Author Comments – Response to Referee 2

Referee comments are marked in black and author responses are marked in blue.

Overall assessment (do not require specific comments by the authors)
The study presents and analyses a rare, high-quality long-term data set of field measurements in an intensively managed ecosystem, led by the colleagues who run the site. The data includes CO2 exchange and management data. The scientific focus is on identifying drivers for interannual variability of gross-primary productivity (GPP) and ecosystem respiration (Reco) in 'normal' and extreme years. The manuscript clarifies nicely that management and especially the aboveground and the complete renewals of the grassland vegetation mark significant short and long-term disturbances putting additional constraints on comparability of periods and years. Weather and management variability and their interactions, which I believe can be expected by the concept of adaptive climate smart management, characterize the challenge, when investigating this ecosystem, compared to, e.g., natural vegetation.

The work creates order in the time series through some useful classification. A period of 20 years is long enough to identify significant environmental trends and weather extremes. The period includes two grassland renewal events and numerous aboveground canopy harvests that define regrowth periods (RPs), still of different length and located in differing seasons. The simplicity and clarity of these decisions to define the perspectives on how to look at the time series and analyse it, is one of the strengths of this work. One sub-period of 5 years, i.e. still ¼ of the investigated period, when part of the study area was subject to a comparative experiment, adds complexity and the way this has been dealt with, raises some questions.

The complexity of the data set requires a complex analysis approach with careful selection of drivers, which I will discuss below. Going for a machine learning approach (XGB) for analysis might be a good choice, as it puts the least constraints on the results compared to alternatives, such as alternative empirical mechanistic modelling approaches. But the different nature of the results, especially the results from the SHAP analysis, are yet a bit difficult to understand and I suggest more explanation and guidance for the reader.

In general, I see this work as a model for such long-term empirical studies in managed ecosystems, for sure a strong scientific contribution to quantitative Biogeosciences.

Dear Dr. Ibrom,
Thank you for your positive and constructive comments! We appreciate the opportunity to improve the manuscript based on your thorough review and feedback. We have addressed your comments and suggestions in the responses below.

Some general critical comments are as follows (please comment and take action, where applicable):

G1: Devising extreme months with a clear and simple Z-score approach on soil physical and atmospheric drivers for drought makes good sense. The classifying variables are well chosen, because a capacitive variable (SWC) and an atmospheric state variable (VPD) are combined to represent the accumulated (SWC deficit) and actual stress (VPD). For the latter, potential evaporation might possibly be an even stronger variable.

Thank you for the comment. There are different approaches to how to represent stress. Based on existing work (Liu et al., 2025; Novick et al., 2024; Shekhar et al., 2023), we think that VPD and SWC are more representative to the actual stress of the ecosystem rather than potential evaporation. Furthermore, from a reader's perspective, VPD is more straightforward and easier to understand as a stress factor compared to potential evaporation.

G2: Machine learning methods are relatively novel, and I believe the interpretation of the results is still a challenge. From my own experience with this text, there is a large risk for a reader, not yet familiar with the SHAP analysis, of miss-interpreting the results from the SHAP analysis, e.g., a "negative effect" as "negative relationship" between a driver (D) and a response variable (R). A negative SHAP value ("negative effect") shows rather only that the contribution from D has made R lower than the reference. This is irrespective of the sign of a relationship: a positive relationship ($dR/dD > 0$) makes R small at low values of D a negative relationship ($dR/dD < 0$) makes R small at high values of D. To avoid misinterpretation by the readers, please consider explaining this possible trap for understanding the text correctly.

We appreciate this constructive input. The explanation of SHAP values needs to consider the value of the feature/driver for each single observation as well, and indeed 'effect' is the appropriate way to describe it. The sign of the SHAP value also depends on the response variable, for example, if NEE is the response variable, negative SHAP value would mean that this driver increases the CO_2 uptake. We are aware that SHAP analysis is still relatively new to the community and readers. To avoid any misinterpretation of the SHAP values, we had already added detailed explanations in the method section and the respective figure captions in the original manuscript. Moreover, we never talked about a negative 'relationship' in the text. Since we see more and more SHAP analyses in the scientific literature (Krebs et al., 2025; Li et al., 2025), we think that our explanations should suffice. In any case, we will carefully check the wording throughout the text to make sure 'effect' is used where appropriate.

G3: However, I claim, for scientific understanding, relationships are more relevant than just effects. By examining the effects further, e.g. by looking into the relationship between the effect and the magnitude of D, you might be able to say something more about the nature of the relationship (see, e.g., D29) and possible interactions between drivers.

Thank you for the comment. We fully agree that 'effects' and 'relationships' are different, and it is important to explain both. In machine learning models, looking at only bilateral relationships between one driver and the response variable cannot be achieved, as in conventional linear additive statistics, the number of drivers and their interactions are strongly limited. PCAs can give a hint on many drivers and their interactions but axes loading does not go so as far as modern, explanatory machine learning can go. For example, while machine learning only looks at one driver at time, based on game theory, we can disentangle many drives. Moreover, since SHAP values for different drivers have the same unit as the response variable (here, flux units), we can still examine the 'relationship' between one driver and the response variable, namely using SHAP dependence plots (Fig. R2-1(a-c) for top three drivers for GPP). For example, with more light (PPFD) or higher soil temperature (TS), SHAP values of these drivers on GPP were generally increasing, suggesting a positive (for soil temperature then saturating) relationship between drivers and the response variables.

We will add partial dependence plots (Fig. R2-1) in the appendix as new Figure A3. Moreover, we will also add a description of this new figure in Section 3.3.

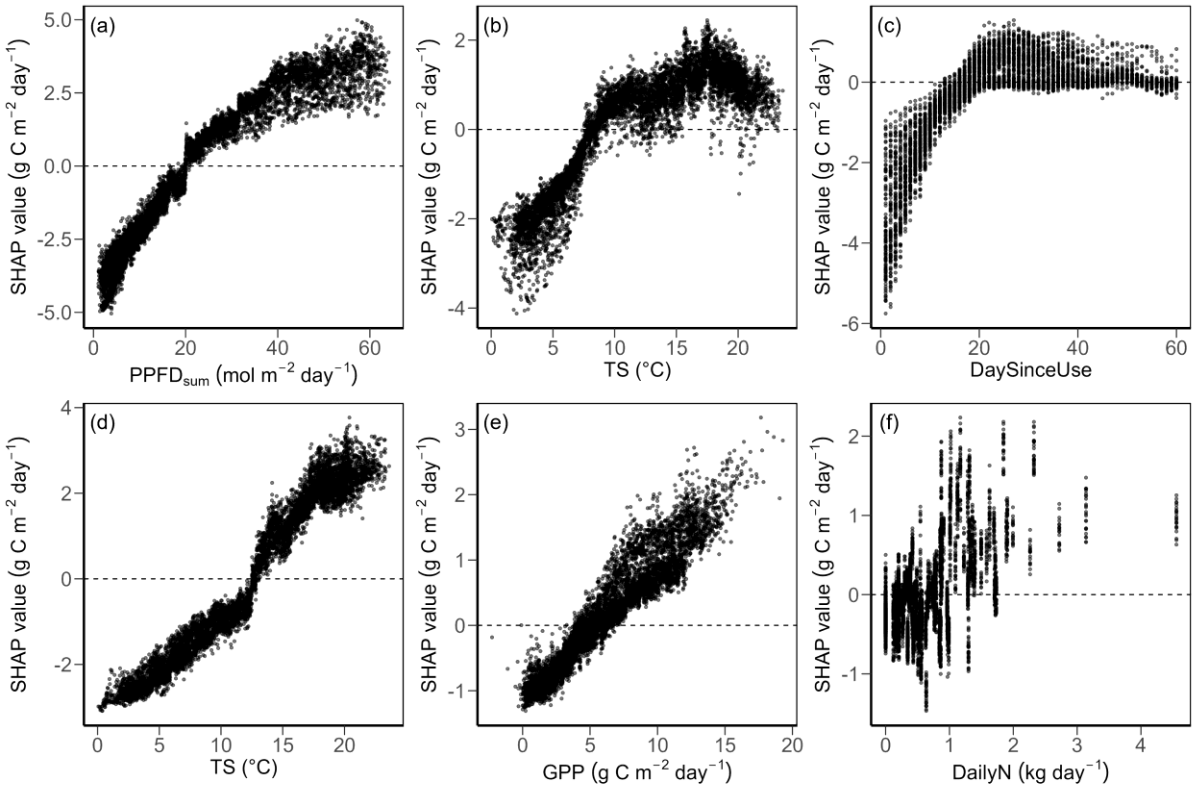

Figure R2-1. New Figure A3: SHAP dependence plots for top three drivers of GPP (upper row) and Reco (lower row)

In addition, we will also modify the following sentences in the method section:

(1) At current line 182 old sentence: "… (… Molnar, 2023). If a feature (i.e., a driver variable) has a positive SHAP value, this feature increases the local prediction relative to the overall mean prediction, and vice versa. Higher absolute…" will be modified to: "… (…Molnar, 2023). The SHAP value of a feature (i.e., a driver variable) represents the effect of that driver, at its specific value, on each individual prediction. A positive SHAP value indicates that this driver increases the local prediction to the overall mean prediction, and vice versa for negative SHAP values. Higher absolute …"

(2) At current line 187 old sentence: "… (… Qiu et al., 2022). In this study, we performed two SHAP analyses (1 and 2) with different foci." will be changed to: "… (… Qiu et al., 2022). The relationship between each driver and its SHAP values can be further explored using SHAP dependence plots. In this study, we performed two SHAP analyses (1 and 2) with different foci."

G4: Although the text introduces it correctly, I first falsely assumed that SHAP analysis 1 was based on RP, rather than on days, while SHAP 2 was on a daily basis. Just to confirm, is it correct that in both cases the SHAP analysis was performed on a daily basis but the daily results from SHAP 1 were then presented as RP averages in Fig. 4? For the next points, I presume that this is correctly understood.

Yes, both SHAP analyses were done on a daily basis. In fact, both SHAP analyses for GPP were based on the same XGBoost model, which was based on daily data. For SHAP analysis 1, we indeed averaged all SHAP values based on regrowth period, and for SHAP analysis 2, we showed daily values with a focus on extreme summer months. We will clarify this in the method section and make it clearer in the respective figure captions. At current line 198, we will add

"… extreme events. Both SHAP analysis 1 and SHAP analysis 2 were based on the same XGBoost model for GPP, which was performed based on daily data."

G5: I was surprised by the apparent lack of coherence between SHAP 1 and SHAP 2 analysis. Doesn't this show how sensitive the results are to the choice of the baseline (and of course the variables)? This should be mentioned when interpreting the results from such analysis.

Thank you for the comment. As mentioned above, these two analyses were based on the same set of variables and one XGBoost model. You are correct that any SHAP analysis strongly depends on the baseline used. For our two SHAP analyses 1 and 2, different baseline datasets were used, to provide insight into two different research questions, namely what are the main drivers under different environmental conditions, as explicitly mentioned in our objective 3.

Here, we provide an analogy: When you want to determine the factors that influence people's height, the results will naturally be different if you compare small infants with all male adults, or if you compare infants with tall basketball players. Similarly, the change of results in our analyses is also rather intuitive to understand. If you compare drivers of only summer GPP vs. drivers of two decades of GPP data (including winters), the variation in daily GPP is more dependent on the light/temperature difference. However, if you compare extreme summer GPP to non-extreme summer GPP with similar light/temperature conditions, you can actually see the effect of other variables (e.g., SWC, VPD) more clearly. So, yes, we fully agree that this nicely shows the sensitivity of such analyses to the chosen baseline. Therefore, we already in the earlier version of the manuscript described the baseline in Section 2.4.2. However, unfortunately, studies using machine learning models and SHAP explainers to understand fluxes have rarely mentioned these baseline datasets.

In order to raise attention to this aspect, we will add two new sentences in the beginning of Section 4.3 (line 430): "Using non-extreme summer months as the background dataset, SHAP analysis 2 focused on drivers of GPP only during extreme summer months, which minimized the confounding effects of strong seasonality in temperature and light conditions. This further improves the interpretability of the results, by isolating and focusing on the influence of extreme-related factors such as SWC and VPD. Across all extreme periods, …"

G6: Possible a priori relationship between response and driver variables: Defining RPs that can be compared across seasons and years seems a very appropriate approach, however, the RPs are of different lengths (see e.g. Fig. 6, where one RP spans over three months in 2018, while others span only over one month in 2022 and 2023). I wonder whether common relationships between driver variables and the RP length cause some artificial interdependence (circular logic) among some driver variables and, and even more worrisome, among some drivers and the response variable GPP.

I mean especially the two drivers DaySinceUse and DailyN, which represent management that both depend RP length. From a farmers perspective, i.e. planning the RP length to reach a certain goal, the RP length is inversely related to productivity and thus response variable daily GPP level.

The particular relationship between DaySinceUse and RP length is that only in cases when RP length is high, DaySinceUse can reach high values. In these cases, high DaySinceUse values coincide with low daily GPP and high RP length.

The particular relationship between DailyN and RP length is a mathematical consequence of the definition of DailyN , i.e. the DailyN value per unit fertilized N will inversely decrease with increasing RP length.

Please comment on the possible effects on the results of the analyses from these interdependencies or consider amelioration by different driver definitions.

Thank you for the constructive comments. The main goal of our XGBoost models was to accurately model GPP and Reco using the available main drivers. Since this grassland is intensively managed, DaySinceUse for mowing/grazing and DailyN for fertilization are two key variables representing management. To avoid any misunderstanding: the length of regrowth periods was never included in the model. In addition, the length of regrowth periods is not necessarily inversely relative to productivity, we rather deal with a cumulative effect. As we can see in the partial dependence plot (Fig. R2-1(c) for GPP), SHAP values for DaySinceUse are normally very negative at the beginning of regrowth periods (DaySinceUse < around 10 days) and then increase. Therefore, higher DaySinceUse values do not mean low daily GPP, but rather the opposite, high daily GPP values. As mentioned above, we will add this partial dependence plot for DaySinceUse together with other drivers in the appendix as new Fig. A3.

For more details about the calculation of DailyN, please see our response below (G9).

Regarding the length of regrowth period (also for comment D12): we have shown the relationship between the length of regrowth period and GPP/Reco below (Fig. R2-2). Overall, with longer regrowth periods, GPP or Reco during the regrowth period gets lower, which is to be expected when comparing spring-summer seasons versus autumn-winter seasons. Within different seasons, longer regrowth period does not always correspond to lower GPP or Reco. We will add more details on this in Section 3.2.

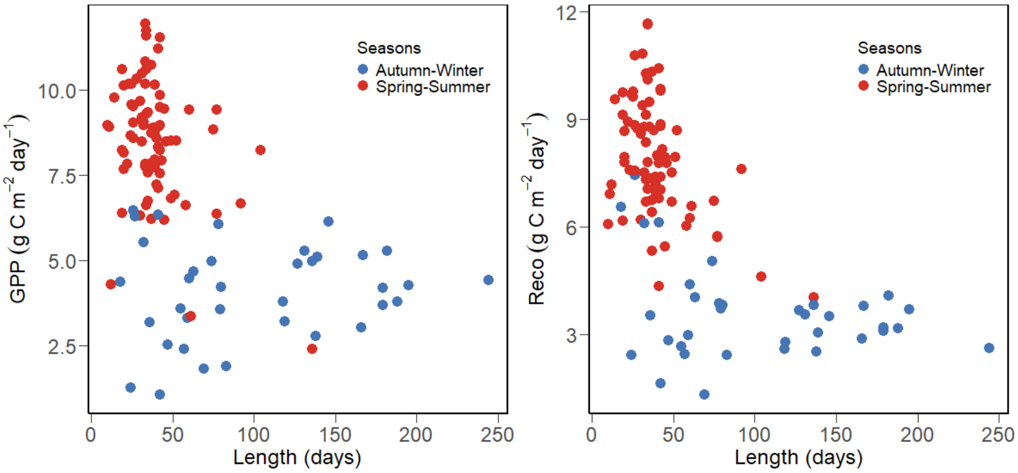

Figure R2-2. Length of regrowth periods vs GPP (left) and Reco (right).

Taken together with comment D19, we will also add statistics about the length at current line 251 as "The average length of the regrowth periods was 58 (± 47) days, with shorter regrowth periods in spring-summer seasons (39 ± 20 days) and longer in autumn-winter seasons (100 ± 60 days). A slight increasing trend ($p = 0.01$) was found in the length of regrowth periods over the 20 years. During spring-summer seasons, the length significantly increased ($p < 0.01$), while no significant trend was found in autumn-winter seasons ($p = 0.87$)."

G7: Critical reflection of using time as a driver: The variable DaySinceUse increases linearly with time until reaching RP length, i.e. it simply represents time as such or canopy age. The analysed relationship is thus the timeline for the development of GPP during an RP.

I wonder, what is the rationale behind using time as a driver? Short development time coincides with high productivity, but the drivers for production are not time, but rather the growth conditions. It will be obvious that the XGB-SHAP analysis will 'turn' time into a driver owing to the a priori decision of the user defining time as a driver.
Please explain the usefulness of time as a driver.

The referee's comment is based on a strong assumption: short regrowth period = high GPP, since a short regrowth period means a low number for the driver DaySinceUse. However, this assumption can be rejected since GPP during the regrowth period does not only relate to the length of the regrowth period, but does also relate to other meteorological and management variables, as seen from our model output (Figure A3). As shown in the partial dependence plot above, short development time (low DaySinceUse) does not correspond to high productivity. The length of the regrowth period is dynamic and can also change with other drivers, for example, farmers could adjust the mowing time during extreme events to "rescue" a regrowth before it dries out during a heatwave or before it gets soaked during a prolonged rainy period.

Nevertheless, it is indeed important in what form one inputs the drivers into the machine learning model and there are different ways of doing that (e.g. as continuous, categorical, binary variable, etc.). The aim behind the "days since" approach is to actually represent the management in a way the model understands. It is **not** exactly equal to time, which would be always increasing, while the DaySinceUse variable goes to zero each time there is a management event. Using "days since" as a driver actually gave reliable results in previous studies (Maier et al., 2022: time since management; Feigenwinter et al., 2023b: days since defoliation), but has also been used in studies by other groups, who used different machine learning models for gapfilling N_2O fluxes (Goodrich et al., 2021: days since grazing).

G8: The above problem raises some fundamental questions about the general meaning of driver and response variables. The term adaptive management suggests that management, e.g. defining the time for harvest, i.e. the RP length, can both be a driver and response – a clear distinction may even be impossible.
What do these uncertainties mean for the interpretation of the analysis?
Is using the term "driver" together with a statistical analysis that is not able to detect cause-effect relationships (just effects) at all appropriate?

Thank you for the comment. As stated above, in our XGBoost models, length of regrowth periods was never included as a driver. The term 'adaptive' clearly indicates that management can both be a driver and a response, which is similar like the chicken and egg problem. Indeed a clear distinction is not possible. However, with no significant trends in GPP during regrowth periods observed, our evidence suggested that adapted management at this site has been able to maintain productivity even with on-going climate change as seen in more extreme events in recent years.

Another example of an environmental factor being both, driver and response, is light. Nobody would argue with the notion that light is driving plant photosynthesis (interacting with many other drivers, from water to nutrients), as is management. With more light (and days since management), photosynthesis increases, the plants grow, and the stand becomes taller and denser. As a consequence, or call it a response, stand structure changes, light attenuation increases and light within the stand is decreasing, limiting photosynthesis compared to before. Thus, light is a driver and a response, as is management.

With machine learning approaches (here XGBoost + SHAP), we can identify top drivers of fluxes and assess their temporal development. These drivers were selected based on their known functional relevance for ecosystem processes. Even with traditional statistical analysis (e.g. linear relationships), a definite causal relationship cannot be detected. Therefore, we argue that the term ‘driver’ is appropriate in the manuscript.

G9: General reflection of the usefulness of DailyN as a driver for daily GPP: If I am right, DailyN is the only driver variable that includes averaging over the RP length. For a SHAP analysis that is based on daily values, the definition of DailyN is counterintuitive and the naming does not reflect what is actually going on (the fertilization is not daily). The authors will agree that N-availability might be the more relevant factor for GPP. N-availability will be larger right after the fertilization event (after reaching the rooting zone) and will decrease maybe not with time but with growth (and leaching, emissions etc.) over RP. Do you see a possibility to define an alternative daily variable “available N” (AN) that parameterizes the decrease from the amount of fertilized N over the length of the RP or even scaling it negatively with GPP (as proxy for growth)?

Thank you for the comment. How to best represent N fertilization and “available N” in any model is indeed an ongoing discussion, also in our group. We agree that the current value can be discussed, but currently we consider this as our best approach. During the majority of regrowth periods, there was only one fertilization event, and normally it happened relatively soon after the mowing. If there was a second fertilization event, we summed up the amount of fertilized nitrogen for that period and calculated DailyN based on the total amount. This way, we consider the supplied N as fully available for growth. To be extremely accurate, one would need to model the available N as suggested, e.g., with an exponential decline, considering many factors like soil properties, precipitation events and soil moisture dynamics, but also microbial activity, potentially delayed responses by soil moisture and soil temperature. Achieving an accurate estimate of this on a daily scale would either need another exhaustive study (with unknown methodology if data were to be collected daily in a real-world setting, not in a lysimeter) or would be based on many untested assumptions. Thus, we cannot solve this issue by changing the regrowth period average to an exponential decay or other dynamic functions without introducing large untested (maybe even untestable) bias. Furthermore, we did not find convincing alternative solutions based on the existing studies in the literature, which are currently still limited. Therefore, we consider our current way of representing real management as best as we can do, but would of course welcome more systematic approaches in future work.

G10: In general, please explain the term adaptive climate smart decisions / management. This is important for two reasons, i) it is used in the interpretation and the conclusions and ii) the definition might help to better understand the nature of the RP, i.e. as depending on certain a priory rules and expectations/ observations on productivity.

Thank you for the suggestion. As also suggested by Referee 1, We will introduce more explanation on this term in current line 69.

G11: The study concludes (L474-L476) adaptive climate smart management as a factor for homogeneous production despite weather trends (likely climate change induced). Is this ‘just’ a plausible speculation or did your study show this? Maybe I overlooked it, I did not find clear evidence in the presented results for this statement that comes up in the discussion, the conclusions and is highlighted in the abstract. A quantitative analysis would examine the interaction between adaptive climate smart management and production, probably in contrast to a plausible BAU scenario.

I wonder whether XGB generated predictions could be used for scenario calculations or whether mechanistic models would be needed to substantiate such speculation. I deem this worth to be clarified in the discussion. I do not suggest such study to be included here. The study is rich enough, but its limitations need careful consideration, i.e. what can be concluded from its results.

Thank you for the constructive comment. With more frequent extreme events in recent years that were observed in our time series, we expected a decreasing trend in GPP during regrowth periods and ultimately lower CO_2 sink strength, unless management was already adapted to these new conditions. The non-significant trend detected in our GPP data clearly showed that the existing management practices were able to maintain productivity, thus suggesting resilience to extreme events through 'climate-smart' management, following the definitions used by science as well as global organizations such as IPCC, FAO and the World Bank. We agree that more "climate-smart" management practices aiming to improve resilience and sustainability in agroecosystems should be tested in the field with experiments or with well calibrated process models (beyond machine learning models), e.g., testing the effect of timing and intensity of certain management practices on productivity and GHG emissions. We have ongoing work in the group using the process-based model MONICA (Nendel et al., 2011) on this exact topic (Kamali et al., submitted).

With all considerations in mind, as also suggested by Referee 1, we will put our argumentation in context and explain this aspect better throughout the manuscript. For the conclusion section, we will modify lines 474-476 as "Moreover, based on two decades of measurements, our evidence suggests that the grassland farmer succeeded in managing the site with relatively stable GPP during regrowth periods, based on climate-smart adaptive management."

General recommendation: I deem the overall quality of the manuscript to be very high and inspiring and maybe its clarity is the reason why it provokes some critical thoughts. I do not claim that this review from reading the manuscript a couple of times, can be assumed to be exhaustive and accurate, as I lack particular knowledge that the Authors probably have. I expect though clarifying responses and look forward to the answers by the authors. It depends very much on these answers, whether minor or major revisions will be necessary.

Detailed comments (please comment and take action, where applicable)

D1: The title includes the word regrowth, which implies biomass production while it is used here as re-establishment or recovery of GPP and Reco. I suggest using the more neutral "grassland CO_2 exchange:" instead of "grassland CO_2 fluxes and regrowth:"

Thank you for the comment. We will change the title into "Drivers of long-term grassland CO_2 fluxes: effects of management and meteorological conditions during regrowth periods".

D2: L 16: consider starting a new paragraph before "CO_2".

We can do this if the journal allows two separate paragraphs in the abstract. We can modify the sentence as "Our results showed pronounced … in CO_2 fluxes, driven by both …".

D3: L 20 and L 25: make a decision on whether the study showed or suggested a relationship between CO_2 exchange and “adapted, climate smart decision making” (see also G11).

We will make sure we stay consistent with ‘suggested’.

D4: L53-54: Define “atmospheric dryness” – explain, why does it not include “reduced precipitation”.

We define “atmospheric dryness” as high VPD, which could be independent from reduced precipitation. We will add this definition in brackets.

D5: L56-L57: While promote productivity makes sense “promote CO_2 fluxes in general” does not– consider rewording.

The studies cited found both increased GPP and Reco. We will modify the sentence to “… that warming can promote grassland productivity and increase CO_2 fluxes (both GPP and Reco) …”

D6: L64: is the word ‘buffer’ appropriate here? ‘mitigate impact of … on …’?

‘Buffer’ was used in the original study. We will change the sentence into “… may mitigate the impact of temperature anomalies on CO_2 fluxes” to improve clarity.

D7: L73 – L77: Please clarify, do you mean the nonlinear, interactive, and highly dynamic “nature of drivers” or rather the nonlinear, interactive, and highly dynamic “nature of responses”? Please consider the difference between “dynamic” and, e.g. “variable”? What would fit better here?

Here, we meant for both drivers themselves and their effects. We will modify the sentence as “Linear models frequently fail to capture the nonlinear, interactive, and highly variable effects of drivers that influence CO_2…”

D8: L84: (objective 1) Do you deem the investigation of something as a scientific objective?

We will use the word “identify”.

D9: L96 – what do you mean with “destroyed”? was just ploughed, or extracted and removed?

The sward was killed by either direct ploughing everything under (2012) or herbicide application (2021). We will modify as “the existing sward was terminated and ploughed…”

D10: L103-L107: If I am right, this is an important decision on how to use and interpret ¼ of the time series. I wonder how this decision has influenced the results. I would like you to discuss the alternative(s), e.g. separating fluxes between parcels and using only the comparable one, parcel B, for this study.

In general, the management regimes (in terms of mowing and grazing dates) were very similar between these two parcels, also during the N_2O mitigation experiment. Differences in management happened in earlier years (2008-2010), and most of the cases were that one parcel was grazed while one was mown, but both around the same time (see management info in Table B2 in Feigenwinter et al., 2023a). In addition, the parcel areas changed twice in the past 20 years, and in earlier years parcel B dominated the footprint area (Figure B3 in Feigenwinter et al., 2023a). Since different wind directions dominate during days and nights, separating fluxes will significantly decrease our data coverage and create bias since we use daily data in the analysis. In our previous study at the same site regarding long-term carbon budgets (Feigenwinter et al., 2023a), we did not separate the parcels either. We have taken the different fertilization regimes during the N_2O mitigation experiment into account when calculating DailyN (lines 115-116) so that the whole field received overall less N fertilization. For all these reasons, we will continue using regrowth periods based on parcel B.

D11: L109-112: Please specify "This" in "This allows" – the logic between the two sentences is not clear (to me). Did you merge shorter periods into one RP? If this was the case, did you check, whether the results in these RP differed from the others?

In earlier years (2008-2010), some parts of the field were first grazed, and some days (e.g., five days) later the other parts were mown. If we account these short periods as regrowth periods, we would have five more regrowth periods than the current 115 regrowth periods. Since all our analyses are based on the entire field, separating the field is not possible (see our answer to D10). Even if we would define these five days between grazing and mowing events as one regrowth period although it only occurred at parts of the field, we could not reliably calculate GPP of the entire field, since we capture the combined signal from the entire footprint area (i.e. areas with short vegetation and areas with high vegetation). Thus, to make this clearer, we will modify the sentence as "If management activities took place only a few days apart within the entire footprint, we defined the later activity as the end of the regrowth period."

D12: L112: Be aware of the impact that averaging over differently long RP has on the meaning of the variable. Is there a negative relationship between the length of the RP and the average GPP or Reco value? Or has the definition of a minimum RP length of 10 days alleviated this relationship. From my distant perspective, if such relationship existed, it would explain relationships between effects from drivers that are (Day[s]SinceUse, dailyN) or are not (PAR, TA, TS, VPD) related to the length of the RP (see also G6).

Thank you for the comment. Please see our response to G6 above.

D13: L 109: Please add a short explanation on how was the position and length of the main "growing season" defined, and how this RP classification affected the analysis. Consider replacing 'middle date' by 'center'

This classification is based on local management, also typical for other central European grasslands, as seen in Figure A2 (which will become Fig. 3a). Usually the first mowing event of the year happens in April and the last mowing happens in mid/end October Thus, this is the "main growing season", mainly used for visualization purposes in Fig. 4, and to test the trend in GPP and Reco in different seasons. We will modify the sentence at line 119 as "If the center of one regrowth period was between April and September (which is the main growing season for agricultural grasslands in central Europe), the period was classified as …". We will also change "middle date" to "center" and change the figure caption accordingly.

D14: L125: consider "atmospheric" instead of "air"

We will change this accordingly.

D15: L128: specify, probably, 'volumetric' SWC

Thank you for the reminder. We will change to “volumetric SWC”.

D16: L130: replace “community guidelines” with scientific references

We will specify this sentence as “Widely used community guidelines (Aubinet et al., 2012; Pastorello et al., 2020, Sabbatini et al., 2018), including …”.

D17: L138: explain why using these percentiles instead of the usual ones 5 %, 50 % , and 95 %?

The 16th and 84th percentiles are statistically equivalent to ± 1 standard deviation (σ) for data that follow a normal distribution. This approach is consistent with the standardized FLUXNET processing pipeline (Pastorello et al., 2020), where these percentiles are used to calculate overall uncertainty.

D18: L147-L149: Please explain this averaging choice considering the vertical distributions of roots and SOM.

Thank you for the comment. We do not know the exact rooting depths of this grassland nor the exact depths of plant water uptake. Rooting and water uptake depths could also differ due to heterogeneity within the field. Furthermore, it is uncertain which soil depths are responsible for ecosystem respiration. With the EC measurements, flux data represent the integrated condition over the entire field. Therefore, we averaged the soil variables across depths to represent the integrated soil conditions over the entire soil profile. We will adjust this sentence as “For soil temperature… averages across all depths were calculate and used in the final analysis to represent the overall soil conditions over the entire profile.”

D19: L154 – L155: If only the GPP Reco averages have been tested, please add information, on whether the lengths of the RP showed a trend. Please specify: in the trend analysis did you exclude extreme months?

We did not exclude extreme months in our trend analysis. We have tested the length of the regrowth periods. Overall, the length of regrowth periods had a slight increasing trend ($p = 0.01$). During spring-summer seasons, the length significantly increased ($p < 0.01$), while no significant trend was found in autumn-winter seasons ($p = 0.87$). We will include this information in Section 3.2. Please also see our response to G6.

D20: L165: Please clarify, what do you mean with “GPP regrowth rate” do you mean dGPP/dt or, in accordance what was explained above, “average GPP rates over RPs”? Then please explain the rationale for choice of a 2nd order polynomial as regression model for the analysis of light response function of GPP.

We meant average GPP during the regrowth periods. For each regrowth period, cumulative GPP and Reco were first calculated and then averaged based on the length of the regrowth period. Following the suggestions from Referee 1, we will adjust the term of ‘GPP/Reco regrowth rate’ to “GPP/Reco during regrowth periods” throughout the text. We chose a second order polynomial to model the saturating relationship between light and GPP.

D21: L168-L172: Specify when adding GPP as driver variable for Reco, why did you still include PPFD and VPD as drivers? Is it correct to say that that all other effects are then residual effects, i.e. effects of a variable on top of its effect on GPP?

We added GPP as a driver for Reco model, since GPP provides the carbohydrates for carbon allocation belowground, for root systems as well as microbes, thus for auto – and heterotrophic respiration. We kept the other drivers in the Reco driver analysis to keep the drivers of the two models (GPP, Reco) as similar as possible.

No, it is not correct to say that these drivers have residual effects after accounting for GPP effects. In the XGBoost model and SHAP analyses, the effects of certain drivers are not comparable to, for example, linear regressions. Instead, the effects are additive based on the nature of SHAP values (Lundberg et al., 2020), see our answer above to G2. Keeping PPFD and VPD in the model will not take away any statistical power of other drivers. We will modify the sentence as "… GPP was added as an additional feature to better represent the carbon supply and allocation for autotrophic and heterotrophic respiration.".

D22: L197 – The sentence does not make sense in the way that in both set-ups the same months (JJA) were selected without further distinction. Is in the sentence describing the second set after "for only the peak growing season" a reference to extreme years missing? Or did I misunderstand anything here?

Sorry for the confusion, and indeed there was some information missing. We will clarify the sentence as "Here, we used only the months during the peak growing season (June, July, August) as the background dataset when no extreme weather condition occurred. We then calculated a second set of SHAP values for only the peak growing season in 2018, 2019, 2022, 2023, as these periods included the majority of recent extreme months.". Regarding the background dataset, please see our answer above to G4.

D23: L181 - L184: Clarify here that an effect is different from a relationship (see G2)

See our response to G2 and G3.

D24: Section 3.1: Would the length of the vegetation period (VP) and the meteorological conditions in the VP - both more relevant for bioclimatological characterization- give a different picture? Consider moving Table 1 in the appendix and focusing only on significant trends here.

Thank you for the comment. In permanent grasslands, the vegetation period, as seen in the regrowth periods, can be year around. Since regrowth periods can differ from year to year and are not aligned to calendar months, we provide the bioclimatological characterization on a monthly basis, which also allows comparison to other sites in Europe. We will move Table 1 to the appendix as the new Table A2.

D25: Figure 2: Nice and clear presentation and reasoning.

Thank you for the positive comment.

D26: Section 3.2: can you provide information about interannual and seasonal variation of the length of RP (e.g. horizontal range bars or replace circle by rounded rectangles in Figure 3). Alternatively you might consider presenting the sums of GPP and Reco over RPs as alternative to the average in the same manner in a second figure. It might show, if I am right, the effects of adaptive management.

We will add statistics on the length of regrowth period in the main text. Please also see our answers above to G6 and D12. Since the new Figure 3 (Fig. R2-3) will be expanded following suggestions from Referee 1, we will not add further information in this figure. The magnitude of GPP and Reco is already depicted in Figure 3, as the fluxes are given by the color code.

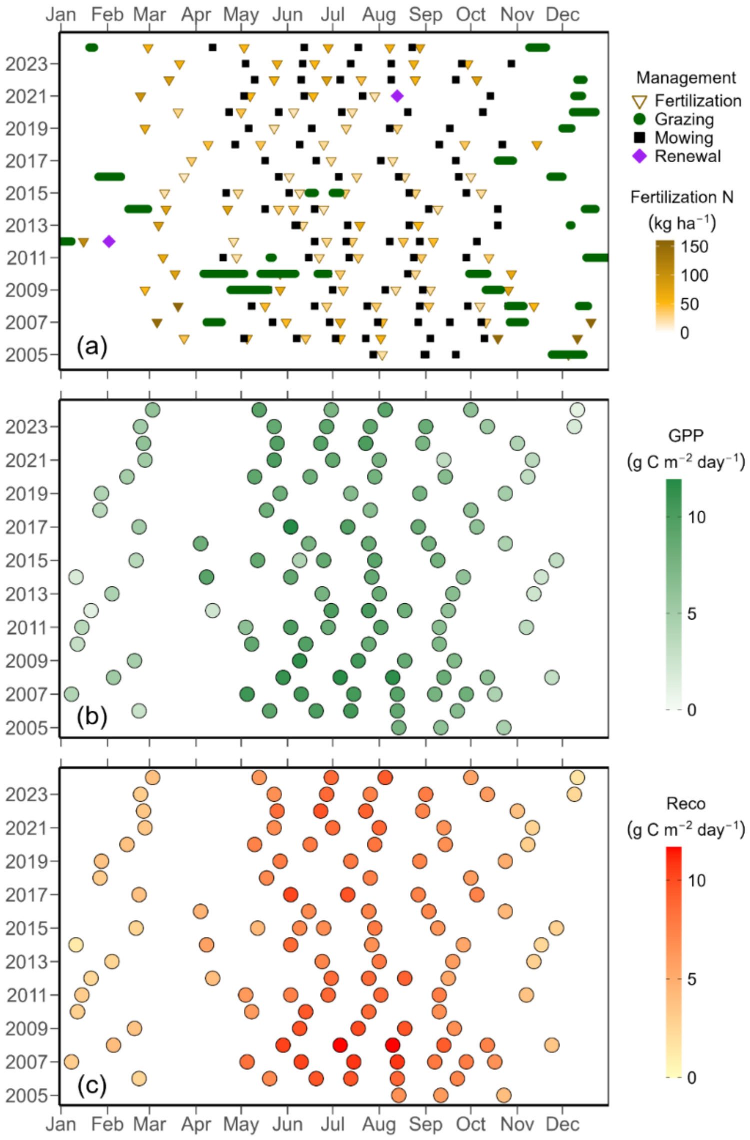

Figure R2-3. New Figure 3 for the revised manuscript

D27: Section 3.3: Fig. 4: I find the choice of the transparent colors confusing because they do not match with the colors of the legend very well. Consider an alternative to show the different season categories if at all necessary. Mark in the figure when the grassland renewal has taken place and which were the months with extreme weather. Then explain that the x-axis variable is not time but number of RP from start of the investigated period.

Thanks for your suggestion, but we thought a lot about the color scheme and how best to convey such complex information as in Figure 4. Therefore, we use the same color for the same driver but more transparent for the fall/winter season (which is shorter than the spring/summer season). For this SHAP analysis 1, we used two decades of data and present the overall picture, thus seasonality is important to depict. We already stated in the figure caption that each stacked bar represents different regrowth periods. To allow comparison with other figures for the same year, we cannot change the x-axis. The number of regrowth periods within each year can be clearly seen. In addition, stats on the regrowth periods will be given in a revised manuscript (see answer to D26). Since this figure is already busy, we cannot put more info in there, such as renewals or extreme weather. For the latter, we provide Figures. 5-6. We will try to improve the legend for different seasons in Fig.4.

D28: Section 3.4: Make sure to mention that the difference between the canopy photosynthesis saturation level and the maximum of a polynomial GPP= f(PAR) have different meanings (see also D20).

We did already specify in line 298 "(GPPmax, i.e. the maximum of the curves in Figure 5)". We now avoid the term saturation to make it more clear and changed the sentence to: "During all periods, daily GPP increased with mean daily PPFD, while maximum GPP (GPPmax, i.e., the maximum of the curves in Figure 5) was reached at a PPFD of about 30 mol m^{-2} day^{-1} (before the renewal years) or at about 45 mol m^{-2} day^{-1} (after the renewal years, all other normal years).

D29: L277-279: 'negative effects of SWC on Reco' is a very good example, how effects may sound counterintuitive. I think it would be good mentioning, recalling that SWC is low during droughts, low values of SWC have caused the low predicted Reco as indicated by the negative SHAP effect values (see also G2).

See our response to G2 and G3. We will modify this sentence into "During extreme summers when low SWC occurred, negative effects of SWC on Reco were more obvious in 2022 and 2023 compared to earlier years…"

D30: Section 3.5: the heading focuses on drivers, but, I believe, the main relevant results are the effects on GPP and Reco.

Second 3.5 is about drivers of GPP in extreme summers, not about Reco.

References

Aubinet, M., Vesala, T., and Papale, D. (Eds.): Eddy Covariance: A Practical Guide to Measurement and Data Analysis, Springer Netherlands, https://doi.org/10.1007/978-94-007-2351-1, 2012.

Feigenwinter, I., Hörtnagl, L., Zeeman, M. J., Eugster, W., Fuchs, K., Merbold, L., and Buchmann, N.: Large inter-annual variation in carbon sink strength of a permanent grassland over 16 years: Impacts of management practices and climate, Agric. For. Meteorol., 340, 109613, https://doi.org/10.1016/j.agrformet.2023.109613, 2023a.

Feigenwinter, I., Hörtnagl, L., and Buchmann, N.: N_2O and CH_4 fluxes from intensively managed grassland: The importance of biological and environmental drivers vs. management, Sci. Total Environ., 903, 166389, https://doi.org/10.1016/j.scitotenv.2023.166389, 2023b.

Goodrich, J. P., Wall, A. M., Campbell, D. I., Fletcher, D., Wecking, A. R., and Schipper, L. A.: Improved gap filling approach and uncertainty estimation for eddy covariance N_2O fluxes, Agric. For. Meteorol., 297, 108280, https://doi.org/10.1016/j.agrformet.2020.108280, 2021.

Kamali, B., Buchmann, N., Feigenwinter, I., Wang, Y., Ewert, F., Gaiser, T.: Navigating the trade-off among biomass production and GHG emissions for smart management of grasslands, Submitted.

Krebs, L., Hörtnagl, L., Scapucci, L., Gharun, M., Feigenwinter, I., and Buchmann, N.: Net ecosystem CO_2 exchange of a subalpine spruce forest in Switzerland over 26 Years:

Effects of phenology and contributions of abiotic drivers at daily time scales, Global Change Biol., 31, e70371, https://doi.org/10.1111/gcb.70371, 2025.

Li, X., Ciais, P., Fensholt, R., Chave, J., Sitch, S., Canadell, J. G., Brandt, M., Fan, L., Xiao, X., Tao, S., Wang, H., Albergel, C., Yang, H., Frappart, F., Wang, M., Bastos, A., Maisongrande, P., Qin, Y., Xing, Z., Cui, T., Yu, L., He, L., Zheng, Y., Liu, X., Liu, Y., De Truchis, A., and Wigneron, J.-P.: Large live biomass carbon losses from droughts in the northern temperate ecosystems during 2016-2022, Nat. Commun., 16, 4980, https://doi.org/10.1038/s41467-025-59999-2, 2025.

Liu, J., Wang, Q., Zhan, W., Lian, X., and Gentine, P.: When and where soil dryness matters to ecosystem photosynthesis, Nat. Plants, 11, 1390–1400, https://doi.org/10.1038/s41477-025-02024-7, 2025.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell., 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Maier, R., Hörtnagl, L., and Buchmann, N.: Greenhouse gas fluxes (CO_2, N_2O and CH_4) of pea and maize during two cropping seasons: Drivers, budgets, and emission factors for nitrous oxide, Sci. Total Environ., 849, 157541, https://doi.org/10.1016/j.scitotenv.2022.157541, 2022.

Nendel, C., Berg, M., Kersebaum, K. C., Mirschel, W., Specka, X., Wegehenkel, M., Wenkel, K. O., and Wieland, R.: The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics, Ecol. Modell., 222, 1614–1625, https://doi.org/10.1016/j.ecolmodel.2011.02.018, 2011.

Novick, K. A., Ficklin, D. L., Grossiord, C., Konings, A. G., Martínez-Vilalta, J., Sadok, W., Trugman, A. T., Williams, A. P., Wright, A. J., Abatzoglou, J. T., Dannenberg, M. P., Gentine, P., Guan, K., Johnston, M. R., Lowman, L. E. L., Moore, D. J. P., and McDowell, N. G.: The impacts of rising vapour pressure deficit in natural and managed ecosystems, Plant Cell Environ., 47, 3561–3589, https://doi.org/10.1111/pce.14846, 2024.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci. Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Sabbatini, S., Mammarella, I., Arriga, N., Fratini, G., Graf, A., Hörtnagl, L., Ibrom, A., Longdoz, B., Mauder, M., Merbold, L., Metzger, S., Montagnani, L., Pitacco, A., Rebmann, C., Sedlák, P., Šigut, L., Vitale, D., and Papale, D.: Eddy covariance raw data processing for CO_2 and energy fluxes calculation at ICOS ecosystem stations, Int. Agrophys., 32, 495–515, https://doi.org/10.1515/intag-2017-0043, 2018.

Shekhar, A., Hörtnagl, L., Buchmann, N., and Gharun, M.: Long-term changes in forest response to extreme atmospheric dryness, Global Change Biol., 29, 5379–5396, https://doi.org/10.1111/gcb.16846, 2023.