In the following, the review comments are included line-by-line in black whereby our answers are identifiable by blue color.

Answers to reviewer Andrew Feldman

1) The major concern I have is that LFMC may not adequately partition only the plant saturation amount (i.e., relative water content, predawn water potential of leaves or xylem, etc.). As a result, the results may have a strong bias in overemphasizing the role of biomass changes because saturation considerations are not independently considered by any of the datasets. Namely, as shown in Table 1 of Konings et al. 2019 referenced in this study, LFMC is still a strong function of AGB. LFMC ultimately does not normalize out effects of saturation alone - a change in dry biomass will still influence LFMC even without a change in plant water status. Ideally, a plant water status parameter should be used that is insensitive to dry biomass. Additionally, LFMC is modeled from MODIS. MODIS is undesirably mainly a function of greenness/biomass (not always strongly a function of plant hydraulics) where use of this dataset relies heavily on how well plant hydraulics and water status are modeled (a highly uncertain modeling process). This is a lynchpin of a study that ultimately relies on partitioning effects of VOD signals into their components, where there is a bias of more certainty in the dry biomass and structure observations and less certainty in those about the plant hydraulic parameters. One insight that suggests saturation may not be well modeled is that there is a large decline in explanation of 8-day versus monthly timescales (Figures S1 and S2). While noise may act more at 8-day timescales reducing its R2 compared to monthly R2 as the authors discuss in lines 460-470, we also know that the short timescales are commonly describing plant saturation dynamics as described in the Konings et al. 2019 and Feldman et al. 2020 papers referenced here. There are also several other published works (see for example: https://doi.org/10.5194/bg-18-831-2021 with error analysis: 10.1109/JSTARS.2021.3124857) based on SMAP VOD that shows the sub-weekly timescale L-VOD variations are indicative of plant rehydration and water loss dynamics, especially in water-limited locations where the authors here tend to find some of the lowest R2 (lines 330-337). I am speculating (with help from Fig. 4 that shows some stronger LVOD sensitivity to LFMC), but this issue could additionally be reflected in lower R2 for L-VOD than for X-VOD where potentially the former is more sensitive to canopy water status, especially in tree canopies. That is assuming the L-VOD and X-VOD retrieval processes are equivalent.

The issue is not enough to prevent publication of the study because there are still very valuable results here and reliable large scale partitioned plant water status information is practically unavailable. However, I strongly recommend that the authors explicitly address this issue beyond few sentences in 500-503. Either consider using a more detailed model output of plant water status (some common large scale DGVMs like LPJ and ED2 now have plant hydraulic schemes to provide canopy water status information) in place of the MODIS LFMC and/or do more to explain the uncertainties of the MODIS LFMC product and how that could bias the results as stated about. For example, I am not familiar with the Yebra et al. 2018 study - I think a summary of how they obtain LFMC and potential limitations relevant to its use in this study are needed.

We thank the reviewer for this very detailed discussion. As pointed out, LFMC represents the ratio between the water mass hold in the leafs and the dry leaf biomass. Hence LFMC is not independent of AGB especially in herbaceous systems. However, in woody systems and forests leaf dry biomass is in

many cases only a small fraction of total AGB. Furthermore, our used AGB dataset accounts only for woody biomass (i.e. does not account for herbaceous and leaf biomass) and hence LFMC and woody biomass are not directly related. The multispectral LFMC dataset represents the relative water content of the canopy, which might not correctly represent the water content of woody parts of the trees and hence might be not sufficient to predict L-VOD. The LFMC retrievals from MODIS are based on physical canopy radiative transfer models. The use of model outputs from hydraulic-enabled DGVMs would be a possible choice to predict VOD, however, this would not really simplify our analysis because we would need to account for biases and inconsistencies between observed and modelled vegetation state variables such as biomass and leaf area index.

We propose to add more details to the LFMC retrieval in section 2.1.2 Predictor data sets:

'Yebra et al. 2018 use three radiative transfer models (RTM) for the retrieval of LFMC corresponding to different LFMC values. The used RTMs was either a coupled version of PROSPECT 1 with SAILH 1 (for grasslands and shrublands) or a coupling of PROPSECT 1 with GeoSail (for forests). Based on forward simulations of the RTMs, look-up tables (LUT) were generated for each vegetation type and the LUT was used to invert LFMC from spectral properties. The results were evaluated with LFMC field measurements whereby the overall model achieved an explained variance of 58% and a RMSE of 40% (Yebra et al., 2018).'

Generally, the uncertainty of the estimated LFMC from the RTM inversion increased with higher LFMC values. In addition, the validation of the LFMC data set is impeded by uncertainties due to difficulties of comparison between measurements on the ground and what is detected by the satellite. Factors, which are contributing to the uncertainty of validation, are for example the temporal matching procedure of in-situ samples and MODIS data, open to closed forest or how much of the understory is exposed to the sensor. However, these factors are difficult to quantify and can be only discussed in a qualitative manner. We propose to add this information in line 478.

In addition, in line 460-470 we will add the discussion related to introduced errors by the VOD retrieval algorithm (LPRM) based on the mentioned papers (Zwieback et al., 2019; Feldman et al., 2021). In particular, we would like to discuss the error shifts from SM to VOD, which are more pronoun for short-term changes as well as for higher VOD values.

2) I think the manuscript should emphasize the time dependence of the main results more. Even though monthly and 8-day timescales are different, the main predictors tend to be different and previous studies do not suggest that this would only be due to noise; the results nicely give evidence for what we have speculated all along that different aspects of the canopy (dry biomass, water, structure, etc.) influence VOD at different timescales. Specifically, the abstract (lines 31-33), overall results (lines 367-373), and much of the discussion are written as an absolute result. However, these statements mainly appear to be only a function of the monthly timescale. Future work should repeat the same analysis as equivalently as possible across different timescales.

We performed our analysis on 8-daily and monthly time scales but included mainly the results from the monthly time scale in the main text. We agree that a more direct comparison of the 8-daily and monthly results would provide a more clear evidence about the influences of vegetation properties on VOD on different time scales. We propose to add Figures representing 8-daily results and briefly discussions regarding differences between the timescales for the global models as well as for the land cover-specific

models within section 3.2.1 and section 3.2.2, respectively. These results may also alter the discussion, which will be updated.

Line-Specific Comments

Line 27: "…level 3 L-band derived…" should it be "…level 3 L-band [VOD] derived…"?

All products were derived by using LPRM. The addressed sentence will be changed to 'level 3 L-band VOD derived from SMOS and SMAP sensors also using LPRM' to clarify that LPRM is used to derive all here used VOD products.

Line 47-52: Please reference some more recent work on the b parameter and other aspects by Kaitlin Togliatti et al.: Togliatti et al., "Quantitative Assessment of Satellite L-Band Vegetation Optical Depth in the U.S. Corn Belt," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 2500605, doi: 10.1109/LGRS.2020.3034174.

We propose to add the following sentences: "The parameter b is usually kept constant which might be insufficient due to its possible dependency on polarization. In addition, the neglecting of the surface soil roughness can lead to an underestimation of VOD, especially when the vegetation do not completely cover the ground (Togliatti et al., 2022)."

Line 118: I think the article would benefit from a more explicit research objective or question. Line 118 is very general and broad and has been done in various forms previously. I suggest adding some more nuance so the reader knows what the authors wish to establish.

We propose to add the following sentence: "Specifically, our objectives are to predict VOD from LFMC, LAI and AGB by using to machine learning regression approaches and to investigate the relationship between VOD and the predictors. This objective goes beyond previous empirical studies that compared VOD with vegetation properties based on bivariate correlations or regressions but not by estimating VOD within multivariate framework."

Line 130: I want to double check that the authors were careful in keeping consistent retrieval processes for all VOD products here. It is a great step to normalize VOD at each frequency (lines 143-146). It is also good they all use LPRM (line 131). However, were there greatly different processes for choosing parameters like single scattering albedo and the modeling of surface roughness? Harmonizing VOD could potentially bias some results in disproportionately influencing certain components of the VOD power spectra (for example, monthly variances of VOD could have been greatly altered while the interannual VOD variance was not influenced).

The merged data from different satellites are all run using the same version of LPRM with principal the same parameterization for the retrievals. The parameter values differs a bit between frequencies (e.g. between Ku-band and L-band), but not over time within the same frequency. Also, C and X-band for example are almost identical in their parameterization. Additionally, in case of the VODCA data set the used time span 2015-2017 is covered mainly by AMSR2 observations and some months of TMI in the Southern Hemisphere. Therefore, the used data should be minimally affected by any bias introduced by merging multiple sensors.

Line 205-206: The authors means that the "AGB dataset is not representative…" not that AGB itself is not representative. AGB is certainly relevant to all plants.

That is correct. We will change it in the proposed way.

Line 212: Please explain what “grid-search” means. Is this describing spatial pixels across the globe or a method within the random forest approach?

Grid-search is a method to find the best hyper-parameters of machine learning model within a given range. RF provides several hyper-parameters, and it is important to find the best one to achieve the best model efficiency. In the grid-search algorithm, the multi-variate space of hyper-parameters is splitted by a grid of parameter combinations and the RF is then trained with each combination of parameters.

We propose to include this information in the text.

Line 218-219: I am not sure these definitions are common knowledge. What exactly is used to determine “minimum samples within a leaf”, “number of estimators”, etc.?

These are hyper-parameters of the random forest, which can be tuned for a better model efficiency. They determine the structure of the random forest model. We used the above mentioned grid search method for a selection of best hyper-parameter values, tested the selection in detail and decided for a final model setup. The description of the hyper-parameters and their impact on a RF model is described on the scikit learn webpage. We propose to add this information and the link in the text.

Line 230-234: It would be helpful to point out what the regressors are (f(x)) and what is being predicted? I am guessing VOD is g(mu) and vegetation structure, leaf, water, etc. observations are the (f(x))? Please clarify.

This is correct. We will specify this equation in the revised manuscript.

Line 236-237: The land cover maps are technically binary “dummy variables” here that tell whether or not to include AGB (values of zero or one). Or are they being used beyond this?

The land cover maps were translated into the fractional coverage of plant function types (PFTs) using the cross-walking approach and hence are representing a numeric value between 0 and 1 for each PFT. They describe how much of a pixel is covered by a certain land cover type. For example, PFT(herb) value of 0.4 says that this pixel is to 40% covered by herbaceous vegetation. The usage of all of them together gives a detailed description of the vegetation structure.

Table 2: A caution that for the “global” regression, one of the land cover “dummy variables” needs to be left out or else it will bias the regression. For example, if you have four binary regressors (tree, herb, shrub, crop), one needs to be chosen not to be in the regression.

Thanks for the hint but the PFTs are not binary values but instead providing a range between 0 and 1 (please refer to the comment above).

Fig. 2 and Fig. S1: A point of major clarification: It may be helpful to walk the reader through what the authors are looking for here. Are you choosing between GAM and RF? Any other specific decisions? I am not sure how to interpret the difference between the short vegetation and tree cover models. If the tree cover models have AGB, they should be expected to have better variance explained by default because they have an additional parameter that the herbaceous pixels do not have. They are also comparing different regions. It may not be a one-to-one comparison. I am not sure how to also interpret the difference between the global model and land cover model (line 289-299). I am still struggling to understand what their difference is from Table 2 and whether it is a one-to- one comparison. One thing

to consider is that models that have more regressors are forced to have higher R2 because adding regressors never reduces the variance explained (at least in least squares regression). These points should be clarified given that they influence the subsequent results.

Figure 2, S1 and S2 show the model performance for all tested models and additional results of the global model grouped by a specific land cover type. The difference between the land cover-specific model and the global models is that land cover-specific models are only trained, tested and validated for a specific land cover class and global models use the PFT fractions as additional predictors and were trained for all pixels, (for PFT description please refer to the answer of your comment on line 236-237). But the R^2 and RMSE of global model grouped by short vegetation is directly comparable to the land cover model (short vegetation) because it compares the exact same data.

For example, for the land cover-specific model treeNE we first filtered the data, which show at least a treeNE fraction (PFT) of > 0.55. This data was then split into train and test data within a cross validation. This model is only valid for pixels dominated by the treeNE land cover type. To be able to compare this model with the performance of the global model, we filtered the original VOD and the predicted VOD from the global model based on the same PFT threshold and used these data to compute the R^2 and RMSE.

We propose to extend the table caption by: 'The land cover-specific models are only trained and tested within a cross validation for pixels which are dominated by certain land cover (threshold PFT fraction > 0.55).' and changing the table line 'Global model (including land cover as predictor)' to 'Global model (including distinct CCI PFT data as additional predictors)'. For the predictors of the global model, we can write: 'AGB + LFMC + LAI + PFT treeNE + PFT treeND + PFT treeBE + PFT treeBD + PFT shrub + PFT crop + PFT herb'. This might clarify that the PFT fractions are included in the global model contrary to the land cover-specific models, which are trained to the data dominated by a specific PFT.

The reviewer is right that more predictors lead usually to higher model performances. In our case the R^2 of the global models increase more than 0.2 compared to the land cover-specific models, not only for L-VOD but also for the shortwave VODs. This is a high improvement by adding vegetation structure information and one important result. At least this is true for RF but not always for GAM, e.g. short vegetation types, as discussed in subsection 3.1.3.

The reviewer is also right, that short vegetation results are not directly comparable with tree results due to different used data. But we can have an approximation within forest classes and short vegetation classes, i.e. for example an approximately comparison of land cover-specific model short vegetation and land cover-specific models herb, crop and shrub. All data of the herb, crop and shrub models together equal the data for the short vegetation model and as an example, the herb model performs always better than the short vegetation model. The same applies to treeAll and treeNE. In some cases, land cover-specific models for short vegetation classes (which use only 2 predictors) outperform land cover-specific models for tree classes (3 predictors), especially for GAM, showing that not always more predictors lead to increased performance.

The three Figures (Fig 2, S1 and S2) are used to illustrate how we structure the result section, i.e. our findings which factors contribute to differing model performances (line 267-270). As reviewer Martin Baur pointed out (RC1, specific comment 4) sections 3.-3.1.4 can not be interpreted without the supplementary. We propose to extend Fig. 2 to a high-level Figure by including results of the GAM as

well as for 8-daily models. Labeling the panels and referencing them in the text will hopefully improve the traceability of the result discussion.

Figure 3 and line 252: I make heavy use of linear regression models and least squares, but tend to not use ML approaches like random forest. The negative R2 are confusing because they are not possible in least squares, fundamentally because they are the square of the correlation coefficient, which is forced to be non-negative. Negative R2 may be because a predictor (a constant like the y-intercept) is left out of the regression and therefore it is not interpreted like other R2. Can the authors clarify?

The reviewer refers to the squared Pearson's correlation coefficient r (r^2):

$$r_{ab} = \frac{\sum_{i=i}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}}$$

It is correct that this metric cannot take negative values.

However, we use R^2 as defined by the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a_i - b_i)^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2}$$

Whereby *a* is the true value and *b* the predicted value. This metric can take negative values if there is a bias between a and b. The coefficient of determination as defined here, typically noted as R^2, equals the Nash-Sutcliffe efficiency (or modeling efficiency, NSE or MEF). We could use the abbreviation NSE to avoid confusion. For clarification we propose to add the formula in the manuscript chapter 2.4.

Figure 4: Is this monthly?

Yes, Fig 4 presents the ALEs for the RF based on monthly data.

Section 3.2.2 and Figure 5: Given that I did not follow the specific differences between the land cover and global model, I am having trouble interpreting Figure 5 compared to Figure 4. Is the only difference that the global model (Figure 4) includes all pixels, but the land cover model are different subsets of pixels? I am guessing the use of "dummy" variable regressors in the global model also has an effect that the land cover models do not? Please clarify.

The global models are applied to all available data and use additionally layer for each PFT including the fraction coverage as predictor. Please refer to the answer to the comment on line 236-237 for an explanation of PFT predictors. Indeed, the land cover-specific models are only applied to a subset of the available data sets which equal the condition of a specific dominant land cover type (defined in table 2). For a further explanation on the differences between global and land cover-specific models, please refer also to the answer to the comment on Fig 2 and Fig S1.

Line 414: Local factors like what?

Examples for local effects are mentioned in line 428-429. We agree, that the examples should be mentioned in line 414. We propose to change line 414: 'The lower global performance of GAM suggests that local factors, *e.g. intercepted or standing water or heterogeneous soil properties*, play a role in the dynamics of VOD, which were not considered and used as additional additive predictors.'

Line 419-422: See my major comment: this could in part be a limitation of not having an independent plant water status predictor. Lower prediction of VOD dynamics in these regions is not expected because VOD and soil moisture errors tend to be smaller in regions with herbaceous vegetation. R2 should therefore tend to be higher in these regions if we have relevant predictors with R2 in forested regions decreasing due to noise. See results from recent VOD error quantification work: see https://doi.org/10.1016/j.rse.2019.111257 and 10.1109/JSTARS.2021.3124857

As this comment is strongly related to the major comment 1, please refer to the related answer.

Line 444: "Varies by wavelength." This is microwave wavelength, correct? (as opposed to a timescale-dependent wavelength related to power spectra of the time series)

Yes, this is correct. We propose to alter the text in line 444: 'varies by *microwave* wavelength'.