Dear Hisashi Sato,

Thank you for your positive feedback, your constructive suggestions and for spotting the mismatch between Eq. 1 and the results/interpretation of residuals. After correcting Eq. 1, our interpretation is consistent. We will also provide more detail on the spatial block CV as suggested. Please find our point-by-point responses below.

Yours sincerely,

Simon Scheiter and Liam Langan

This manuscript presents a novel and carefully designed framework for diagnosing DGVM deficiencies by modeling data-model residuals with generalized additive models (GAMs). Rather than simply quantifying model performance or mapping model-data differences, the study attempts to identify the environmental drivers underlying these residuals and to translate them into priorities for future model development. I consider this shift from descriptive model evaluation to process-oriented diagnosis to be an important methodological contribution. The statistical framework is generally rigorous, particularly the use of spatial block cross-validation and variance partitioning, and the manuscript is overall well written and logically organized. I therefore enjoyed reading this study. My comments below are intended primarily to improve the clarity of the methodology and to moderate several interpretations that, in my opinion, go somewhat beyond what the analyses directly support.

Thank you for the positive feedback.

___ Major Issue ___

(1) I appreciate that the authors evaluated the models using spatial block cross-validation rather than conventional random cross-validation. This is an important strength of the manuscript, as it provides a more realistic assessment of model generalization in the presence of spatial autocorrelation. However, the implementation of the spatial block cross-validation is currently insufficiently described, making it difficult to evaluate or reproduce the analysis.
In particular:

[a] The manuscript states that blocks of 100 km2 were used. This appears inconsistent with the 0.5° resolution of the analysis, and I wonder whether the intended block size was 100 km rather than 100 km2. Please clarify.

Indeed, this is a mistake in the methods section. We did not use 100 km2 or a 100x100 km grid, but we rather used a grid with 50 rows and 50 columns, representing hexagons with around 1.4° size. This size covers multiple 0.5° grids of our aDGVM simulations. We used the 100km approach originally but then changed to the 50 rows approach without adjusting the Methods. We will correct and clarify this in the manuscript.

[b] More generally, additional details of the blockCV implementation would be helpful. For example, how were the blocks constructed (regular grid or otherwise)? How was the block size selected? Was it chosen based on the spatial autocorrelation range of the residuals or by another criterion? How were blocks assigned to the five folds?

We divided the study area across Africa into a grid of 50 vertical spatial tiers using hexagonal blocking combined with a 5-fold random assignment. Statistically, hexagonal blocks are highly robust for spatial partitioning as they mitigate edge effects and provide uniform distances between block centers compared to standard square grids. Given the geographic extent of Africa, this setup yields an average block dimension of approximately 1.4°, which is roughly three times larger than our 0.5° simulation grid. This spatial buffer ensures that training and testing datasets are sufficiently separated, effectively breaking the spatial autocorrelation structure of the underlying raster data and providing a rigorous, unbiased assessment of model performance. We will revise the manuscript to be more specific about these points.

Because spatial block cross-validation is a key methodological component of this study, a clearer description would substantially improve the manuscript's transparency and reproducibility.

Thank you for this comment, as described above, we will clarify the methods section describing the spatial block cross-validation.

(2) Sign consistency between residual definition and partial-effect interpretation (Section 3.1, Figs. 4-5)

Residuals are defined as r = aDGVM - RS (Eq. 1), so a positive residual means overestimation and a negative residual means underestimation. Checking Fig. 4 against this definition, several interpretations in Section 3.1 appear to have the sign reversed, and this is not limited to a single instance:

This is indeed another mistake, thank you for spotting this. In the main text, we defined the residuals as r = aDGVM – RS, while in the R scripts for our analyses, we did it the other way round, i.e., r = RS – aDGVM. This implies issues with the interpretation of residuals but also in Figs. 1 and 2. In the figures and analyses, positive residuals represent underestimation in aDGVM, negative residuals represent overestimation in aDGVM.

We will correct Eq. 1 in the manuscript. To facilitate interpretation, we will state in the text and figure captions that positive residuals represent underestimation in aDGVM, negative residuals represent overestimation in aDGVM.

bio12: partial effect is positive at high values (implying overestimation), but the text states this indicates "underestimation of biomass in wetter regions." Notably, bio18 is discussed in the same sentence with the opposite sign (negative at high values), yet both are used to support the same conclusion.

With the corrected Eq. 1, our statement is true and the positive residuals imply underestimation in wetter regions.

Burned area: partial effect decreases (becomes negative) with increasing burned area, implying underestimation, but the text states this indicates "overestimation of woody biomass and cover."

With the corrected Eq. 1, the negative residuals imply overestimation with increasing burned area as stated in the text.

HFI: partial effect is negative at high HFI, implying underestimation, but the text concludes the model "overestimates vegetation in heavily human-modified landscapes."

Positive residuals imply overestimation after correcting Eq. 1.

Population density and N deposition: partial effects are positive at high values (implying overestimation), but the text states aDGVM "underestimates biomass and tree cover" in these regions.

Similarly, with the corrected equation our statements are correct.

I may be missing an intended sign convention. Still, since this pattern recurs across nearly all disturbance and anthropogenic predictors in this paragraph, I would ask the authors to systematically re-check every sign attribution in Section 3.1 against Eq. (1), rather than correcting individual sentences.

Correcting Eq. 1 implies that our statements in the discussion are correct. However, as suggested by the referee, we will carefully check the respective text in the Results and Discussion to make sure that all statements are correct.

This matters beyond wording: these sign interpretations directly support the development priorities in Section 4.3 (e.g., claims about the fire module or under-representation of grazing/land degradation). If signs need to be reversed, the corresponding conclusions in the Discussion (and possibly the Abstract) may need revision as well. A supplementary table listing, for each predictor, the sign of the effect at high/low values, alongside the stated over- or underestimation conclusion, would help verify consistency.

We agree that these mismatches also influence priorities in Section 4.3. As correcting Eq. 1 seems to fix issues regarding the interpretation of over- vs underestimation, our conclusions in Section 4.3 should be correct. Nonetheless, we will check the text carefully. We will also add a table with the signs of the effect at high or low values. We agree that this would be helpful and facilitate the interpretation of the partial effect plots.

___ Minor Technical Issues___

(1) The manuscript contains an unresolved cross-reference ("Table ??") in Section 2.4. Please correct the table reference.

Thank you for spotting this, these results refer to Table S1. We will check the ms to make sure that all references are correct.

(2) Please clarify the interpretation of the effective degrees of freedom (edf) of the spatial smooth in Section 3.1. A large edf reflects the complexity of the fitted spatial surface, but does not by itself indicate that a large fraction of the residual variance is explained by spatial structure. This conclusion seems to be more directly supported by the variance partitioning analysis presented later.

We thank the reviewer for this clarification. We agree that the edf reflect complexity but not variance explained by the GAMs. We will revise the text in 3.1 to clarify this difference: “High edf values indicate complex and fine-scale spatial surfaces, and the associated F-statistics confirm that spatial structure remains statistically significant after accounting for climate, fire, and anthropogenic factors.”

(3) The first paragraph of Section 3.1 could be clarified. The statement that "spatial patterns of residuals predicted with GAMs agreed well with residuals derived from aDGVM and remote sensing" initially gave the impression that the manuscript was referring to the residuals remaining after fitting the GAM. However, this sentence actually compares the original data-model residuals with the residuals predicted by the GAM. Since the following sentences then switch to discussing the residuals of the fitted GAM, I suggest making this distinction more explicit to improve readability.

The first paragraph of Section 3.1, l. 189 is indeed about residuals from data and aDGVM (Figs 1, 2), and the statement “spatial patterns of residuals predicted with GAMs agreed well with residuals derived from aDGVM and remote sensing”, l. 215 compares residuals from data vs aDGVM to residuals predicted with the GAM. We will revise the manuscript to clarify those differences and make clear where we refer to data-aDGVM residuals or GAM residuals.

(4) In Section 3.2, stating that variance partitioning "confirms" the contribution of unmeasured spatially structured processes seems too strong. This analysis quantifies the spatial smooth's unique contribution (11.0% AGB, 9.6% TC) but cannot identify its cause. Notably, this is in tension with Section 4.4, where the authors acknowledge that spatially structured uncertainty in the remote-sensing products "may generate patterns in the GAM residual surface that reflect measurement error rather than aDGVM process deficiencies." The same spatial signal therefore appears to be interpreted somewhat differently in the two sections. I suggest softening "confirming" to "suggesting" in Section 3.2, cross-referencing Section 4.4, and applying similarly cautious wording to "The third hypothesis is also confirmed" in Section 4.1, which relies on the same evidence.

We agree that this statement is strong and that our wording/interpretation in different sections of the ms is not fully consistent. Following the recommendation by the reviewer, we will replace “confirming” by “suggesting” and carefully revise the results section and the discussion to make sure that the interpretation of our results agrees in both section.