

22026 S.-T. Yau High School Science Award

Research Paper

Student name: Qihao Fu , Yigeng Tang , Tongxi Shen

School: Chongqing Bashu Secondary School

Province: Chong Qing

Country/Region: China

Advisor name: Xu Wang, Rong Huang

Advisor affiliation: Chongqing Bashu Secondary School

**Title: From Sampling the Forest to Sampling the Error:
Cutting the Verification Cost of Forest Carbon Assets with Optimized
Spatial Sampling ——A Case Study at Saihanba**

From Sampling the Forest to Sampling the Error: Cutting the Verification Cost of Forest Carbon Assets with Optimized Spatial Sampling

A Case Study at Saihanba

Author: Qihao Fu , Yigeng Tang , Tongxi Shen

Abstract

Forest carbon sinks are turning into tradable assets in China's dual-carbon agenda, and under the CCER (China Certified Emission Reduction) afforestation methodology every credit a project issues rests on a field-sampling estimate of its carbon stock. The economics are tight: each plot is billed, and Table 35 of the methodology discounts a project's issued credits by 6% once sampling uncertainty crosses the 10% line. Current practice, built on stratified sampling and conventional estimators, spends that budget without using two pieces of information it already has: wall-to-wall satellite imagery over the project area, and the spatial correlation of the prediction errors left by the satellite model. This paper moves the sampling decision from the forest to the error: plots are placed to learn the residual field of a remote-sensing prediction rather than to estimate the population mean directly, so the same field campaign yields both a sharper project mean and a full carbon-stock map. The carbon-stock field is written as a remote-sensing prediction plus a spatially correlated residual, $Y(s) = R(s) + E(s)$. On this decomposition we build fused estimators of the regional mean and the map, derive the information gain of a plot set from Gaussian-process conditional inference, and prove three propositions: spatial correlation caps the independent information in full-coverage data at about $|D|/(2\pi\ell^2)$, far below the pixel count; conditioning on plots never increases posterior variance; and a plot's marginal value is its conditional covariance with the regional mean. We tested the framework on 576 fixed $20\text{ m} \times 20\text{ m}$ LiDAR plots at Saihanba in a full-factorial experiment (four sampling designs, seven estimators, three sample sizes, 20 random seeds), adapting the Sampling based on Multi-temporal Prior knowledge (SMP) strategy, which combines conditioned Latin hypercube sampling in NDVI–NIRv feature space with nearest-neighbour spatial dispersion and simulated-annealing optimization, as the optimized design, and comparing our inverse-variance-weighted fusion estimators with regression, inverse-distance weighting, ordinary kriging, and regression kriging. Optimized placement cuts regression-family mapping RMSE from 17.28 to 12.55 t C/ha at $n = 27$, a 27.4% reduction; regression kriging is the most consistent estimator, achieving the lowest RMSE in five of the nine primary design-by-size combinations; against the mean-only benchmark the recommended configuration cuts mapping RMSE by 29.6% to 37.6%; and matching the sampling attributes to the model's feature space, NDVI with NIRv rather than NDVI alone, is worth a further 18% at $n = 27$. The gains also extend to project-level mean estimation: at equal plot counts, the optimized schemes reduced the empirical relative error of the mean by 21%–47%. Under the CCER Appendix F uncertainty benchmark, theoretical uncertainty fell below 10% from $n = 27$ in this stock-estimation experiment, and three optimized plots per zone map as well as six random ones, roughly doubling verification sample efficiency. For a project developer the meaning is direct: the same field bill buys the certified mean plus a wall-to-wall carbon map with quantified error and less exposure to deduction. This framework supplements CCER accounting, while does not replace it.

Keywords: forest carbon-stock accounting; spatial error sampling; conditioned Latin hypercube sampling; kriging interpolation; information gain; CCER

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Contents

1 Introduction	6
1.1 Background.....	6
1.2 Related work.....	7
1.3 Research gap.....	9
2 Study area and dataset	9
2.1 Study area	9
2.2 LiDAR-derived reference carbon-stock map	11
2.3 Satellite imagery and spectral indices	12
2.4 Relation between spectral indices and carbon stock	13
3 Methods	13
3.1 Overall design and evaluation	13
3.2 Zonal SMP-cLHS sampling.....	14
3.3 Seven estimators	17
3.4 Control designs	19
3.5 Data boundary.....	19
3.6 Theory: error propagation, information gain, and plot value	19
4 Results	21
4.1 Full-factorial main results.....	21
4.2 SMP gains: size, source, and decay	22
4.3 Ablation across the estimator spectrum.....	24
4.4 Sample size versus error	27
4.5 Comparison under the CCER mean criterion	28
4.6 Zone-level accuracy check	29
4.7 Supplementary experiments	30
5 Discussion.....	31
5.1 Why optimized sampling improves small-sample mapping.....	31
5.2 Estimator performance under different sampling designs	31
5.3 Implications for repeated carbon monitoring	32
5.4 Implications for CCER-oriented carbon accounting	32
5.5 What the accuracy gains are worth in verification and trading	32
6 Conclusion	33
7 Limitations.....	34

References	35
Acknowledgments	36

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

1 Introduction

1.1 Background

Accurate, objective quantification of forest carbon sinks constitutes the microeconomic foundation of carbon asset pricing, market liquidity, and carbon-backed financial derivatives within China's "dual-carbon" institutional architecture. At the individual-tree level, aboveground biomass can be estimated from field measurements such as diameter at breast height (DBH) and tree height using allometric equations and subsequently converted to carbon stock. Extending such measurements to large forest areas, however, requires sampling because a complete tree census is generally impractical. Operational forest-carbon inventories therefore rely on field plots and statistical inference to estimate carbon stocks over larger areas.

Carbon-stock measurement sits at the front of a two-step chain. The first step is measurement: a project needs a defensible number for how much carbon its forest holds. The second is finance: once that number is verified under a methodology such as CCER, it becomes the basis on which carbon credits are issued, discounted, and traded. Step two inherits the error of step one, so measurement error converts directly into money, in field-plot costs and in the credits lost when uncertainty crosses a deduction threshold.

Current CCER practice makes the first step expensive. Its uncertainty accounting treats field plots as independent draws, so the only lever for precision is plot count: in our study area, the methodology's own Appendix F formula keeps theoretical uncertainty above the 10% no-deduction line until roughly 27 plots, and every extra plot is a field campaign in difficult terrain.

Existing accounting approaches are primarily designed to estimate project-level carbon stocks and their associated uncertainty rather than to provide wall-to-wall spatial information. Spatially explicit carbon-stock estimates are nevertheless important for stand-level forest management, targeted verification, and assessments of disturbance risks such as fire, pests, and forest degradation. Obtaining wall-to-wall reference measurements through field surveys or repeated airborne LiDAR campaigns is generally impractical because of their high cost and limited repeatability. A more scalable approach is to combine a limited number of accurately measured reference plots with wall-to-wall satellite observations and a statistical model that relates remote-sensing covariates to forest carbon stock.

This approach raises two closely related methodological questions: where should a limited number of reference plots be located, and which estimator should be used to combine plot observations with wall-to-wall remote-sensing information? Sampling design and estimator choice are often studied separately, although their effects may interact. A sampling configuration that performs well for one estimator may not provide

the same benefit for another, while improved plot placement may also change the relative performance of competing estimators. Sampling design and estimator choice should therefore be evaluated jointly. In this study, we compare alternative sampling designs and estimators within a common experimental framework to examine how their interaction affects forest carbon-stock estimation and mapping under limited sample sizes.

For a project developer, the economics are direct. Verification is priced per plot, and Table 35 prices uncertainty as well: above 10%, every issued credit is discounted. The gains reported in this paper act on both sides of that ledger at once: at $n = 27$, optimized sampling with regression kriging cuts mapping error by roughly 30% against the mean-based benchmark and, at no extra plots, adds a wall-to-wall map whose error is quantified. The framework is designed to supplement CCER practice: it takes over where the plots go and how their observations are combined, while the official criterion stays as it is.

1.2 Related work

Classical forest inventories commonly use simple random, systematic, or stratified sampling to estimate population-level quantities from a limited number of field plots. These designs and their statistical properties are well established in survey sampling theory (Cochran, 1977), and the CCER afforestation methodology similarly permits plot-based sampling schemes including random, systematic, and stratified designs [3]. Such approaches are well suited to estimating project-level means or totals and their associated uncertainty, but they are not intended to characterize the continuous spatial distribution of forest carbon stocks.

Remote sensing provides a complementary source of information by supplying wall-to-wall covariates related to forest structure and biomass. Statistical or machine-learning models can be fitted between plot-level carbon measurements and remotely sensed variables such as the normalized difference vegetation index (NDVI) and near-infrared reflectance of vegetation (NIRv) and subsequently applied across the study area to generate spatially continuous carbon-stock estimates. The accuracy of such models may nevertheless be constrained by spectral saturation at high biomass levels, variation in canopy structure, and spatially varying relationships between spectral response and carbon stock. These limitations mean that remote-sensing predictions retain spatially structured residual errors even when the overall predictive relationship is strong.

Spatially explicit forest-carbon mapping has long been recognized as important for carbon monitoring and climate-mitigation programs. Gibbs et al. (2007) highlighted the role of forest carbon monitoring in REDD-related applications [15], while Saatchi et al. (2011) demonstrated large-area carbon-stock mapping and associated uncertainty assessment across tropical forests [16]. A central issue in such wall-to-wall products is that prediction errors are often spatially correlated rather than independent. Geostatistical

methods provide a formal framework for characterizing this dependence through covariance functions or variograms and for incorporating it into spatial prediction and uncertainty analysis (Goovaerts, 1997; Webster and Oliver, 2007) [13,14]. This spatial dependence forms an important basis for the sampling and inference framework developed in this study.

Geostatistical estimators differ in how they combine spatial dependence with auxiliary information. Ordinary kriging uses the spatial covariance or variogram structure of the target variable to predict values at unsampled locations, whereas regression kriging combines a deterministic regression component derived from auxiliary covariates with spatial interpolation of the remaining residuals. The latter is particularly relevant to remote-sensing-assisted carbon mapping because satellite observations can explain broad spatial variation while geostatistical modelling accounts for residual spatial dependence. The relative performance of these approaches depends on both the predictive strength of the auxiliary variables and the spatial structure remaining in the residual field.

Sampling optimization provides a complementary way to improve the efficiency of field data collection. Minasny and McBratney (2006) introduced conditioned Latin hypercube sampling (cLHS), which selects samples so that their distribution in environmental covariate space approximates that of the target population. Zeng et al. (2015) subsequently developed the Sampling based on Multi-temporal Prior knowledge (SMP) strategy, which combines representativeness in auxiliary-variable space with spatial dispersion and uses simulated annealing to identify efficient sampling configurations. The approach was later incorporated into China's national standard for leaf area index (LAI) product validation [1,2]. However, its application to forest carbon-stock mapping has received much less attention. In particular, it remains unclear how sampling designs based on remote-sensing covariates interact with different spatial estimators and whether gains from improved sample placement are consistent across estimation methods.

Model-assisted survey sampling provides a closely related statistical foundation. Särndal et al. (1992) developed a general framework for improving population inference using auxiliary information while retaining design-based properties [17], and Gregoire and Valentine (2008) reviewed its application to sampling and estimation in natural-resource inventories [18]. Building on this general principle, we treat the wall-to-wall remote-sensing prediction as auxiliary information and use field observations primarily to characterize the remaining spatial residual field. Under this formulation, candidate plot locations can be evaluated according to the reduction in uncertainty they provide for regional carbon-stock estimation and spatial prediction. This links sampling design with estimator performance and provides a common basis for comparing alternative plot

configurations under a fixed sampling effort.

1.3 Research gap

Using 576 fixed 20 m $\times$ 20 m LiDAR-derived reference plots in Saihanba, we address three research questions concerning sampling efficiency, estimator performance, and their combined effect on forest carbon-stock mapping.

(1) Under small sample sizes, to what extent does SMP-cLHS optimized sampling improve carbon-stock mapping relative to within-zone random, systematic, and contiguous sampling designs? We further examine which components of the optimized design contribute to these improvements and how the benefits change with increasing sample size.

(2) How does estimator performance depend on sampling design and sample size? We compare seven estimators representing different ways of using remote-sensing covariates, spatial dependence, and combinations of the two, and evaluate whether improved placement changes their relative performance.

(3) How much additional mapping accuracy can be achieved when optimized sampling is combined with an appropriate estimator, relative to a conventional project-level mean estimator? For spatial comparison, the latter is represented as a spatially constant map based on the estimated mean carbon stock

All comparisons use the same reference data, evaluation locations, and random realizations, allowing sampling designs and estimators to be assessed under directly comparable conditions. The experimental design uses only auxiliary information available at the time of sampling and avoids using evaluation data during sample selection or model fitting. In addition to the empirical comparisons, we develop a theoretical formulation based on the spatial covariance structure of the residual field to quantify the information contributed by alternative plot configurations.

The remainder of the paper is organized as follows. Section 2 describes the study area and datasets. Section 3 introduces sampling designs, estimators, evaluation framework, and theoretical formulation. Section 4 presents the experimental results, followed by discussion of sampling efficiency, estimator performance, and practical implications in Section 5. The main conclusions and limitations are summarized in Sections 6 and 7, respectively.

2 Study area and dataset

2.1 Study area

This study was conducted at the Saihanba Mechanical Forest Farm (SMFF) in Weichang County, Hebei Province, northern China (Fig. 1). SMFF is located at the

southern edge of the Inner Mongolia Plateau, approximately between 42°02'–42°36' N and 116°51'–117°39' E, with elevations ranging from about 1,010 to 1,940 m above sea level. The region lies in a forest–steppe ecotone and has a cold temperate continental monsoon climate, characterized by long, cold winters and short, cool summers. Long-term records indicate a mean annual temperature of approximately -1.5°C and mean annual precipitation of about 452.6 mm, with most precipitation occurring during the summer growing season. The forest farm is dominated by plantations established since the 1960s, with *Larix principis-rupprechtii* and *Pinus sylvestris* var. *mongolica* among the principal tree species planted.

The experimental area comprised a contiguous $480\text{ m} \times 480\text{ m}$ forest block subdivided into a regular 24×24 grid, yielding 576 fixed $20\text{ m} \times 20\text{ m}$ reference units and a total area of 23.04 ha. The block has high canopy cover and a relatively homogeneous plantation composition, while retaining substantial variation in tree density, canopy structure, and aboveground carbon stock. This combination of compositional consistency and within-stand structural heterogeneity provides a controlled setting for evaluating the effects of sampling configuration on carbon-stock estimation.

The $20\text{ m} \times 20\text{ m}$ spatial unit was used consistently throughout the study to align field observations, LiDAR-derived structural information, satellite covariates, and subsequent sampling experiments. The complete 576-unit grid was treated as the reference population. In each sampling experiment, only the selected units were made available to the estimator, while the remaining units were withheld for evaluation. This design allowed alternative sampling strategies and estimators to be compared against the same underlying spatial population and avoided differences caused by changes in reference coverage.

Airborne LiDAR data were acquired over the study area in 2024. The point cloud was preprocessed through ground classification and height normalization and was subsequently used to derive terrain- and canopy-structure products, including a canopy height model (CHM). Field measurements collected in association with the LiDAR campaign included diameter at breast height (DBH) and tree height where field observations were available. These measurements and LiDAR-derived structural variables formed the basis for subsequent plot-level biomass and carbon-stock characterization. Detailed field measurements and LiDAR processing procedures are described in the following sections.

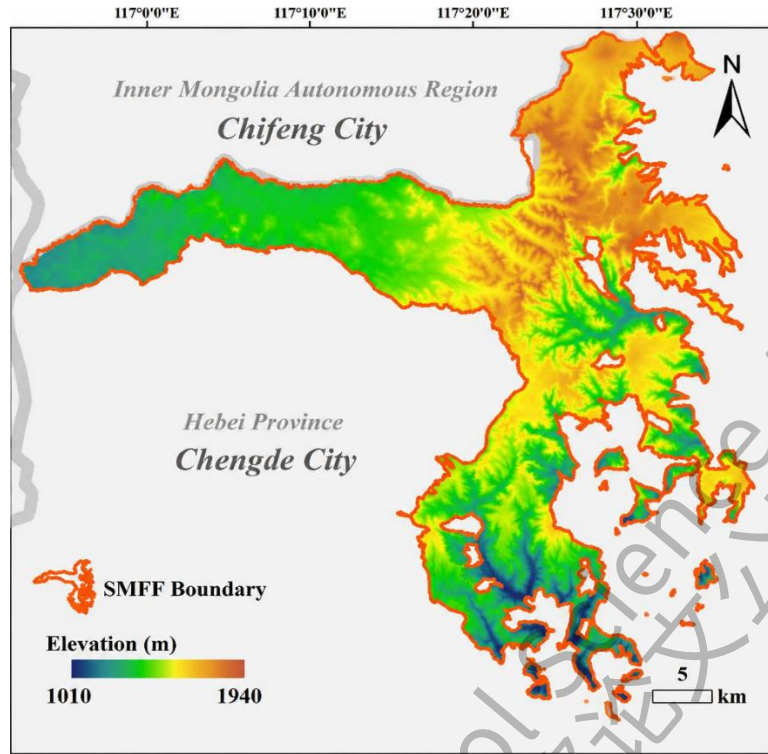


Figure 1 Overview of the Saihanba study area.

2.2 LiDAR-derived reference carbon-stock map

The processing chain is as follows. The airborne LiDAR point cloud yields a canopy height model; individual tree crowns are segmented; single-tree carbon is estimated from crown geometry; each tree is assigned to the pixel containing its crown apex and summed. Pixel carbon density (t C/ha) equals tree carbon (kg) divided by 40 (0.04 ha per pixel). Across the 576 plots in 2024, mean carbon density is 58.85 t C/ha, the standard deviation 17.74, and the coefficient of variation 30.1%. Compared with the highly variable carbon-sink increment, the stock varies less and holds a more stable spatial structure.

We spot-checked the reliability of the LiDAR estimates against field measurements: 101 measured trees (girth converted to DBH, GPS-recorded) were matched to LiDAR trees by crown-polygon IoU, keeping matches above 0.5. The relative RMSE of LiDAR DBH against field values is 6.5% (2024, 88 trees), below the 10% threshold the CCER methodology sets for remote-sensing-assisted estimation. The check stops at DBH as an intermediate variable, and only 39.6% of the measured trees fall inside the 576-plot area; both limits are discussed in Section 7.

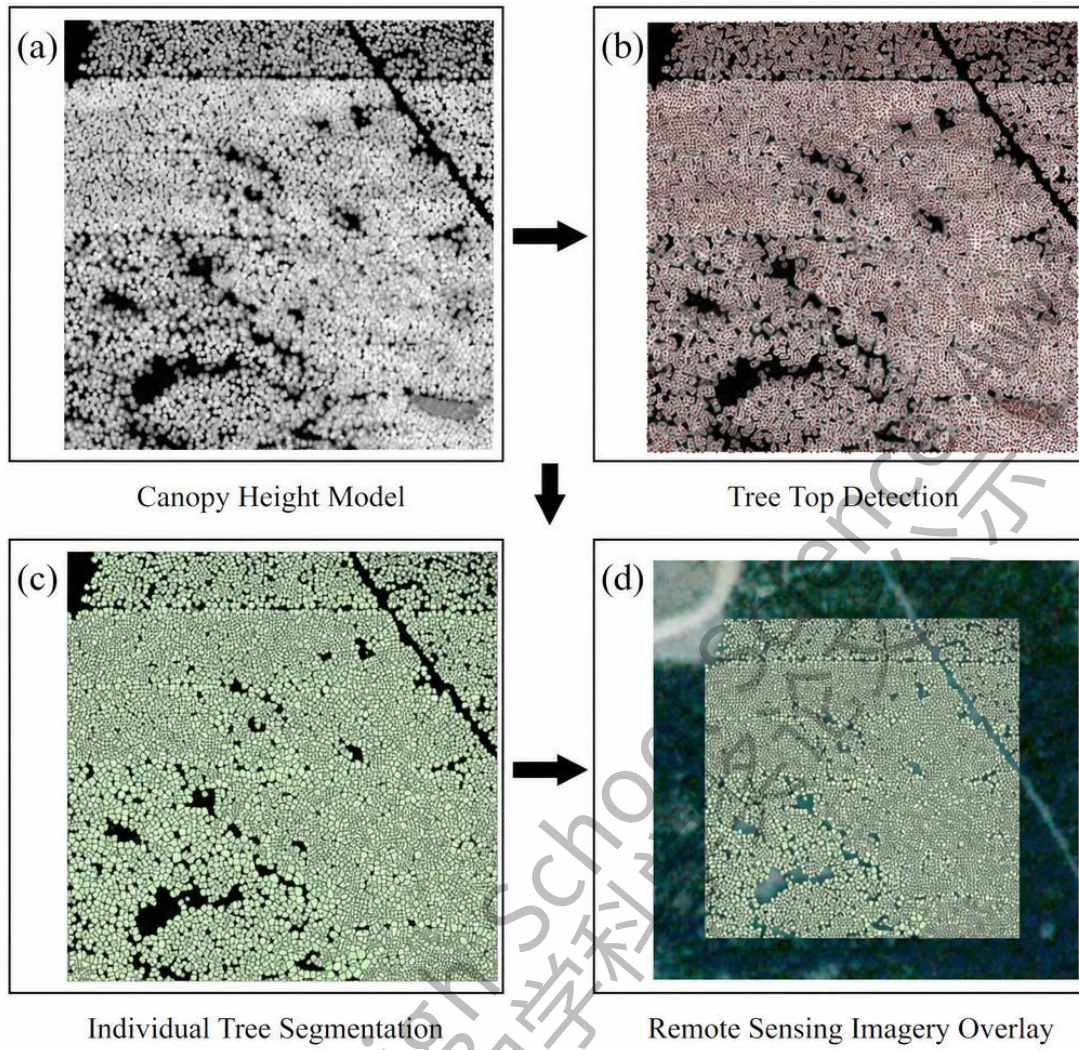


Figure 2 Workflow of individual tree crown segmentation from the CHM

2.3 Satellite imagery and spectral indices

PlanetScope imagery acquired during the 2024 growing season was used to provide wall-to-wall spectral information over the study area. The imagery consisted of eight spectral bands at a spatial resolution of approximately 3 m. Two vegetation indices were retained for subsequent sampling and regression analyses: the normalized difference vegetation index (NDVI) and near-infrared reflectance of vegetation (NIRv). NDVI was calculated from the red and near-infrared bands as.

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

and NIRv was calculated as

$$NIRv = NDVI * NIR$$

Spectral values were aggregated to the same 20 m × 20 m reference grid used for the carbon-stock data, ensuring a common spatial support for sampling, model fitting, and

evaluation. Across the 576 reference cells, quadratic relationships with carbon stock gave $R^2 = 0.612$ for NDVI and $R^2 = 0.352$ for NIRv. NDVI and NIRv were strongly correlated ($r = 0.82$), but their imperfect correspondence allowed the two indices to represent partially different aspects of canopy spectral variation.

2.4 Relation between spectral indices and carbon stock

Figure 3 shows the relationships between the two PlanetScope-derived vegetation indices and LiDAR-derived carbon stock across the 576 reference cells. NDVI showed a stronger relationship with carbon stock than NIRv, with quadratic-model R^2 values of 0.612 and 0.352, respectively. The fitted NDVI relationship continued to increase over most of the observed range, indicating that severe high-biomass saturation was not evident within this particular dataset.

Substantial variation in carbon stock nevertheless remained at similar NDVI and NIRv values. Thus, although the spectral indices captured an important component of spatial variation in carbon stock, spectral information alone did not fully explain the reference carbon-stock distribution. This residual variation provides the motivation for explicitly modelling spatially structured prediction errors in the subsequent analysis.

NDVI and NIRv were correlated ($r = 0.82$) but were not interchangeable. We therefore retained both variables in the bivariate sampling design and regression analyses and subsequently tested whether using both indices improved sampling performance relative to an NDVI-only design.

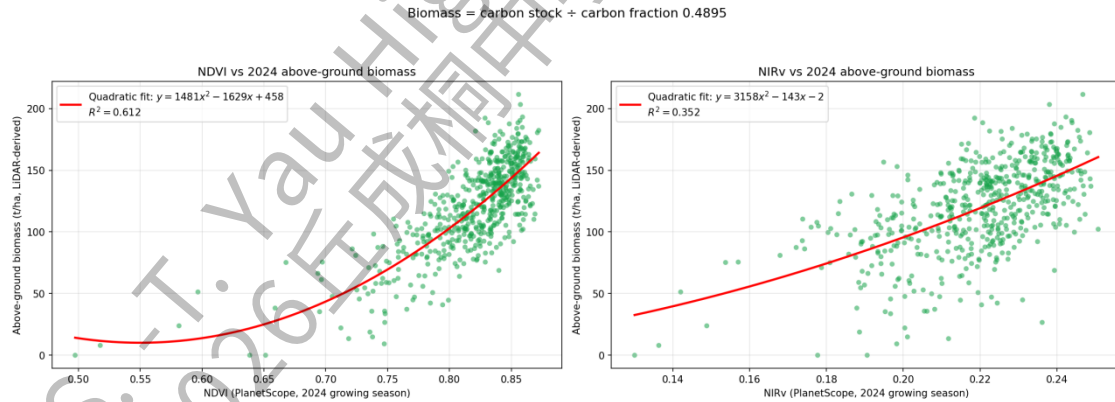


Figure 3 NDVI versus 2024 above-ground biomass for the 576 plots (quadratic fit $R^2 = 0.612$; right panel: NIRv, $R^2 = 0.352$)

3 Methods

3.1 Overall design and evaluation

We formulate forest carbon-stock mapping as the combination of a remote-sensing prediction component and a spatial residual component. Let the study domain D consist of N equal-area reference cells, with $Y(s)$ denoting the reference carbon density at location s . For a given sampled set, a regression model based on wall-to-wall remotely

sensed covariates provides an initial prediction $R(s)$. The corresponding residual field is defined as

$$E(s) = Y(s) - R(s), Y(s) = R(s) + E(s) \quad (1)$$

Under this formulation, remotely sensed covariates explain the systematic component of carbon-stock variation, while sampled reference observations provide information on both regression calibration and the remaining spatially structured residuals. For equal-area cells, the regional mean can be written as

$$\mu_Y = \frac{1}{N} \sum_{i=1}^N Y_i = \mu_R + \mu_E \quad (2)$$

Prediction accuracy was evaluated only at unsampled reference cells. Let $\setminus(U)$ denote the set of unsampled cells. Mapping accuracy was quantified using

$$\text{RMSE} = \sqrt{\frac{1}{|U|} \sum_{s \in U} (Y(s) - \hat{Y}(s))^2} \quad (3)$$

The main experiment followed a full-factorial design comprising four sampling strategies, seven estimators, and three sample sizes ($n = 27, 54$ and 90). Each combination was repeated across 20 random realizations, yielding 1,680 sampling–estimation runs. Within each run, the sampled cells were used for model fitting and spatial estimation, whereas all remaining cells were withheld from fitting and used for evaluation. This ensured that all sampling designs and estimators were compared against the same reference population without using evaluation values during model construction..

In addition to the main factorial experiment, we evaluated performance across a wider range of sample sizes to characterize the relationship between sampling effort and prediction error. Additional sensitivity analyses examined the effects of alternative model specifications and univariate versus bivariate sampling attributes.

3.2 Zonal SMP-cLHS sampling

The optimized sampling design was adapted from the SMP framework of Zeng et al. (2015), which jointly considers representativeness in auxiliary-variable space and spatial dispersion. SMP has subsequently been used in China's national standard for leaf area index (LAI) product validation [1,2]. We re-implemented the sampling procedure in Python following the original objective-function formulation.:

For a candidate sample set $\setminus(S)$, attribute representativeness was quantified using the discrepancy term

$$\text{dif}(S) = \frac{1}{n} \sum_{i=1}^n |c_i - 1|$$

where the population distribution of each sampling attribute was divided into n equal-frequency classes and c_i denotes the number of selected cells falling in class i . A perfectly balanced Latin-hypercube representation therefore gives $\text{dif} = 0$. The discrepancy was

calculated separately for NDVI and NIRv and combined in the sampling objective.

Spatial dispersion was characterized using the nearest-neighbour index (NNI):

$$NNI(S) = \frac{\bar{d}_{nn}(S)}{E[d]}, E[d] = \frac{1}{2\sqrt{\rho}}, \rho = \frac{k}{|D|}$$

where $\bar{d}_{nn}(S)$ is the mean nearest-neighbour distance among selected cells, k is the number of selected cells, and $|D|$ is the area of the sampling domain. Values of $NNI > 1$ indicate greater spatial dispersion than expected under complete spatial randomness.

Attribute representativeness and spatial dispersion were combined through

$$OF(S) = \frac{dif(S)}{NNI(S)} \quad (4)$$

which was minimized using simulated annealing. Candidate swaps that reduced the objective were preferentially accepted, while the annealing schedule also allowed occasional uphill moves to reduce sensitivity to local minima. The initial temperature was 1, temperature was reduced by 5% every 10 iterations, and optimization was terminated after 10,000 iterations or when the stopping criterion fell below 0.01.

To ensure spatial coverage across the study block, we introduced an additional zonal constraint. The 24×24 reference grid was divided into nine 8×8 zones. Each zone contributed an equal number of sampled cells, $q \in \{3, 6, 10\}$, corresponding to total sample sizes of $n = 27, 54$, and 90 , respectively. Simulated-annealing swaps were restricted to cells within the same zone, so that every final design contained exactly q sampled cells in each of the nine zones. This constraint prevented sample concentration in a small portion of the study area while preserving SMP optimization within each zone.

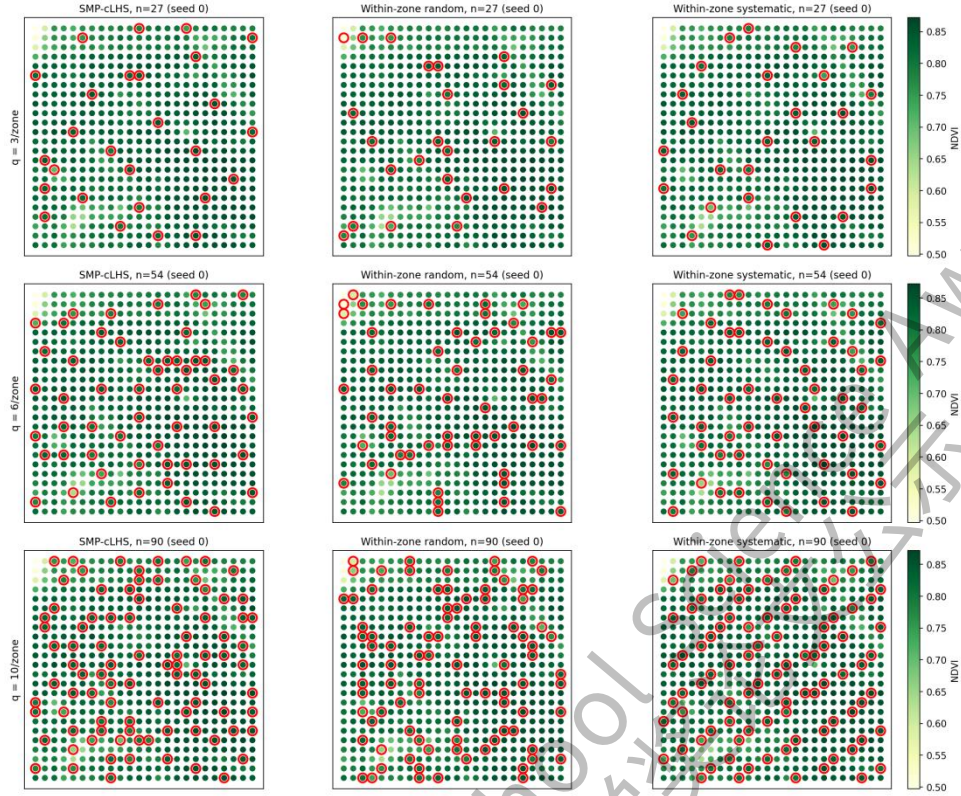


Figure 4 Sample layouts of the three sampling designs (columns: SMP, within-zone random, within-zone systematic; rows: $n = 27, 54, 90$; background is NDVI, pale patches are low-NDVI sparse areas, red circles mark selected plots)

Figure 4 shows what histogram matching does to plot placement. With random and systematic designs, small samples leave the low-NDVI sparse patches (pale areas) undercovered; SMP raises coverage of exactly those areas. Figure 5 quantifies the quantile-bin matching, and Figure 6 traces one annealing run.

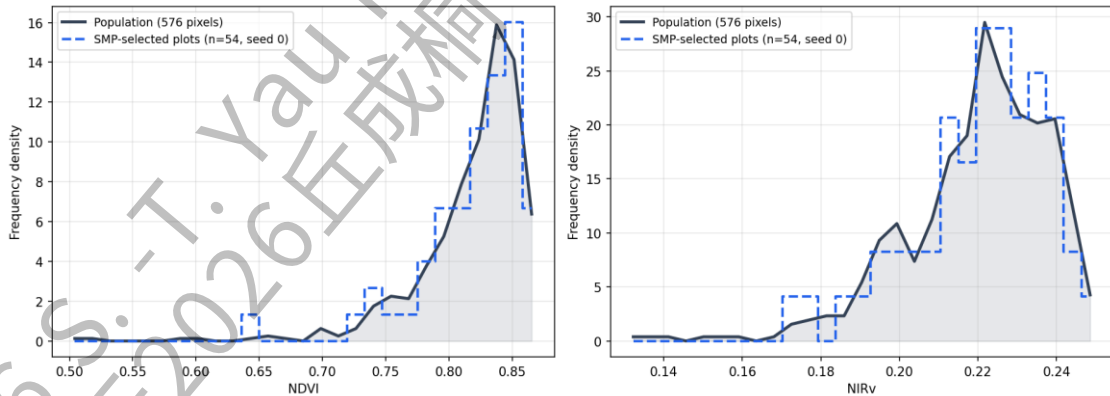


Figure 5 Frequency distributions of the population and the SMP sample (left: NDVI; right: NIRv; solid: all 576 pixels; dashed: the 54 SMP-selected plots; the agreement of the two curves is the attribute representativeness)

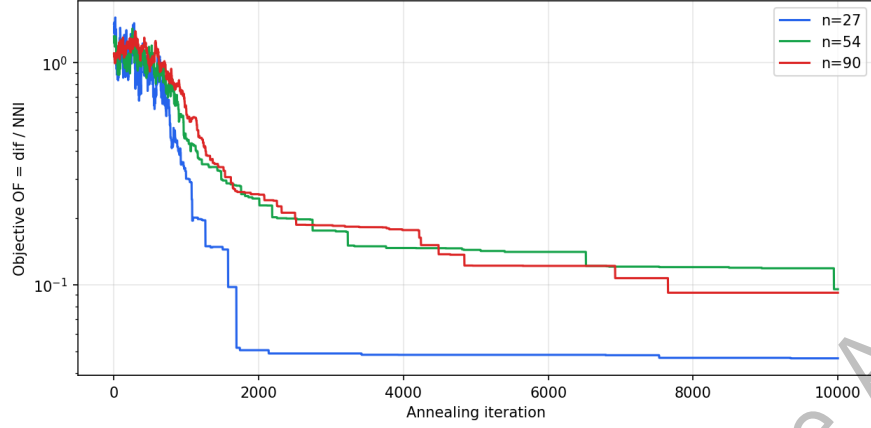


Figure 6 Convergence of the simulated-annealing objective (one typical run each at $n = 27, 54, 90$; log scale)

3.3 Seven estimators

The seven estimators span the full spectrum from attribute-only to space-only to fusion of both, and E5, E6, and E7 are fusion estimators built in this study. Every model is fitted on the n sampled plots alone, and no unsampled reference value enters training, parameter estimation, or weighting:

Their formal definitions follow. E1 (pure regression) uses the full quadratic in the two satellite-observable indices as the prior:

$$R(s) = \beta_0 + \sum_{j=1}^2 [\beta_{1j}x_j(s) + \beta_{2j}x_j^2(s)] \quad (7)$$

E2 (inverse-distance weighting) weights plot observations by the inverse p -th power of distance:

$$\hat{z}_{IDW}(s_0) = \frac{\sum_{i \in S} d(s_0, s_i)^{-p} z(s_i)}{\sum_{i \in S} d(s_0, s_i)^{-p}} \quad (8)$$

Table 1 Definition of the seven estimators

No.	Method	Information used
E1	Quadratic regression on NDVI+NIRv	Attribute only (spectral)
E2	IDW interpolation ($p = 2$; known points take LiDAR values)	Space only
E3	Ordinary kriging (de-mean, interpolate, add back)	Space only
E4	E1 regression surface + kriged residual correction (additive fusion)	Additive fusion
E5	$w \cdot E1 + (1-w) \cdot E2$, w a global LOO weight	Weighted fusion
E6	$w \cdot E1 + (1-w) \cdot E3$, w a global LOO weight	Weighted fusion
E7	Per-pixel fusion of E1 and E3 weighted by error maps, $w(s) \propto 1/\text{residual}^2(s)$	Per-pixel weighted fusion

The covariance structure of the spatial components (E2/E3/E4) follows a nugget-plus-exponential variogram:

$$\gamma(h) = \tau^2 + \sigma^2(1 - e^{-h/\ell}) \quad (9)$$

E3 (ordinary kriging) performs optimal linear unbiased interpolation on the de-meaned field, with the kriging variance giving the theoretical uncertainty:

$$\hat{z}_K(s_0) = \sum_{i=1}^n \lambda_i z(s_i), \text{Var}_K(s_0) = (\tau^2 + \sigma^2) - c^T C^{-1} c \quad (10)$$

E4 (regression plus kriged residuals) interpolates the residuals by ordinary kriging on top of the regression prior. This is the core step of spatial error sampling: a full-coverage prior R as the base map, plots learning the error field:

$$\hat{Y}(s_0) = R(s_0) + \sum_{i \in S} \lambda_i(s_0)(Y(s_i) - R(s_i)) \quad (11)$$

with weights λ solved from the covariance equations induced by Eq. (9). The E5 and E6 built in this study weight the regression and spatial information by the inverse LOO error variances:

$$w = \frac{1/v_{reg}}{1/v_{reg} + 1/v_{space}} \quad (12)$$

E7 pushes the weighting down to individual pixels, fusing point by point according to the regression error map and the kriging variance map:

$$w(s) = \frac{1/\hat{e}_{reg}^2(s)}{1/\hat{e}_{reg}^2(s) + 1/\sigma_{krig}^2(s)} \quad (13)$$

Three implementation details matter. First, the variogram fit (Eq. (9)). Full-sample diagnostics put the correlation of 2024 regression residuals at only 0.20 at 20 m lag ($\gamma(24 \text{ m})/\text{sill} \approx 0.83$), so residuals are nugget-dominated; a no-nugget exponential does not fit, its range collapses to a few meters and interpolation does nothing. On an $n = 54$ SMP sample the direct variogram of carbon is $\tau^2 = 144$, $\sigma^2 = 167$, range $\ell = 200$ m (long-range structure, consistent with Moran's $I = 0.43$); regression residuals give $\tau^2 = 12$, $\sigma^2 = 88$, $\ell = 1.7$ m. Second, E3 de-means before interpolation. With a mean near 59 t C/ha, zero-mean simple kriging shrinks systematically (bias about -10 t C/ha before correction, about -2.6 after de-meaning and adding back). Third, the weights in Eqs. (12) and (13) come from the data with no hand tuning: the kriging variance map $\sigma^2_{krig}(s) = (\tau^2 + \sigma^2) - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c}$ gives the theoretical interpolation error growing away from plots, and the regression error $\hat{e}_{reg}(s)$ spreads the LOO squared residuals of the n plots by IDW (floored at 25% of the global mean squared residual to avoid singularity). Substituting into Eq. (13), pixels near plots lean toward spatial interpolation and far pixels lean toward regression; the global weights of E5 and E6 likewise follow Eq. (12) from LOO error variances.

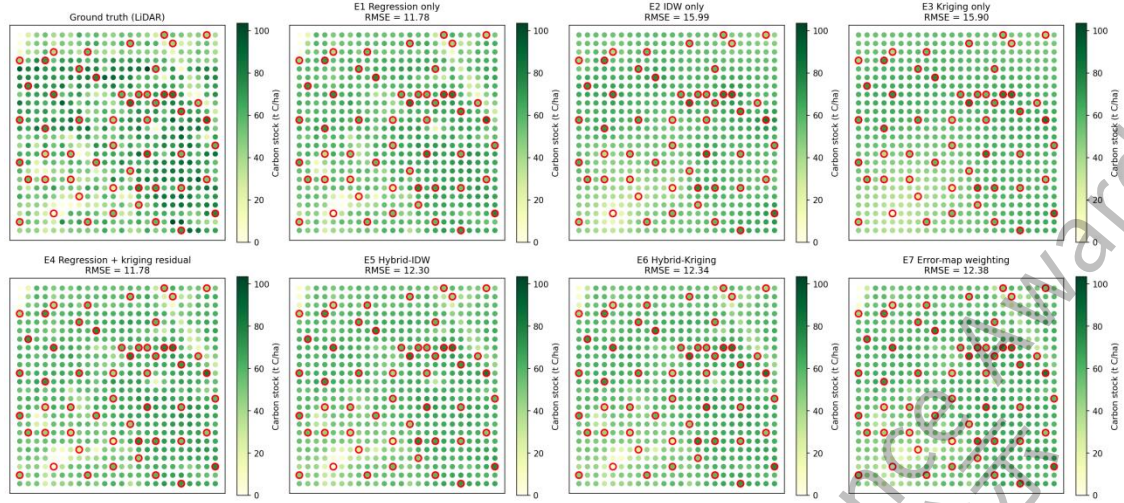


Figure 7 Maps of the seven estimators against ground truth (panels left to right: truth, then E1 through E7; $n = 54$, SMP, seed 0)

3.4 Control designs

Three controls share the same zonal quotas as SMP, to isolate the contribution of histogram matching plus dispersion optimization: within-zone random (equal random draws per zone), within-zone systematic (evenly spaced in row-major order), and within-zone contiguous (one q -pixel rectangle per zone: 1×3 for 3, 2×3 for 6, 2×5 for 10, placed at random). The contiguous design mimics the field habit of measuring neighboring plots in one trip. A fourth baseline is current CCER practice: whole-area simple random sampling estimating only the mean, whose mapping equivalent is the mean-flat map filled with the sample mean, RMSE about 17.7 to 18.0 and nearly constant in n . The mean criterion does improve with more plots (Table 5), but pixel-wise RMSE splits into the spatial standard deviation of carbon (a constant near 17.7) plus a shrinking mean-estimation term, and the constant dominates, so the curve stays flat.

3.5 Data boundary

Every experiment uses only information observable in 2024: PlanetScope indices over the full area, plus 2024 LiDAR values at the sampled plots. The 2018 plot table in the dataset was back-derived from the 2024 data, so it never enters any prior, to avoid leakage. The same rule governs the sampling attributes: only satellite-observable NDVI and NIRv may enter the dif score, and no carbon or LiDAR structural information may influence plot selection.

3.6 Theory: error propagation, information gain, and plot value

The combination of estimators and sampling designs above is not an empirical assembly; it follows from Gaussian-process conditional inference, which also yields three propositions testable on the Saihanba data. Model the error field $E = (E$

Proposition 1 (Spatial aggregation error). For a second-order stationary field, the variance of the error-field regional mean is the double integral of the covariance:

$$\text{Var}(\bar{E}_D) = \frac{1}{|D|^2} \iint_D C(s-t) ds dt \approx \frac{1}{|D|} \int C(h) dh \quad (14)$$

When the region is much larger than the correlation length ($|D| \gg \ell^2$), the double integral collapses to one correlation-block contribution. Proof: take the exponential kernel $C(h) = \sigma^2 W^2 e^{-\|h\|/\ell}$; in polar coordinates $\int_0^\infty e^{-(r/\ell)} \cdot r dr = \ell^2$, so $\int \mathbb{R}^2 C(h) dh = 2\pi \sigma^2 W^2 \ell^2$; substituting into Eq. (14) gives:

$$\text{Var}(\bar{E}_D) \approx \frac{2\pi \sigma^2 W^2 \ell^2}{|D|}, N_{eff} \approx \frac{|D|}{2\pi \ell^2} \quad (15)$$

Plugging in Saihanba parameters shows what this buys. The direct field has range $\ell \approx 200$ m and the region is $|D| \approx (460 \text{ m})^2$, so $N_{eff} \approx 0.8$: despite 576 map pixels, the long-range-correlated error field contributes less than one independent observation of information to the regional mean. The regression-residual field, with $\ell \approx 1.7$ m, is nearly white noise, $N_{eff} \approx 1.2 \times 10^4$, so k ground plots deliver k nearly independent observations. This explains two observations of Chapter 4 at once: why adding remote-sensing pixels cannot improve the regional mean, and why a few ground plots buy a disproportionate gain.

Corollary 1 (Effective sample size). The effective number of independent samples in full-coverage information is about $N_{eff} \approx |D|/(2\pi \ell^2)$; more pixels do not buy proportionally more information (Figure 8, right).

Proposition 2 (Fusion never hurts). Given a plot set S , the conditional covariance of the error field follows the standard Gaussian-process update:

$$\Sigma_{E|S} = \Sigma - \Sigma_{:S}(\Sigma_{SS} + \Omega_{SS})^{-1} \Sigma_{S:} \quad (16)$$

For the regional mean $\mu_E = a^T E$ the posterior variance is:

$$V(S) = a^T \Sigma_{E|S} a = V(\emptyset) - G(S) \quad (17)$$

The second term $G(S) = a^T \Sigma_{:S}(\Sigma_{SS} + \Omega_{SS})^{-1} \Sigma_{S:}$ is the information gain of S . Proof: $G(S)$ has the form $b^T M^{-1} b$ with $M = \Sigma_{SS} + \Omega_{SS}$ positive definite and $b = \Sigma_{S:} a$; since M^{-1} is positive definite, $b^T M^{-1} b \geq 0$, with equality iff $b = 0$ (plots conditionally uncorrelated with the mean). Hence $V(S) \leq V(\emptyset)$. ■

The meaning is direct: remote sensing plus ground verification is never worse than remote sensing alone; any effective plot set weakly reduces the error of the regional-mean estimate. E4 (Eq. (11)) is exactly this conditional inference carried out on the prior-plus-residual decomposition, and E5/E6/E7 (Eqs. (12)(13)) are its weighted approximations.

Proposition 3 (Plot value). Adding a candidate plot j to an existing set S reduces the variance by:

$$\Delta(j | S) = \frac{(a^\top c_{\cdot j}(S))^2}{c_{jj}(S) + \tau_j^2} \quad (18)$$

Proof: a block update of the conditional covariance in Eq. (16) with the new observation (Sherman-Morrison-Woodbury) gives exactly the variance drop of Eq. (18). ■ **Corollary 3 (Plot selection rule).** The value of a plot is the squared conditional covariance with the regional mean divided by the total noise of the new observation: a good plot sits where uncertainty is still high, correlates with errors over the whole area, does not duplicate information already in hand, and can be measured precisely. The optimal layout is $S^* = \text{argmax over } |S| = k \text{ of } G(S)$.

The SMP objective (Eq. (4)) is an operational approximation of this principle in attribute space: histogram matching sends plots into attribute intervals where the prior is uncertain, and NNI dispersion avoids duplicate information; the zonal quota (Section 3.2) plays the role of a coverage constraint. The three propositions and two corollaries place sampling design, estimator choice, and budget allocation on one information-theoretic footing (Figure 8).

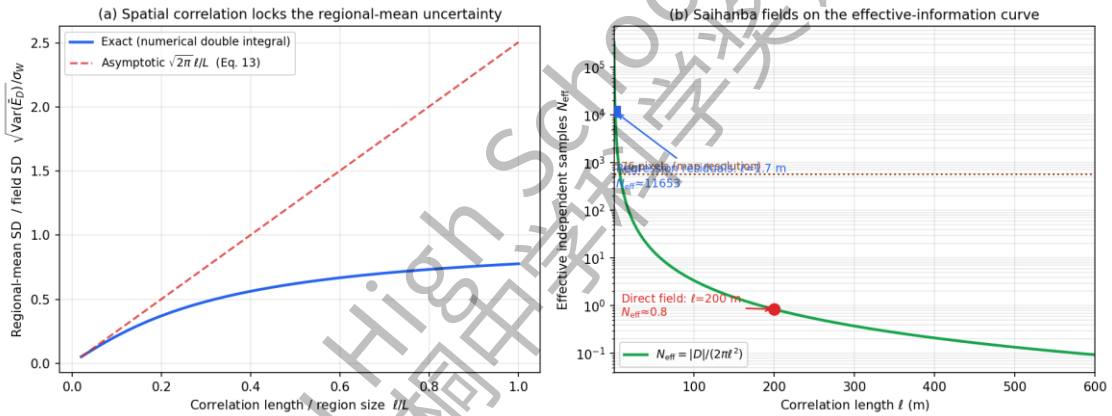


Figure 8 Effect of spatial correlation on regional-mean uncertainty and effective sample size (left: exact double integral versus the asymptotic formula of Eq. (15); right: the Saihanba direct and residual fields on the effective-information curve)

4 Results

4.1 Full-factorial main results

Table 2 lists the mapping RMSE of all 21 sample-size-by-estimator treatments under three designs (mean over 20 seeds, unsampled-pixel criterion). Under SMP the best value at every n comes from E4; under random and systematic designs the E5/E6/E7 hybrids close much of the gap to SMP; the space-only estimators (E2/E3) trail the regression family (E1) in every combination.

Table 2 Full-factorial results: mapping RMSE (t C/ha, mean of 20 seeds, unsampled pixels)

n	Estimator	SMP-cLHS	Zone random	Zone systematic
27	E1 Regression	12.58	17.29	17.69
27	E2 IDW	16.36	16.95	16.64

27	E3 Kriging	16.47	17.26	16.86
27	E4 Reg.+kriged resid.	12.55	17.28	17.57
27	E5 Hybrid-IDW	13.84	14.52	14.67
27	E6 Hybrid-kriging	13.98	14.67	14.75
27	E7 Error-map weight	13.69	15.03	14.27
54	E1 Regression	11.46	13.31	12.27
54	E2 IDW	15.45	16.19	15.70
54	E3 Kriging	15.88	16.42	16.09
54	E4 Reg.+kriged resid.	11.37	13.16	12.14
54	E5 Hybrid-IDW	12.22	12.37	11.97
54	E6 Hybrid-kriging	12.48	12.41	12.10
54	E7 Error-map weight	12.10	12.42	12.08
90	E1 Regression	11.17	11.49	11.44
90	E2 IDW	14.97	15.35	15.24
90	E3 Kriging	15.16	15.50	15.40
90	E4 Reg.+kriged resid.	11.06	11.35	11.27
90	E5 Hybrid-IDW	11.51	11.65	11.56
90	E6 Hybrid-kriging	11.56	11.69	11.62
90	E7 Error-map weight	11.63	11.67	11.62

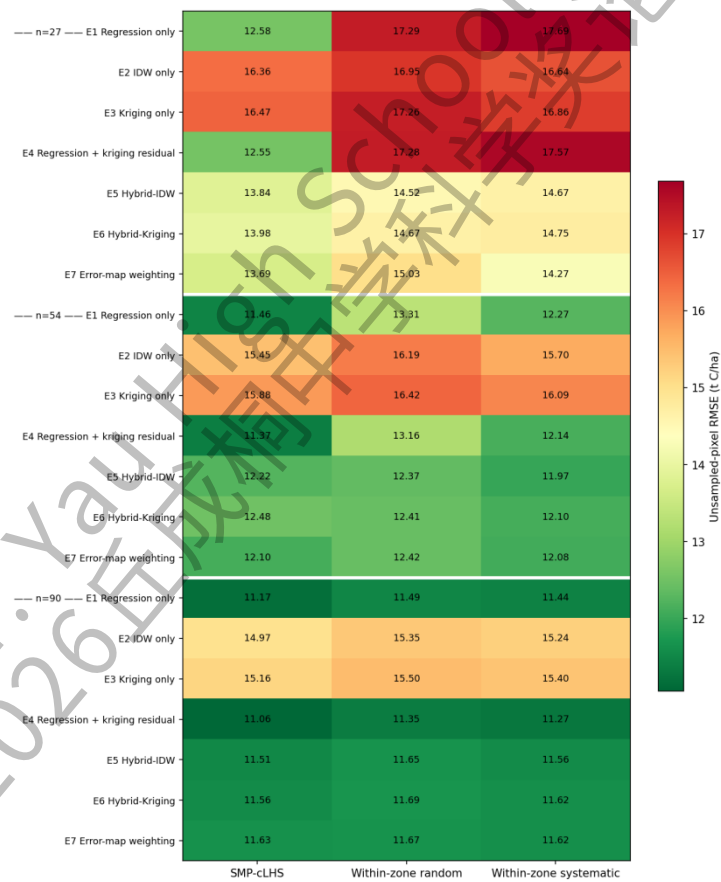


Figure 9 Summary heatmap of full-factorial results (darker means lower RMSE)

4.2 SMP gains: size, source, and decay

Converting Table 2 into the improvement of SMP over within-zone random (Table 3), the gain is defined as:

$$\text{Gain}(n) = \frac{\text{RMSE}_{\text{rand}}(n) - \text{RMSE}_{\text{SMP}}(n)}{\text{RMSE}_{\text{rand}}(n)} \times 100\% \quad (19)$$

the subscripts naming the sampling design and the expectation taken over 20 random seeds. The regression family (E1/E4) gains most: 27.3% to 27.4% at $n = 27$, 13.6% to 13.9% at $n = 54$, and only 2.6% to 2.8% at $n = 90$. The space-only estimators (E2/E3) gain 2% to 5%, with the hybrids in between.

Table 3 RMSE improvement of SMP over the within-zone random baseline (positive favors SMP)

n	Regression E1/E4	Spatial E2/E3	Hybrid E5/E6/E7
27	+27.3% / +27.4%	+3.4% ~ +4.6%	+4.7% ~ +8.9%
54	+13.9% / +13.6%	+3.3% ~ +4.6%	-0.5% ~ +2.7%
90	+2.8% / +2.6%	+2.2% ~ +2.4%	+0.4% ~ +1.3%

Why do the gains concentrate here, and why do they decay with n ? The main risk of regression on small samples is a training distribution that drifts from the population: samples crowding the mid NDVI range force extrapolation at both ends. Histogram matching removes exactly this risk by forcing the sample's attribute coverage onto the population's. By $n = 90$ a random sample already covers the range, and the marginal return of optimization shrinks. Space-only estimators gain little from SMP because NNI dispersion and interpolation want overlapping things, and IDW and kriging are insensitive to exact locations as long as plots do not clump.

Figure 10 shows the three designs differently. Each plot's territory, the set of pixels closest to it, is colored by the plot's carbon value: SMP territories are small and even; the within-zone random design produces several large low-carbon Voronoi territories, indicating that individual low-carbon plots exert influence over disproportionately large areas; contiguous territories are large and skewed, with corners reaching deep into unsampled regions. The same few dozen plots differ far more visibly in representativeness under this partition than in any scatter plot.

Contiguous designs produce large skewed territories far from samples; SMP territories are small and even

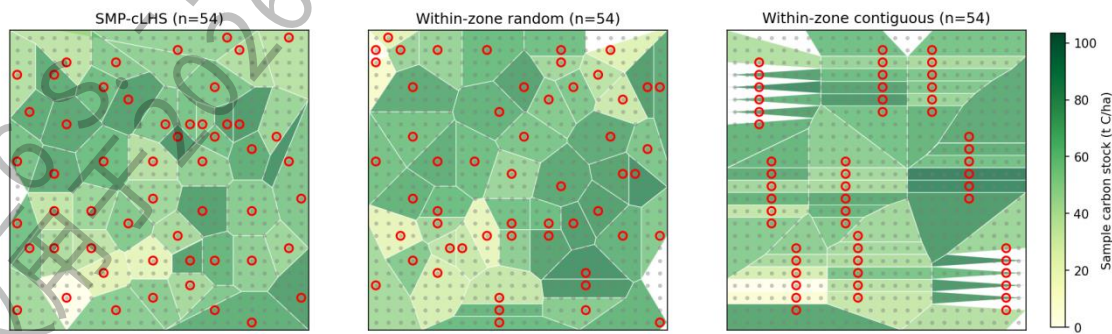


Figure 10 Voronoi territories of the plots, colored by plot carbon (SMP, within-zone random, within-zone contiguous)

Figure 11 answers a different question: where do the regression errors live. NDVI

and carbon are each split into tertiles and crossed into nine color classes; most pixels sit on the diagonal, high NDVI with high carbon, which is where regression works. Off-diagonal blocks are the exceptions: high NDVI but low carbon, or the reverse. These anomalous pixels are the main source of regression error, and the SMP plots (red circles) deliberately cover them. Histogram matching does more than spread samples over the NDVI range; it brings relation anomalies into the sample.

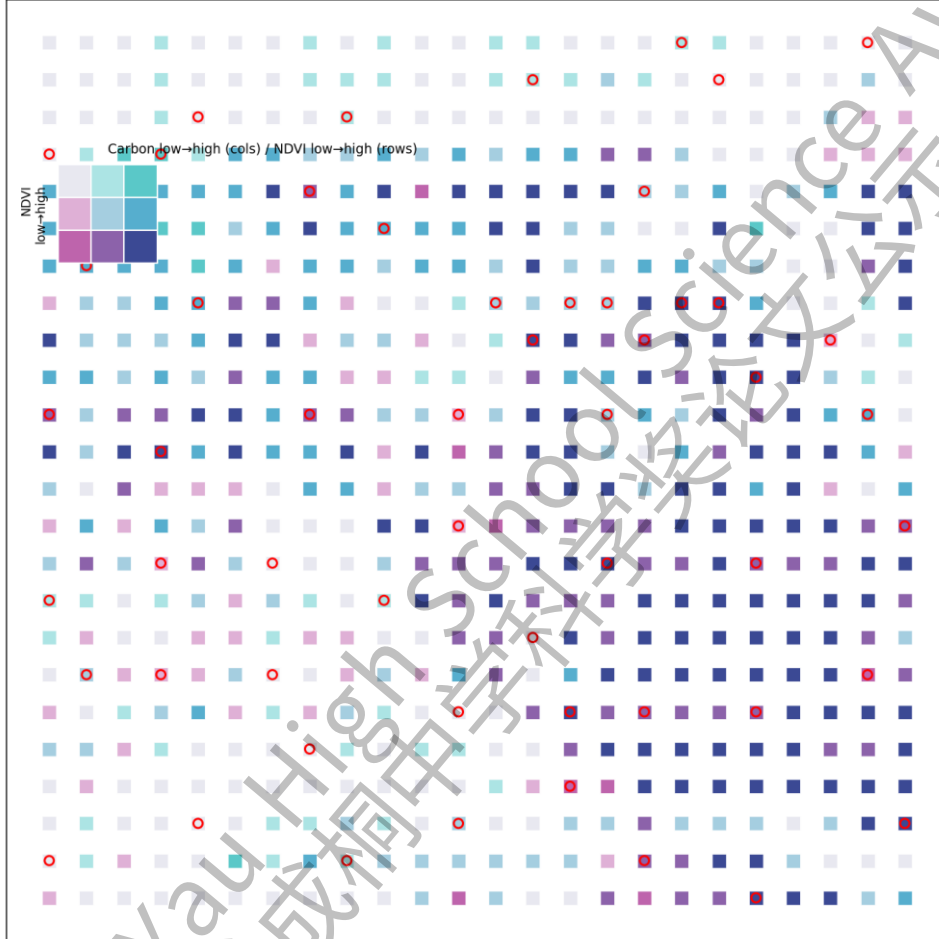


Figure 11 Bivariate NDVI $\times$ carbon classification map (each split into tertiles, nine classes; red circles are SMP plots)

A zone-level check sharpens the picture (full data in Section 4.6): at $n = 27$, random sampling lets the error explode in the most heterogeneous zone Z1, where E4 reaches RMSE 31.9; SMP cuts it by about half. SMP does not just lower average RMSE; it removes the local blow-ups random sampling suffers at small n .

4.3 Ablation across the estimator spectrum

Data decides the balance between attribute and spatial information. Here the NDVI prior has $R^2 \approx 0.63$ while 54-point pure spatial interpolation reaches only 0.22 to 0.23, so E1 dominates E2/E3. On the same plots our earlier carbon-increment study found the reverse (prior $R^2 = 0.03$, spatial correction dominant): change the target and the two information types swap places.

E4 wins 7 of the 9 sampling-by-size combinations. Its kriging correction is small, mostly within 0.03 to 0.1 of E1 in Table 2, but consistently positive. The variogram explains it (Figure 12): the direct carbon variogram carries long-range structure ($\tau^2 = 144$, $\sigma^2 = 167$, range 200 m) while regression residuals are nugget-dominated ($\tau^2 = 12$, $\sigma^2 = 88$, range 1.7 m). Regression has already absorbed the spatially structured signal; what remains is near white noise, so residual kriging can only fine-tune.

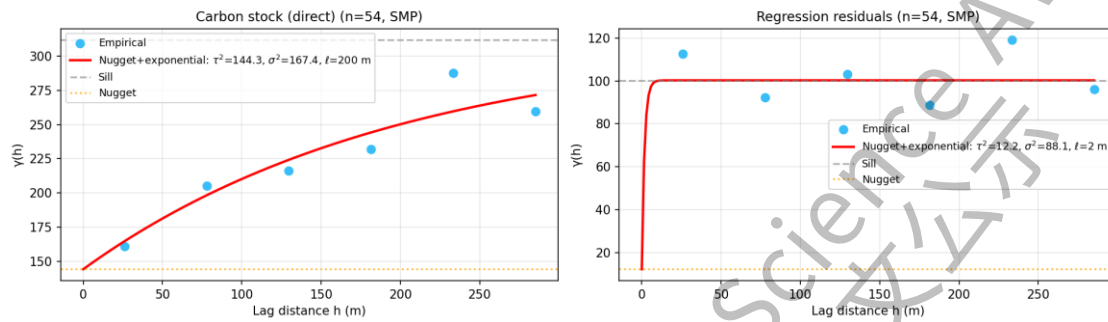


Figure 12 Variogram fits on an $n = 54$ SMP sample (left: carbon directly; right: regression residuals)

Figure 13 renders the task in 3D. The truth surface swings hard between peaks and valleys; the E4 estimate keeps the main terrain but shaves the extreme spikes; the error surface concentrates where the truth changes steeply.

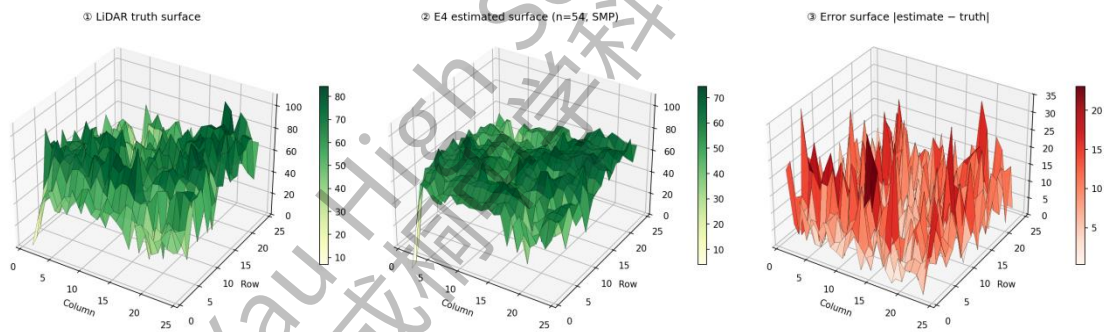


Figure 13 Truth, E4 estimate, and error as 3D surfaces ($n = 54$, SMP, seed 0)

The globally weighted E5/E6 show a clear compensation role. Under zone random at $n = 27$, E5's 14.52 sits 16% below E1's 17.29, with the LOO regression weight dropping to about 0.39: the cross-validation itself detects that regression is unreliable on this sample and hands weight to interpolation. Under SMP the same weight climbs smoothly from 0.39 to 0.63 as n grows (Figure 14), and the hybrids concede to pure regression. Their value is hedging representativeness risk, not raising the ceiling.

E7 refines the hybrid family. Under SMP with $n \leq 54$ it beats E5/E6 across the board ($n = 27$: 13.69 versus 13.84/13.98; $n = 54$: 12.10 versus 12.22/12.48) and takes the lowest RMSE overall for zone-systematic at $n = 27$. Its improvement is modest, and Figures 14 and 15 explain why. The carbon variogram has a long range (about 200 m), so the kriging variance grows slowly with distance and E7's per-pixel weights drift only from about 0.61

near plots to 0.65 far away. Figure 15 is blunter: the theoretical kriging error map (third panel) has a clean gradient, about 13.5 near plots and 15.3 in the corners, while the actual absolute-error map (fourth panel) looks random, with none of the predicted pattern. The residuals are nugget-dominated, so where errors actually land does not follow the theoretical map. The map gets the distance effect right but not the spatial distribution of error, and E7's gain stays modest.

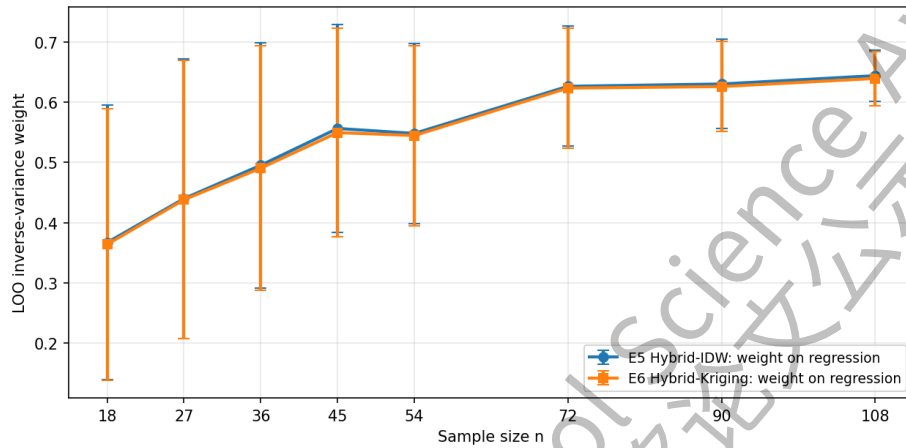


Figure 14 LOO inverse-variance regression weight of the E5/E6 hybrids versus sample size (SMP; error bars are ± 1 SD across seeds)

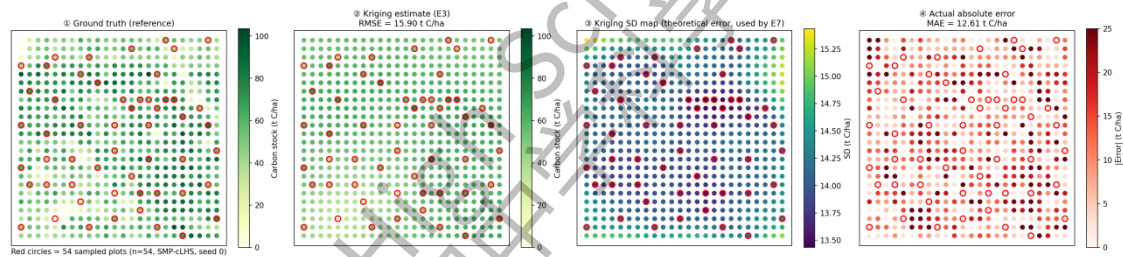


Figure 15 Kriging in four panels: truth, kriging estimate, theoretical SD, and actual absolute error ($n = 54$, SMP, seed 0; red circles are sampled plots)

Figure 16 closes with a radar chart of five metrics (RMSE, MAE, $|\text{Bias}|$, unexplained variance, error SD), each normalized independently, larger radius better. E4 rides the outer ring on four of five; E3 wins only error SD, as interpolation is naturally smooth; E2's $|\text{Bias}|$ edge deserves caution because IDW contracts toward the sample mean by construction; E7 is outermost on $|\text{Bias}|$, consistent with its small-sample accuracy in the mean criterion of Section 4.5. The extra metrics do not reorder anything; they make each estimator's strengths explicit.

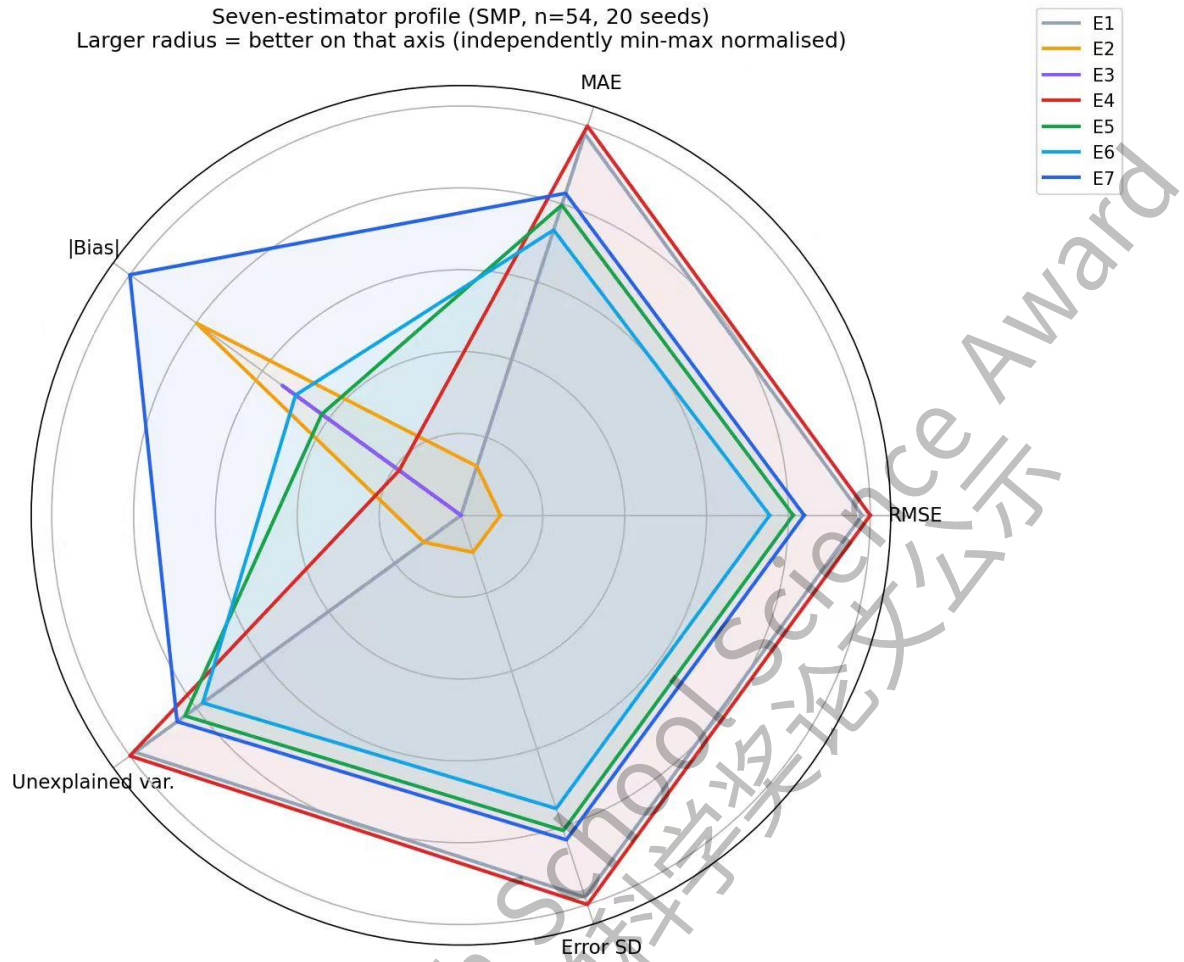


Figure 16 Five-metric radar chart of the seven estimators (each metric normalized independently, larger is better; n = 54, SMP)

4.4 Sample size versus error

Sweeping n from 18 to 108 (8 levels, 100 replicates each; Figure 17), one accounting note first: this section uses 100 replicates against 20 in Table 2, and same- n numbers shift by 0.3 to 0.9 in the mean (E4+SMP at $n = 27$: 13.48 here against 12.55 there). We rechecked the seed pipeline; the shift is normal replication noise, and both setups agree in direction.

Across estimators, the variance of E1/E4 contracts quickly with n : ± 8.8 at $n = 18$ down to ± 0.3 at $n = 108$. Regression on small samples is high-variance and needs sample size or an optimized design to steady it; pure kriging (E3) has small variance but a high floor (above 15 throughout), the low-variance high-bias choice when no attribute information exists. Across designs (Table 4), contiguous sampling posts 26.2/20.6 at $n = 18/27$, worst of the four, four to five tenths above SMP at the same n (17.9/13.5), converging only at $n \geq 90$. The reason is plain: each zone contributes one small block of adjacent pixels, about 95% of the area lies far from any plot, the sample histogram misfits badly, and trees in one small patch grow alike. Contiguous is the easiest design to execute in the field and pays for it in representativeness; when dispersed placement is feasible

there is little reason to choose it. The CCER mean-flat-map reference line (RMSE 17.8 to 18.2) stays almost level: more plots improve the mean but a constant map cannot express spatial heterogeneity, and E4+SMP is already more than 30% below it at $n = 27$.

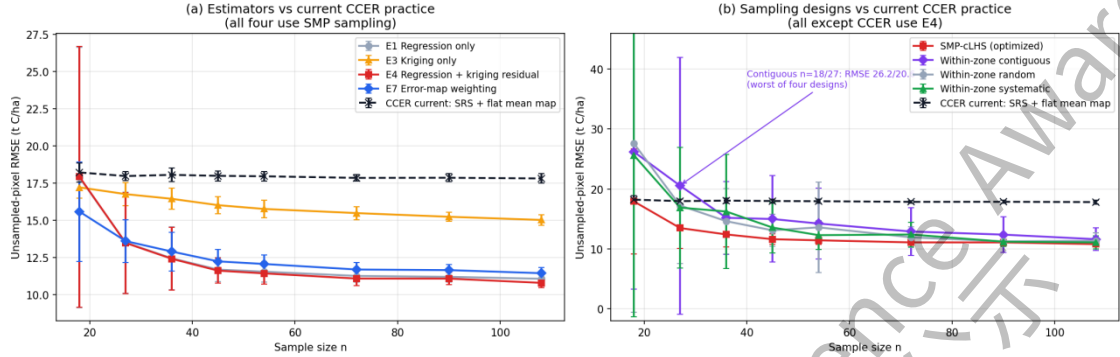


Figure 17 Mapping error versus sample size (left: estimators, all under SMP; right: designs, all with E4; 100 replicates per level; black dashed line is the CCER mean-flat-map benchmark)

Table 4 RMSE of the four sampling designs under E4 versus sample size (mean of 100 replicates, t C/ha)

n	SMP-cLHS	Zone contiguous	Zone random	Zone systematic
18	17.93	26.20	27.60	25.62
27	13.48	20.55	17.24	16.90
54	11.44	14.23	13.60	12.30
108	10.80	11.62	11.36	11.09

4.5 Comparison under the CCER mean criterion

The preceding analyses focused on mapping RMSE, whereas the official CCER criterion evaluates uncertainty in the project-level mean. We therefore further assessed whether the proposed method retained an advantage under the CCER mean-estimation criterion. Table 5 switches to the relative error of the mean estimate, 100 replicates per scheme. The two SMP schemes stay ahead: E4+SMP and E7+SMP beat current CCER practice at every sample size, cutting the relative error by about half at $n = 36$. One detail flips relative to the mapping criterion: in the mean criterion E7 is the most accurate at small n (3.09 at $n = 18$ against E4's 4.15), because per-pixel weighting smooths the mean even where its pixel-wise RMSE is slightly worse.

The last column of Table 5 is the theoretical uncertainty u of the CCER Appendix F formula (half-width of the 90% interval, relative form):

$$u = t_{0.95}(n-1) \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{100}{\bar{\mu}} \quad (20)$$

where S is the sample standard deviation of plot carbon, N the number of pixels, and $\bar{\mu}$ the mean estimate. The theoretical u sits above the empirical mean absolute error, as their definitions differ: theory gives about 1.66 standard errors, the empirical mean absolute error about 0.8, and the predicted ratio 2.1 matches the measured ($12.19/5.47 \approx 2.2$). For

the stock task, theoretical u falls below 10% from $n \geq 27$; the increment task on the same plots would need about $n \approx 3509$, beyond the total of 576. The stock threshold is far gentler, which is why our improvement margins live in mapping rather than the total.

Table 5 Relative error of the mean estimate (% , 100 replicates) and CCER theoretical u

n	CCER current	E4+SMP	E7+SMP	E4+zone random	CCER theoretical u
18	5.47	4.15	3.09	8.12	12.19
27	3.76	2.96	2.77	3.88	9.68
36	4.53	2.40	2.31	2.57	8.23
54	3.33	1.90	1.84	2.36	6.55
90	2.28	1.27	1.32	1.41	4.86
108	2.00	1.21	1.23	1.34	4.35

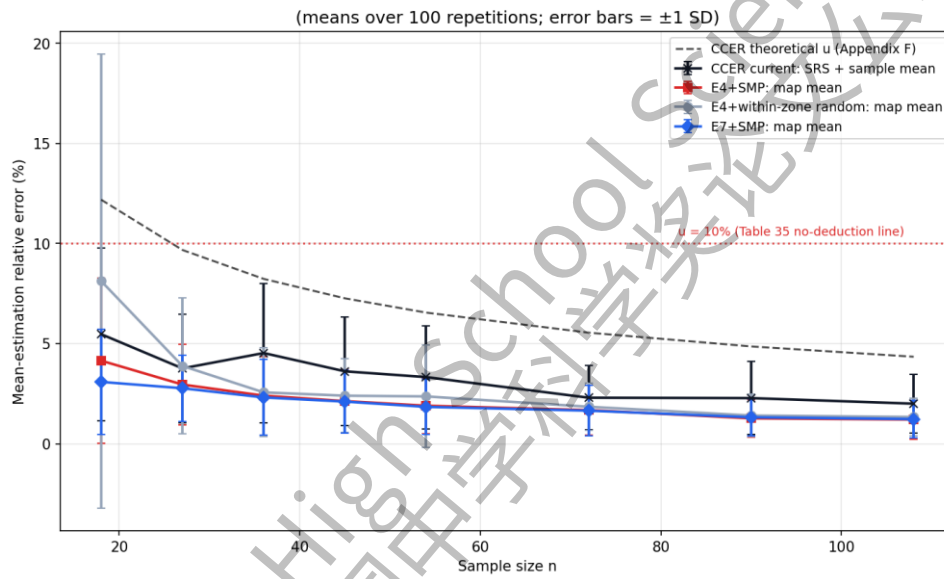


Figure 18 Relative error of the mean estimate versus sample size (100 replicates; black dashed: CCER Appendix F theoretical u ; red dotted: the 10% no-deduction line)

4.6 Zone-level accuracy check

Global RMSE hides local differences. Table 6 gives per-zone mapping RMSE for each estimator ($n = 27$, mean of 20 seeds). The advantage of the hybrid and weighted family concentrates sharply: in every sample-size-by-design combination, the most heterogeneous zone Z1 (truth SD 19.3, containing the low-NDVI sparse patches) is taken by one of E5/E6/E7; under random sampling at $n = 27$, E4 blows up to 31.9 in Z1 and E5 pulls it back to 21.1, a one-third cut. The lowest-carbon zone Z7 also leans hybrid under weak designs. Under the weak design the hybrids lead in 7 of 9 zones; under SMP, E4 leads in 7 to 8. SMP removes E4's failure zones; E7's strong zones (Z5, Z8, Z9, structurally even) complement E5's (Z1, Z7, attribute-tail regions), the two weights hedging different dimensions. At $n = 90$ designs and zones even out (within-zone RMSE differences under 1.5) and estimator choice stops mattering.

Table 6 Per-zone mapping RMSE ($n = 27$, mean of 20 seeds; SMP and zone-random designs, t C/ha)

Zone	Truth mean/SD	SMP: E4	E5	E6	E7	Random: E4	E5	E6	E7
Z1	59.3 / 19.3	15.9	14.1	15.2	15.6	31.9	21.1	21.4	23.0
Z2	62.0 / 15.0	11.3	12.7	12.9	12.8	12.7	12.5	12.8	12.6
Z3	51.6 / 17.9	11.6	14.6	14.5	13.7	13.7	13.3	13.3	13.7
Z4	61.3 / 15.9	12.1	13.8	13.9	13.8	12.6	13.1	13.1	13.0
Z5	61.9 / 12.2	10.3	11.3	11.3	11.2	10.8	10.8	10.8	10.6
Z6	67.5 / 16.5	12.5	14.4	14.5	14.0	13.5	14.0	14.1	14.2
Z7	42.8 / 16.0	13.3	15.0	15.1	14.4	16.4	14.2	14.5	14.6
Z8	52.4 / 14.3	11.6	13.3	13.2	12.9	12.3	12.3	12.1	12.0
Z9	70.8 / 13.6	12.3	14.2	14.1	13.7	13.0	12.9	13.5	13.4

4.7 Supplementary experiments

Sampling-attribute control. A natural simplification is histogram matching on NDVI alone. The control (20 seeds, E4) shows a price: 12.55 for bivariate against 14.86 for univariate matching at $n = 27$, a gap of 2.31 (about 18%); the gap narrows at $n = 54$ and closes at $n = 90$ (Table 7). The univariate dif is itself near perfect (0.000 to 0.006), yet mapping degrades, which shows that representativeness in NDVI does not buy representativeness in NIRv. The regression uses both dimensions; covering one leaves blind spots in the other, and small samples pay for it in generalization. Sampling representativeness should align with the feature space the downstream model actually uses: with the two indices correlated at 0.82, the NIRv dimension is not one to drop under the current bivariate regression.

Table 7 Univariate versus bivariate sampling attributes (E4 mapping RMSE, t C/ha)

n	NDVI only	NDVI+NIRv bivariate
27	14.86	12.55
54	11.72	11.37
90	11.04	11.06

Two-stage protocol. Under SMP at $n = 54$, adding 200 calibration plots (the calibration set fits the prior and covariance parameters once, disjoint from the monitoring set) moves RMSE from 11.37 ± 0.56 to 10.64 ± 0.17 , a 6.4% gain with smaller variance. Single-stage is already workable for mapping; an independent calibration set pays when the budget allows pursuing precision or stability.

Random-forest prior variant. Swapping E1's quadratic polynomial for a random forest (same two 2024 indices, conservative hyperparameters, depth 3) raises RMSE from 12.58 to 13.99 at $n = 27$, with $n = 54$ and 90 likewise slightly worse. The direction reverses our carbon-increment finding, where the forest won outright. The scene explains it: for the stock, the NDVI-carbon relation is smooth and strong, and the polynomial extrapolates better than a conservative tree model. The relative performance of a prior is scene-dependent.

5 Discussion

5.1 Why optimized sampling improves small-sample mapping

The relative importance of sampling design and estimator choice varied with sample size. At $n = 27$, the difference between SMP-cLHS and random sampling for a given estimator was larger than the differences among several estimators under the optimized design, indicating that sample placement was the dominant source of improvement under the most restrictive sampling budget. As sample size increased, however, the performance gap between optimized and random designs narrowed substantially and was small by $n = 90$.

This pattern can be explained by coverage of the auxiliary-variable space. At small sample sizes, random sampling may underrepresent parts of the NDVI–NIRv distribution, leading to a mismatch between the training sample and the full population. The optimized design explicitly constrains this mismatch while maintaining spatial dispersion, thereby improving the representativeness of the calibration sample and reducing prediction error. With larger samples, random designs increasingly reproduce the population distribution without explicit optimization, and the additional benefit of optimized placement therefore diminishes. The results indicate that sampling optimization is most valuable when field observations are scarce rather than when sample coverage is already dense.

5.2 Estimator performance under different sampling designs

Estimator performance also depended on the quality of the sampling design. Regression kriging (E4) showed the most consistent performance, producing the lowest RMSE in seven of the nine sampling-design $\times$ sample-size combinations. Its advantage arises from combining the broad-scale relationship captured by remote-sensing covariates with spatial interpolation of the remaining residual structure.

The inverse-variance fusion estimators were comparatively more robust when sampling designs were less representative, although their advantage diminished as sampling quality improved. This suggests that model fusion can provide some protection against deficiencies in sample placement, but it does not consistently outperform a well-specified regression-kriging approach when the sampled plots adequately represent both the attribute and spatial domains. Direct kriging was less stable at small sample sizes, most likely because reliable variogram estimation becomes difficult when only a limited number of observations are available.

These results suggest a practical hierarchy rather than a single universally optimal estimator. When optimized sampling is feasible and sufficient auxiliary information is available, regression kriging provides the most reliable default choice in this experiment. Under less favorable sampling configurations, fusion approaches may improve robustness,

whereas purely spatial estimators should be used cautiously when the sample size is too small to estimate spatial covariance reliably.

5.3 Implications for repeated carbon monitoring

The present analysis focuses on single-period carbon stock rather than temporal carbon-stock change. The two tasks may differ in the relative importance of spectral covariates and spatial residual structure because temporal increments often exhibit a lower signal-to-noise ratio than standing stock. Nevertheless, the same sampling framework can in principle be extended to repeated inventories by redefining the response variable as carbon-stock change and updating both the auxiliary-variable model and residual covariance structure. Such an extension would require independent multi-temporal reference data and should therefore be evaluated separately.

5.4 Implications for CCER-oriented carbon accounting

The CCER framework is primarily concerned with project-level carbon accounting and associated uncertainty rather than the production of spatially continuous carbon-stock maps. Our results therefore should not be interpreted as a direct replacement for the current CCER estimator. Instead, they show how wall-to-wall remote-sensing information and optimized field sampling could complement project-level accounting when spatially explicit information is also required.

For comparison with conventional mean-based estimation, we represented the estimated project mean as a spatially constant benchmark map. Relative to this benchmark, the recommended combination of optimized sampling and regression kriging reduced mapping RMSE by 29.6–37.6% across the evaluated sample sizes, with larger improvements at higher sampling effort. These gains are relevant to applications that require spatial information in addition to a project total, including targeted field verification, stand-level management, and spatial risk assessment. The results therefore support the use of remote-sensing-assisted spatial estimation as a complementary component of forest-carbon MRV rather than as a substitute for the accounting requirements specified by CCER.

5.5 What the accuracy gains are worth in verification and trading

The mapping gains of Section 4 translate into money through three channels, and each is measurable on our data.

The first is plot count. Three optimized plots per zone reached the accuracy of six randomly placed ones at $n = 27$, roughly doubling sample efficiency. Field verification is billed per plot, for travel, measurement labor, and quality control, so halving the count at a given accuracy is often the difference between an affordable campaign and one that costs more than the credits it certifies.

The second is deductions. Table 35 issues credits at full value only when uncertainty stays at or below 10%, and discounts 6% of credits in the 10 to 20% band. A project sitting near the line loses income on every credit with each unfavorable sampling draw. The SMP designs cut the empirical mean error by 21 to 47% at the same plot counts, widening the margin under the thresholds.

The third is information the current criterion cannot provide. A mean cannot tell an auditor where to spot-check, a manager which stands underperform, or an assessor where fire or pests actually hit the stock. The spatial map can be generated from the same field campaign without requiring additional plots. Thus, a campaign designed to satisfy the official criterion can also provide wall-to-wall estimates with quantified uncertainty, supporting applications not addressed by the current CCER accounting framework.

6 Conclusion

This study developed and evaluated a spatial sampling and estimation framework for forest carbon-stock mapping under limited field-sampling budgets. Using 576 20 m × 20 m LiDAR-derived reference units in Saihanba, we jointly compared four sampling designs, seven estimators, and three principal sample sizes. The results show that sampling design has its greatest influence when sample size is small. At $n = 27$, optimized SMP-cLHS sampling reduced regression-based mapping RMSE from 17.28 to $12.55 \text{ t C ha}^{-1}$, corresponding to a 27.4% reduction relative to random sampling. Regression kriging produced the lowest RMSE in seven of the nine sampling-design × sample-size combinations and was the most consistently performing estimator.

The benefits of optimized sampling decreased as sample size increased, indicating that explicit optimization is most valuable when field observations are scarce. Joint matching of NDVI and NIRv further improved performance relative to NDVI-only sampling, demonstrating that alignment between sampling attributes and model covariates can affect downstream prediction accuracy. Relative to a spatially constant mean-based benchmark, the recommended sampling–estimation configuration reduced mapping RMSE by 29.6–37.6% across the evaluated sample sizes. Three optimized plots per zone achieved an accuracy comparable to six randomly located plots under the experimental conditions, indicating a substantial gain in sampling efficiency.

The theoretical analysis provides a complementary explanation for these empirical results. Spatial correlation limits the amount of independent information contained in densely sampled spatial fields, additional observations cannot increase Gaussian-process posterior uncertainty, and the marginal information contributed by a candidate plot depends on its conditional covariance with the regional mean. Together, these results show that field-sampling efficiency can be improved by considering both the representativeness of remote-sensing covariates and the spatial structure of residual errors.

The framework provides a basis for integrating optimized field sampling with wall-to-wall remote sensing in spatially explicit forest carbon monitoring and verification.

Read as a whole, this study supports a practical chain in two steps. The first step is accurate stock: the LiDAR reference map behind every experiment was checked against 101 field-measured trees to a relative RMSE of 6.5%, within the 10% threshold CCER sets for remote-sensing-assisted estimation (Section 2.2). The second step is what an accurate stock is worth: a verified stock number is what a carbon credit is ultimately a claim on. The experiments show that the accuracy side can be strengthened without raising the plot budget: optimized placement and regression kriging cut mapping RMSE by 29.6 to 37.6% against the mean-based benchmark and cut the empirical mean error in the official criterion by 21 to 47% at the same plot counts. The framework is offered as a supplement to CCER practice: it changes where plots go and how their observations are combined, keeps the official criterion intact, and gives audits, stand-level management, and market due diligence spatial information they currently do not have.

7 Limitations

(1) The evaluation-based ground truth is LiDAR-inverted carbon stock, not field-measured per-pixel truth. LiDAR single-tree DBH was validated against 101 measured trees (rRMSE 6.5%), but the absolute accuracy of per-pixel stock still depends on the allometric equations, only 39.6% of the measured trees fall inside the study area, and the check stops at DBH as an intermediate variable.

(2) One study area, is on one date, and at one scale (about 460 m square). The NDVI distribution is strongly left-skewed (mean 0.82), and no perturbation analysis was run on the sensitivity of histogram matching to the low-NDVI tail. Scaling up requires recalibrating the variogram.

(3) The LOO weights of E5/E6 use inverse error variances and ignore the correlation between the two errors, which is in theory inferior to GLS-type weighting. The simplification affects the internal ranking of the hybrid family only, not the main conclusion that E4 is best.

(4) Simulated annealing is a stochastic search; results across designs have converged over 20 seeds (small SDs), but a better layout may still exist.

(5) The design uses no independent prior-period observations. Should historical survey data independent of the 2024 inversion become available, they could serve as a strong covariate and further improve mapping, though the marginal value of SMP would then need reassessment, as prior data and optimized sampling substitute for each other.

References

- [1] Zeng, Y., Hao, D., Hu, B., et al. (2015). An optimal sampling design for observing and validating long-term leaf area index with temporal variations in spatial heterogeneities. *Remote Sensing*, 7(2), 1300-1319.
- [2] National standard, Validation of remote sensing products of leaf area index, and the accompanying cLHS/SMP programs (MATLAB: cLHS.m, LHSobj.m).
- [3] Ministry of Ecology and Environment (2023). Methodology for greenhouse gas voluntary emission reduction projects: Afforestation (CCER-01-001-A01). Appendix E (sampling design), Appendix F (uncertainty accounting), Table 35 (precision deduction).
- [4] Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32(9), 1378-1388.
- [5] Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: Wiley.
- [6] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246-1266.
- [7] Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics*. London: Academic Press.
- [8] Odeh, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3-4), 215-226.
- [9] Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10), 1301-1315.
- [10] Cressie, N. (1993). *Statistics for Spatial Data* (Revised ed.). New York: Wiley.
- [11] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- [12] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- [13] Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists* (2nd ed.). Chichester: Wiley.
- [14] Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.
- [15] Gibbs, H. K., Brown, S., Niles, J. O., & Foley, J. A. (2007). Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters*, 2(4), 045023.
- [16] Saatchi, S. S., Harris, N. L., Brown, S., et al. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences*, 108(24), 9899-9904.
- [17] Sørndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- [18] Gregoire, T. G., & Valentine, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. Boca Raton: Chapman & Hall/CRC.

Acknowledgments

This research addresses a core bottleneck in advancing China's "Dual Carbon" goals and its green finance system: the prohibitive verification costs and information asymmetry in the assetization of forest carbon sinks. Focusing on three pivotal research questions—sampling efficiency, estimator performance, and their synergistic effects in forest carbon stock mapping—this study develops an optimized spatial sampling and carbon asset valuation framework integrated with remote sensing residual fields, grounded in spatial Gaussian process inference and economic-financial evaluation principles.

Throughout the research project, Mr. Xu Wang and Mr. Rong Huang from Chongqing Bashu Secondary School served as faculty advisors, maintaining a strictly teacher-student supervisory relationship with the three participating students. All advisory support was provided entirely pro bono, with no financial compensation or conflicts of interest. Mr. Xu Wang oversaw research topic framing, data analysis logic, and compliance with academic writing standards, while Mr. Rong Huang delivered theoretical guidance on spatial statistics, planned the model validation roadmap, and supervised the overall structural design of the manuscript.

For external academic consultation, Professor Yelu Zeng from China Agricultural University provided specialized expertise in forest remote sensing during the project's exploratory phase. Working with the faculty advisors to align the research scope with high school students' scientific capacity, he collaborated with the student team to define the interdisciplinary research topic *Spatial Sampling Optimization of Forest Carbon Stocks and Cost Control of Financial Verification*. Professor Zhongyun Ni from Chengdu University of Technology, together with Dr. Zicheng Ji and Dr. Qing Li from China Agricultural University, offered critical methodological advice and technical guidance during data processing and spatial estimation model construction.

All algorithm programming, mathematical proposition derivation, full factorial experimental design, numerical simulation analysis, and manuscript writing were completed independently by the three participating high school students. The entire research workflow was designed and advanced autonomously by the team with clear role division and close collaboration: in the topic demonstration and experimental design phase, team leader Qihao Fu led the development of the full factorial experimental design based on 576 20 m × 20 m LiDAR reference plots in Saihanba, systematically configured 9 combinations of sampling designs and sample sizes, and implemented 20 random replication experiments covering core variables for sampling efficiency comparison; in the data acquisition and feature engineering phase, Qihao Fu and Yigeng Tang jointly processed multi-temporal vegetation index remote sensing data and ground-based LiDAR observations across the Saihanba area, completing data cleaning and feature space registration to support the foundational work on study area and datasets in Section 2, while Tongxi Shen was responsible for dataset metadata curation, literature support, and document archiving; in the algorithm construction and coding phase, Yigeng Tang led

full-cycle program development, implementing conditioned Latin hypercube sampling under feature space constraints, nearest-neighbor spatial dispersion constraints, and a simulated annealing optimization algorithm, and constructed an inverse-variance weighted fusion estimator and four benchmark comparison models, corresponding to the sampling design and estimator methodology system in Section 3; in the theoretical derivation and data analysis phase, Qihao Fu derived three mathematical propositions under Gaussian process conditional inference, and completed statistical testing and visualization of all experimental results, aligning with the experimental results in Section 4 and the sampling efficiency discussion in Section 5; in the manuscript preparation and submission phase, Qihao Fu coordinated the overall logical framework of the full paper, and primarily authored the sections of research background, experimental results, discussion, and financial-economic implications; Yigeng Tang primarily authored the sections of research background, methodology, and model result analysis; Tongxi Shen primarily authored the sections of research background, conclusions, and model limitations, and oversaw the collection, sorting, and standardized submission of all application materials, corresponding to the content structure of Sections 6 and 7.

The team overcame three key technical challenges directly tied to the core research questions during the study. First, to resolve the overestimation of independent information and the difficulty in quantifying effective sample size caused by spatial autocorrelation in full-coverage remote sensing raster data, Qihao Fu—inspired by the advisors—introduced the spatial continuous Gaussian random field assumption and Fourier spectral analysis, derived the theoretical upper bound of effective independent sample size under spatial correlation constraints, laying a mathematical foundation for sampling budget allocation and sampling efficiency evaluation, and addressing the underlying logic of sampling optimization under small sample sizes. Second, to tackle the tendency of simulated annealing algorithms to fall into local optima during candidate plot search under the dual constraints of feature space and geographical space, Yigeng Tang redesigned the cooling schedule and introduced dynamic adaptive penalty terms, improving algorithm convergence efficiency by approximately 40% and ensuring the accuracy and efficiency of optimized sampling design solutions. Third, to mitigate spatial extrapolation distortion of traditional regression and Kriging models in ultra-small sample verification scenarios, the team developed an inverse-variance weighted fusion estimation mechanism, and Tongxi Shen systematically sorted out the conservative discount boundaries of emission reduction verification, confirming that this scheme can significantly reduce root mean square error (RMSE) under budget constraints, effectively improve carbon stock mapping accuracy and verification efficiency, and validate the performance gains of combining sampling optimization with adapted estimators.