Thank you for giving us the opportunity to revise our paper.   You will see from the tracked changes version that we have made significant changes to the manuscript in line with the reviewer's comments (see responses), particularly simplifying and clarifying the description of the remote sensing and expanding the presentation of the clustering results.

We have made one one modification to our initial response to the reviewers and response below is not slightly different to the response already posted. This is identified by purple text.

Reviewer 1

Many to the reviewer for the helpful comments

General comment

This manuscript tackles an important and timely question: whether multi-year satellite data on phenology (based on NDVI/EVI/NDWI processed with TIMESAT) can classify soil degradation states in smallholder grasslands and meaningfully relate to on-the-ground soil condition. The study spans two Kenyan landscapes and couples a remote-sensing classification (2013–2018) with field sampling at 90 sites (Oct–Nov 2019) for a suite of physical, chemical, and microbial variables (0–10 cm). The overall conclusion—that only microbial biomass C (and to a lesser extent bulk density) consistently aligns with the remote sensing classes—has practical implications for monitoring and restoration. However, several aspects of the methodology need clarification or strengthening before the evidence can fully support the claims.

Major concerns.

The paper mixes sensors (Landsat TM/OLI and Sentinel-2) and resamples to 10 m, but the harmonization/preprocessing steps are not fully described.

Response:

We thank the reviewer for their comments. We have now added the steps used to harmonise Landsat TM/OLI and Sentinel-2 imagery. Also, provided is a table that summarise the sensor by acquisition dates. New text below.

'To analyse the structural characteristics of grasslands supporting smallholder communities in Kuresoi and Lower Nyando, we implemented a time-series seasonal analysis that classified the stages of degradation on grasslands. We used 35 satellite image scenes from the archives of the European Space Agency (ESA 2016) and United States Geological Surveys (https://earthexplorer.usgs.gov/) (Table 1). The selection of different sensors was necessary to: i) fill missing dates from the Sentinel collection which had higher spatial resolution but shorter temporal resolution and ii) to maintain consistency in annual seasonal sampling between 2013 and 2018. The final satellite imagery was from Landsat-Thematic Mapper (TM) L2, Landsat Operational Land Imager (OLI) L2 and Sentinel-2 sensors L2A. Level 2 images are Analysis Ready Data (ARD) which are atmospherically corrected surface reflectance data and therefore free from the effects of haze and water vapour. Landsat-TM and OLI images were acquired with a spatial resolution of 30m, while Sentinel-2 images had a spatial resolution of 10m. The decision to select and process high-resolution imagery is due to the focus on smallholder dairy farms which are associated with grazing lands that are often less than 1ha and therefore easier to detect with higher resolution imagery. For LandSat-TM scenes, we downloaded blue (band 1), red (band 3), near-infrared (band 4), and shortwave infrared (band 6) spectral bands from USGS earth explorer repository. For Landsat-OLI scenes, we downloaded blue (band 2), red (band 4), near-infrared (band 5) and shortwave infrared (band 6). For the Sentinel-2 scenes, we downloaded blue (band 2), red (band 4), near infrared (band 8) and shortwave infrared (band 11). We loaded the individual bands into RStudio using the raster package. All Landsat images were resampled to 10m with Sentinel-2 images as reference. We resampled the same spatial resolution because TIMESAT 3.3 requires all image scenes to have the same spatial resolution when creating raster stacks and before model fitting. No further image enhancements were applied because TIMESAT algorithm reduces negative biases arising from cloudiness by fitting the model to the upper envelope of the vegetation/water index data (REF). Despite these corrections, TIMESAT is unable to reduce negatively biased residuals related to surface anisotropy and sensor defects. However, we did not detect the effects of both during our analysis. Afterwards, we calculated NDVI values in each pixel by dividing the difference with the sum of near-infra red and red bands (Equation 1).To derive EVI values in each pixel, we applied correction factors and divided the difference between near-infrared and red bands with near-infrared band (Equation 2). We calculated NDWI in each pixel by dividing the difference with the sum of near-infrared and shortwave infrared (Equation 3).

Table 1: Summary of dates of acquisition of Landsat and Sentinel-2 imagery used for the determination of Normalized Difference Vegetation Index, Enhanced Vegetation Index and Normalized Difference Water Index.

2013	2014	2015	2016	2017	2018
NA	2014/01/25[1]	2015/01/12[1]	2016/01/08[2]	2017/01/12[2]	2018/01/22[2]
2013/04/28[1]	2014/04/01[1]	2015/04/02[1]	2016/04/27[2]	2017/04/02[2]	2018/04/10[1]
2013/06/17[1]	2014/06/18[1]	2015/06/07[1]	2016/06/06[2]	2017/06/11[2]	2018/06/11[2]
2013/08/18[1]	2014/08/21[1]	2015/08/11[2]	2016/08/25[2]	2017/08/20[2]	2018/08/05[2]
2013/10/05[1]	2014/10/25[1]	2015/10/25[1]	2016/10/29[1]	2017/10/29[1]	2018/10/03[1]
2013/12/24[1]	2014/12/11[1]	2015/12/29[2]	2016/12/23[2]	2017/12/28[2]	2018/12/18[2]

[1] Landsat Thematic Mapper (TM) or Operational Land Imager (OLI) imagery
[2] Sentinel-2 imagery
Note: Landsat images were resampled to 10m resolution.

The specific land cover ESA product used for masking is not named or discussed in terms of accuracy/limitations for these mosaics.

To avoid sampling and analysing non-grassland areas, we masked urban, forest and water bodies. We used the ESA Climate Change Initiative (CCI) land cover vector layer (2015). We apply the masks during site selection for field visits and after the analysis of vegetation using time-series satellite images.

Terminology should be standardized (e.g., “Normalized Difference Vegetation Index,” and clarify that your NDWI formulation uses NIR–SWIR, i.e., Gao-type, to avoid confusion with the original NDWI.)

All mentions of normalised differential vegetation index have been replaced with normalised difference vegetation index.

Degradation states are defined from average distributions of TIMESAT metrics and then selected by visual consistency with Google Earth, without an independent accuracy assessment. At minimum, the manuscript should report a quantitative agreement/uncertainty analysis for the remote sensing maps.

This area of the paper was confusing and we hope it is now much clearer. We used Google earth to remove locations that coincided with recently cultivated areas (<10 years) and/or road tracks

Locations that coincided with recently cultivated areas (<10 years) and/or road tracks were removed following examination using Google Earth (2008 - 2018)

With only ~35 scenes over 2013–2018 (≈ 6 per year) and no explicit treatment of cloud cover impacts on phenology fits, TIMESAT-derived timing metrics are likely uncertain.

Moreover, remote sensing labels summarize 2013–2018 whereas field sampling is in 2019 a gap that can be consequential in smallholder systems. These choices plausibly weaken soil–RS correspondence. Several sections are overly detailed (lab methods) while key methodological choices (RS preprocessing, TIMESAT parameters) are terse.

We have now significantly expanded the description of the remote sensing work. See above. Given that we are publishing in a soil science journal we think that SOIL readers will be interested in the methods we used for soil analysis so do not propose to cut them.

The mismatch in timing of the remote sensing and the field programme are because the field programme had to be planned on the basis of the remote sensing, hence the last year we could use was 2018 for a 2019 field programme. We have added two sentences to the discussion to explain this.

‘Additionally, the fact that the RS was used to plan the soil survey, meant that the RS images did not coincide with the survey dates. Given that we were using the RS data to consider seasonal shifts in vegetation indices over six years (Table S1) we do not think that an additional year of data would have changed our findings.’

Please also state whether a research permit/ethical clearance was obtained.

Yes, all research was ethically approved.

L78-80: Reporting a single stocking rate (1–2 cattle ha^{-1}) without nuance is misleading; please contextualize it by describing the different production systems.

We have updated this with more detail and supporting references in the revised manuscript.

‘Smallholder systems in the highlands of Kenya have a range of stocking rates, typically expressed in Tropical Livestock Units (TLU) per hectare. For the for the Kenyan highlands between 1 and 1.4 TLU ha-1 are reported depending on the nature of the system (Bebe et al 2003), for Murang’a County to the south east of our study area 3-6 TLU ha-1 (Ortiz-Gonzalo et al, 2017) and for dairy cattle in Kiambu County to the west of our study area an average of 2.1 TLU ha-1 (Were et al 2025). ‘

L82-83; The discussion of soil degradation is overly simplified. Even if not the central objective, the manuscript should briefly address the complexity of degradation processes and site-specific drivers at the study locations

We wonder if the reviewer wrote this comment before reading the next paragraph which is devoted to degradation processes and describe the complexity of both the drivers and the process. We would argue that the introduction is not the place to describe all of the site specific drivers.

l172: Use ‘difference’ rather than ‘differential’ here.”

Corrected

L310: This is an isolated citation.

Deleted

Reviewer 2

Many thanks to the reviewer for the helpful comments

This study investigates the use of satellite-derived vegetation indices to identify soil degradation in Eastern Kenyan grasslands. The authors classify fields into three categories (equilibrium, transition, and degraded) and validate these classifications against a large set of laboratory soil analyses of topsoil samples collected from a subset of the sites. Due to limited initial findings, a PCA and k-means clustering approach was applied to all the soil data, identifying two primary clusters which represent degraded and non-degraded soils.

While the development of rapid, remote-sensing-based degradation assessments is of great practical importance, there are significant concerns that need to be addressed. The major points are outlined below under "General comments" with more details provided in the "Detailed comments".

General comment

While the authors' objectives and overall workflow are clear, the manuscript suffers from a lack of general clarity and technical detail.

There is an inconsistency in the terminology used throughout the paper, particularly regarding the degradation and soil property classifications (see detailed comments). Furthermore, while the authors provide an extensive list of laboratory-measured soil properties, reasons for their selection and the analysis remain unclear and confusing for the reader.

Thanks for this comment. We have added justification for the inclusion of the soil parameters into the text of the 'soil sampling and analysis' section

The description of the remote sensing methodology is incomplete. Vital information is missing regarding the selection (e.g. specific sensor details, image selection criteria, total number of images used, reasons for the temporal coverage) and the pre-processing (e.g. cloud masking is likely crucial in the humid tropics) of satellite imagery.

Following the comments of referee 1 we have now significantly improved the section on the remote sensing approach.

Statistical analyses also require further explanations and details: It is unclear on what data the tests were performed (e.g. was there data splitting by site? What were explaining and response variables?) and whether model assumptions were verified before performing ANOVA. Additionally, the authors utilized pairwise t-tests to compare three levels of a factor (degraded, transitional, and equilibrium); a post-hoc test such as Tukey's HSD could be more statistically robust and appropriate?

We thank the reviewer for the helpful comment! We agree that the section would benefit from further details on the testing procedure, for which we have revised the paragraphs in sections "Statistical analysis of field data". We also agree that some improvement is needed for the testing procedure. Therefore, we made the following changes.

The analyses in this study were done on the data from Kuresoi and Nyando respectively. Initially, we used ANOVA to test the difference between degradation classes for all variables. We have now replaced the ANOVA with a non-parametric test (Kruskal-Wallis) when the samples failed the normality test (Shapiro-Wilks). This only applies to a few variables, but it does result in two more variables (H2ONO3, Inorganic_P) being significant for Nyando.

To investigate the details of the differences, we initially used the two-sample t-tests. This is because we considered the follow-up comparison between two classes to be no longer a multiple comparison problem, as the inflation in Type-I error caused by multiple comparisons has already been taken care of by ANOVA or its non-parametric equivalent. However, we agree that the Tukey's HSD test and its non-parametric equivalent would be a better choice for reporting the result in Table 2, due to the consistency between them and the first stage of the analyses. Therefore, we have updated the letter-based display with the results from Tukey's HSD test or pairwise Wilcoxon test for the non-parametric case.

The results section generally lacks structure. Several results mentioned in the text are not supported by data in the manuscript or the supplementary material (see detailed comments below). Some visualizations would help the reader to better understand the results (e.g. boxplot and score plot; see below).

We address these comments below

The discussion section requires major rewriting. It is currently a mix of introduction and a repetition of the results, also often drifting into unrelated topics. Furthermore, the discussion does not cite enough literature (in my opinion, 10 references in the entire discussion section is not enough to discuss the results obtained (there are numerous high-quality articles published in the field)).

Ultimately, the study demonstrates that while laboratory analysis successfully identifies soil degradation status, the proposed remote sensing method does not. Given that the efficacy of laboratory analysis to identify soil degradation is already well-established, the value of this paper lies in a deeper exploration of why the remote sensing approach underperformed. The authors should focus their discussion on comparing their remote sensing results with existing studies to identify specific challenges or limitations.

Detailed comments

L24: Not all soils across Sub-Saharan Africa are degraded! Please rephrase.

We don't actually say this, but we have rephrased the opening line to add clarity.

Now reads as 'Soils across sub-Saharan Africa are exposed to extensive degradation processes, which can reduce their ability to produce crops and support livestock. '

L30: What is equilibrium, transition, and degraded soil? The reader has absolutely no idea what is meant. Please introduce the terms.

We have clarified this in the abstract as follows:

'classify grasslands as productive grazing lands (equilibrium), unstable and unproductive (degraded) and grazing lands that followed a variable trend in vegetation productivity (transition),'

L40: What are stable and transient soil variables? Please explain.

We have clarified this in the abstract:

'into stable (those that are slow to change) and transient (those that change rapidly in response to a changing pedological environment) soil variables'

L125-158: It would be more useful to have a description of the sampled sites instead of the hydrological catchments / basins. Also, since topography is a major factor influencing soil degradation this would be crucial to add.

Given the number of sample sites it is impractical to provide a description of each site. Figure 1 illustrates the topography using a 30 m DEM. We have added elevation ranges and mean slopes derived from the DEM to the field area description (L125-158).

L156-158: The degradation classification should be properly introduced, and the terms should be kept consistent throughout the manuscript. It is not clear to the reader what equilibrium, transition, and degraded soil really means (see above).

We agree that there is a lack of consistency in the manuscript and that the terms could have been better introduced. We have now reworked a paragraph (below) to introduce the terms we have decided to use and have checked the document for consistency. We believe our choice of productive, transition and degraded as the classes will have more resonance with SOIL's readership.

‘Approach to degradation classification

Our approach to defining degradation states following the concepts developed in the ecological literature (Briske et al., 2003). Productive grasslands are those where biomass productivity is and returns rapidly following dry seasons. On the other hand degraded grasslands are those with lower peak biomass and only slowly recover following drought. Those grasslands which exhibit characteristics of both productive and degraded grasslands and termed transitional. These states are defined using an   analytical approach using remote sensing images of both study sites which is set out in the following sections This spatio-temporal analysis covered a period of 5 years (2013 – 2018).’

L198-204: It is not relevant which parameters were calculated if they were not used within this study.

We are not sure what the reviewer is getting at here. The parameters are used in table 1. Examples of model parameters used include maximum index values (MV), and End of Season Values (EoV).

L171-225 and Table 1: It is not clear to the reader what kind of models were trained. It is also unclear why models were compared or what their underlying parameters are. Moreover, I recommend moving any data or outcomes presented in this section to the results (if it is relevant).

The difficulty with moving these results to the results section is that we subsequently use the model to decide where we are going to sample, a procedure that we need to describe in the methods. We therefore propose to leave this section where it is as it will be easier for the reader. However, if the Editor is not happy with this we will move it.

L235-236: This visual inspection seems arbitrary.

We agree the visual inspection might seem arbitrary. However, we assessed the benefits of including several model outputs from the time-series analysis in the sampling decision process and determined that models without landscape consistency (e.g., classifying large forest-patches as productive/equilibrium) did not provide additional insights for the landscape classifications before field sampling.

L241-243 and L246: What is “the status of the locations” and “land use history check”? Please add more details on this. The long-term history of the grassland is crucial to understanding soil degradation.

The sentence with ‘status of locations’ now reworded to ‘and use history check’ refers to the removal of land that had been previously tilled (see above). We have now edited this section to make it clearer for the reader.

‘Locations that coincided with recently cultivated areas (<10 years) and/or road tracks were removed following examination using Google Earth (2008 - 2018). Locations that had signs of recent cultivation or tillage lines were excluded. In October-November 2019, the locations were visited to remove sample locations that were inaccessible or when landowners denied access. In total, 90 sites were selected for study (Figure 2).’

L255–310: The soil sample analyses section is currently disorganized and lacks critical information. Sample processing (e.g. drying, sieving, grinding) and citations for laboratory methods are missing. To improve clarity, the authors should remove any variables that were not further investigated within this study. I recommend presenting the measured soil properties in a table, for example categorized by their subsequent classification (stable/transient).

We are a little confused by this comment. Our soil methods section follows the standard approach for methods in soil science papers, see for example the recent paper in SOIL (https://soil.copernicus.org/articles/12/55/2026/). Almost all the methods are referenced, if not then the machine used is referenced and more detail given. The paragraph ‘Description of data’ clearly identifies the stable or transient variables and the reasons why they were considers transient or variable. We have also added this information to Table 2.

We can put all this information in a table if the Editor requires it, but it would be unusual to do so.

L312: Which soil variables?

Not clear what the reviewer means here. We have added ‘soil’ in front of variables to make it clear that we are talking about soil.

L314: Why and how were variables classified and what does transient and stable mean? The information is missing completely. Literature is needed to justify such a classification too.

We disagree that this information is missing completely since there is an entire section devoted to it in the paper. However, we have now re written this to add clarity and introduced some new references.

‘For testing and clustering analysis, we focused on a total of 28 soil variables measured from the soil samples collected from the 0-0.1m depth in Kuresoi and Nyando, respectively. These variables were grouped in relation to relate to their rate of change in response to degradation as either stable (changes over multi-year time periods) or transient (changes over seasonal time periods) soil variables

Bulk density and soil hydraulic properties change over multi-annual,timescales (Berisso et al., 2012), as can contents of C (Tully et al., 2015), N (Sun and Chen, 2025) and P along with pH (Tully et al., 2015), and thus were considered stable. Other soil physical variables (sand percentage, silt percentage, and clay percentage; and aggregate stability) were also considered stable. Although there is little literature evidence to support this, we reason that textural changes will be in response to weathering or prolonged soil erosion, both processes which take place over multi-year periods and that aggregate stability is strongly related to soil texture and organic matter (Kemper and Koch, 1966).
In contrast, soil biological parameters, including enzyme activities, microbial biomass, and rates of nutrient mineralisation, respond rapidly to change in response to seasonal changes environmental conditions (Cordero et al., 2023) and therefore soil enzymes (PHO, GLC, NAG, XYL, CBH, PER, POX, URE), water extractable NO_3, and KCl-extracted NH_4, microbial C, ‘microbial N, total dissolved C, organic dissolved C, mineralisation and nitrification were considered transient.

Kemper, William Doral, and Elbry James Koch. 1966. *Aggregate stability of soils from Western United States and Canada: Measurement procedure, correlations with soil constituents* (Agricultural Research Service, US Department of Agriculture).

Sun, S. , and S. S. Chen . 2025. “ Extensive Decline of Soil Nitrogen and Its Drivers in the Lake Victoria Basin of Tropical Africa (1996–2015).” *Land Degradation & Development* 36, no. 17: 5911–5926. https://doi.org/10.1002/ldr.70045.

Skinner, R.J. and Todd, A.D. (1998), Twenty-five years of monitoring pH and nutrient status of soils in England and Wales. Soil Use and Management, 14: 162-169. https://doi.org/10.1111/j.1475-2743.1998.tb00144.x

Tully K, Sullivan C, Weil R, Sanchez P. The State of Soil Degradation in Sub-Saharan Africa: Baselines, Trajectories, and Solutions. Sustainability. 2015; 7(6):6523-6552. https://doi.org/10.3390/su7066523

L325: Why are there missing observations? Specify. Number of samples should be added to the overview table (see comment above).

The number of samples has been added to the overview table as requested.

The reasons are multifaceted. Much of the work was carried out during the covid pandemic resulting in problems with access to laboratories, samples becoming expired and and in some cases lost or damaged in transit. We do not think the reader needs to know reasons , rather what is important is how this situation was dealt with which is explained.

L330-337 and Table 2: It is not clear to the reader how the models were trained (which data, which explaining and response variables). Also, have the model assumptions been met? The dataset is unbalanced, has this been considered? Why was a t-test applied to a factor with three levels (degradation status). A Tukey's HSD test might be more appropriate in this case..?

For the analyses of the field sampling data with the degradation classes, we first split the data by areas, Kuresoi and Nyando. Then, for each area, ANOVA was applied to each variable using the degradation class labels as the explanatory variable. We have added some explanations to the main manuscript. We have also updated our testing procedure by including the non-parametric tests (Kruskal-Wallis and Wilcoxon) and the addition of Tukey's HSD test. We have provided a detailed response to this in the general comments above.

Here we would like to provide some more discussion on the use of non-parametric tests as opposed to ANOVA when some of the model assumptions (e.g., normality) are violated, and explain why we did not replace ANOVA for non-parametric Kruskal-Wallis tests on all variables. In our case, running the Shapiro-Wilks test suggests that the majority of the groups pass the normality test (21 out of 28 for both Kuresoi and Nyando). Similarly, the majority of the soil variables have similar variances across groups, but some show larger differences. Overall, there are some violations to the classic assumptions of ANOVA in some soil variables, but the majority are still suitable for a classic ANOVA. Non-parametric tests, such as the Kruskal-Wallis test, do have advantage in situations where the normality assumption is not met, but a loss of information when we convert the observed values into ranks is inevitable. In addition, there are different opinions with respect to how robust ANOVA is when there is violation to the assumptions, for example, Lantz (2013) and Blanca et al (2017). So far there hasn't been a conclusive answer to this question (as many of the analyses on the robustness of ANOVA were based on simulation studies, not theoretical proof). Therefore, the new approach we take here is to apply ANOVA to variables that passed the normality test, and to apply Kruskal-Wallis test only on those who do not pass the normality test.

Lantz, B., (2013) The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology*, 66, 224–244.

Blanca, M., Alarcón, R., Arnau, J., Bono, R., Bendayan, R., (2017) Non-normal data: Is ANOVA still a valid option? *Psicothema*, Vol. 29, No. 4, 552-557

L337: What is the "content of the mean differences"?

This refers to how the degradation classes differ for each variable. This essentially means pairwise group comparison. We have revised text in the main manuscript for clarification.

L343-344 and L349-360: Why was the Gaussian mixture model compared to the k-means clustering? What does an application of a model "for reference" mean? This is again very arbitrary, and the entire paragraph is irrelevant and should be removed. L355-360 sounds like a justification for why k-means clustering was selected in the end. In my opinion this is not relevant.

As this clustering analysis is exploratory, that is, we do not know the truth, nor do we take the ERU labels as the "truth", we applied different clustering methods to the data sets, to see how their results differ and whether one result makes more sense than the other (in terms of separating the sites). We agree that this information needs not to be presented in the main content. We now have removed the related content.

L384-386: It is not clear to the reader how the t-test was used to determine the separation of populations.

The two-sample t-test was carried initially to obtain additional information on the mean difference of different soil variables between the two clusters. It is used when we determine the better clustering method in this case. As the comparison of clustering methods has been removed from the main manuscript, we removed this part as well.

L388-391: It is not necessary to describe the used R functions. The R packages and their versions should be sufficient.

We have now removed the details of R functions.

L 394-395: It would be helpful to the reader to visualize the results of the variables which differ significantly between the degradation classes in a graph (for example in a boxplot).

Thanks for the suggestion! We have added the boxplots overlaid with observations to the manuscript. See below

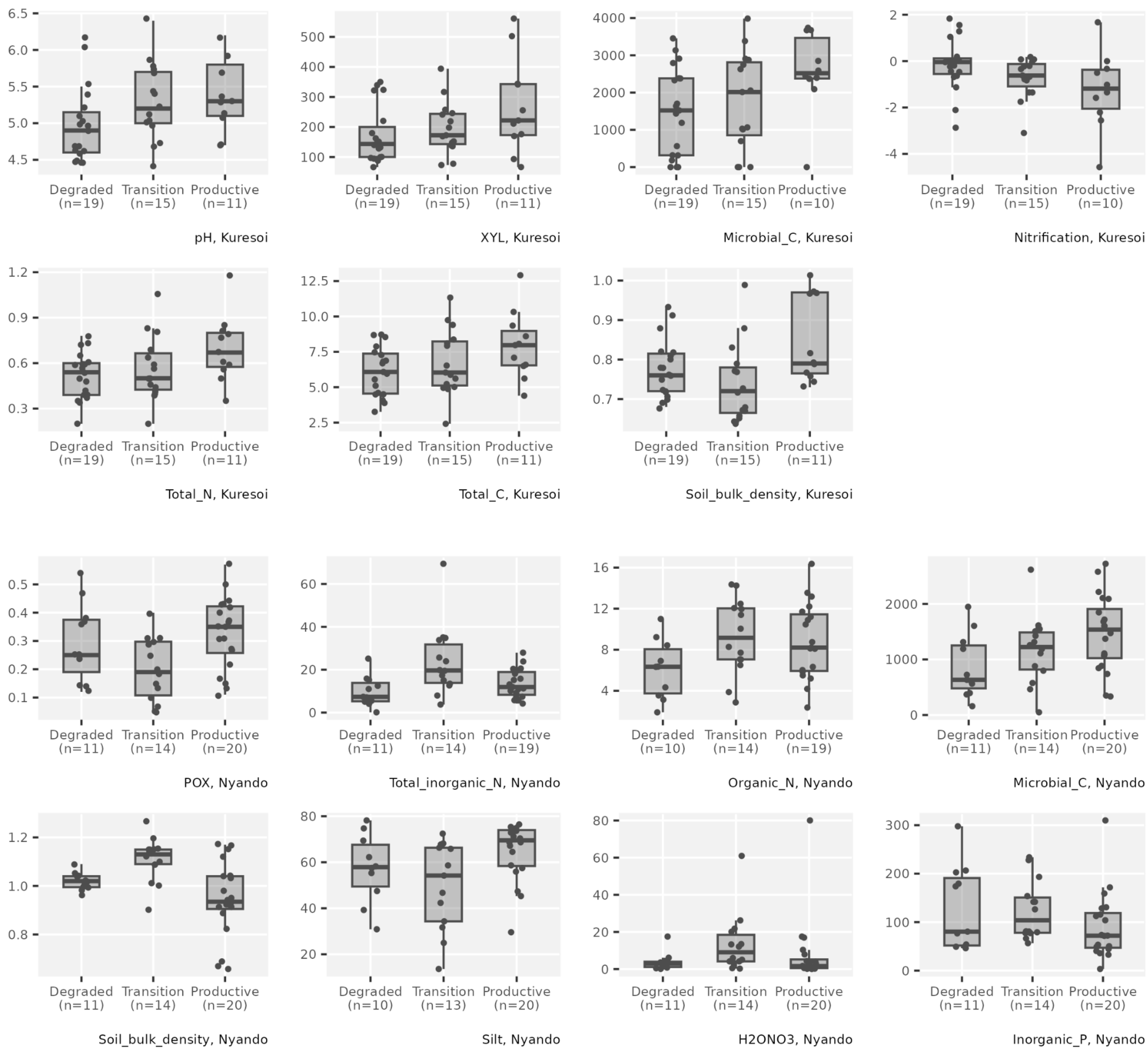

Figure **: Boxplots of the soil variables with significant difference between degradation classes based on ANOVA or Kruskal-Wallis test from Kuresoi and Nyando respectively. The dark grey dots are the observations overlaid on top of the boxplots.

Please correct line numbers (they should be continuous but randomly restart here).

L10-30: Numerous soil properties which are reported in this section are not presented in the manuscript (for example inorganic N, organic N, pH, clay, bulk density). When you report results, always refer to the table or figure where the reader can find these results (this is completely missing with one exception of the density plots). Moreover, it would be nice to have a graph showing the results of the principal component analysis and the k-means clustering.

Thanks for the advice! We have strengthened the link between the tables, figures and the text.

For the result of k-means clustering, we would like to point to Figure 3 and 4 and Table 3 for the presentation of the clustering results as density curves and in numbers. The reason we only presented the density plots of a few selected variables is because, although all stable or transient soil variables were considered in the analysis, not all of them are well separated by the K-means method. Hence we only presented the result of five selected variables that are well separated by the clustering algorithm. As for the principal component analysis, it is used for the purpose of dimension reduction, so that the result from k-means clustering will be more stable. For this reason, the features (for example the variances and the loadings) of the principal components themselves are not of particular interest to the clustering analysis. Therefore, we decided not to present the details of the principal components in the main manuscript.

Table 3: It would be interesting to see which sites have been correctly classified and which didn't (for example using a confusion matrix). This table is otherwise irrelevant.

Yes, that is good advice. We have now extended Table 3 to also include the count of sites in the two clusters under different degradation classifications.

Table 3: Number of stable and transient sites allotted to two clusters at Kuresoi and Nyando using K-means clustering.

	Kuresoi		Nyando	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Stable variables				
Degraded	8	9	2	3
Transition	8	7	4	6
Productive	2	7	4	12
Total	18	23	10	21
Transient variables				
Degraded	4	10	2	8
Transition	6	9	8	6
Productive	6	3	5	13

Total	16	22	15	27

L23-25, Figure 3, and Figure 4: Use the same terms for soil properties in figures and text. Otherwise, the reader cannot easily relate the text to the figures. In these two figures it is not clear what the p-value indicates. Also, the number of samples in each cluster needs to be added.

We have updated the x-axis labels in the density plots in Figure 3 and Figure 4 to be consistent with the variable names in Table 2. We have also removed the p-value from the plots. The p-value is the result from the t-test used to compare different clustering methods. Since the comparison has been removed from the main manuscript, we also removed the p-value information in the plots.

L26: What are degradation labels?

L26-28: This belongs to the discussion and a reference is missing.

We agree that this should be in the discussion

L44-53: This is a repetition of the introduction.

We have changed the wording to avoid repetition

L71-72: Completely unrelated sentence to the topic.

We disagree. Soil C and N status is often used as a degradation indicator. However, this could been worded better and has been revised. Now reads:

'Apart from microbial biomass C, there was little consistent agreement between the remotely sensed classification with field-based soil variables (Table 2). . Some variables considered good proxy indicators for soil health and are correlated with other important soil functions (Lal, 2016), such as C and N concentrations , C:N ratio and pH, were statistically significant for one site, but not the others.'

L75: This is the first time rainfall appears and seems very arbitrary.

We discuss erratic rainfall in the introduction. We are discussing problems with remote sensing soil degradation so it seems reasonable to discuss it here.

L85-96: This is repetition of your results. This does not belong to the discussion.

Agreed. Deleted

L97: Salinisation? Again, very arbitrary.

Not at all . We are discussing high pH values and the possible reasons for this.

L105: Since when were these ratios used? These properties have not been introduced and occur the first time.

The properties are given in table 2 and although we do not go through every non significant property we do mention the significant ones. Here we point out that the there is a lack of corroboration from the C, N P ratios hich seems reasonable. We added a reference to Table 2 to help the reader.

L96-108: Some discussion, but no reference at all.

Reference added.

L108-109: Repetition

Repetition deleted

L110-117: This part of the discussion seems interesting. But again, literature is missing and it needs to be extended.

We will extend the discussion.