Methodology for GHG and Co-Benefits in Grazing Systems V1.0

Response to Reviewer Hugh Sturrock

Reviewer's Information

Specialization:	Soil ecology, biogeochemistry, rangeland ecology

---

Authors response to reviewer

Please find below a detailed response to each comment. We indicate where in the text and how each comment was addressed.

Reviewer's Comments	Authors Response	Reviewer´s Decision (PASSED/ REQUIRES FURTHER REVIEW with comment)
In 1.2 the second approach is described as 'a spatial interpolation based on proxies from GIS data layers'. I think the wording is confusing. Isn't it just spatial interpolation? I would reword to 'The second approach uses spatial interpolation'.	We reworded the text accordingly.	*-passed-*

The section ‘Temporal boundaries’ discusses timings of ‘monitoring rounds’ however, it isn’t clear at this point what a monitoring round actually is. Related – in Figure 2, the legend discusses timings of ‘sampling rounds’. Are these the same as monitoring rounds?	Good catch! we were using it interchangeably, but we have now left “sampling round” in most of the text to make it more clear. monitoring round would be more comprehensive and is now only used in sections where we refer to the full set of activities including the GIS analysis	*-passed-*
In the sample size calculator explanatory document, the table showing the different characteristics to consider has ‘topography’ twice. It is also labelled ‘proximities to SOC’ instead of ‘proxies’. It strikes me that these site characteristics could be determined in a more standardized way using GIS/remote sensing. Perhaps this could be integrated into a future version of this protocol?	We just corrected those errors. We agree about integrating the option to carry out a more standardized approach, the aim of the calculator and this qualitative assessment is to reduce the barrier of implementation, but better approaches should always be allowed. Currently there´s no specific procedure for determining these characteristics, so GIS could be used if the monitoring team has the skill set. I think if there's a more standardized procedure or tool to do this, it could be cited as an alternative, but I would avoid imposing the use of remote sensing for any step where a more simple or economic approach can be implemented.	*-passed-*
In the section about ancillary data (3.1.1.2), I think the constraints are too strict, particularly when predictors are used to generate spatial predictions. With appropriate statistical approaches, a broader set of data can be useful. For example: - I. If data from outside the project area are from a different time period, then time can be included in the model itself. If spatial interpolation is being used, then it is possible to conduct spatio-temporal kriging. Similarly, if predictors are used (e.g. satellite imagery), then the corresponding image in time can be used in the model. i.e. if soil data were taken 5 years ago on a	We deleted this section, as it was confusing. The intention behind this section was to allow for projects to use data from other farms if there had been data collection in the past and they were not fulfilling the sample size requirements as independent projects. So it worked similar to an option of grouping projects. We agree that the requirements were so strict that it is very unlikely this would at all happen. Also, because we are now being more rigorous in that the projects would need to accomplish the minimum number of samples required, it simply does not make sense any more.	*-passed-*

neighbouring farm, then the spectral values at that time point can be extracted and used in the model. - II. As for the temporal dimension, if the climatic zone is accounted for in the model, then data from a different zone can still be useful. I can see this might be an issue if only spatial interpolation is used though. - III. As for II, if management practices are accounted for in the model, then data from farms with different management practices can still be useful. - IV. As for II and III - V. I think this is probably OK to include as there are a number of different ways soil samples are taken.	Ancillary data for the GIS analysis is allowed and some examples are provided in section 3.3.1.1.	
For the section on 'Extracting spectral values at sample points' (3.3.1.1), it isn't clear why imagery at a single time point within 4 months should be used. We know that sequestration of C within soils occurs over long time period. Using remote sensing to look at variations in spectral indices over time (e.g. variation in NDVI over the past 12-24 months) is likely to be useful. I would at least suggest this as an extension.	We have now updated this to allow for a whole year (+/- 6 months around the sampling date). The idea behind this is that correlations should be linked in time with the sampling period. Although the changes in soc stocks might not be too large within the year, we have some examples of pretty high SOC sequestration rates, above 4 ton CO2eq/ha/y in which cases the changes in one year could be pretty substantial.	*-passed-*
I always get nervous when outliers are removed. I think it should be clear that this is only done in exceptional circumstances. Given that removal of outliers almost always improves model fit, there is a large incentive to find a reason to remove data points.	We agree with this, so we further clarified in the section 3.3.1.2. that "Outliers can be removed, but the methods used and a justification must be included in the report. Outlier removal should be performed **only under exceptional circumstances** by using standard statistical techniques (...)"	*-passed-*
I don't think it's a good idea to constrain values of predictions to the maximum observed in the data. This makes it impossible to detect hotspots (or coldspots) of SOC that are missed in the soil data but	This has been a topic of internal debate for a while and we are open to discuss and make changes if there is consensus. The logic behind this is to lean more	*-passed-*

could be important in estimates of overall SOC stock. I'm guessing that rule is there to avoid erroneous extrapolation into feature space not experienced by the model, but I think the drawbacks of including this rule are a greater issue.	on the conservative side for the total quantification of SOC stocks. Monitors should attempt to cover the whole spectra of potential soc stock values when sampling. Stratification, use of GIS and pre-sampling are some of the recommended steps, along with achieving the minimum number of samples and expanding to an optimal whenever it's possible. But, because beyond the sampled range of values there's uncertainty about the relationship to the satellite data, any predictions of values outside the sampling range would be based on the assumption that the same relationship holds. If this is not true, there´s a risk of overestimation that from our perspective would be more harmful for sales and on the risk of getting bad reviews or critiques from the VCMs. Conservative measures tend to be well received by the VCMs in general, and because the ultimate goal here is to sell credits, we think it worth keeping it for now. But we are happy to further discuss with the reviewer, and also this seems to be a good topic to touch on during the public comment phase, so we can have this discussion in the eyes of the public directly, if the reviewer agrees.	
When evaluating model predictive performance, which metric should be optimized? It might be that the model with the best R^2 and the best RMSE are different.	This was now clarified in the text, and so it reads: "Once all models have been scored, the model with the highest predictive accuracy should be selected to predict SOC% throughout the project area, **prioritizing smaller errors over higher R^2 values when there is divergence**."	*-passed-*

For clarity, I wouldn't repeat the section on modeling twice (I think they are word for word replicas?). I would have a single section on fitting models whether you are fitting to SOC or SOC stock.	we expect people to choose one or the other so we prefer to keep this duplicated to ease the user to follow the method. But we agree that is a duplication here.	*-passed-*
Why in equation 6 is coarse particle fraction unitless? Shouldn't it be percentage?	Yes, thank you for noting. It was corrected accordingly.	*-passed-*
I understand that you are proposing using CV as a way to assess model predictive performance. But, this doesn't really tell you how well the model works at locations the model has never seen. I would leave space or recommend that an independent test dataset is collected.	By using leave-one-out CV, the model is trained and tested using independent data sets for each iteration. But we agree in that it doesn't reflect exactly the uncertainty from the whole project area. We now renamed that section as the Error estimation and added a second option to estimate uncertainty, with a clarification on the differences in both approaches in terms of reliability and expertise requirements. Please review section 3.6.4.	*-passed-*
In calculating final SOC stocks section 3.3.3. , I'm not following this suggestion. 'Alternatively, please be sure to correct the spatial resolution of pixels to match units of tonnes per hectare before calculating the sum.' In my mind, this would mean creating pixels that are 1ha in area. Is that what you mean? The alternative is to calculate the mean t/ha across all pixels and multiply this by the area (in hectares) of grassland across the whole project site.	We have simplified this section following the reviewer's suggestion. It now reads: *To ensure only grasslands are included in the final soil organic carbon stock estimate, the grasslands mask created in Section 2.1.1 should be used to estimate final stocks. The QGIS zonal statistics tool (or equivalent tool) can be used to estimate the average SOC stock (ton/ha) across all the pixels within the Net Grasslands mask, which then should be multiplied by the total number of hectares of net grasslands area calculated in section 2.1.1.* *The resulting number is the final soil organic carbon stock estimate for the sampling monitoring round, in tonnes.*	*-passed-*

I may have missed it, but it seems as if each monitoring/sampling round is treated independently to the previous rounds which seems like this misses an opportunity. As written, when fitting models/interpolating only the data from that monitoring round should be used to generate estimates. Almost certainly, data from previous rounds can help inform predictions in any given round, as long as the model used accounts for the time at which observations were made (which can be done using the dynamic remote sensing data, or a 3D kriging model (e.g. x, y, time). The only thing to make a decision on will be whether to update the baseline estimates retrospectively at this stage. I think this would be a good idea as you will most likely get the best estimate at any given time point if data from across all sampling periods is used.	This might be a potential upgrade to be further discussed with the reviewer. Because this development idea would demand a major upgrade and some testing, we would like to explore this without stopping the current review and update process. Our current assumption is that correlation between samples and bands would vary from year to year within the same project area , but as we said, we are open to review our assumption and explore other options as well.	*-passed-*
In section 3.6.4, I take it nRMSE is normalized by average (i.e. RSME / mean of observed values)? If so, I would specify here for clarity.	It is now specified in the text.	*-passed-*
In this section (3.6.4) you specify 'When estimating the final uncertainty for the estimated change in carbon stocks between two periods, the total absolute uncertainties (not normalized) from both periods would be summed before normalizing.' I'm not sure I follow what this means. Do you mean the RMSE should be calculated using the error between observed and predicted values across all time periods and the normalized (divided) by the mean of all the observed values across all time periods? I would include equations here to make things explicit.	We corrected the text now, so it's more clear that we are referring to normalization by sample average at each point in time.	*-passed-*

For Equation 16, is $E_{fertilizer}$ the same as FEt_{-0} in equation 9? If so, I would keep the mathematical notation common across equations to avoid confusion.	Corrected in the text.	*-passed-*
Section 5. For the selection of control areas for 5.1 and 5.3, if option B (target properties) is used, could this not open up opportunities for corrupt practices – e.g. projects could select sites within the same ecoregion/soil type that they know have large patches of bare soil. A somewhat similar, but less strong, argument could be made for using option C. What is the downside of focusing on option A only? Alternatively, could you not create a baseline using historic EO data (e.g. mean NDVI 3 years prior to project start date)?	The idea behind this is to generate a benchmark area that will work as a dynamic baseline. Using the historical data from satellite remote sensing could be tricky, as climate conditions might be hard to disentangle from the management impact. So we prefer to allow for options that could work as a dynamic baseline instead. Option A was the original only allowed possibility in the older version of this method, but we decided to expand to other options as well. Noting that options B and C might be less rigorous or on risk to be questioned, we decided to increase the requirements, so the minimum control or ideal benchmark area is now set to 1,000 ha. We decided this min. area intuitively, so we will bring this up during the public comment period to adjust it to what seems reasonable for the different stakeholders.	*-passed-*
For the vegetation index section, step 2 of scoring the vigor is calculating the highest NDVI/EVI value per pixel. I would avoid using the highest value, as this can be skewed by cloud/haze in the image. I would use something like the 75^{th} percentile or the median.	We agree with this suggestion, and it has been changed to the 75th percentile.	*-passed-*

For 5.2.2 – should this be conducted every year? If so, how do you convert yearly scores into a single score for the scorecard (Table 6)? The same is for 6 – animal welfare.	The co-benefits MRV will be done following the same frequency as the SOC MRV. Realizing that this was not explicit anywhere, we now added one line in the Co-benefits introduction at the beginning of the document, in section 1.4.	*-passed-*
Section 6. As for 5.2.2, it isn't clear how often this should be conducted?	Addressed in the previous comment.	*-passed-*
Section 9. What happens in the case where the model used to generate predictions is proprietary/SaaS? Should the verifier use the same tool/platform to generate the predictions?	This is a good question, we are not opinionated here . But one thing that occurs to me is that it would be good to enable access to a future verifier when the Monitor pays and signs in for a tool or platform that is proprietary/ SaaS. I would suggest that such a platform provides the verifier access as part of the "package".	*-passed-*
The methodology is feasible, however, clearly projects will need access to expertise in GIS and statistical modeling. Without these expertise, the methods used to generated predictions could wrongly implemented and generate erroneous results.	We now added the skills required to run it in the beginning of the Methodology doc, in section 1.2.: *1.2. Minimum skills required from the Monitor* *This Methodology requires that the Monitoring Team has the following skills and experience:* ● *soil sampling experience* ● *moderate to strong quantitative spatial analysis and remote sensing skills (e.g. GIS, Google Earth Engine, geostatistics)* ● *some experience in GHG accounting and in*	*-passed-*

	environmental monitoring. A professional in the area of agronomics, environmental science, soil science or biology is recommended.	
I think the general approach of generating estimates using soil sample data, remote sensing and statistical modeling is sound. As mentioned, I have some questions about whether the proposed way of measuring uncertainty actually relates to a statistical difference and would love to see either some evidence that two estimates made at two different time points with a CV error of <20% actually corresponds to a statistical difference between the two, and/or to allow projects to take a more statistical approach to uncertainty/change (as discussed above).	We would like to change into measuring uncertainty at the stock map instead of sample error, which is the metric that we are using as a proxy to uncertainty right now. In an attempt to improve this, we now renamed the uncertainty section as 3.6.4. UNCERTAINTY/ERROR DEDUCTIONS section, in which one we clarify the main differences between both and add an option to estimate uncertainty and keep the option to estimate the error. Please review section 3.6.4. We are planning to open this discussion very actively during the public comment, and expect to rn a webinar about this with some experts including yourself to debate on the best approach to take here and discuss feasibility with project developers and monitors implementing or potentially interested in implementing the protocol.	*-passed-*
I'm slightly confused by the section on stratification. Stratification is primarily used as a tool to support design-based estimates of outcomes. This protocol is focused on generating model-based estimates. There are more suitable survey designs for spatial interpolation/modeling than stratification. See work by Diggle (e.g. Diggle and Lophaven (2006)) and Heuvelink.	We are intentionally biasing the capture of variation in feature space using methods such as conditioned Latin Hypercube since our predictive modeling approach (using regressions or ML) assumes we have sampled feature space variation well. But we also added your references as another bullet , for when	*-passed-*

For simplicity, I would recommend/suggest a spatially disaggregated (inhibitory) design with some close pairs of points. This allows good coverage over space and over landscape variability, as well as enables good estimates of covariance parameters (for kriging) both of which are beneficial for spatial modeling. Peter Diggle and Søren Lophaven. "Bayesian geostatistical design." Scandinavian Journal of Statistics 33.1 (2006): 53-64 Heuvelink, Gerard BM, Dick J. Brus, and Jaap J. de Gruijter. "Optimization of sample configurations for digital mapping of soil properties with universal kriging." Developments in soil science 31 (2006): 137-151	interpolation is the end goal approach.	
Is there good evidence to suggest that sampling to 10cm depth is adequate? I would have thought a minimum of 30 cm would be more appropriate?	This has been addressed and we now mandate 30cm min depth. Please see section 3.1.4., point 2)"The minimum sampling depth is 30 cm. Justification must be provided if sample depth is shallower than 30 cm10-30cm."	*-passed-*
Its not clear to me what the guidance is on sub-sampling/compositing. Given the reliance on interpolation/modeling, I think subsampling is a bad idea, unless the subsamples are from essentially the same point in space (i.e. very close to each other). In the extracting samples section it mentions scenarios where sub-samples are taken >4m apart.	Good catch! subsamples are supposed to be taken within the 4m radius, otherwise they are considered a separate sample. We deleted the point in section 3.1.4 and clarified the 4m radius requirement	*-passed-*
Given a recent report on the very wide variability of lab results, I would recommend the same lab is used through the project lifetime.	We agree with the reviewer. We now added a note in section 3.2.1. *<u>Note</u>: It is strongly recommended that the same analytical procedures and service laboratory are used*	*-passed-*

	across all sampling events, to reduce additional errors from calibration bias or changes in the analytical techniques.	
- Similarly, the accuracy of bulk density estimates affects estimates of SOC stock. I would therefore strongly recommend that the same approach to measure BD is taken across all sampling events.	We agree with the reviewer. We now added a note at the end of section 3.2.1.1. *Note: It is strongly recommended that the same approach to measure bulk density is taken across all sampling events.*	*-passed-*
- Given the relatively small changes in SOC stock that are likely to occur, why is the threshold for worrying about gravel context set to 15%? In theory this means that soils with 1% gravel and 14% gravel are treated as having no gravel, which means that any variation within this amount can impact SOC stock estimates. I would be tempted to be much stricter with this criteria/threshold.	The main reason for minimizing the requirement for gravel content assessment is to reduce sampling efforts to the extent possible. We reviewed the current literature more in depth to adjust the threshold. We concluded, based on our findings, that we should require that rock or gravel content be assessed "if soil has >10 % gravel or the stones are >2 cm …" Some extracts and citations justifying this below: - *Soil organic C concentration can have a greater effect on the SOC stock variance than bulk density and rock fragment volume in non-stony soils (Goidts et al., 2009, Schrumpf et al., 2011). However, Schrumpf et al. (2011) reported that rock fragment volume fraction is more important than SOC concentration to SOC stock variance in soils with rock fragment concentrations greater than 20%.* - In https://www.sciencedirect.com/science/article/abs/pii/S0016706120325921	*-passed-*

	- *If soil has >10 % gravel or the stones are >2 cm conventional bulk density readings will be inaccurate, as most coarse fragments have bulk densities of 2.2–3.0 g/cm3 (McKenzie et al., 2002).* Coughlan, K., Cresswell, H., & McKenzie, N. (2002). *Soil Physical Measurement and Interpretation for Land Evaluation* . CSIRO PUBLISHING. Retrieved from: https://www.perlego.com/book/1468472/soil-physical-measurement-and-interpretation-for-land-evaluation-pdf (Original work published 2002)-also cited in: https://www.soilquality.org.au/factsheets/bulk-density-measurement We corrected the previous requirement and added the citation in section 3.2.1.2. GRAVEL CONTENT	
- For the SOC stocks calculation approaches, I'm not sure what merit there is in approach A. Given that SOC and BD estimates are available at each point, I would have thought approach B would provide a better approach (i.e. less error) and doesn't require the collection of any additional data.	We decided to keep option A for now as we are still in our learning curve and don't have enough use cases to make sure option A wouldn´t be preferred in some conditions, i.e. if soc% correlates better than stocks.	*-passed-*
In terms of sample size for baseline, presumably it could/should be possible to take an initial sample, fit models to understand their predictive performance, and then take more samples if the performance is deemed too low? This also has relevance for the point related to prior data – if prior data/estimates are available that could allow for fewer samples to be taken at baseline, an	We agree about integrating the option to carry out a pre sampling and test models performance in advance to adjust sample sizes. The aim of the calculator is to simplify this initial estimation by providing a conservative estimate and reduce costs, but there's a tradeoff that the monitors should evaluate. We are	*-passed-*

initial assessment of model performance might help establish whether more samples are needed. I think it would be worth allowing/mentioning this.	now making it more explicit in section 3.1.1. that the sample calculator can be used in the absence of a better approach for the estimation of the sample size, and that in the case of implementing any alternative approach, it should be backed up by a robust scientific citation.	
In section 3.6.4, would it not be preferable to use a sliding scale to calculate the deduction from uncertainty as opposed to the classification/step scheme. Under the current protocol, if you have an uncertainty of 20% there is no deduction, but an uncertainty of 20.5% there is a 10.25% deduction. Similarly, an uncertainty of 30% incurs a deduction of 15%, whereas 30.5% uncertainty incurs a deduction of 22.8%. You could use a formula such as UD = U * (2*(U/50)) Which could be used when U>20% if you want to ensure UD=0 when U<20%.	We appreciate this suggestion and we included this as an upgrade in section 3.6.4.	*-passed-*
I would like to suggest another way to estimate uncertainty of change. Instead of using cross-validation to calculate a discount, could there be space to allow projects to calculate probability distributions or standard errors around the estimate of SOC stock for each time period? This can be done using the spatial modeling/interpolation methods proposed (e.g. using conditional simulation or quantile regression). With this information it would then be possible to statistically test whether SOC stock is higher than baseline, i.e. estimate the probability that the difference is real. These estimates of probability could then be used to calculate	We have been digging into methods and discussing this with several members of our community, including the reviewer. Given this is a hard decision due we need to balance expertise requirements and scalability, we are going to open an active debate in our social platform Hylo and we are planning to have another instance (for example a live webinar with experts) to decide on the best solution. In parallel we are doing some tests internally with some datasets to test	*-passed-*

the deduction (i.e. if there is an 80% chance the difference is real this would equate to a 20% uncertainty). This would allow for a more statistically robust measure of change. The only issue I can see here is that this is a fundamentally different way to estimate uncertainty. Projects could, therefore, estimate uncertainty both ways and choose the method that produces the lower uncertainty estimate. If it were me, I would only use the more statistically robust way, but I can appreciate that this might involve a higher level of statistical expertise.	ourselves what's the complexity behind some of the options. Being humble about our own limitations in geostatistics, we hope that this decision will be based on the outcomes from a more collective discussion. For now, we have reworded the Uncertainty section as "3.6.4. UNCERTAINTY/ERROR DEDUCTIONS". In this section we now clearly distinguish between both concepts, and highlight the pros and cons in each case. The Uncertainty section involves conditional simulations. The error section uses the former LOOCV nRMSE. We then allow for any of both options, as long as it is clearly reported with the corresponding word. We believe that, in order to improve trust in the results, those users that can implement the more complex analysis will choose the Uncertainty option. We are leaving a benchmark in this section 3.6.4 of the Methodology to allow for changes depending on this reviewer's opinion and from public comments debates. All the debates, webinars and documents that emerge from this will be held public. The reviewer is welcome to participate during the public comment period as well.	
For 5.2.1 seems to allow projects to essentially propose their own metrics/scoring. While this is flexible, for projects not familiar with these metrics, it will involve a lot of back and forth in discussions which can slow things down. I would lay out a default approach using something like Google's dynamic world data layer (https://dynamicworld.app/) as an input and have space for projects to develop/suggest their own alternative should	We now added the option to pull data from the Dynamic world dataset. But in terms of landscape metrics, it's hard to generate a unique protocol given that the relevant landscape metrics will change depending on the keystone species or umbrella species endangered in a specific area and their habitat requirements.	*-passed-*

it be more appropriate given the local context.		
I'd like to see something on additionality. How can the projects demonstrate that these changes were not going to occur in the absence of the credit generated? The use of control areas for NDVI/BSI is a good start. There may be some analytical things that could be done (as per the use of control areas for NDVI/BSI), but there might also be some things projects could present to demonstrate that there are no other incentives (i.e. government) available to make these changes.	This is addressed at the Credit class level. Additionality in this credit class is defined as the outcomes from practices that are not business as usual. We don't use the financial definition of additionality, as we think that project developers and land stewards that have been trying to outperform management in previous years without any incentives have even more merit and shouldn´t be excluded. Projects must demonstrate that they are doing additional management than business as usual, and that there ´s no double counting.	*-passed-*

Reviewer´s response through email:

The responses look good to me.

The two main things I would like to see further discussion on are:

1) Inclusions of historic data when generating predictions for a given time period. The argument that relationships with predictors may be different at different time periods is valid, but this can be captured as part of the modeling process. The inclusion of data from across time and space will almost certainly improve the precision of the estimates at any given time period.

2) Estimating uncertainty in SOC/stock estimates for a given time point. As I've mentioned before, I think it would be preferable to take a more statistically rigorous approach to measuring change, which would require generating uncertainty estimates for each time period instead of calculating cross validation error.

-Authors will follow up on these during the public comment period-