Authors' response

We would like to thank the two reviewers for their constructive comments. The location of changes made to the manuscript is stated in this response and can be seen in the submitted document (with tracked-changes). In addition to this response and the revisions made to the manuscript we also intend to submit an article to the EGU Biogeosciences division blog. This article will plainly describe our methodology and discuss the role of earth observation for grassland management inference and ecosystem biogeochemistry modelling.

First reviewer

[**Comment #2**] *L400: please double check the usage of GCD<0 and GCD>0 in this section. For example, GCD<0 sometimes means mostly-grazed, but means mostly-cut in other cases.*

[**Response to comment #2**] We ensured GCD is used correctly in the text

[**Comment #4**] *Table 1: The NBE can not be obtained from the values presented in the table. It would be helpful to give values for all the components of NBE and ΔSOC. In addition, it is not clear what is the meaning of C flux into soil. Does it include litter and manure?*

[**Response to comment #4**] Table 1 does not present the total/sum annual NBE/NEE but the mean across the ~2000 simulated individual fields (i.e. area average). Also, the estimates for each simulated field show the mean predicted since our model-data fusion framework is probabilistic. For these reasons, one should not expect the area-average annual NBE to be equal to : the area-average annual NEE + Bc + Bg - Manure. We have added the area-mean (+\- SD) predicted value for manure-C to Table 1. "C flux into soil" has now been removed from Table 1. The term was used in our initial submission to describe the C that flows from the litter to the SOC pool but was replaced with Δsoc, which is more informative and easy to understand.

[**Comment #5**] *L434: It is not clear what are included in the "high inputs of C to soils", litter + manure? Does manure from refinement included in this study? If not, it should be mentioned and discussed. Because it will cause an underestimation of C input for grassland.*

[Response to comment #5] Figure 1 shows the C pools and fluxes simulated by the model. Manure from refinement cannot be inferred from earth observation data and relevant data (agricultural stats/census etc) cannot be spatially disaggregated in robust ways. As described in section 2.1.2 of the MS, at every time-step (i.e. week) manure is simulated as being produced by grazing animals in proportion to the simulated grass consumed. The grazing livestock-produced manure is immediately deposited to the soil (i.e. enters the soil litter pool). We discuss the fact that simulated manure production/deposition is based on inferred livestock density in the limitations section (4.5) of the MS.

[Comment #6] L600: I would think it would be helpful to use the meaningful parameters' name (e.g., PNUE, or LCA) rather than the Code of parameters (e.g., P10, and P15) across the manuscript.

[Response to comment #6] Some model parameters have very large names that cannot be abbreviated. We follow a convention when referring to model parameters the use of parameter codes is preferred because it helps us avoid using very long sentences at certain parts of the MS. We understand, however, that having to look at Table A1 is not easy for the reader, this is why we use abbreviations for those 2-3 that are frequently mentioned in the MS.

Second reviewer

[Comment #1] *Net Carbon flux - My main concern is that the manuscript does nothing to convince the reader that the net carbon fluxes (NEE and NBP) can be inferred by assimilating only leaf area index (LAI) data. This outcome seems counter-intuitive. I can accept that assimilating LAI can provide better estimates of GPP and possibly Ra. However, it is far from clear that this will give the correct results for Rh and hence NEE or NBP. I understand Rh in the model to be driven primarily by a temperature response and the amount of soil carbon. If I have understood the manuscript correctly the soil organic carbon is set by using data from the SoilGrids data base and the initial value is not tuned as part of the data assimilation. So, in essence, this analysis is attempting to improve the temperature response of Rh based only on observations of LAI. There is an "EDC" that constrains the rate of change of the SOC pool (EDC #3), which is not unreasonable, but I am not convinced this necessarily helps get the values of the parameters that control Rh correct. The main thrust of the paper is the carbon budget of GB grasslands, so I think it is incumbent on the authors to provide some evaluation, otherwise it is really only model*

output with no indication of how trustworthy it is. I have skimmed the two cited publications by the lead author on this subject and, as far as I can tell, the only comparison with NEE is at a single site (Easter Bush). Also, in that study, the methodology had notable differences from the current one (no EO data, different list of EDCs and so on). Some validation of the net fluxes is required.As a minor point in reference to the above, I also notice that EDC #2 constrains the size of the SOC pool (Table A2) but the initial size of the SOC pool is not one of things tuned, according to Table A1. Presumably this EDC isn't used in assimilation in this study?

[**Response to comment #1**] This is the main comment of the second reviewer and we would like to provide a thorough response.

My main concern is that the manuscript does nothing to convince the reader that the net carbon fluxes (NEE and NBP) can be inferred by assimilating only leaf area index (LAI) data. This outcome seems counter-intuitive. I can accept that assimilating LAI can provide better estimates of GPP and possibly Ra. However, it is far from clear that this will give the correct results for Rh and hence NEE or NBP

We argue that a quantitative study, which focuses on a specific type of ecosystem in order to provide estimates at high resolution (spatial/temporal) and across a large domain, cannot be validated against flux data just as a field or landscape scale study can. This is because there are no measured C flux data to compare predictions with at the pseudo-national scale. We clearly state and highlight in the MS that the credibility of model estimates, at the resolution/scale of our study, depends on (1) model calibration/validation within the domain of application; and on (2) whether or not observations are used to —even partly— validate model predictions. In this respect, we have used two datasets to calibrate and validate the DALEC-Grass model: (i) the most extensive ground-measured, managed grassland-specific dataset of C pools and fluxes available in the UK (Easter Bush site) and (ii) a shorter measurements dataset produced by using different state-of-the-art CO2 measuring instruments (Crichton site). Based on this fact, we argue that we are using a calibrated and validated model whose parameter priors reflect the biogeochemistry of a typical UK managed grassland (dominated by perennial ryegrass, with some clover, that has been a grassland for years/decades). In terms of the use of observations, this study is the first in the UK that uses observational data on a key aspect of grassland C cycling (aboveground biomass volume) for the purposes of quantifying C pools and fluxes. Because of that, we argue that this study produces more credible results than previous, relevant quantitative studies.

If I have understood the manuscript correctly the soil organic carbon is set by using data from the SoilGrids data base and the initial value is not tuned as part of the data assimilation.

We would like to thank the reviewer for pointing out that the initial SOC pool size parameter was missing from Table A1 (now added). The size of the soil organic carbon (SOC) pool of every simulated field is an optimisable model parameter. The prior range of SOC ranges between +/- 10% of the spatially-corresponding SoilGrids value. Using SoilGrids data to set an initial value for each field's SOC pool allows us to control the model's predictive uncertainty. This is important because the size of the SOC pool is the largest source of uncertainty around grassland C cycling estimates.

I have skimmed the two cited publications by the lead author on this subject and, as far as I can tell, the only comparison with NEE is at a single site (Easter Bush). Also, in that study, the methodology had notable differences from the current one (no EO data, different list of EDCs and so on). Some validation of the net fluxes is required.

In the first of these two cited publications we have used 11 years of daily-measured data from two variably-managed grassland sites in Scotland, UK in order to refine the parameter priors and validate the predictions of DALEC-Grass. This data included : soil surface respiration, above and below-ground biomass, ground-measured leaf area index and chamber and eddy-covariance-based NEE measurements. Beyond the Europe/grassland-focused studies already cited in the MS we do not know of other relevant recent studies that provide measured/modelled grassland net C flux estimates, and which could be used for further validation of our results. We would be glad to include more observational studies/ data on UK grassland NEE if this reviewer can point them out. Considering the uncertainty around field-measured C flux data, it is generally believed that permanent grasslands in the UK (and NW Europe in general) are almost C neutral (i.e. NEE = ~0). The results of our study are in agreement with this statement. We would like to highlight again the fact that, in contrast to the majority of model-based studies on grassland C fluxes at large scale, our predictions are not "completely unvalidated" as observational LAI data are assimilated and thus estimated aboveground biomass/C is being validated. Moreover, we are particularly interested in seeing studies discussing/presenting the impact of the 2018 heatwave on grassland NEE. We have cited a number of studies that do this using model predictions and measured point data extrapolations. Unfortunately, we could not find any UK measurements-based studies discussing the impact of the 2018 heatwave. We anticipate

such studies to be published soon. In this regard, data from monitored managed grasslands in the UK show the positive response in NEE (reduction in C sinking) that our simulations are predicting (http://nora.nerc.ac.uk/id/eprint/525106/1/N525106PO.pdf)

[**Comment #2**] *Use of EO LAI - Despite requests from previous reviewers it is still not clear how the two different EO LAI data sets are used, and the justification for using them both is not well made. I read the relevant sections several times and it is still not clear. It appears that CGLS data are used as a driver to quantify "vegetation reduction" and the Sentinel-2 data is used for assimilation. I suggest a complete rewrite of these sections to include a much clearer description of how this is done. In addition the use of the GCLS data is not well justified. The argument seems to be that it is too spatially coarse to represent a field, but have sufficient spatial resolution to detect grazing or cutting. I personally do not understand this. The choice is apparently driven by a better temporal resolution (10-days) than Sentinel-2, but the combined Sentinel-2 instruments actually have a shorter revisit time than this, so the only advantage appears to be that the GCLS data are gap-filled. But (a) won't the gap filling itself reduce the ability of the data to represent grazing/cutting? and (b) why not gap fill the Sentinel-2 LAI data? Can the authors provide a better justification for using both data sets?*

[**Response to comment #2**] Indeed the main reason for using the CGLS data is the 10-day temporal resolution. This is important when considering that there are ~25 cloud free Sentinel2-based images per year; and even fewer images in coastal UK areas. However, the temporal resolution of the satellite-based data is not the only reason for using the CGLS data. Firstly, CGLS data are produced using images retrieved by a different satellite system (originally Proba-V and since more recently Proba-V + Sentinel-3). This means that we constrain the model-estimated LAI (thus aboveground biomass/C) using information from two different systems, which is, per se, more robust than relying on a single system. Secondly, we agree that we could have gap-filled cloud-free Sentinel-2-based LAI data points to produce continuous LAI time series. However, had we gone down that road we would have developed and tested a method very similar to that used by CGLS. This is because, grassland vegetation volume changes within a year in ways that are much less predictable and visible than e.g. crop and timber harvesting. Therefore, the “best” way to interpolate between scarce LAI data points is to use past/historical and/or neighbouring grassland-specific pixel data; which is the method used to produce the CGLS data. Moreover, relying on the freely-available, well-documented, and continuously-maintained and updated CGLS data is better than using any in-house and partly-validated method/data.

We believe that we have extensively revised the MS to explain how and why we are using the CGLS data in this study in our previous revision. This revision included adding Figure 3, which we believe clarifies how CGLS and Sentinel-2 data are used (when/where they are used). We cannot see how re-revising the relevant text can further clarify things.
We would like to add at this point that we see the spatial and temporal resolution of the EO data as critical to the accuracy of the predictions of the model-data fusion algorithm. For this reason we are working on developing a robust, grasslands-tailored and reproducible method to interpolate Sentinel-2 based vegetation indices (LAI in particular). This is still work in progress but our initial testing shows that we will be able to stop using the CGLS data in the near future. This work is pending further validation using a larger ground-truthing dataset (see https://datashare.ed.ac.uk/handle/10283/4086 for more details).
In conclusion, it is not unreasonable for the reader to wonder why we are using the EO data product that we use. However, the main arguments for using the CGLS data in addition to Sentinel-2 data (i.e. temporal resolution, validation and maintenance of the CGLS data) are presented in the MS. We do not think that dedicating more text on EO data choices is necessary; especially when considering (1) that our response to reviewer comments is public, (2) that we intend to further discuss EO data for grasslands in a blog article and (3) that our previous recent publication presents/discusses the pros/cons/effectiveness of using the CGLS data.

[Comment #3] *SHAP values - This is a more minor point than the previous ones, but the Random Forest approach appears to have been used solely for the purpose of obtaining SHAP values. A potential issue with this is that the SHAP values tell us about the sensitivity of the machine learning model to its feature space and not necessarily about the mechanistic model. Consequently, it can result in misleading conclusions if one is trying to infer things about the model that has been emulated, for example when two or more features are correlated. There are other techniques that work directly on models to perform sensitivity analyses and given the model used in this paper is sufficiently computationally efficient to perform MCMC calibration, it would seem an odd choice to emulate it just to back out these sensitivities. Furthermore, the fact that the correlation analysis (Fig 7) provides very similar information tell us it has not added a great deal to the analysis. Given that, I suggest removing the parts about SHAP values.*

[Response to comment #3] Indeed because DALEC-Grass is mechanistic we can explain its behaviour and the logic behind its predictions. In general, the SHAP method is used here in the same way that the correlation analysis is. However, we believe that building a machine

learning (ML) model and using SHAP to present its sensitivities is interesting and has to be included because SHAP offers a quantitative assessment that is clearer compared to that of the correlation analysis. We also believe that this RF + SHAP section of the MS provides a brief test using ML + SHAP as a method for creating an emulator, quantifying its predictive ability (R2) and its sensitivities. We believe that ML + SHAP has the potential to be used to assess and apply an ML emulator in order to extrapolate site-scale model-data fusion-based GHG flux estimates. While this is beyond the scope of our study and do not discuss it we believe that keeping the RF + SHAP section in the MS will be useful small addition to the relevant literature.

[Comment #4]

L170: Do you really mean that you calculate the likelihood from the RMSE? Yes, RMSE is used as the log likelihood.

L190: The sentence here is a bit odd. I don't understand why it's relevant to state that the data are processed from top-of-atmosphere reflectance. Presumably they are corrected to surface reflectances prior to estimating LAI? Indeed, we have reworded this sentence in the revised MS.