The objectives are very clear and well-motivated. The topic fits for publication in HESS. The manuscript is quite technical but provides interesting results concerning the use of a machine learning algorithm (Random Forest) for gap filling and downscaling of a vegetation related parameter (enhanced vegetation index). The paper needs some additional work to make it more understandable to non-ML experts and to better justify some hypotheses.

We want to thank the reviewer for reviewing our manuscript, the reviewer's comments have provided us with valuable insights and suggestions that will improve the quality of this manuscript for future readers. The updated line numbers refer to the marked-up version of the manuscript in this case.

General Comments:

Meteorological data: what about radiation (as discussed line 424-427) ? Only potential evapotranspiration (or evaporation, L150) is considered. This could be a problem, especially under drought conditions. Please comment.

We agree with reviewer that solar radiation is important in determining vegetation productivity at the global scale. We state in lines 135 -138 that the guiding principle for feature selection was capturing variables that are important in determining a plant's water and energy budget. The rationale for including potential evapotranspiration was because it captures variations in solar radiation with the added influence of wind speed and relative humidity on rates of evaporation. The grow of plants is also directly correlated with their ability to evaporate water in the photosynthesis process. Finally, potential evapotranspiration is positively correlated with air temperature and solar radiation (Thornthwaite 1948, Monteith 1965, Priestley and Taylor 1972), and thus the information of solar radiation is already captured by including *pet* as a feature even during drought conditions. We have included this information in line 143-144 with: "pet was included as it is directly correlated to air temperature and radiation (Thornthwaite 1948, Monteith 1965, Priestley and Taylor 1972) and the photosynthesis potential of plants and thus can account for a plural of other variables."

References:

Thornthwaite CW. 1948. An approach toward a rational classification of climate. Geographical Review. 38(1):55– 94

Monteith JL. 1965. Evaporation and environment. Proceedings of the 19th Symposia of the Society for Experimental Biology; New York: Cambridge University Press. pp. 205–233.94.

Priestley CHB, Taylor RJ. 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review. 100(2):81–92.

Provide more explanation on how you moved from 15 vegetation types to 8. What are the criteria?

We agree with the reviewer that this is necessary and followed the reviewer's suggestion and have now added this information in lines 164 - 168. The information reads as follows, "*It is important to note that the RF was supplied with the remainder 15 unique land cover types; however, these were collapsed into eight broader classifications for brevity and clarity in the results, discussion and visualisations. Grasslands, wetlands, croplands, urban and mixed did not require grouping and represent the accompanying class in accordance with the Geosphere-Biosphere Programme*

classification scheme. Forests refers to the grouped class which containing evergreen needleleaf, evergreen broadleaf, deciduous needleleaf, deciduous broadleaf and mixed broadleaf forests. Shrubland refers to the grouped class containing closed and open shrubland; whereas savannas refer to the grouped class containing woody savannas and savannas."

L215-223: This part needs to be clarified for non-expert in RF. Describe the two hyper parameters, the search space (why 1-40, 1-20 and not 1-15 for example), what are the criteria for 12 15 as optimal? Fig. 2a only?

We have double checked and can confirm that the reference to Fig. 2a is correct; Fig 2b refers to the split strategy for selecting data for training (which is referenced in line 230). We agree with the reviewer that the text was rather short and technical and have added additional explanations in regards to what these two parameters do to make it more accessible to the general reader, however we do still realize it is rather technical, but this is something that is difficult to avoid. Please see the new information in lines 235 – 241: "*The parameters that were not subjected to hyper-parameter optimisation were set as follows: the squared_error criterion was used to measure the quality of the splits in branches, the maximum number_of_features considered in each split was set at auto which instructs the algorithm to consider all features when considering a split. The minimum samples_per_leaf_node, which determines the minimum number of samples required in a leaf node, was set at the default value of 1. The minimum samples_per_split was also set at the default value of 2, which means a split will only be considered if each branch left and right of an internal node has at least two samples in it.*".

Regarding the explanation as to why the values 12 and 13* are optimal, we do believe that this is already in the text (please see lines 219-221), which states that the risk of over-fitting increases as these values increase. In addition, we have also included that the computational load also increases as these values increase which is additional motivation to have these parameters set to optimal values. Lastly, computational time was the main motivation for setting the search space to 1-40 and 1-25 for the number_of_estimators and maximum_depth. The computational time required for hyper-parameter tuning is determined by the number of iterations that needs to be completed and the bounds 1-40 and 1-25 yielded acceptable practical computational times. We have added this argument to the manuscript, please see lines 212-218. The section of text now reads: "*The Maximum_depth determines the maximum depth of the decision tree and the number_of_estimators determines the number of decision trees used. The search space used for the number_of_estimators and Maximum_depth was 1-40 and 1-25, respectively. The upper bounds of the search space were largely determined by computational considerations, increasing the upper limits beyond 40 for number_of_estimators and 25 for Maximum_depth would result in impractical computation times. Nonetheless, even with this constraint, increasing the Maximum_depth and number_of_estimators past 12 and 13, respectively, yielded only marginal increases in test scores (Fig. 2a).*"

**We noticed that we ambiguously refer to 15 and 13 as optimal settings for number_of_estimators, we have checked the code and confirm that 13 was the correct optimal setting. The manuscript is updated accordingly.*

L235-240: again, this is too technical. It looks like a manual for the python software. Provide more information. Explain your choice, what is 'Auto', why 'Auto', what means 1 and 2 for minimum and maximum samples_per_leaf_nodes?

We agree with the reviewer that this section was too technical and have thus added additional context regarding the function of these parameters, however it will always remain a bit technical as it refers to parameters in algorithms. Please see lines 238 – 241 for the new text: : "*The parameters that were not subjected to hyper-parameter optimisation were set as follows: the squared_error criterion was used to measure the quality of the splits in branches, the maximum number_of_features considered in each split was set at auto which instructs the algorithm to consider all features when considering a split. The minimum samples_per_leaf_node, which determines the minimum number of samples required in a leaf node, was set at the default value of 1. The minimum samples_per_split was also set at the default value of 2, which means a split will only be considered if each branch left and right of an internal node has at least two samples in it.*".

Discussion: "RF can accurately predict evi at unseen geographic locations when trained on relatively few data - only 6% in this case.". Is it not a question of data quantity, but rather a question of representativity. If all 'typical situations' or all classes are sampled in the 6%, it is fine. If some 'typical situations' or classes (like under-represented land cover types) are not sampled, the prediction might be poor for these situations or classes, even if the sample size is larger.

We agree with the reviewer that this is interesting information and have added this to the manuscript. The main point here is that in the global sense only 6% is required and that it seems to be enough to sample even underrepresented landcover types. We have increased clarity relating this in the lines referred to by the reviewer and have directed the reader to the relevant section in the discussion. Please see lines 307 – 309: "*The results here show that RF can accurately predict evi at unseen geographic locations when trained on relatively few data; here training the RF on only 6% provides a representative sample of global distribution of evi values (see section 4.4 for further discussion on the influence of data representativity).*"

§ 4.3 This paragraph is too descriptive and I missed the discussed (except some technical issues). Does it mean that the method failed?

We thanks the reviewer for this comment. We have included an additional line making the point that even though the RF has more issues with fitting extreme values in the distribution than other machine learning methods it still produced acceptable results. One of the advantages of RF is also that they are more easily trained and more capable of dealing with non-linearities in the data. Using more advanced machine learning models will probably lead to better results for some aspects, but with having negative impacts on computational costs and potential non-linearities that are less well captured by other methods. Please see lines 339 -345: "*Nonetheless, given that the majority of the grid cells exhibited positive anomaly correlations, the ability to predict vegetation status under drought is still a positive result in accordance with previous research (Prodhan et al. 2022, Hauswirth* et al. 2021*). Although, provided that more sophisticated machine learning models tend to predict extreme values more accurately than the RF used here (e.g., Kladney et al. (2024)) future studies should aim evaluate their feasibility and applicability to predict vegetation status under drought conditions at the global scale. Yet in comparison with RF, the more complex algorithms have larger*

computational requirements during training of the model and are less capable of capturing potential non-linearity's ."

Minor comments

Typo line 99:vegitation instead of vegetation

Thank you, this has been corrected. Please see lines 98.

L153 Tp or tp, please choose.

We have corrected this in the updated manuscript, please see lines 147.

L317 "for at least three reasons" but only two are mentioned in the next sentence. Please rephrase.

We agree and have changed the word three to two, please see lines 346.

§ 4.5 sounds like a conclusion.

We agree and have moved it to the conclusion, please see lines 431 – 440.