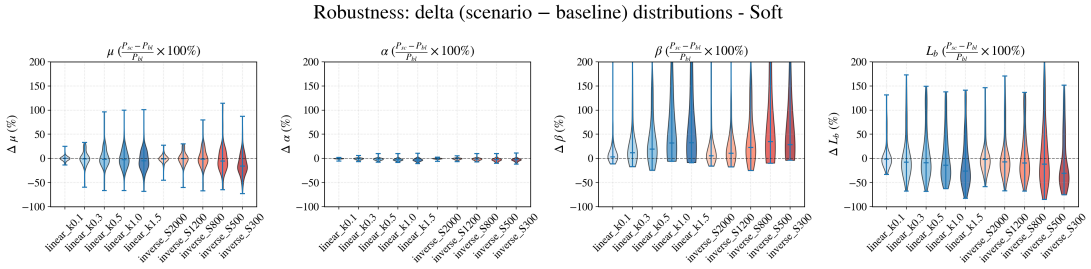
(a) cci_v7_2010



(b) cci_v7_mean



(c) GlobBiomass in tropical regions



(d) Randomly selected tropical high biomass landscapes

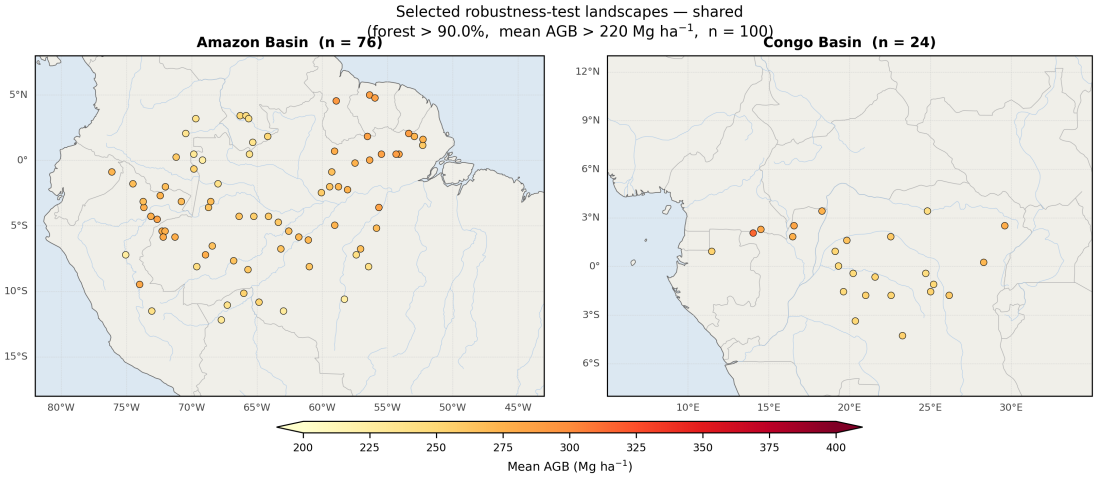


Figure S16. Parameter sensitivity and directional divergence under progressive un-saturation transforms. The multi-panel violin plot displays the relative percentage changes ($\Delta P = \frac{P_{sc}-P_{bl}}{P_{bl}} \times 100\%$, where sc denotes the recovery scenario and bl denotes the unperturbed baseline) of four DRPs across five linear and five inverse recovery configurations (intensity increasing from left to right). Rows (a) and (b) present replication benchmarks using the ESA

CCI Biomass v7 (2010 single-year) and the CCI 7-year temporal mean datasets, respectively. Row (c) highlights the specific subset of 100 randomly sampled tropical high-biomass tiles (Row d) within the Amazon and Congo basins under strict thresholds (forest cover $> 90\%$ and AGB > 220 Mg ha^{-1}). Consistent response patterns emerge across all products (including GlobBiomass results in Figure 9): α maintains stability centered at 0% with narrow distributions; μ shows moderate sensitivity; while β and K_b exhibit severe sensitivity marked by expanding variances. Crucially, the coupled directional shift, where the mean of β shifts upward ($+\Delta$) and the mean of K_b shifts downward ($-\Delta$), is reproducible across all products and is visibly amplified within dense tropical high-biomass landscapes (row c).