Remote sensing-based forest canopy height mapping: some models are useful, but might they provide us with even more insights when combined?

Responses to reviewers (please see the revised manuscript)

Nikola Besic[*1], Nicolas Picard[2], Cédric Vega[1], Jean-Daniel Bontemps[1], Lionel Hertzog[1], Jean-Pierre Renaud[1,3], Fajwel Fogel[4], Martin Schwartz[5], Agnès Pellissier-Tanon[5], Gabriel Destouet[6], Frédéric Mortier[7,8], Milena Planells-Rodriguez[9], and Philippe Ciais[5]

[1]IGN, ENSG, Laboratoire d'inventaire forestier (LIF), 54000 Nancy, France

[2]Groupement d'Intérêt Public (GIP) Ecofor, 75116 Paris, France

[3]Office National des Forêts RDI, 54600 Villers-lès-Nancy, France

[4]Department of Computer Science, École Normale Supérieure, 75230 Paris, France

[5]LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris Saclay, 91191 Gif-sur-Yvette, France

[6]UMR SILVA, INRAE, AgroParisTech, Université de Lorraine, 54280 Champenoux, France

[7]CIRAD, Forêts et Sociétés, 34398 Montpellier, France

[8]Forêts et Sociétés, University of Montpellier, CIRAD, 34090 Montpellier, France

[9]CESBIO, Université de Toulouse, CNES/CNRS/INRAE/IRD/UPS, 31401 Toulouse, France

[*] *Corresponding author: nikola.besic@ign.fr or n.m.besic@gmail.com*

Dear reviewers, thank you very much for your very stimulating questions, comments and suggestions. We did our best to address them appropriately in the revised version of the manuscript, and by doing so to improve the quality of the presented research.

The remarks and comments are enumerated and denoted as **R1**/**R2** + **C** + **number** and *the response to each of them* contains the reference to the modified section of the manuscript which you will find denoted using the blue font color in the annotated version of the revised manuscript.

The figures framed in blue were equally modified:

- either substantially (Fig. X),

- or slightly (Fig. X).

Reviewer 1 (R1)

This manuscript is based on remote sensing data and deep learning technology to carry out high-resolution mapping of forest canopy height, which has positive value for the application of remote sensing technology in forestry.

In summary, major revisions are recommended. Specific suggestions are as follows:

R1.C1: The title should be modified to be more precise and concise;

The title has been revised to address the comments from both reviewers (see R2.C7).

However, we chose to retain the paraphrased aphorism from George Box, as we believe it not only reflects the content of the paper but also has the potential to attract GMD readers who may not have a specific interest in forestry.

$\Longrightarrow$ *See Title: Page 1.*

R1.C2: References should not be included in the abstract, and the abstract should be revised to be more direct and clear;

References previously included in the original abstract have been removed, and several formulations have been revised to address the reviewers' feedback (see R2.C8, R2.C9, R2.C10, R2.C11, R2.C12, and R2.C13). Overall, the abstract has been substantially revised.

$\Longrightarrow$ *See Abstract: Page 1.*

R1.C3: In the introduction, a summary of the deficiencies of previous work should be added to introduce the work of this manuscript and increase logical coherence;

To the best of our knowledge, no similar exercises have been reported in the literature for remote sensing-based models/products.

Nevertheless, we were careful to evoke in the introduction an alternative to the Bayesian Model Averaging (BMA) approach presented, the Simple Model Averaging (SMA) method, both to enhance logical coherence and to establish a benchmark for comparison throughout the manuscript.

The shortcomings in applying BMA, observed in other fields (e.g., meteorological models), particularly its inability to effectively address bias and the saturation of individual models, were already highlighted in the abstract, discussion, and conclusions. Our study confirms these issues, which are now also discussed in the introduction.

$\Longrightarrow$ *See Introduction: Page 3.*

R1.C4: The structure of the manuscript should be modified to include five sections: Introduction, Data and Methods, Results and Analysis, Discussion, and Conclusion;

The structure of the manuscript has been modified to include the following sections: Introduction, Data: Models' descriptions, Data: Reference dataset description, Method description, Results and Analysis, Discussion, Conclusions.

For the sake of clarity, we found it important to present the models, reference data, and applied methods in separate sections.

$\Longrightarrow$ *See entire manuscript.*

R1.C5: Although there is a section for "Results and Discussion," the discussion content is very limited. It is recommended that the discussion section be separated into an independent section and that more content be added;

As recommended, the 'Results and Discussion' section has been split into two distinct sections: 'Results and Analysis' and 'Discussion.' The 'Discussion' section has been significantly expanded, with two main areas of focus: a more in-depth analysis of the observed effects and a comprehensive discussion of the limitations of the presented approach.

$\Longrightarrow$ *See sections 5 and 6: Pages 9 - 20.*

R1.C6: From Figure 4, the scatter plot does not show a good pattern, which may indicate that the analysis method in this paper is not ideal. Please analyze the problems in it;

Instead of simply presenting the correlation coefficient and the ideal regression line, we have fitted the most appropriate regression line to the data.

We do not believe that the patterns observed in these scatter plots compromise the validity of the employed approach or the derived conclusions. Namely, while there are some discrepancies between the models at low altitudes or in flat terrain, the fitted regression lines and Pearson correlation coefficients indicate a clear tendency: as terrain elevation and slope increase, the variation among models becomes more pronounced.

Although a non-linear curve might fit the data slightly better, the overall message conveyed would remain unchanged.

$\Longrightarrow$ *See Fig. 4: Page 14.*

R1.C7: Please add more discussion content. It is recommended that the authors analyze and compare the advantages and limitations of this work from multiple perspectives. In addition, it is recommended to add a future outlook.

That's a valid point. We have significantly expanded the discussion in the revised manuscript, which now extends over three complete pages without figures.

Additionally, the response to comment R2.C1 provides a detailed examination of the limitations, which has had significant influence on the enhancements made to the revised discussion section.

$\Longrightarrow$ *See section 6: Pages 17 - 20.*

Reviewer 2 (R2)

Overall comments

The paper "Remote sensing-based high-resolution mapping of the forest canopy height: some models are useful, but might they be even more if combined?" by Besic et al. is an interesting study of various regional or global-level

canopy height models, their performance across an entire country (France) and whether a mixture of these models contains additional information that could be exploited to improve predictions. The general approach to validate canopy height, based on a fully independent, field-based data set (French NFI data) convinces me, as it ensures good spatial coverage, but also allows the authors to extend the evaluation to forest structure metrics that cannot be properly assessed from remote sensing data only, such as Lorey's height. They thus can assess the link between what is seen from above and what is seen from below, with important implications for biomass assessments.

Overall, the study is timely and innovative. As the authors note, there is an increasing amount of (poorly validated/flawed) global canopy height models, which will likely be used for a variety of purposes, such as biomass assessments, disturbance monitoring or the validation/calibration of vegetation models. It is thus crucial to systematically assess these products, provide guidelines on their strengths and weaknesses and find ways to mitigate errors. I thus think that the study at hand would be of great interest and value to readers of Geoscientific Model Development.

I do, however, have a few major concerns that the authors need to address before I could recommend the article for publication. I will detail them in the next few paragraphs and then provide line-by-line comments below.

Model validation

R2.C1: The authors describe the mixture model derived via Bayesian Model Averaging (BMA) as providing a superior performance compared to each of the individual canopy height models, but the article currently does not provide enough evidence for this statement. As far as I understand, BMA is essentially fitting a higher-order mixture model to reference data, where, for every predicted site, the individual models get assigned weights based on performance at that site. The weights are model parameters, and I imagine that BMA is as prone to overfitting as any other modelling approach, particularly when done only in a "Bayesian-flavoured" way without fully propagating uncertainty. I also did not understand whether or how the site-specific weights are regularized (e.g., hierarchically nested within the overall weights?). I would therefore expect such a mixture to automatically perform better, because it allows to pick models based on local predictive accuracy, but it is not clear to what extent this approach will be fitting noise and to what extent it will pick up on systematic differences between models.

To make this analysis convincing and assess how well the model captures systematic patterns, the authors need to employ a cross-validation approach, ideally a spatial cross validation approach (as in Ploton et al. 2020, Nat. Comms.). One suggestion would be to do a spatial leave-one-out cross validation where 1,000 NFI plots are selected at random. For each NFI plot, the authors would then remove the NFI plot itself and all other NFI plots within a specific radius around the NFI plot (I would suggest 100 km to account for a decent range of spatial autocorrelation), calculate the BMA weights from the remaining training NFI plots outside the 100 km radius, and then predict with the trained BMA mixture to the validation NFI plot. This would be repeated 1,000 times. The comparison between prediction and observed value would then allow the calculation of a RMSE/MBE that should be relatively robust to overfitting and spatial autocorrelation between training and validation data. If computational costs are low, this could, of course, also be done from all 5,475 NFI plots.

If computational costs are higher, a slightly less computationally intensive alternative would be to split the data set into spatial folds such as the 91 sylvo-ecological regions (I'll call them SERs here), and perform a leave-one-SER out validation. So all NFI plots within one SER would be predicted from the BMA mixture trained on the other 90 SERs. Ideally, the authors would also account for spatial autocorrelation in this approach by removing the SERs directly adjacent to the validation SER, so the training data set would usually be 80-90 SERs, and the validation a single SER, buffered by its neighbouring SERs.

This is a very valid point! Fig. 5 in the manuscript indeed illustrates the goodness-of-fit of the obtained mixtures with respect to the data used for its fitting. The goal of this figure was specifically to show that the obtained mixtures make sense, meaning that the weights analyzed in the preceding subsections are meaningful. Indeed, it does not assess the predictability of the implemented BMA method, which the reviewer justifiably requests us to evaluate.

However, implementing the reviewer's suggestion exactly as proposed appeared to be not entirely feasible. Specifically, to apply the BMA outside the reference plots, one would need to determine how to transform the estimated weights into ones that can be applied outside the plots. What we believe the reviewer is suggesting, and what most people applying the BMA would do, is to calculate the mean, by either excluding a specific domain (geographical exclusion) or using all SERs except the test one. The challenge here is that our findings indicate substantial variability in model weights across the studied area, as shown in Fig. 3 of the revised manuscript. Even when averaged by sylvo-ecological regions (SERs), the weights differ significantly from one SER to another. Since SERs are used for illustration rather than as stratification criteria, the weights remain heterogeneous within these regions. Despite some ecological homogeneity within SERs, forest height and other factors impacting remote sensing data quality and methodologies can vary considerably. As a result, averaging weights from areas outside the validation zones (whether from plots within a 100 km radius or from another SER) would likely produce suboptimal results for those zones.

Nevertheless, as the reviewer correctly points out, enhancing the credibility of our analysis must involve focusing more on the stability and representativeness of the local weights, which are indeed essentially allowing for fitting a higher-order mixture model to reference data. We have considered two approaches to address this, both utilizing a k-fold cross-validation framework:

- *We analyzed the stability of the estimated local weights in a k-fold cross-validation framework, where $k-1$ folds are used for training and 1 fold for test. Each local weight (corresponding to an NFI plot) is thus estimated $k-1$ times, meaning the higher-order mixture model is fitted $k-1$ times, each with a slightly modified reference sample ($\frac{1}{k-1} * 100\%$ of the total sample). The stability of local weights across these $k-1$ estimations is illustrated through the coefficient of variation, averaged by SER, in Fig. R1 for different height types and for two values of k: 5 and 10 (the $k=5$ portion is shown in Fig. 6 of the revised manuscript). The results show relatively low coefficients of variation, with an interesting trend of increasing variation in mountainous regions, which aligns with other findings we present. These results strengthen the credibility of our conclusions based on local weight analysis and demonstrate the robustness of the BMA mixing approach, which consistently converges to very similar mixtures despite changes in the input data sample.*

- *The situation becomes more complex when dealing with test plots, specifically in assessing the interpolation or extrapolation capabilities of the BMA. As partially discussed earlier, we cannot rely on the spatial continuity*

of our weights, as many factors influencing the model mixture — such as topography, management practices, ownership, and tree species, vary significantly across space. This suggests that a random selection of points, as advocated by Wadoux et al. (2021); Meyer and Pebesma (2022), is more appropriate for verifying interpolation and extrapolation that the geographical exclusion principle (Ploton et al., 2020). In this context, the test fold should consist of a random set of NFI plots distributed across the study area. However, this raises the question: which weights should be applied to these test plots?

- *Should we calculate the average of all known local weights (from the training set) and apply them? Given the observed variability, this approach does not seem ideal, as we demonstrate that model behavior evaluated at the national scale is not representative across all regions.*
- *Should we calculate the average by SER, which reflects some level of homogeneity in forest stands and was chosen as a criterion for illustration (i.e., for averaging local weights)? We explored this option, but as mentioned earlier, the significant variability in factors influencing model performance within these regions makes them unsuitable for true stratification.*

The key question, when addressing the interpolation or extrapolation capability of the BMA as applied in this article, is the choice of a stratification criterion. Specifically, how do we stratify the derived weights to make them applicable outside the reference data sites? This question is not fully answered in this paper, despite some formulations in the original version that may have implied otherwise. We have corrected this throughout the manuscript, while also suggesting possible avenues for tackling this important issue.

One example is presented in Fig. R2, where we show 5-fold cross-validation results with the BMA applied to test plots, using the average of local weights within clusters of points with similar estimated height distributions among models. This stratification allows the BMA to outperform both the SMA and the best individual model for all reference heights considered. It proves to be more effective than SER or other variables we tested (e.g., dominant tree type, dominant tree species, and even vertical structure). However, it does not demonstrate extraordinary predictive capabilities, as it only slightly increases the R^2 compared to the best-performing model.

Another stratification criterion which appears to be promising is the ownership type. However, ownership type maps, used as a proxy for forest management strategies, are not widely available or lack satisfactory quality and spatial resolution. To the best of our knowledge, there are only a few spatialized products that could potentially be useful for such stratification. One of them is spatially resolved at $100\ m$ for the year 2015 (Lesiv et al., 2022), while another is available at $250\ m$ resolution for the period 2001–2020 (Xu et al., 2024).

Another effort made was the modification of the method, specifically in Eq. 8, where we added the weight value in both the numerator and denominator. This adjustment, as mentioned in the discussion section, causes the method to converge more towards the dominant models, resulting in a greater contrast in the weight distribution among models without improving the goodness-of-fit presented in Fig. 5. It does provide a slight improvement in predictive skills (Fig. R3), yielding scores similar to those shown in Fig. R2, but when using SER as the stratification criterion, which is more practical than the previously described clustering based on height distribution among models. However, the trade-off is a significant decrease in the stability of the estimated weights, as illustrated in Fig. R4, which represents a version of Fig. R1 for the implementation containing w

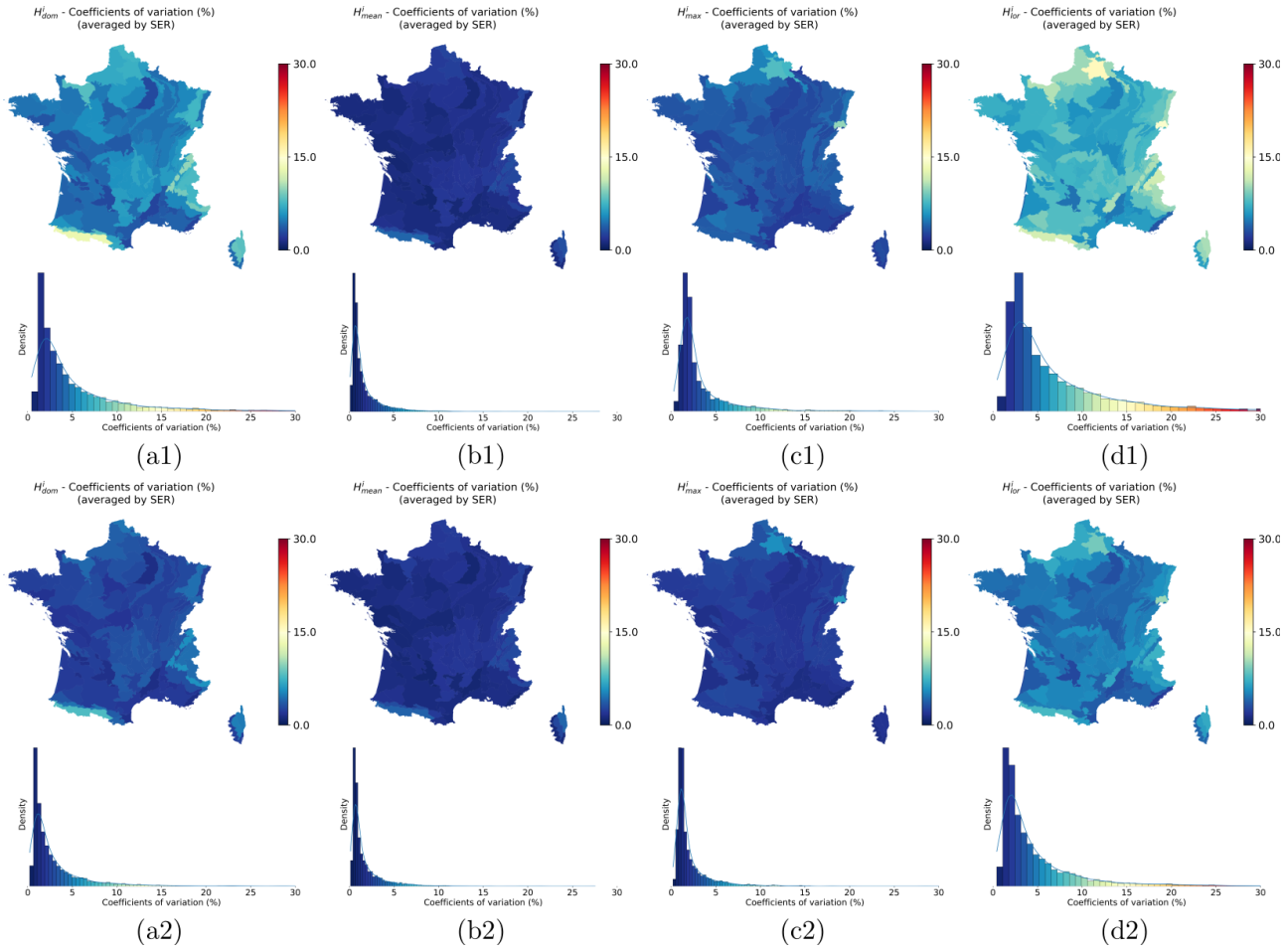

Figure R1: Coefficient of variation of local weights (averaged by SER for the purpose of clearer illustration), as well as their respective densities, derived across K-fold cross validation (above (1) - $K = 5$, below (2) - $K = 10$), with the reference being: (a) the NFI H^i_{dom}, (b) the NFI H^i_{mean}, (c) the NFI H^i_{max}, (d) the NFI H^i_{lor}.

in Eq. 8. Given the importance of the stability of these weights for the analysis presented in the article, we did not find it prudent to pursue this alternative, despite its slightly improved predictive capabilities.

In light of all of this, we have carefully revised any language that might suggest the proposed method enables spatialized multi-scale model fusion, which requires predictive skills that are not demonstrated in this paper. This study focuses on analysis and offers an interesting perspective for the fusion of spatially continuous estimates, which will be explored in future work. Essentially, we faced a choice between analysis and synthesis, as outlined here and in R2.C34. In this article, we chose to prioritize the opportunity for a detailed analysis over predictability. Future research should address the latter, focusing on the critical issue of spatially estimating uncertainty (Lu et al., 2024), especially between reference data sites, by utilizing the BMA framework through one of the previously discussed avenues. An alternative approach to addressing this issue would be to replace the NFI plots with a continuous reference, such as LiDAR HD-derived canopy height maps (CHM). However, before these CHMs can be considered as reliable as NFI field measurements, they must be carefully processed to account for heterogeneities arising from the use of different sensors and data acquisition across various seasons.

$\Longrightarrow$ *See entire manuscript, in particular subsection 5.5 and section 6: Pages 15 - 20.*

R2.C2: In both cases, the authors could assess the quality of predictions both against simpler model

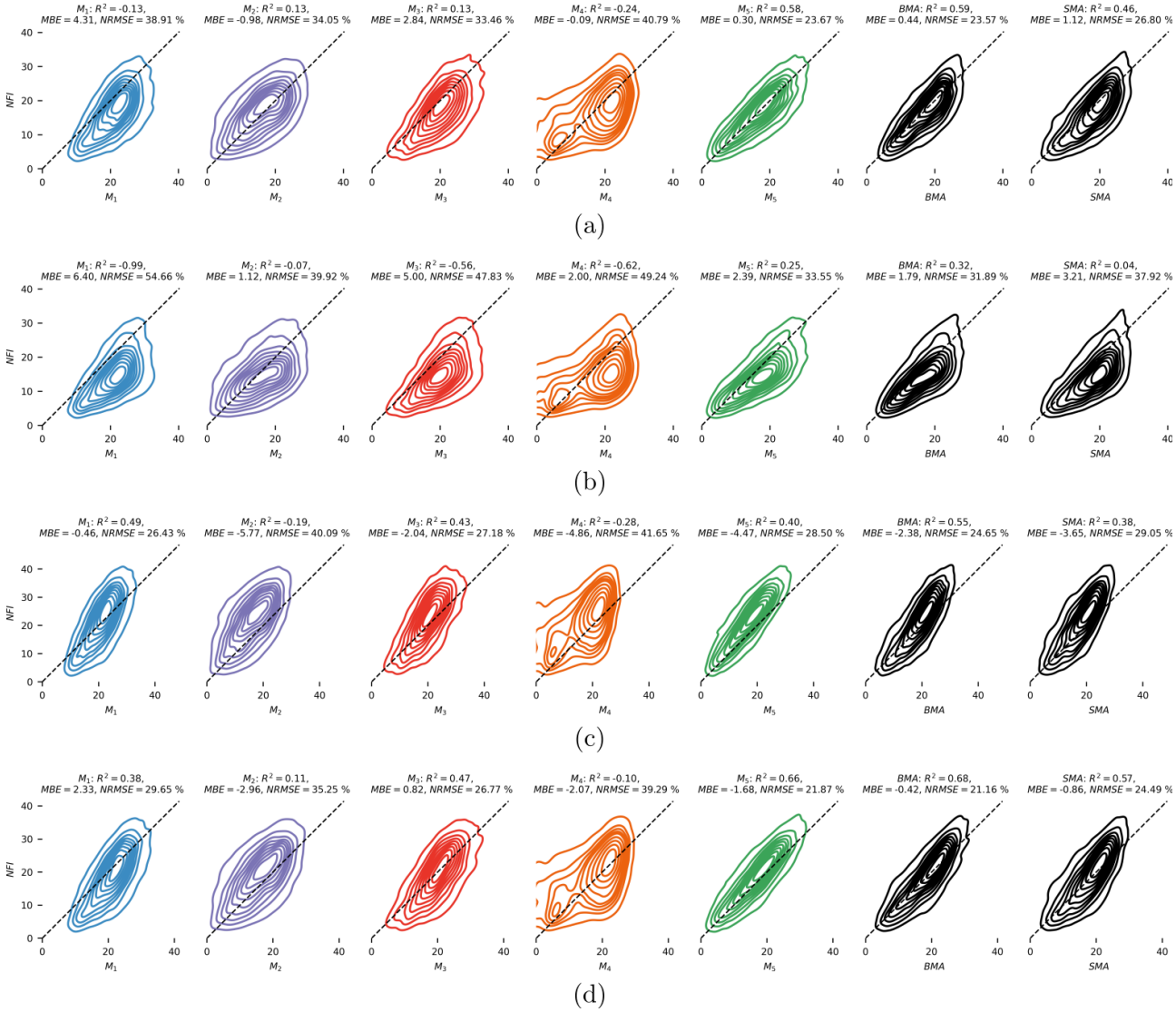

Figure R2: Kernel density estimate (KDE) plots comparing, using independent validation plots (5-fold cross validation), individual models and their BMAs and SMAs with the four employed NFI references: (a) H^i_{dom}, (b) H^i_{mean}, (c) H^i_{max}, (d) H^i_{lor}.

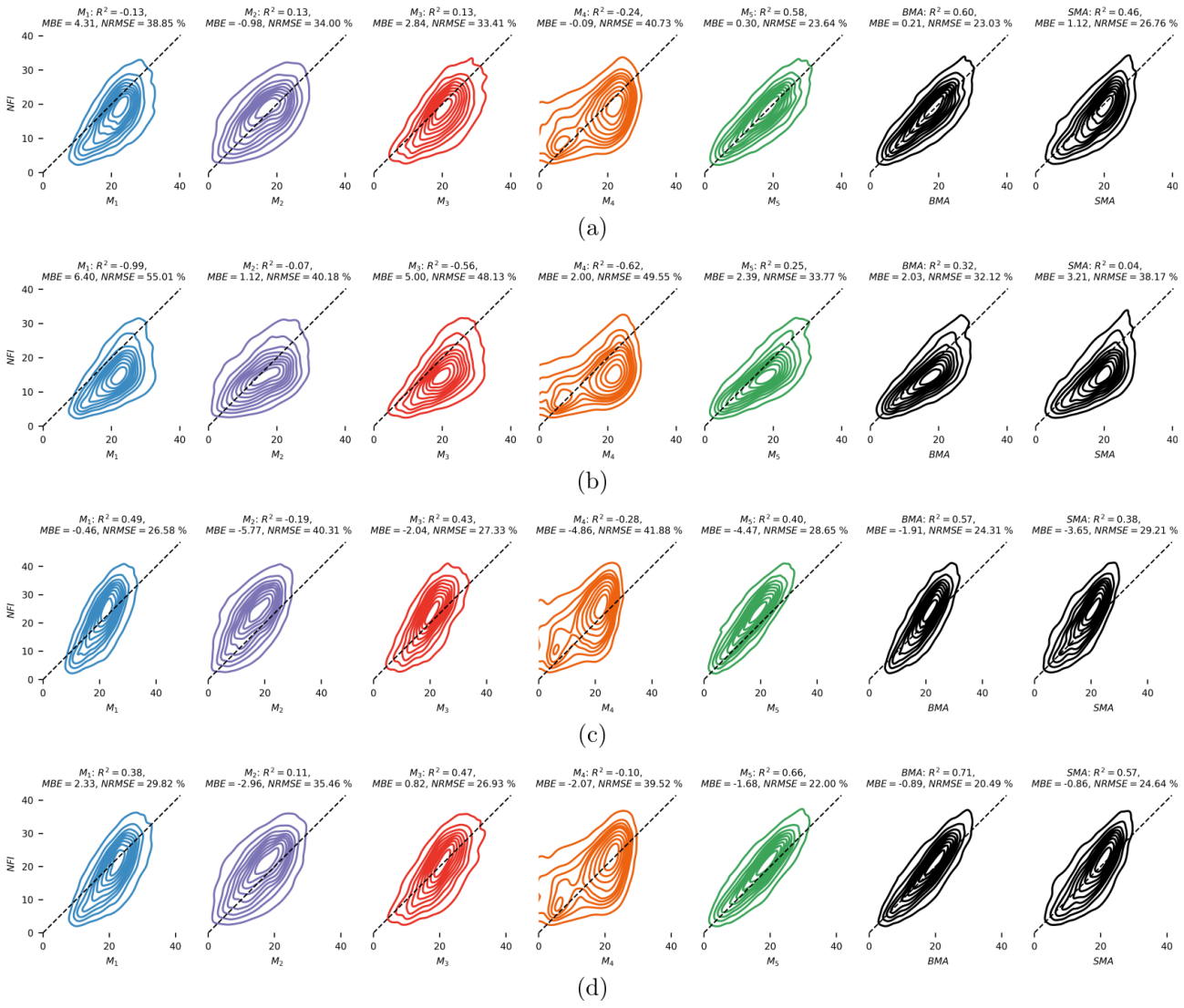

Figure R3: Kernel density estimate (KDE) plots comparing, using independent validation plots (5-fold cross validation), individual models and their BMAs (for the alternative version of the E-M algorithm - "with weights" in Eq. 8 of the manuscript) and SMAs with the four employed NFI references: (a) H^i_{dom}, (b) H^i_{mean}, (c) H^i_{max}, (d) H^i_{lor}.

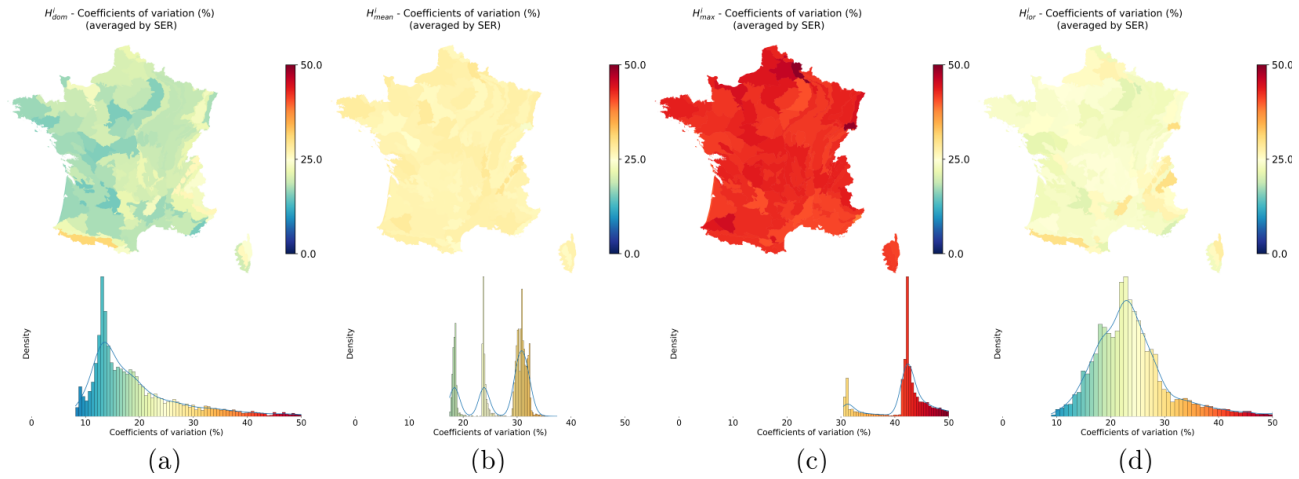

Figure R4: Coefficient of variation of local weights (averaged by SER for the purpose of clearer illustration), as well as their respective densities, derived across 5-fold cross validation (for the alternative version of the E-M algorithm - "with weights" in Eq. 8 of the manuscript), with the reference being: (a) the NFI H^i_{dom}, (b) the NFI H^i_{mean}, (c) the NFI H^i_{max}, (d) the NFI H^i_{lor}.

averaging (SMA), as well as the individual models.

We have included the SMA in all the comparisons discussed in the previous response.

⟹ *See Fig. 5: Page 16.*

R2.C3: I generally would also suggest moving the performance evaluation (5.5) to the front of the Results section, i.e. to make it 5.1. It is super helpful for readers to understand the quality of the BMA as well as the quality of each individual model before delving into the details of the BMA weights. In addition, I would like to see a figure of paired M1-M5 canopy height values for all NFI plots, i.e. scatter plots of height predictions from M1 against M2, M1 against M3, etc., and a table with their respective correlations. It would also be great to see a supplementary figure that shows the overall height maps across France from model M1-M5 as well as the consensus product, and areas around a few sample NFI plots, with predictions from M1-M5 at small spatial scale (locations do not have to be exact, of course, just to get a visual impression).

As suggested by the reviewer (both here and in R2.35), we have included mutual comparisons of individual model height estimates at NFI plots, along with the Pearson's correlation coefficient (see Fig. R5 in this document, and Fig. A1 in the revised manuscript). Additionally, we have incorporated overall height maps across France for models M_1 - M_5, as well as zoomed-in views of three prominent forest regions: the Vosges Mountains, Sologne Forest, and the Landes of Gascogne (see Fig. R6 in this document, and Fig. A2 in the revised manuscript).

As explained in response to R2.C53, we have chosen to keep the performance evaluation in subsection 5.5. The two aforementioned figures are included in Appendix A, which is referenced throughout the manuscript before presenting the estimated BMA weights.

As mentioned in response to R2.C1 and outlined in the revised manuscript, we do not address the interpolation/extrapolation issue in this study, and therefore have chosen not to advance the consensus spatialized product.

⟹ *See Appendix A: Pages 21 - 23.*

Generalizability/Limitations of the study

The authors make quite a few sweeping statements that I believe are not warranted by the study.

R2.C4: The most important limitation of the study is that it assesses the canopy height models only in France. I appreciate that this, by itself, is an enormous effort, and as stated above, I think it is super important, but the authors need to recognize much better the limitations of this approach. Forests in France may be (with emphasis on "may") representative of temperate forests in Europe, and include a few challenging areas in terms of topography (Pyrenees, Alps, Corsica), but their species composition and structural complexity is nowhere near representative of forest canopies globally. A statement such as "Thus, we pinpoint (high) mountainous regions as the primary challenge for ongoing model advancements" is not warranted. It may be the primary challenge in Europe, but even this is not 100% clear to me. I have not conducted a systematic study of canopy height products and urge the authors to correct me where I err, but when I informally compared 2-3 of the global height products to ALS-derived heights (mean and max)

at tropical sites that I know well, the predictions were sometimes catastrophically bad. There was evidence of strong saturation of predicted heights around 30-40 m in some products, while other products were highly impacted by cloud cover and processing artefacts (tiling). If I had to guess the biggest challenges in developing global canopy height models, I would say they are (1) the accurate representation of tall forest canopies (lots of biomass), (2) the accurate representation of tropical forests (also lots of biomass) (3) the accurate representation of mountainous areas, with a combination of the three (tall tropical forests on slopes) likely the gold standard to evaluate what models can or cannot predict. The study at hand only really can show (3).

A very valid point, which has led to revisions throughout the manuscript and is specifically addressed in the discussion section:

"It is important to note that limiting this study to France may lead to an overly optimistic evaluation of the models and their combinations. This is likely due to the availability of higher-quality training data in Europe compared to e.g. most tropical regions (e.g., fewer clouds, superior ALS data, and clearer topographic visibility in less dense forests) and boreal regions where no GEDI data is available. Additionally, the range of forest types, while extensive by European standards, is narrower compared to tropical forests, which are also characterized by denser stands and greater biomass. Consequently, remote sensing-based forest attribute mapping faces significant challenges in tropical forests, which are not adequately addressed in this study focused on temperate European forests."

Furthermore, as reviewer correctly states and as noted in the revised Sec. 6, after analyzing the kernel density estimate plots in Fig. 5 (of the revised manuscript), it can be observed that the predicted heights, particularly maximum and Lorey's heights, tend to saturate around 30 m. The issue of model saturation and the limitation of the BMA in addressing it has been discussed at several points throughout the revised manuscript.

$\Longrightarrow$ *See entire manuscript, particularly section 6: Pages 17 - 20.*

R2.C5: These caveats also extend to assessments of global canopy height models elsewhere. The authors describe the existing model performances as "impressive" (51), and they also describe the 3 m resolution of one of the canopy height models as "impressive" (110), but they do not provide evidence for either claim. The fact that models can be produced at these impressive scales does not mean that they are good models. Surely, there will be important improvements over the next few years, and I am excited about these, but in the current state, all global models come with substantial flaws. In particular, the resolutions given by models often seem to be only "nominal". I.e. most 10 m resolution canopy height models (or lower) are so highly smoothed/averaged towards the mean that I doubt that they can resolve structures below 50-100 m resolution accurately. So 50-100 m is probably a more accurate description of most model's actual resolution. I am sympathetic that this is part of modelling and I do not expect these global efforts to provide excellent local predictions, but apart from the sheer amount of computing and data involved, "impressive" goes too far. Just as an example, in France, the article calculates the RMSEs of models as lying between 4 and 9 m, depending on metric and model (Figure 5), which probably corresponds to relative RMSEs of ca. 30-60%. This is massive uncertainty in probably one of the better-modeled systems.

The best models for each height have relative RMSE values ranging from 22% to 36%, as shown in the revised

Fig. 5 of the manuscript. However, we agree that the original version overemphasized the performance of the models, and we have addressed this in the revised version (please also refer to R2.C17, R2.C18, and R2.C22).

$\Longrightarrow$ *See entire manuscript, in particular sections 1 and 2: Pages 2 - 6.*

Presentation/Writing style

R2.C6: Finally, the presentation/writing style of the article is a bit involved and the syntax structure makes it often difficult to read. Where possible I have tried to give suggestions to simplify sentences or adjust phrases (e.g., not using "the metropolitan France", but "metropolitan France", or maybe just France?), but I would urge the authors to go through the article again and to try to simplify some of the more complex sentence structures. The article may also profit from being proofread by a native speaker.

We made every effort to rephrase the problematic sections in the revised version, following the reviewer's thoughtful suggestions.

Regarding the mentioned example, we find it important to specify metropolitan France, as French territory also includes five overseas regions/departments, one of which is French Guiana (with 8 million hectares of tropical forest).

$\Longrightarrow$ *See entire manuscript.*

Detailed comments

R2.C7 - Title: just "forest canopy height" instead of "the forest canopy height"? And "might they be even more so if combined?" [I think you need a "more so" in English]

The suggested changes have been implemented. The title has been revised to better reflect all the feedback and enhancements suggested during the review.

$\Longrightarrow$ *See Title: Page 1.*

R2.C8 - 2-3: The sentence is a bit long and sometimes repetitive (e.g., "can" and "potential" mean the same thing). Here's a suggestion for a slight simplification: "This has resulted in the availability of multiple, sometimes conflicting, sources of information, which can lead to confusion, but also makes it possible to learn about forest attributes through the joint interpretation of multiple models."

The recommended modifications have been applied.

$\Longrightarrow$ *See Abstract: Page 1.*

R2.C9 - 9: just "forest height" instead of "the forest height"?

We have incorporated the suggested change.

$\Longrightarrow$ *See Abstract: Page 1.*

R2.C10 - 10-11: The "In this contribution" sentence seems a bit difficult to understand. Do you mean: "In this contribution, we evaluate models with respect to their probabilities of correctly predicting measurements at NFI plots."?

Exactly right. We have revised the phrase as recommended.

⟹ *See Abstract: Page 1.*

R2.C11 - 14: What is "dominant prediction"?

We have adjusted the wording, which now specifies: "the different models inadvertently predict different types of canopy height".

⟹ *See Abstract: Page 1.*

R2.C12 - 15: "allowing us to come to understand" could be shortened.

We have revised the phrase to state: "enabling us to grasp".

⟹ *See Abstract: Page 1.*

R2.C13 - 16: "systematically appear to be statistically significant factors" is unclear to me, systematic and statistically significant are relatively clear statements, but then "appear" qualifies this again; reformulate?

We have adjusted the wording, which now specifies: "systematically emerge as statistically significant factors".

⟹ *See Abstract: Page 1.*

R2.C14 - 42: I would remove "while having a limited lifespan"; essentially having a short lifespan explains why GEDI has low sampling density and few recurring shots that overlap with each other.

Acknowledged and done.

⟹ *See section 1: Page 2.*

R2.C15 - 43-46: Please rephrase, it's unclear what "the particular acquisition configurations" refers to.

We have revised the phrase to state: "particular acquisition setups".

⟹ *See section 1: Page 2.*

R2.C16 - 47: what is a "forest stand factor"?

We have adjusted the phrase to "forest structure properties", representing in this particular case: average stem height, size and number density, proportion of canopy and understory cover.

⟹ *See section 1: Page 2.*

R2.C17 - 51: "showing often impressive performances in constructing links"; as outlined above, I fundamentally disagree with this statement. What would be impressive performance? Please provide evidence for this. Personally, I have not been that much impressed by the AI-inferred canopy height models I have seen. They usually do reasonably well in open systems and they seem to also predict some shorter-statured forests in the temperate zone well, particularly in areas where they were calibrated with high-quality ALS data or where topography is not too challenging and GEDI-based inferences are more

reliable. But even in areas where they get into the right ballpark of mean canopy height, they seem to homogenize forest structure a lot, overstate their resolution (a target resolution of 1 m or 10 m is very different from a 1 m or 10 m resolution in ALS assessments, for example, and usually corresponds more to 100 m resolution). Most importantly, I have seen them repeatedly fail to provide acceptable predictions in tall and heterogeneous canopies or under cloud cover, i.e. the tropics. This is an introduction, so I don't expect a full discussion of these issues here and models will undoubtedly improve in the future, but to call the existing models "impressive" is overstating their performance.

Acknowledging reviewer's arguments, we decided to adjust the phrase to "AI methods have played a significant role in achieving this, by constructing links between lidar derived forest attributes, such as canopy height, and broad coverage images."

$\Longrightarrow$ *See section 1: Page 2.*

R2.C18 - 52: "not faultless"; this goes back to my statement above; I think this is a severe understatement. The models are currently quite faulty and error prone and this is why articles such as this one are so important!

Acknowledging reviewer's arguments, we decided to adjust the phrase to "far from being faultless".

$\Longrightarrow$ *See section 1: Page 2.*

R2.C19 - 52-60: I think some of these things could be stated much more succinctly. E.g.: There is a basic question of whether remote sensing data captures all necessary information (answer is: no). Second, some remote sensing data come with huge uncertainties and biases. Third, there are more general modelling issues.

We have revised the phrases to state: "Firstly, electromagnetic interactions in remote sensing data cannot theoretically explain all forest attribute variability. Even if they could, the data would still be prone to imperfections from lidar (Roy et al., 2021; Schleich et al., 2023; Tang et al., 2023; Yu et al., 2024) or imaging sources (Teillet et al., 1982; Joshi et al., 2017; Mutanga et al., 2023), and from modeling choices and parameterizations."

$\Longrightarrow$ *See section 1: Page 2.*

R2.C20 - 74: This needs to be discussed later, but testing the models in France gives overly optimistic assessments of them, because a) models have likely had better training data in Europe than, for example, in the tropics (fewer clouds, better ALS data, topography is more easily visible in less dense forests), b) the range of forest types is heavily restricted.

The revised discussion section addresses this important and excellent point raised by the reviewer, by basically rephrasing his suggestion.

$\Longrightarrow$ *See section 6: Page 20.*

R2.C21 - 75-76: I like the idea of using field data to verify remote sensing predictions and it does allow you to calculate quantities otherwise not available (Lorey's height). But I was wondering: why did you not

also use the IGN's lidar data as a second, independent validation data set? This is not a major issue, but it would seem to make sense to pair the NFI plots with these data, no?

In this article, we chose to use the NFI data because it allows for defining various types of height references (maximum, dominant, mean, and Lorey's). Additionally, the IGN's lidar data are not yet fully available over the entire country.

However, we clearly identify the use of Lidar HD data as a reference as a future direction for this research. As emphasized in both the discussion and perspectives, this approach is particularly promising because it would enable the direct synthesis of spatialized mixtures. In other words, with a continuous reference like Lidar HD-derived CHM, the issue of uncertainty between reference points, which we encounter in the current study, would no longer arise, even when applying the analysis suited version of the E-M method (R2.C1).

$\Longrightarrow$ *See section 7: Page 21.*

R2.C22 - 100: For me, all these resolutions are always a bit tricky to interpret. Most global models look much blurrier than an ALS-derived estimate at the same resolution. E.g., I recently compared the Lang et al. 2023 model to ALS-derived canopy height estimates in Italy, and the effective resolution of the global product seemed much lower than 10 m. Similarly, Meta's 1 m canopy height model only nominally has a resolution of 1 m, but often looks like 100 m resolution. Could you include somewhere a statement that the resolutions that are given are "nominal resolutions"?

While we understand your point, which is perhaps to a degree observable in Fig. A2 in the revised manuscript, it is challenging to label these resolutions as 'nominal.' We believe this difficulty arises primarily for two reasons: first, the remote sensing data used does not capture forest stand variability as effectively as ALS; second, the developed model may not adequately translate the variability in the remote sensing data into canopy height variability.

$\Longrightarrow$ *See Fig. A2: Page 23.*

R2.C23 - 110: again I would very much argue against "impressive". As stated above, I have strong doubts about claimed or "nominal" resolution vs. actual or "effective" resolution.

We have removed the adjective 'impressive' in response to your concerns.

$\Longrightarrow$ *See section 2: Page 5.*

R2.C24 - 117: "covering metropolitan France", no "the" needed.

We have removed the article in response to your suggestion.

$\Longrightarrow$ *See section 2: Page 5.*

R2.C25 - 116: What's "Linear Forest Regression"?

In response to your feedback, we have clarified that 'Linear Forest Regression' refers to an algorithm that combines Random Forest and Multiple Linear Regression. This clarification has been added to the manuscript and to Figure 1, replacing the previously confusing terminology.

$\Longrightarrow$ *See section 2: Page 5.*

R2.C26 - 130-151: As stated above, I really like this approach, and it's great that you evaluate different height types, especially Lorey's height.

Thank you for appreciating the idea. We are also pleased to witness the "comeback" of Lorey's height, which is likely tied to the growing body of evidence supporting the increased resilience of heterogeneous stands to the impacts of climate change, and we are glad to somehow contribute to this discourse.

R2.C27 - 152-221: One overall question: how well does this approach/BMA deal with highly correlated input layers? E.g., if two of the tested models provide similar predictions across France, how do weights get assigned? Will the weights get assigned more or less randomly, and both come out as similar, or could it happen (as with highly correlated predictors) that one model artificially gets a much higher weight than the other model by accident? How could you test for this?

In another project where we applied BMA to an ensemble of forestry models, we observed the second of the two mentioned effects: when redundancy exists between two models, one model receives a very high weight while the other receives a very low one. Similar to redundant predictors in Multiple Linear Regression, a very small change can completely reverse these weights.

An analogy can perhaps be drawn with the some issues found in factorial ANOVA with unbalanced data (Smith and Cribbie, 2022). For example, the order of the models could pose a problem in cases of observed redundancy, though this was not observed in the current study.

R2.C28 - 156: yes, I like this! But you should also mention somewhere that there is also the opposite problem. If a few models are particularly bad/biased/noisy, then this might get propagated in your analysis, no?

Indeed, total uncertainty can be inflated if a poor model is included. However, the aim of this exercise is not to evaluate just any model, but to incorporate credible efforts that use different methods and data to model reality, with the potential to complement each other. Only in this way does the concept of total uncertainty become meaningful.

$\Longrightarrow$ *See section 4: Page 7.*

R2.C29 - 157: You need to explain a bit more what BMA is. I have worked with Bayesian models, but never done model averaging, and I don't understand what deterministic vs. non-deterministic BMA would be (how do you optimize model parameters of already calibrated models?) and how you would even apply that in this context. Would you have to rerun the AI models?

We have made efforts to clarify certain aspects of the BMA, particularly the point raised by the reviewer.

For relatively 'simple' models, such as biomass allometric equations (Picard et al., 2012), the idea is to simultaneously optimize the equation coefficients and the BMA weights. However, AI-based models are not as well-suited to this approach, as the simultaneous optimization of parameters and hyperparameters across different models is difficult to implement within the BMA framework. This is one of the reasons why we opted for a more 'deterministic' approach in the paper, which is now more clearly explained in the manuscript.

$\Longrightarrow$ *See section 4: Page 7.*

R2.C30 - 173: From a practical point of view, I understand the weighting by models. But I don't understand the theoretical underpinnings. Is it correct to call Pr(Mk — H) the posterior probability of model Mk and sum to 1 across the 5 tested models? I can easily imagine a situation where all 5 models that you test are actually highly improbable models, and that there are much better models out there that we have not found yet, so the existing probabilities should not really sum to 1? Of course, the relative probabilities do not change, and that's what's important for your aim, so this is mostly a conceptual/terminological question, but it would be interesting to get your take on this.

An interesting point! Stating that the posterior probability of model M_k represents the likelihood that the model 'generated' the reference observations can indeed become meaningless if we only consider improbable models — that is, if we fail to include 'all possible' models. However, the mathematical framework, as I understand it, implicitly assumes that 'all possible' respectable models (as noted in R2.C28) are taken into account. We have made an effort to subtly but clearly address this valid remark in the revised version.

$\Longrightarrow$ *See section 4: Page 7.*

R2.C31 - 175: How justified is this assumption? Do you have any evidence for this?

This assumption is almost universally made in the literature when applying BMA to continuous variables. Alternatively, one could attempt using a log-normal distribution, though it failed to converge in our case with the E-M algorithm. The results are indeed somewhat conditioned or constrained by the validity of the assumptions we make for the posterior distribution, with the Gaussian being the 'safest' choice, largely due to the Central Limit Theorem.

$\Longrightarrow$ *See section 4: Page 7.*

R2.C32 - 185-186: Yes, but are you fitting an extra model, and introducing new parameters (the weights), so creating a more complex model, no?

Indeed, that's correct. It's important to highlight this at this point - that a better fit comes at the expense of increased complexity.

$\Longrightarrow$ *See section 4: Page 8.*

R2.C33 - 206: Could you explain why z_{ki} is treated as "missing data"?

Our assumption is that z_{ki} are treated as 'missing data' because they represent latent or unobserved variables that are inferred during the estimation process.

R2.C34 - 213-214: The authors need to explain this.

The methodological decision in question is discussed in the revised discussion section:

"As for the methodological decision mentioned in Sec. 4.1 not to include the updated weights in the convergence as suggested by Raftery et al. (2003), we also tested the opposite approach, which is more common in the literature. However, this did not improve the goodness-of-fit shown in Fig. 5, but only increased the contrast between the local

weights of individual models, favoring the overall dominant model. Therefore, we opted to present an analysis that leads to at least equally good higher-order mixtures while highlighting the potential contributions of models that are not overall dominant."

This information might be particularly important for potential readers who may want to apply our code.

$\Longrightarrow$ *See section 6: Page 21.*

R2.C35 - 222: As part of the Results/Discussion, I would definitely like to see one paragraph/figure where you just compare the predictions of the different models. I think all readers would like to know how much the 5 different models actually disagree/agree with each other in predicting canopy height. This could take the form of a simple correlation plot between predicted canopy heights for each pair of models. It would also be ok to put this figure into the supplementary. In addition, I would like to see a whole map of France, with height predicted from M1-M5, and a few sample locations across France (mountain/non-mountain, etc.) in the Supplementary, where the predictions of M1-M5 are shown. This would greatly help the readers understand the relative strengths and weaknesses of the models. Ideally, as pointed out earlier, this would be juxtaposed with CHMs derived from IGN's ALS scans, but this is more of a bonus, and I won't insist on that.

As suggested by the reviewer, we added the two supplementary figures (Fig. R5 and R6 in this document, Fig. A1 and A2 in the manuscript), which now form Appendix A in the revised manuscript, with the original Appendix A now becoming Appendix B.

$\Longrightarrow$ *See Appendix A: Pages 21 - 23.*

R2.C36 - 224: Since I don't fully grasp the weighting approach, are the overall weights somehow tied to the local weights and provide some form of regularization (e.g., as in a hierarchical Bayesian model)?

The overall weights represent the average of the local weights (i.e., 'missing data'), as outlined in Eq. 9 of the revised manuscript.

R2.C37 - 232-257: Overall, I like this paragraph, and it is an interesting comparison!

Thank you for your appreciation of this subsection.

R2.C38 - 235: "relatively significantly" is odd. Please rephrase. What do you mean? Also, when looking at the graphs, the deviations do not seem so large. With the exception of M5, all model weights generally seem to fall between 0.15 and 0.25, so relatively close to 0.2. None of the models is clearly worse than the others, none of the models is clearly better.

When we said 'relatively significantly,' we meant that, although a 5% deviation from the Simple Model Averaging ($\frac{100\%}{5} = 20\%$) may not seem substantial, it can be quite impactful in this specific context. This is evident in the revised Fig. 5 (modified in response to another reviewer's comment — see R2.C3) and in Fig. R2 in this document, where such deviations lead to notable changes in the height estimates between the BMA and SMA.

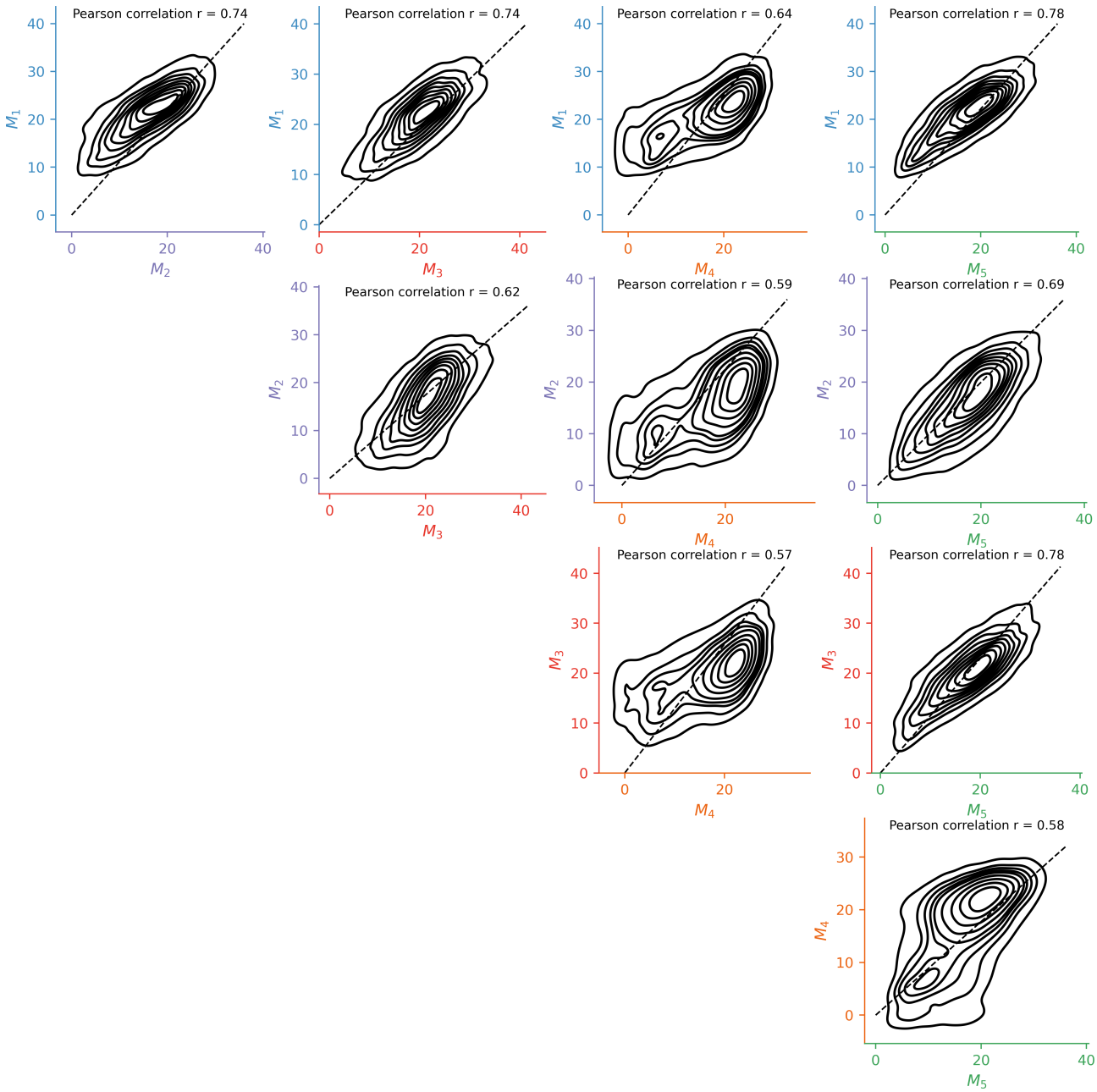

Figure R5: Kernel density estimate (KDE) plots mutually comparing individual models.

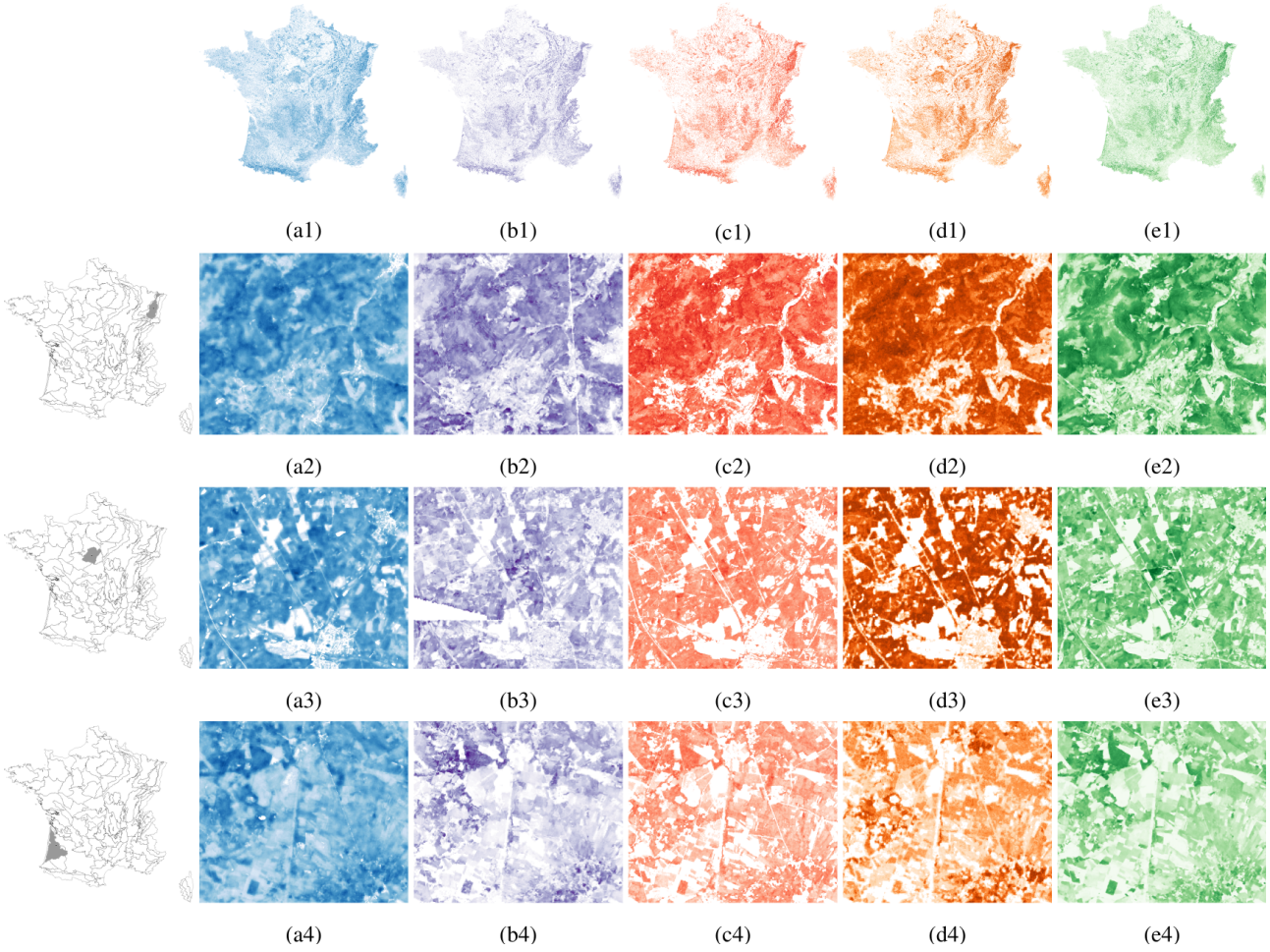

Figure R6: Forest canopy height maps produced by: (a) M_1 (Lang), (b) M_2 (Liu), (c) M_3 (Morin), (d) M_4 (Potapov), and (e) M_5 (Schwartz), illustrated for: (1) metropolitan France, (2) a zone in the Vosges Mountains, (3) a zone in Sologne Forest, and (4) a zone in the Landes de Gascogne. The leftmost column indicates the location of the highlighted zones within metropolitan France.

R2.C39 - 236: Simple Model Averaging (SMA) has not been properly introduced, and needs its own short section. Did you take the median or average?

Indeed, this has been corrected in the revised version - SMA in the context of our article reflects the average value.

⟹ *See subsection 5.1: Page 10.*

R2.C40 - Figure 2: Could you potentially just leave the bar "WITHOUT imputations" blank, i.e., no internal black lines? I think it would be visually easier to understand the difference, at first sight I did not recognize it.

Noted and addressed.

⟹ *See Fig. 2: Page 10.*

R2.C41 - 249-254: I agree with the overall statement, and think it's important to clearly define what canopy height models are actually predicting. But my impression is that this is slightly overstating the results. Yes, there are some differences, and M5 seems to do slightly better at predicting dominant and Lorey's height, but we are still talking about a weight of 0.25-0.27, compared to other models' weights of 0.15-0.2. For me, this is much less clear than what you describe. I further have doubts about biomass estimates relying on several height descriptors at the same time. Evidently, these different height estimates will covary between them, and using highly covarying/correlated predictors for biomass models is probably not going to improve model predictions massively.

Interpreting such results is rarely straightforward. However, we maintain that these differences are significant and demonstrate that some models are better suited for predicting specific height types than others.

Concerning the reviewer's doubts of biomass predictions using several height descriptors, Picard et al. (2012) have shown that BMA can improve the estimation of biomass from different allometric equations. We hypothesize that using BMA on models linking different height descriptors (albeit covarying) and biomass could similarly lead to improved biomass predictions.

R2.C42 - 255-258: I don't understand – there is an impact of imputation, but it's unclear to me whether this is an improvement or not. It may just be that the imputation routine shares some similarities in predicting to new data with the AI models, so is this a real effect? This may become clearer under spatial cross validation though!

That's a valid point. However, we argue that the imputation represents an improvement, as the mean or dominant height cannot be accurately estimated without these imputations. There are simply too few measured tree individuals per sampled plots to obtain a reliable mean or dominant height value.

R2.C43 - 260: "metropolitan France", not "the metropolitan France".

Noted and corrected.

⟹ *See subsection 5.2: Page 10.*

R2.C44 - 260-262: I don't understand what is being said here? Why is the homogeneity of the 86 regions important for the averaging?

In response to R2.C1, we have reformulated this section. Specifically, the 86 SER regions are used solely for illustrative purposes, as it is quite unwieldy to present results for over 5000 individual NFI plots.

$\Longrightarrow$ *See subsection 5.2: Page 10.*

R2.C45 - Figure 3: I know that this figure is already quite complex, but is there a way of visualizing or describing the average NFI-based metric per local region somewhere, i.e. add a 7th column where the average dominant height, average mean height, average maximum height and average Lorey's height is mapped? Because from a biomass perspective, we clearly care most about model performance in areas with tall/complex canopies, and I suspect that this could shift the evaluation of model weights a bit.

That's a good point! As suggested, we revised Fig. 3 by adding the 7th column containing the average height in question by SER.

We have also integrated the following comment: "Given that height can serve as a proxy for volume/biomass, and recognizing that forests denser in terms of biomass can be more challenging to monitor via remote sensing, we also present in Fig. 3 the average height values by SER. However, this analysis did not reveal any significant impact of average height values by SER on the weight distribution between models, suggesting either that density is not critical enough in temperate forests, or that none of the models stands out in addressing it.".

$\Longrightarrow$ *See subsection 5.2, particularly Fig. 3: Pages 10 - 12.*

R2.C46 - 268-270: Yes, that's a great point! You could also mention Corsica here, where the Potapov model also seems to do better for these variables. I don't understand the Landsat argument. You mean that coarser resolution approaches average out local topographic errors? Also, maybe rephrase "disturbing effects".

Yes, our hypothesis is that the coarser resolution could paradoxically be beneficial by smoothing local topographic errors. This is consistent with the findings of Riano et al. (2003), who worked with Landsat Thematic Mapper for vegetation classification and suggested that smoothing the slope in rough terrain improves the performance of topographic corrections.

The suggested corrections have been implemented.

$\Longrightarrow$ *See subsection 5.2: Page 12.*

R2.C47 - 288-290: I don't fully understand the meaning of "within variance". Could you make clearer what it means maybe already in the methods section, but also here. Does it mean how much model prediction quality/weights vary across sites/regions within the same model?

Within variance represents a weighted average of the estimated model uncertainties (standard deviation squared, σ from Eq. 10). The weighted average is calculated locally (at positions i) using the missing values (z_i). Essentially, we have one estimate of σ per model for the entire study area, and the weighted average varies from one point to

another according to the changing local weights. Less "suited" models (lower weights) have higher uncertainty which makes that the two somewhat compensate leading to the observed spatial consistency.

We have hopefully clarified this explanation in both the methods section and the results and analysis section.

$\Longrightarrow$ *See section 4: Page 8.*

R2.C48 - 291-296: This is important, because a lot of well-preserved forest area worldwide is usually located in mountainous terrain due to accessibility issues. It's an issue if models are performing worst in these areas.

Indeed, we have emphasized this point pretty explicitly in the revised version.

$\Longrightarrow$ *See section 6: Page 19.*

R2.C49 - 297-298: I don't understand this point, or, if I understand it, I don't agree. If within-model variance is larger than between-model variance, it probably just means that the models are all pretty similar in their predictions, no? That does not mean that the predictions are reliable per se. If we evaluate 5 models that are all using more or less the same input data and comparable extrapolation approaches, they could all be similarly bad at predicting something and would then have a very low between-model variance. For example, 4 out of 5 models here use GEDI shots as input, so I would already expect a lot of homogeneity in predictions just from that.

We do not claim that the predictions of individual models are more reliable when the between-model variance is lower than the within-model variance. Rather, we assert that the resulting mixture is valid when the divergence falls below the level of the combined uncertainty, i.e., that it makes sense to fit a higher-order mixture model of the considered models.

The contrary would have been, as stated in the revised manuscript: "An exceeding between variance would have indicated that the models are structurally too different from each other, making their combination ineffective. In such a case, the assumption stated at the beginning of Sec. 4, that the considered models have the potential to complement each other, would have been disproved."

Certainly, an excessively high within-model variance would also be concerning, as it would suggest that the assumption made at the beginning of Sec. 4 regarding "respectable performance" does not hold. Acknowledging the skepticism surrounding the models' quality, we believe that the evaluations, illustrated in Figures 5 and A2 of the revised manuscript, along with the articles presenting these models, validate their inclusion in the current study. This implies that the resulting mixture is not merely a valid combination of any models, but rather a valid mixture of models that possess the potential to effectively describe forest canopy, albeit with some limitations, as discussed in the Introduction and throughout the entire manuscript.

$\Longrightarrow$ *See subsection 5.3: Page 13.*

R2.C50 - 316-319: Yes and no! I fully agree with your assessment that mountainous regions are an important, and one of the primary challenges for model advancements. In Europe they may well be the "primary challenge", but this needs to be qualified. This study evaluates models only in France and thus

does not represent the variety of global forest types. I strongly suspect that the "primary challenge for ongoing model advancements" are actually tall canopies, in particular in the tropics, where cloud cover and forest structural complexity make predictions much more volatile. I would then see predictions in mountainous terrain as a strong "secondary challenge" globally. The absolute gold standard for model evaluation would be tall tropical forests with strong topographic gradients.

Yes, this point has been addressed in our response to R2.C4.

$\Longrightarrow$ *See section 6: Page 20.*

R2.C51 - 337-340: I don't think it's accurate to state that "classes" cause variations in the models. They are linked to these variations, or are predictors of them, but what causes the variations is more tricky to say. Most likely specific tree species occur in specific environments (e.g., high altitude) which are trickier to predict.

Indeed, as mentioned in this section, it is challenging to disentangle the effects of tree species from topographic influences. We have further clarified this point in the revised version.

$\Longrightarrow$ *See section 6: Page 19.*

R2.C52 - 341-344: I was surprised by this. So this would indicate that low forest canopies are worse-predicted.

This suggests that model discrepancies are greater for regular low forests, other low stands, or irregular structures, compared to regular high stands. As stated in the revised manuscript:

"This is not surprising, as low structures tend to be highly heterogeneous. As a result, they are poorly captured by models that target a specific height proxy. This effect could potentially be reinforced by the fact that all models exhibit some degree of saturation (Fig. 5). Therefore, in contrast to lower heights, the models tend to 'converge' in the case of higher stands, which cannot exactly be corrected by fitting a higher-order mixture model."

$\Longrightarrow$ *See section 6: Page 19.*

R2.C53 - 358: In my opinion, this paragraph should be the first of the results paragraphs, as it evaluates whether the mixture model is actually any good at modelling.

As discussed in R2.C1 and in the revised manuscript, the primary focus of this paper is the analysis. Therefore, we chose to prioritize presenting and analyzing the weights. After careful consideration of this suggestion, we decided to retain the discussion of the fitted mixtures towards the end, as it naturally leads into the perspectives of this work, which is also emphasized in the revised discussion section.

R2.C54 - 359: Broadly, the authors are calibrating a higher-order model that assigns weights to different underlying models. This approach is prone to the same modelling issues as any model, e.g., overfitting the data, and needs to be a) better described – I currently have no information as to how R2, MBE and RMSE have been calculated, and b) it needs to be done with a formal cross validation strategy. This should take the form of a spatial cross validation, as suggested in Ploton et al. 2020. One way could be that the

authors select 1000 NFI plots, remove all other NFI plots within a 100 km radius, and then predict the left-out plot's height structure from the mixture model calibrated on the remaining NFI plots. Another way could be that the authors divide France into spatial folds, e.g. the 96 sylvo-ecological regions, and predict each region from the other 95. Ideally, here I would also leave out all sylvo-ecological regions that are directly adjacent to the validation plot to account for spatial autocorrelation. The resulting MBE/RMSE could then be compared against the original model performance, as well as SMA, for example.

Kindly refer to R2.C1.

$\Longrightarrow$ *See entire manuscript, in particular subsection 5.5 and section 6: Pages 15 - 20.*

R2.C55 - 367: Is this actually possible? In my (very coarse) understanding, there is often a bias-variance tradeoff in modelling, so that, beyond a certain point, reducing one often comes at the expense of increasing the other.

According to two cited papers (Bao et al., 2010; Erickson et al., 2012), actions can be taken to reduce both bias and variance. Often, individual models are bias-corrected before averaging (using methods such as linear regression, additive bias correction, or the cumulative distribution function (CDF) approach). However, one could also envision correcting a mixture of non-bias-corrected models.

As stated in the revised manuscript, there is perhaps another avenue to explore: "One alternative approach that might yield less biased mixtures, and is worth exploring, would be to adapt the well established employed E-M algorithm by incorporating Restricted Maximum Likelihood Estimation (REML) instead of Maximum Likelihood Estimation (MLE) (Pinheiro and Bates, 2000)."

$\Longrightarrow$ *See section 6: Page 20.*

R2.C56 - 375: Without proper spatial cross validation I would not believe this "super" model yet.

As outlined in R2.C1, we exercised much greater caution in using such formulations in the revised version. In fact, we no longer employ them in this revised version.

$\Longrightarrow$ *See entire manuscript, in particular subsection 5.5 and section 6: Pages 15 - 20.*

R2.C57 - 380: How accurate is it to actually call this Bayesian model averaging, if it's only "Bayesian flavoured"?

We used the term 'Bayesian-flavored' to more accurately convey the nature of the method, which is widely recognized as being an BMA approach. This term is frequently used in the model averaging literature (e.g. Dormann et al., 2018).

R2.C58 - 386: cf. my objections above. Mountainous regions are a "real challenge", and probably the most important one in France (and likely most of Europe), but that does not mean it's true worldwide, and this needs to be acknowledged here!

Acknowledged and done.

$\Longrightarrow$ *See section 7: Page 20.*

R2.C59 - 391-393: "more so if combined"; epistemologically, I am not sure, whether I fully agree with this statement. Yes, combining models can make them more useful, but you could also argue that the authors are essentially fitting a much more complex model that picks the best prediction at every location, so we are gaining predictive power in return for a lot of new parameters. But to trust the results, we need proper spatial cross validation.

In response to R2.C1 and several other comments, this has been adequately revised in the updated version.

$\Longrightarrow$ *See entire manuscript.*

R2.C60 - Figure 5: I would like to see relative RMSE here as well.

Normalized RMSE (with respect to the mean value) has been added to each panel in Fig. 5.

$\Longrightarrow$ *See Fig. 5 - Page 17.*

References

L. Bao, T. Gneiting, E. P. Grimit, P. Guttorp, and A. E. Raftery. Bias correction and bayesian model averaging for ensemble forecasts of surface wind direction. Monthly Weather Review, *138(5):1811 – 1821, 2010. doi: 10.1175/2009MWR3138.1.*

C. F. Dormann, J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, S. Ciuti, J. Elith, K. Gerstner, J. Guelat, P. Keil, J. J. Lahoz-Monfort, L. J. Pollock, B. Reineking, D. R. Roberts, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, S. N. Wood, R. O. Wüest, and F. Hartig. Model averaging in ecology: a review of bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs, *88(4):485–504, 2018. doi: 10.1002/ecm.1309.*

M. J. Erickson, B. A. Colle, and J. J. Charney. Impact of bias-correction type and conditional training on bayesian model averaging over the northeast united states. Weather and Forecasting, *27(6):1449 – 1469, 2012. doi: 10.1175/WAF-D-11-00149.1.*

N. Joshi, E. T. A. Mitchard, M. Brolly, J. Schumacher, A. Fernández-Landa, V. K. Johannsen, M. Marchamalo, and R. Fensholt. Understanding 'saturation' of radar signals over forests. Scientific Reports, *7(1):3505, Jun 2017. doi: 10.1038/s41598-017-03469-3.*

M. Lesiv, D. Schepaschenko, M. Buchhorn, L. See, M. Dürauer, I. Georgieva, M. Jung, F. Hofhansl, K. Schulze, A. Bilous, V. Blyshchyk, L. Mukhortova, C. L. M. Brenes, L. Krivobokov, S. Ntie, K. Tsogt, S. A. Pietsch, E. Tikhonova, M. Kim, F. Di Fulvio, Y.-F. Su, R. Zadorozhniuk, F. S. Sirbu, K. Panging, S. Bilous, S. B. Kovalevskii, F. Kraxner, A. H. Rabia, R. Vasylyshyn, R. Ahmed, P. Diachuk, S. S. Kovalevskyi, K. Bungnamei, K. Bordoloi, A. Churilov, O. Vasylyshyn, D. Sahariah, A. P. Tertyshnyi, A. Saikia, Z. Malek, K. Singha, R. Feshchenko, R. Prestele, I. H. Akhtar, K. Sharma, G. Domashovets, S. A. Spawn-Lee, O. Blyshchyk, O. Slyva, M. Ilkiv, O. Melnyk, V. Sliusarchuk, A. Karpuk, A. Terentiev, V. Bilous, K. Blyshchyk, M. Bilous, N. Bogovyk, I. Blyshchyk,

S. Bartalev, M. Yatskov, B. Smets, P. Visconti, I. Mccallum, M. Obersteiner, and S. Fritz. Global forest management data for 2015 at a 100 m resolution. Scientific Data, 9(1):199, May 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01332-3.

K. Lu, S. Bates, and S. Wang. Quantifying uncertainty in area and regression coefficient estimation from remote sensing maps, 2024. URL https://arxiv.org/abs/2407.13659.

H. Meyer and E. Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Communications, 13(1):2208, Apr 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29838-9.

O. Mutanga, A. Masenyama, and M. Sibanda. Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects. ISPRS Journal of Photogrammetry and Remote Sensing, 198:297–309, 2023. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2023.03.010.

N. Picard, M. Henry, F. Mortier, C. Trotta, and L. Saint-André. Using Bayesian Model Averaging to Predict Tree Aboveground Biomass in Tropical Moist Forests. Forest Science, 58(1):15–23, 02 2012. doi: 10.5849/forsci.10-083.

J. C. Pinheiro and D. M. Bates. Mixed-Effects Models in S and S-Plus. Springer, New York, NY, 2000. ISBN 978-0-387-22747-4. doi: 10.1007/b98882.

P. Ploton, F. Mortier, M. Réjou-Méchain, N. Barbier, N. Picard, V. Rossi, C. Dormann, G. Cornu, G. Viennois, N. Bayol, A. Lyapustin, S. Gourlet-Fleury, and R. Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11(1):4540, Sep 2020. doi: 10.1038/s41467-020-18321-y.

A. E. Raftery, F. Balabdaoui, T. Gneiting, and M. Polakowski. Using bayesian model averaging to calibrate forecast ensembles. Technical report, University of Washington, 2003. Technical Report no. 440.

D. Riano, E. Chuvieco, J. Salas, and I. Aguado. Assessment of different topographic corrections in landsat-tm data for mapping vegetation types. IEEE Transactions on Geoscience and Remote Sensing, 41(5):1056–1061, 2003. doi: 10.1109/TGRS.2003.811693.

D. P. Roy, H. B. Kashongwe, and J. Armston. The impact of geolocation uncertainty on gedi tropical forest canopy height estimation and change monitoring. Science of Remote Sensing, 4:100024, 2021. ISSN 2666-0172. doi: 10.1016/j.srs.2021.100024.

A. Schleich, S. Durrieu, M. Soma, and C. Vega. Improving gedi footprint geolocation using a high-resolution digital elevation model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16:7718–7732, 2023. doi: 10.1109/JSTARS.2023.3298991.

C. E. Smith and R. Cribbie. Factorial anova with unbalanced data: A fresh look at the types of sums of squares. Journal of Data Science, 12(3):385–404, 2022. ISSN 1680-743X. doi: 10.6339/JDS.201407_12(3).0001.

H. Tang, J. Stoker, S. Luthcke, J. Armston, K. Lee, B. Blair, and M. Hofton. Evaluating and mitigating the impact of systematic geolocation error on canopy height measurement performance of gedi. Remote Sensing of Environment, *291:113571, 2023. ISSN 0034-4257. doi: $10.1016/j.rse.2023.113571$.*

P. Teillet, B. Guindon, and D. Goodenough. On the slope-aspect correction of multispectral scanner data. Canadian Journal of Remote Sensing, *8(2):84–106, 1982. doi: $10.1080/07038992.1982.10855028$.*

A. M.-C. Wadoux, G. B. Heuvelink, S. de Bruin, and D. J. Brus. Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling, *457:109692, 2021. ISSN 0304-3800. doi: $10.1016/j.ecolmodel.2021.109692$.*

H. Xu, B. He, L. Guo, X. Yan, J. Dong, W. Yuan, X. Hao, A. Lv, X. He, and T. Li. Changes in the fine composition of global forests from 2001 to 2020. Journal of Remote Sensing, *4:0119, 2024. doi: $10.34133/remotesensing.0119$.*

Q. Yu, M. G. Ryan, W. Ji, L. Prihodko, J. Y. Anchang, N. Kahiu, A. Nazir, J. Dai, and N. P. Hanan. Assessing canopy height measurements from icesat-2 and gedi orbiting lidar across six different biomes with g-liht lidar. Environmental Research: Ecology, *3(2):025001, apr 2024. doi: $10.1088/2752-664X/ad39f2$.*