General assessment

This manuscript presents a potentially valuable annual stem-density dataset for China at 30 m grid spacing for 1985–2022. Its main contribution combines inventory-based calibration, long temporal coverage and inclusion of trees outside forests. I appreciate the spatially blocked validation, separate hold-out evaluation and acknowledgement of high-density underestimation.

However, the evidence supports spatial prediction more clearly than annual changes across the full reconstruction period. Temporal transferability, dependence on upstream products, non-forest performance and uncertainty require stronger assessment. I recommend major revision, prioritizing these issues rather than additional methodological complexity.

Major comments

1. Validate temporal transferability and changes

The reference inventories cover 1999–2018, whereas the product spans 1985–2022. A pooled model reduces discontinuities between period-specific models but does not establish temporal transferability. Agreement among scenarios sharing observations and predictors is not independent validation.

Please conduct leave-one-inventory-cycle-out or forward/backward transfer tests, reporting period-specific and regional bias alongside RMSE and R^2. Where repeated plots are available, assess observed versus predicted density changes against a no-change baseline. Filippelli et al. (2024; https://doi.org/10.1016/j.srs.2024.100119) provide a relevant framework. For earlier years, assess departures from the training domain and use historical evidence where available; otherwise, clearly qualify the hindcasts and their applications.

2. Clarify reference samples, zeros and measurement dates

Each inventory cycle contains exactly 670,279 observations. Please distinguish unique locations, repeated measurements and plot-period records, and explain the resulting Scenario 3 sample after selecting 25% from each cycle. Summarize sample counts and zero/non-zero proportions by cycle, climate zone and land-cover class.

Although zero-density plots are defined in Section 2.2.1, clarify how their status was established and whether any zeros were assigned without field measurement. Distinguish genuine zeros from missing observations and report performance for non-zero plots separately, particularly given the large alpine sample. Also explain how actual survey dates were matched to predictors rather than treating five-year cycles as single years. Where necessary, assess sensitivity to cycle-end-year matching and document relevant coordinate uncertainty.

3. Assess dependence on forest age, canopy cover and land cover

Forest age, LULC and canopy cover dominate the reported feature importance. Please test how strongly the reconstructed trajectories depend on these products through controlled ablations, using comparable spatial partitions and tuning. Evaluate national trends and non-forest estimates as well as spatial accuracy. Clarify forest-age encoding outside its valid domain and the nearest-year substitutions for early auxiliary data. Document whether upstream calibration samples overlap with the NFI hold-out observations and, where feasible, assess sensitivity to plausible input errors. Fig. 5 provides useful cross-sectional evidence but does not independently validate mapped temporal changes.

4. Validate trees outside forests separately

Pooled validation does not establish performance in cropland, grassland, shrubland, impervious surfaces and wetlands. Please report class-specific sample sizes, non-zero prevalence, bias and MAE/RMSE, including results for positive non-forest observations. Where feasible, assess tree-presence omission and commission errors. Independent counts should use compatible diameter thresholds. High-resolution imagery or UAV data can supplement validation, but visible crowns are not equivalent to all stems with DBH ≥5 cm. Also distinguish trees in impervious pixels from the broader urban-tree population.

5. Distinguish ensemble spread from predictive uncertainty

The five models share substantial training data, predictors and assumptions. Their standard deviation measures sensitivity to training composition and may remain small despite shared biases. Please apply the narrower interpretation given at lines 337–338 consistently throughout the manuscript. Complement ensemble spread with stratified hold-out residual statistics and assess coverage if prediction intervals are provided. Explain how uncertainty in national totals and changes was calculated, accounting for spatial and temporal dependence. Spatial-block resampling and sensitivity to dominant input products would strengthen the assessment without requiring a complete error budget.

6. Clarify spatial support and prediction consistency

Field plots cover approximately 667–800 m², whereas optical predictors are averaged over 8,100 m². Please state whether the same 3×3 averaging is applied during wall-to-wall prediction. Different transformations during training and inference could cause a predictor-distribution mismatch. Compare plausible extraction supports using unchanged validation partitions, particularly for forest edges and sparse non-forest trees. Distinguish 30 m grid spacing from effective spatial detail, and document predictor resampling, alignment and missing-value handling.

7. Quantify density-dependent bias and strengthen product comparisons

In Fig. 6 and lines 533–617, the final Model 3 hold-out panel reports R^2=0.68, RMSE=165.44

stems ha^{-1} and a fitted slope of 0.68, whereas Table 3 reports development cross-validation statistics. Please keep these evaluation levels distinct and quantify residual bias by observed-density range, forest type and climate zone. A slope below one indicates range compression, not a directly interpretable percentage bias in the national total.

For Figs. 10–11, specify the area-weighted aggregation and masking procedures, not only grid harmonization. Control for reference year and stem definitions as far as possible. In particular, the publication year of the static Crowther product should not be treated as a uniform observation year. If NFI diameter records permit, evaluate a DBH ≥10 cm subset to improve comparability. Differences from a LiDAR-derived map may arise from subcanopy detection, but higher predictions do not by themselves establish greater accuracy. Evaluate products against common compatible reference observations before making superiority claims.

For the SAR comparison, the similar 2008 results and larger 2018 difference suggest period-dependent benefits. Use a matched ablation with the same samples, partitions and comparable tuning to support claims specifically attributed to SAR, rather than differences among separately optimized pipelines.

Specific and minor comments

1. Define the mapped variable consistently as stem density for DBH ≥5 cm, including treatment of multi-stem individuals. Avoid implying that density changes alone identify recruitment, mortality or carbon gains.

2. Identify the exact CLCD, CAFA and ChinaTCC versions, including post-2019 CLCD data. Reconcile SAR availability with the years used and distinguish Scenario 1 inputs from those of the released product.

3. Consolidate the repeated spatial-partitioning descriptions in Sections 2.3.3–2.3.4. In Fig. 6, define the fitted regression, R^2 calculation and uncertainty in feature-importance scores; importance should not be interpreted as explained variance or causality.

4. Clarify the source of ages in Fig. 5 and qualify interpretations of cross-sectional age patterns, particularly for sparsely sampled old plantations.

5. Provide annual numerical values underlying Figs. 7–8, clarify their land-cover coverage and check unit conversions. Add local examples to Fig. 9 to demonstrate practical spatial detail. In the discussion, prioritize demonstrated validation limitations over generic suggestions to replace algorithms.