Response to referee #2 comment

Referee comment on "Estimating dry biomass and plant nitrogen concentration in pre-Alpine grasslands with low-cost UAS-borne multispectral data – a comparison of sensors, algorithms, and predictor sets" by Anne Schucknecht et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-250-RC2, 2021

We thank referee #2 for the constructive comments. Please find below, how we want to address the raised issues (referee comments in italic) in a revised version of the manuscript.

Main issue 1 (regarding motivation for comparing two UAS sensors)

The motivation for comparing these two UAS sensors is not clear. Are these two types of sensors are popularly used in UAS remote sensing studies? How the findings from the two sensor comparison are relevant to other studies and the UAS remote sensing community? Overall, SEQ and REM sensors are very similar. These two sensors have similar pixel resolution, similar wavelengths in green (550/560 nm), red (660/668 nm), and red edge (735/717 nm). Furthermore, the manuscript pointed out that SEQ performed well for predicting plant nitrogen concentration, while REM had a better performance for predicting dry biomass. However, it is not clear why these two sensors had such different performances in the current manuscript. The analysis and explanation for sensor performance on dry biomass and nitrogen predictions need to be strengthened.

We agree with the referee that the motivation of comparing two multispectral sensors is not well addressed in the current manuscript version. Several studies applying low-cost UAS sensors for vegetation mapping/monitoring used the Parrot Sequoia sensor and some highlighted certain quality issues (e.g. Olsson et al, 2021; Poncet et al. 2019). We wanted to test if the Micasense RedEdge-M is a good alternative in the low-cost segment and if it has a better performance due to the additional blue band and slightly different central wavelengths/band widths in the other bands. In the revised version of the manuscript we will include some sentences about our motivation to compare two multispectral sensors and strengthen the discussion about the two sensors (also considering the next part of referee comment).

In Table 2, you labeled 790nm as near infrared. However, we usually refer to 700-800nm as red edge, while wavelengths beyond 800nm as near infrared. From the soil-vegetation radiative transfer modeling view, red edge wavelengths are vital for vegetation chlorophyll content and nitrogen content retrieval. The near infrared is more sensitive to the vegetation canopy structure such as leaf area index and total biomass. From my interpretation, SEQ has two red edge bands and could potentially get better results for nitrogen concentration retrieval, but not dry biomass as lacking information in near infrared. Meanwhile, REM has information on near infrared which is good for biomass retrieval.

Thanks for the valuable comment. The forth band of the Parrot Sequoia sensor is denoted from the manufacturer as NIR band. It covers the wavelength range of 770 to 810 nm (central wavelength of 790 nm), therefore being in the transition between red edge and NIR according to the referees definition. To foster the comparison with other studies using the Sequoia sensor, we will keep the band names as used by the manufacturer. But we will include a paragraph in the discussion chapter on the differences in the “NIR” band between the Sequoia and RedEdge M sensor and potential implications of it, building on the explanations of the referee and literature information.

Main issue 2 (regarding motivation for selecting GBM and RF)

The motivation for selecting Gradient Boosting Machines and Random Forest is also not clear. Why not other more popular machine learning or statistical approaches, such as partial least-squares regression, LASSO, Ridge, or Neural Networks?

We share your concern, thanks. Actually, we explored many other algorithms, however the focus of the paper is not comparing a vast number of the ML algorithms, we constrained ourselves to the two selected algorithms, based on the understanding that those are confirmed to work as good as the others as we cited in Section 2.3.1, Caruana and Niculescu-Mizil, 2006; Fernández-Delgado et al., 2019, 2014; Orzechowski et al., 2018. We will make it clearer in the corresponding section by adding more references about the quality of the chosen algorithms.

The purpose of applying machine learning algorithms is not only to achieve good model predictive performance. Many machine learning algorithms like random forest can help to identify the relative importance of each feature input. This feature importance analysis is very necessary to understand the relationship between feature inputs and the predicted variables. However, such analysis is missing in this study. I strongly recommend further feature importance analysis to identify scientific linkage among input variables and the predicted variable to strengthen the manuscript result interpretation.

Thank you for the comment. Actually we diagnosed variable importance using both of the algorithms as mean variable importance (= feature importance) for all tested models is currently provided in the supplementary material (Table ST4) and partly addressed in chapter 3.4.2. (focussing on the most important observations). However, we did not discuss them in depth. Following the referees suggestion, we will provide more information and discussion on the variable importance.

Main issue 3 (regarding single date acquisition of UAS and field data)

The UAS multispectral data were collected from one single flight in each site. How robustness of these results across different growth stages and dates is uncertain?

We are aware of the limitation of a single field campaign. The study presented in the manuscript tests the general approach of grassland trait estimation with an application of immediate grassland trait mapping. To increase the general validity or the results of a mono-temporal campaign, we selected several sites differing in species composition, current growth height and nutrient status, which allowed us to compile a dataset of variable grassland traits. We will include this aspect in the material and method section and add a paragraph in the discussion chapter to address the applicability of the study and requirements for other phenological stages (multi-temporal applications).

Main issue 3 (hyper-parameter tuning)

Machine learning parameter tunning is a very necessary and common step to implement model training. However, this manuscript highlights the hyper-parameter tunning as one major research question. The innovations of this study need to be strengthened.

We admit that the importance of ML parameter tuning was pronounced in many of previous studies, but in practice, we observe many ML applications in geo- and environmental-science applications still omitting a comprehensive calibration procedure. In this paper, we want to show that how much progress actually can be achieved in a practical application, for suggesting that this has to be a mandatory step in similar modelling studies (e.g. 10% performance loss). We will clarify this aspect in the discussion section.

Minor issues

There are many abbreviations in Figure 2. The caption should add explanations of these abbreviations for readers.

We will add explanations of the abbreviations in Figure 2.

The reflectance values in Figure 4 look quite different from the two sensors. Do you have ground reflectance collection to validate your reflectance?

Unfortunately, we do not have ground reflectance values for validation. It is important to note that Figure 4 shows the spectral profiles of selected samples for better readability of the plot. An example for all samples is given in the Figure below for NIR reflectance vs. DM (see a comment from referee #1). The figure shows a positive, but not very strong relationship between NIR reflectance and DM as well as the different reflectance value range of the two sensors. Possible reasons for the different pattern between the REM and SEQ sensor could be the different spectral and radiometric properties, radiometric calibrations and changes in acquisition conditions.

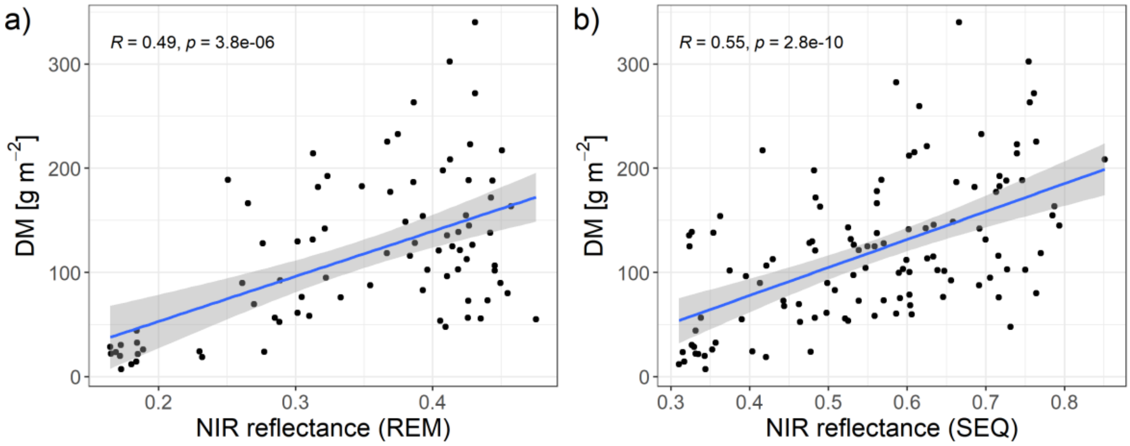

Figure 1. Scatterplots of NIR reflectance vs. DM for the REM sensor (a) and SEQ sensor (b) with linear model fit (Spearman correlation coefficient and p-value indicated in the plot). Note: there are less data points for REM as there were no flights with this sensor at the Eschenlohe site.

The manuscript mentioned that mountain regions have frequent cloud occurrences to argue the weakness of Copernicus satellite missions. However, UAS data collection under cloudy environment also has data quality issues. The manuscript may need to discuss such potential issues and mitigation strategies.

We agree with the referee that studies utilizing UAS data have their own challenges and drawbacks. Therefore, we will include a section in the discussion chapter that will discuss the challenges with UAS and low-cost sensors in general and the limitations of our study in particular.

Most parts of the manuscript used nitrogen concentration. However, Figure 6 used nitrogen content in the (c) and (d) subplots.

We will correct Figure 6.

The same issue of nitrogen concentration on Figure 7.

We will correct Figure 7.

Figure 8 (d) has clear shadows. The reflectance from these shadows needs to be either corrected to real surface reflectance to quantify vegetation traits or simply removed. I don't think the current estimates for areas in tree shadows are right.

The estimation of DM and N concentration can just be used for vegetated grassland, not affected by shadows. To make this clearer for the reader, we will mention that the plant traits estimates are only valid for un-shaded pixels and vegetated grassland areas, and mask out all non-vegetated and shaded areas in the maps of Figure 8.

The figure panel design of Figure 8 is strange. We normally put RGB into the first subplot. You have paired maps for DM and N. These paired subplots could be in one row.

We will rearrange the sub-figures as proposed.