Discussion of “What matters when? Temporal development of drivers and sources of nitrous oxide emissions in winter wheat”

Author (Turco et al.) response to Referee 1 comments

In the following, *reviewer comments are given in blue italics*, author comments are given in normal font.

General Comments:

This manuscript is a thorough and insightful exploration into the drivers of soil N2O emissions. Building on a well-researched field, the paper employs strong and novel methods to demonstrate this idea through use of spatio-temporally integrated flux measurements and frequent measurement of crop growth dynamics to estimate soil nitrogen dynamics, plant nitrogen uptake, and N2O emissions over time. The experiment appears well conducted and to have yielded a wealth of data, to which the authors add investigative statistical methods which contribute a strong picture of the drivers of N2O and how these drivers shift over time, which I found particularly insightful. The manuscript is also clearly written, if overly descriptive and verbose at times.

Thanks for the positive feedback. We will shorten the text where appropriate.

However, I have key concerns regarding the methodological rigor and the resulting interpretations in three key areas. First, there is significant opacity regarding the data handling process during machine learning, including a high risk of temporal data leakage during model validation. Second, the methods used to classify 'background' vs. 'hot-moment' emissions are tenuous and lack a mechanistic basis. Third, while the authors identify Gross Primary Productivity (GPP) as a preeminent negative driver of emissions in their discussion of results, the exclusion of soil mineral N concentrations from driver analysis creates a risk of omitted variable bias, in which GPP may act as a proxy for N exhaustion. To justify their major claim that GPP is a suppressor of N2O emissions and that N synchronicity is a path toward N2O mitigation, the authors should make efforts to decouple GPP from simple substrate limitation in their statistical investigation of drivers.

We will address these points in detail below.

With a tighter text and these improvements to statistical methods, I am confident this manuscript will represent a strong and novel contribution to understanding pathways to agricultural GHG mitigation.

Major Comments:

L55-56: Is this sentence referring to nitrogen cycling or crop growth?

Thanks for pointing out this ambiguity. It did refer to N_2O emissions. This will be clarified in the revised version of the manuscript.

L80-84: This is true, strictly speaking. However, it is also possible to infer the dominant N-transforming process based on the rate of N2O production, as hot moments have been shown to be overwhelmingly the result of denitrification. Relatedly, a recent work used ML to differentially model nitrification-dominant and denitrification-dominant emissions (Lussich et al., 2026) https://doi.org/10.1002/jeq2.70126

We thank the reviewer for this suggestion. While it is generally accepted that high N_2O fluxes often reflect denitrification-dominated conditions, this inference does not necessarily hold across all sites or years. For this reason, we consider stable isotope approaches essential for robust source attribution and have therefore retained this methodological distinction. We consider machine learning approaches complemented by stable isotope approaches very powerful, also for future studies. In this regard, we recognize the value of innovative ML approaches like Lussich et al. (2026), so we will add a sentence and the reference to the revised manuscript.

L84-90: I'll use this passage to illustrate a broader observation that characterized much of this manuscript. The authors frequently provide detailed explanations, such as the principles and mechanics behind stable isotope analysis. While thorough, these explanations are often characterized by a poor economy of words, include details which are not directly pertinent to the narrative at hand, and add up to a manuscript of considerable length (I count nearly 13,000 words). This work would significantly benefit from greater concision, illustrating only the most pertinent details and taking advantage of the expected level of familiarity of a biogeoscientist audience.

We will shorten the text throughout the manuscript, improving logical flow while providing enough details so that biogeochemists not familiar with highly specialized topics, like SP, still understand our approach and results.

Table 1: slurry was applied the day after mineral N fertilizer was applied. Is this typical practice in the region? This creates perfect C and N chemistry to produce N2O. A study focused on understanding N2O mitigation opportunities, this is not an efficient fertility management decision. Can slurry be applied in fall, before wheat planting?

We thank the reviewer for highlighting the peculiarity of this management practice. This practice can occur in integrated farms (i.e., farms combining livestock and arable crop production), which are common in Switzerland, and is compliant with Swiss agricultural regulations. We agree that this practice sets favorable conditions for denitrification. We will provide additional context to clarify this management practice and discuss its relevance in detail in the Discussion section.

L162-163: Why were two separate outlier detection methods used to filter erroneous fluxes?

We applied different outlier detection methods sequentially to leverage their complementary strengths, as each method identified erroneous fluxes that were not captured by the others. However, not all methods were required for every flux (e.g., for NEE, only the Hampel filter was applied). For N_2O fluxes, 503 outliers were identified using the rolling standard deviation filter, 47 using the Hampel filter, and no outliers were detected with the rolling z-score method. All steps of the outlier detection procedure are transparently documented in our project repository.

This aspect will be clarified in the revised manuscript. The sentence will read: "Outliers were further removed by sequentially applying one or more of the following filters, as needed: rolling standard deviation, Hampel, and rolling z-score, as these methods capture different types of anomalies.".

L173-184: Regarding the RF gap-filling: was there any analysis of error propagation? Or an assessment or justification of the accuracy of RF as a gap-filling method? Accurate gap filling N2O emissions is still a topic of much research, and has resulted in mixed success. One of the papers cited here reports R2 of gap-filling for N2O between 0.6 and 0.76, using only 15% of 'missing values' as the test set, in comparison to ~47% of the time period missing in this study. Other work has

reported lower R2 values from 0.42 (Taki et al., 2019, 10.1139/cjss-2018-0041) to 0.66 (Goodrich et al., 2021, 10.1016/j.agrformet.2020.108280), also using just 15% of the data as 'missing values' to be filled. The other paper the authors cited here performed no analysis on the accuracy of the RF gap-filling method used. The success of ML gap-filling has also been shown to be related to the length of gaps (Taki et al., 2019), yet there is no discussion of typical gap length or the impact of lengths beyond reporting the aggregate percentage of missing values. While I acknowledge that flux gap-filling is a major challenge which has yet to be solved, and that work must go on in the meantime, nevertheless I feel it is important to acknowledge the limitations of gap-filling high resolution N2O flux data and the effect that these limitations might have on this study.

We thank the reviewer for this careful and constructive assessment. We fully agree that the flux community has not yet converged on a "best" gap-filling procedure for non-CO_2 eddy covariance fluxes. While systematic comparisons of multiple gap-filling approaches are available for CO_2 (Moffat et al., 2007; Vekuri et al., 2023; Zhu et al., 2023), comparable studies for N_2O fluxes are still lacking. In addition, reported model performances vary widely depending on site conditions, data availability, gap structure, and evaluation strategy.

We appreciate the reviewer's suggestion to provide explicit performance metrics for the Random Forest (RF) gap-filling approach. In the revised manuscript, we will therefore include quantitative performance indicators (R^2, MAE, and RMSE) based on an independent test set.

As the reviewer notes, gap length is also an important factor. Because long gaps are particularly difficult to fill reliably, we aim to minimize their occurrence through frequent monitoring of the Swiss FluxNet data stream and rapid troubleshooting when problems arise. For example, while Taki et al. (2019) tested linear vs. ANN techniques for filling gaps of up to 20 days, we used RF to fill much shorter gaps, typically less than a day. In the revised manuscript, we will therefore add a brief description of the gap-length distribution, in addition to the overall fraction of missing values, immediately after the data coverage section. A new table (see Table R1.1 below) providing gap statistics for each flux time series will be included in the appendix of the revised manuscript.

A full error-propagation analysis and/or a dedicated sensitivity analysis of model performance across different gap-length classes would certainly be valuable. However, this is beyond the scope of the present study and would substantially increase the length of the manuscript. Instead, we will add a table in the Appendix showing cumulative fluxes under different u* filtering scenarios, following community best practice after Pastorello et al. (2020). In addition, we will expand the discussion to address the choice of the gap-filling method and the associated uncertainty, particularly in relation to gap length, while acknowledging the important contributions of Taki et al. (2019) and Goodrich et al. (2021) in this area.

Table R1.1. Summary of data gaps in EC flux time series. Gap statistics are reported for two quality-control selections: QC0, including only best-quality records, and QC0–1, including best- and medium-quality records. Gap length is expressed in time steps (1 step = 30 min), where a "gap" is a contiguous sequence of missing records. Total number of gaps, median, third quartile (i.e., 75th percentile), and maximum gap length per time series are shown, as well as the percentage of gaps that are shorter than one day, the number of gaps that are longer than three days, and the number of gaps that are longer than one week.

Flux	QC level	# gaps	Median (steps)	Q3 (steps)	Max (steps)	% < 1 day	# > 3 days	# > 1 week

NEE	QC0–1	1,782	2	4	232	99.8%	1	0
NEE	QC0	1,947	2	5	233	99.7%	1	0
N_2O	QC0–1	2,329	1	3	441	99.6%	5	1
N_2O	QC0	2,011	2	5	459	99.5%	5	1
CH_4	QC0–1	2,950	2	4	458	99.7%	5	1
CH_4	QC0	1,636	4	10	459	97.6%	6	1

L183-185: It seems here like the authors estimated the background flux level by just excluding 30-day post-fertilization flux data with assumption that fertilization effect lasts only for 30 days. First of all, I don't think this assumption is correct and your own data did not support the assumption. Fertilization can have long-lived impacts on N2O emissions, particularly during dry periods, offseason, etc. As well, not all hot moments are fertilizer-driven, as those like Claudia Wagner-Riddle's group have shown. Moreover, not all fertilizer-driven N2O emissions are caused directly by fertilizer at all: the excess N added to the soil system by fertilization may be temporarily captured by plant or microbial biomass and mineralized months or even years later, thus driving emissions that would not occur in a natural, unfertilized system but which nevertheless occur distantly from any fertilization event. The authors' data shows this: the large peaks, as big as post-fertilization, from mid-June to August is well beyond 30-day post-fertilization period and unlikely to happen in a truly unfertilized control treatment. Moreover, the SHAP analysis (Fig 4) also shows that fertilization effect was a dominant factor for almost two-months after fertilization.

I have serious reservation about this method of distinguishing between HM and BG emissions by simply excluding fluxes within 30 days of fertilization. This potentially inflated the background emissions and perhaps underestimated emission factors. Authors might consider alternative methods of contextually distinguishing background emissions from hot moment emissions based on outlier detection. Please see Ackett et al., 2025; doi.org/10.1029/2025JG008953.

After careful consideration of the methodological concerns raised, we decided to remove the background-flux estimation from the revised manuscript. More importantly, there was likely some misunderstanding caused by our use of potentially confusing terminology, particularly "background emissions." Our intention was <u>not</u> to distinguish between hot moments and background emissions, but rather to estimate emissions directly related to N fertilization for the calculation of the N_2O emission factor according to IPCC methodology. However, as the reviewer correctly pointed out, the effects of N fertilization can persist over extended periods, and in the absence of an unfertilized control plot, no method can reliably separate fertilization-induced emissions from those not related to fertilization.

Therefore, in the revised manuscript, we will calculate an apparent emission factor simply as the cumulative N_2O-N emission divided by the amount of fertilizer N applied. This approach is consistent with most eddy covariance studies on N_2O in the literature (e.g., Cowan et al., 2020; Feigenwinter et al., 2023; Lognoul et al., 2019; Maier et al., 2022). We will revise the wording throughout the manuscript accordingly and make it explicit that this does not correspond to the IPCC Tier 1 method, since non-fertilization-induced emissions are not subtracted when calculating the emission factor. As a result, the estimated emission factor may be inflated.

Therefore, we will reduce the emphasis on the emission factor in the manuscript and focus on total N_2O-N losses during the growing season and on yield-scaled emissions.

L199-201: Spatially aggregated flux measurements are captured across a field using EC, yet a single point measurement is used for soil moisture content and temperature? Soil moisture is also highly spatially heterogenous. This seems a potentially noteworthy limitation.

We agree that soil moisture (and to a lesser extent soil temperature) can be spatially heterogeneous, and that using a single point measurement to represent conditions within an EC footprint is therefore a potential limitation. In our study, however, soil moisture and temperature were primarily used to capture temporal dynamics (wetting-drying and seasonal trends) as drivers of flux variability, rather than to quantify spatial patterns or absolute footprint-mean states. We therefore expect that while absolute values can differ between locations and sensors, the event timing and temporal trajectories are broadly consistent at the field scale during this period. This expectation is supported by an independent sensor profile located a few meters away (using a TEROS 12 sensor ; METER Group), which shows very similar temporal patterns in 5-cm soil water content compared to the profile used in the manuscript (Fig. R1.1).

Moreover, the use of a single soil profile (or a small number of profiles) to characterize soil state variables alongside EC fluxes is the standard configuration in EC studies, where the primary aim is often to relate flux dynamics to representative environmental time series rather than to resolve within-field spatial variability (Feigenwinter et al., 2023; Lognoul et al., 2019; Maier et al., 2022).

In the revised manuscript, we will explicitly acknowledge this representativeness issue as a limitation and clarify that our conclusions rely mainly on temporal co-variation between fluxes and soil drivers, not on footprint-mean soil moisture estimates.

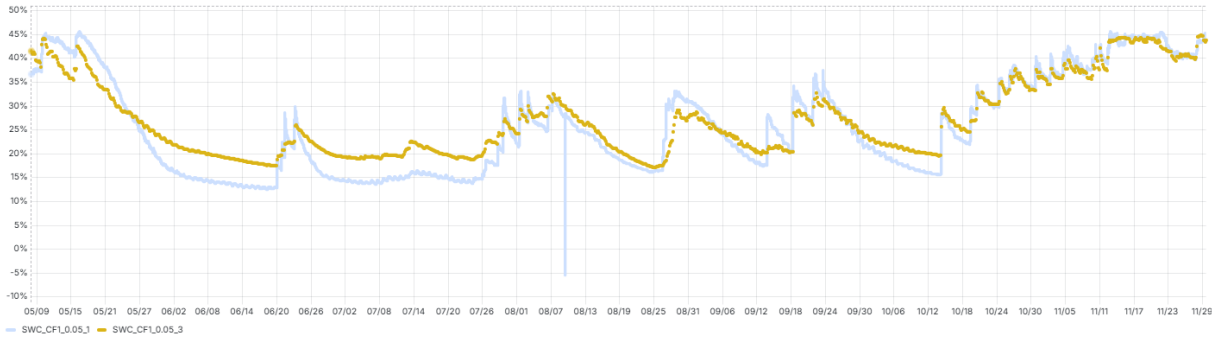

Figure R1.1. Soil water content (SWC) at 0.05 m depth from two different sensor profiles between May and November 2023. Light blue indicates SWC from the sensor used in the manuscript (5TM, Decagon Devices, USA); yellow shows SWC from an independent sensor a few meters away (TEROS 12; METER Group).

L303-305: Again, the persistence of elevated N availability is highly variable and can be much longer than 30 days.

We agree with the reviewer that the persistence of elevated N availability, and its effects on N_2O emissions can be highly variable and may extend beyond 30 days in some cases.

The 30-day cumulative N input was therefore not intended to represent the full duration of fertilization effects, but rather to serve as a pragmatic engineered predictor of recent fertilization intensity. Similar 30-day windows are commonly used in field studies for event-

based reporting of fertilizer-related emissions (Cowan et al., 2019, 2020), although we acknowledge that such an approach may miss longer-lived or legacy effects.

In the revised manuscript, we will clarify this interpretation. The updated paragraph will read: "To represent recent fertilization inputs, we used the cumulative fertilizer N applied over the preceding 30 days (kg N ha^{-1}). This window was chosen to capture the short-term post-application period commonly considered in event-based reporting, but it does not account for longer-lived or legacy fertilization effects (Bouwman, 1996; Cowan et al., 2020; Qian et al., 2025)."

L318-325: A more complete description is needed about the data splitting process is needed in order to verify its validity. In this custom time-block method, how large were the time chunks? Was the model trained on data further in the future than the test data? When working with a single time series, the most correct method of cross validation is to use a method similar to that employed by the TimeSeriesSplit function in scikit-learn. By this method, the timeseries is split into n+1 chunks, and the model is trained on all prior data within the timeseries, including an initial runup chunk (i.e. an expanding window). This ensures that the model is never trained on future data, which would constitute a form of data leakage.

We thank the reviewer for this careful comment. In the revised analysis, we will avoid any potential data leakage by reserving the last 20% of the time series (chronologically) as an independent hold-out set used only for final model evaluation. Feature selection and hyperparameter tuning will be performed exclusively on the first 80% of the data using 5-fold cross-validation (k-fold). Under this revised setup, model performance on the hold-out test period is R^2 = 0.35 and RMSE = 1.46. As shown in Fig. R1.1, the revised model underestimates the highest fluxes and overestimates the lowest fluxes while successfully reproducing the overall temporal dynamics.

We agree that an expanding-window TimeSeriesSplit approach is appropriate for cross-validation in time series. However, given our single-year dataset and the limited number of high-flux events, TimeSeriesSplit resulted in training folds that were too short and insufficiently representative of the full range of conditions, leading to unstable tuning and unreliable estimates of driver effects. In addition, since forecasting was not our objective, we used k-fold cross-validation within the training period to obtain more robust hyperparameters while preserving an independent, time-ordered test set.

For the SHAP analysis, we will train the final model on the full dataset, because the goal is to explain the patterns in the observed dataset rather than to forecast unseen future values. The SHAP results (Fig. R1.2) remain consistent with the previous model, though slightly smoother, reflecting the more generalized model obtained under the revised tuning strategy.

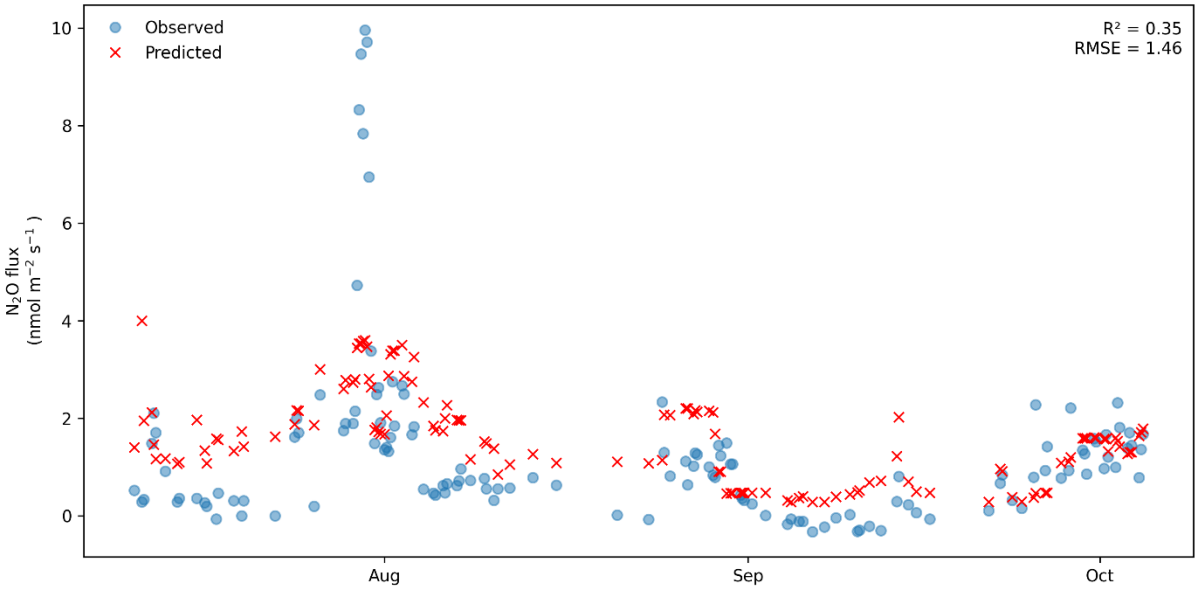

Figure R1.1. Observed and predicted N_2O fluxes at the cropland Oensingen for the hold-out test set (last 20% of the time series; not used for training or tuning). The blue circles represent the values predicted by the XGBoost model, while the red crosses indicate the observed values. The coefficient of determination (R^2) and root mean square error (RMSE) for the test set are given.

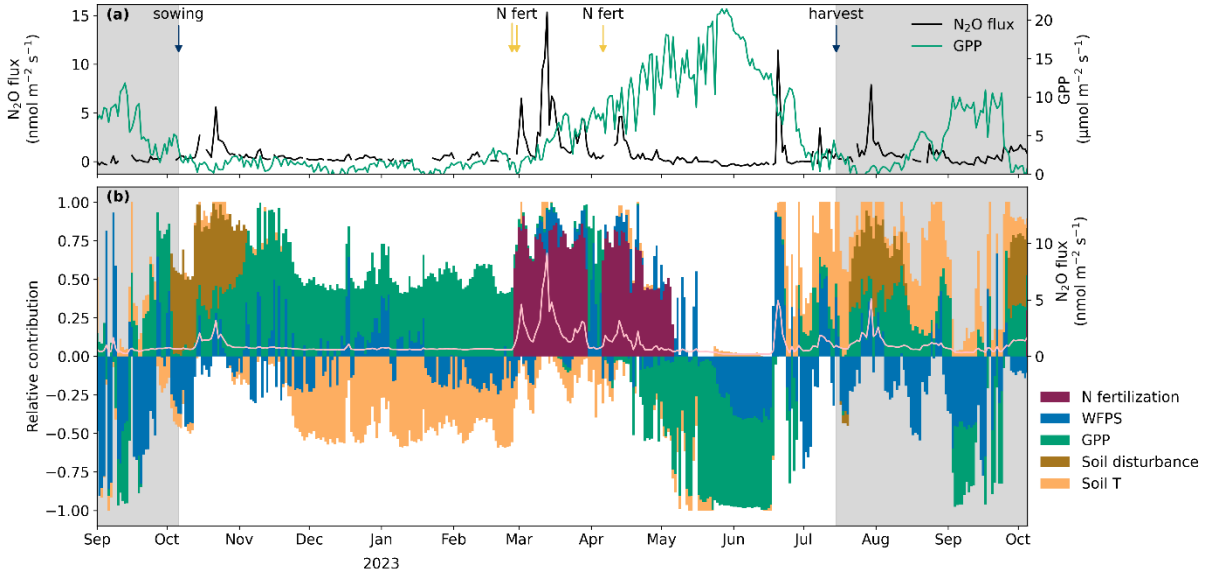

Figure R1.2. Temporal dynamics of N_2O fluxes, gross primary productivity (GPP), and relative driver contributions to N_2O fluxes at the Oensingen cropland from September 2022 to October 2023.

``*L323-325: RMSE and R2 both give disproportionately large influence to hot moments by way of heavily weighing large residuals. Consider using an evaluation metric that evenly weights residuals of all sizes, like MAE, to give a more balanced evaluation of model performance across the distribution of values.*

Thank you for the suggestion. We will add MAE as an additional performance metric.

L328-331: I have searched through the document and could not find a description of what the authors here call the "final model," nor what the "test set" was. There seem to be a lot of mixing of

terminologies regarding data handling and model evaluation, leading to a very opaque picture of the actual methods used. Starting at the beginning of this section:

"Following variable selection, model hyperparameters were optimized using 10-fold cross-validation. To account for temporal autocorrelation and avoid overfitting, while also providing representative coverage of the measurement period, we employed a custom time-block strategy. This approach involved an 80/20 split between training and validation…"

This seems plainly contradictory. If a 10-fold cross validation were used, then across each fold 90% of the data would be used for training and 10% for validation, yet the authors claim an 80/20 split. My best guess might be that the 10-fold cross validation might be a separate process exclusively for hyperparameter tuning, for which no details were provided, and then an abrupt shift to a new data handling process involving an 80/20 split of some kind takes place? Yet descriptions of each process are incomplete and poorly differentiated in their purposes, leading to a jumbled passage.

"with the validation set comprising randomly selected, non-overlapping time blocks that together represented 20% of the available data."

The authors here claim that time blocks were randomly selected, but a timeseries should not be randomly split. The authors seem to acknowledge this idea with their reference to temporal autocorrelation, but their description of their methods here does not give confidence that they have properly dealt with this challenge.

"This splitting strategy was consistently applied throughout the modeling workflow, including during cross-validation and final model evaluation."

This does not make sense to me. There have been no clear definitions of what constitutes the "cross-validation" and the "final model evaluation."

"Cross-validation results showed a R2 of 0.60 and a RMSE of 1.1 nmol N2O m-2 s-1 on the validation set, while the training set showed a R 2 of 0.98 and a RMSE of 0.29 nmol N2O m -2 s-1. The final model, trained with early stopping (10 rounds) to prevent overfitting, achieved a R2 of 0.70 and a RMSE of 1.14 nmol N2O m-2 s-1 on the test set (Fig. A2)."

This passage invites more confusion. Beyond the unclarity regarding CV vs final evaluation, the authors describe model performance on "the test set," "the validation set," etc. Was there a singular test set, as the language here suggests, or are the authors instead referring to the average model performance on holdout data across the 5 or 10 folds?

We thank the reviewer for this careful reading of the Methods section and apologize for the confusing and partly inconsistent description of the modeling workflow.

The apparent contradiction between "10-fold cross-validation" and an "80/20 split" arose because the original procedure was not a standard k-fold design. Instead, it consisted of repeated blocked resampling, in which approximately 20% of the data were assigned to validation in each iteration through randomly placed contiguous blocks. As a consequence, validation sets could overlap across iterations, and individual observations were not necessarily evaluated exactly once.

In addition, the previously described test-set evaluation did not constitute a fully independent nested evaluation framework. We recognize that this was a methodological weakness in the original workflow, and our terminology in this section was not sufficiently precise.

As explained above, we have revised the modelling framework and will rewrite Sect. 3.5 accordingly.

L333-349: In contrast to the preceding passage, this description of SHAP values is exhaustive, and may be excessive.

We provided a more detailed explanation of SHAP because it is still unfamiliar to many readers in the biogeosciences community, although its use is increasing. Given the central role SHAP plays in interpreting our results, we aimed to ensure accessibility for non-expert readers. Nevertheless, we will shorten this section in the revised manuscript by removing redundant details and focusing on the information necessary for result interpretation.

L360-384: This passage is a lengthy textual description of figure 2, which I'm not sure is necessary given the length of the manuscript.

Thanks for the suggestion. We will shorten this text in the revised version of the manuscript.

L452-453: The average background flux estimate of 0.44 nmol m^{-2} s^{-1} is equivalent to ~11 g N2O-N/ha/d, which is substantially (3-4 times) larger that globally estimated background emissions of ~2.5 to 3 g N2O-N/ha/d. This could be due to the reasons I explained above.

As explained above, given the concerns raised about our method for estimating background emissions, we have removed the background-emissions calculation and the associated text from the revised manuscript.

L462-466: "In contrast, the cover crop acted as a net CO 2 source, emitting 108 g CO2-C m-2"

The cover crop phase was only measured for 2 months, which was influenced by cultivation after wheat harvest. Is this the typical cover crop growth duration? If not, then this is misleading information that provides a snapshot of the rotation. Cover crop active growth phase is often known to sequester C (NEE sink) and negligible N2O emissions.

As the authors correctly point out in their discussion, it is not the cover crop per se that is acting as an emitter of CO2, but rather this is the response to harvest of the winter wheat and the bare soil. It is tricky, but I wonder if there is a better way to convey that the cover crop is actually reducing emissions, not magnifying them, even though these emissions are taking place during the cover crop's establishment.

Thank you for highlighting this. We agree that the phrasing "the cover crop acted as a net CO_2 source" can be misleading without appropriate context. In our study, the summer cover-crop phase itself lasted only ~2 months (late July–September), following directly after the winter wheat harvest and soil cultivation, a period when CO_2 exchange is strongly influenced by residue decomposition and soil disturbance (as discussed in Section 4.2). We will revise the text to make explicit that the reported net CO_2 release reflects NEE during this short, post-harvest cover-crop establishment phase under disturbed conditions, not an intrinsic effect of cover crops in general. At the same study site, a previous model–observation comparison by Emmel et al. (2018) showed that, when left bare and without a cover crop, the field acted as an even larger source of CO_2.

At this site, establishing a summer cover crop between two winter cereals is not standard practice: this was the first time the farmer implemented such a short cover crop. Previous cover crops at the site were mainly autumn–winter (e.g., Phacelia) and lasted longer (~6 months). We

will clarify this management context in the Results and Discussion and refer readers to Emmel et al. (2018) for additional rotation history.

L480-483: in L452, the authors mentioned background flux of 0.44 nmol m $^{-2}$ s $^{-1}$. This is confusing as it seems different subset of data are being used in places to estimate background flux.

We agree that using both "baseline" and "background" terminology was confusing. Here, "baseline flux" refers to the mean model-predicted N2O flux for the reference ("background") dataset used for SHAP, which excluded periods immediately following management to represent typical conditions. In the revised manuscript, we will clarify this wording and use consistent terminology. Moreover, because we will remove the background-flux calculation used for EF estimation (see response above), this source of confusion will be eliminated.

L486-509: SHAP values are best interpreted by relating the direction/magnitude of impact alongside the magnitude of the factor value. For example, the authors write: "N2O emissions were mainly suppressed by WFPS." The fact that soil moisture had a large negative impact on flux predictions, as described here, leaves a lot of room for interpretation. Were these negative influences related to high or low soil moisture?

Thank you for pointing this out. We agree that SHAP impacts must be interpreted in relation to driver values. We will add SHAP dependence plots in the Appendix (Fig. R1.3) and revise the text to explicitly link the sign and magnitude of SHAP values to the corresponding range of drivers (e.g., high vs. low WFPS).

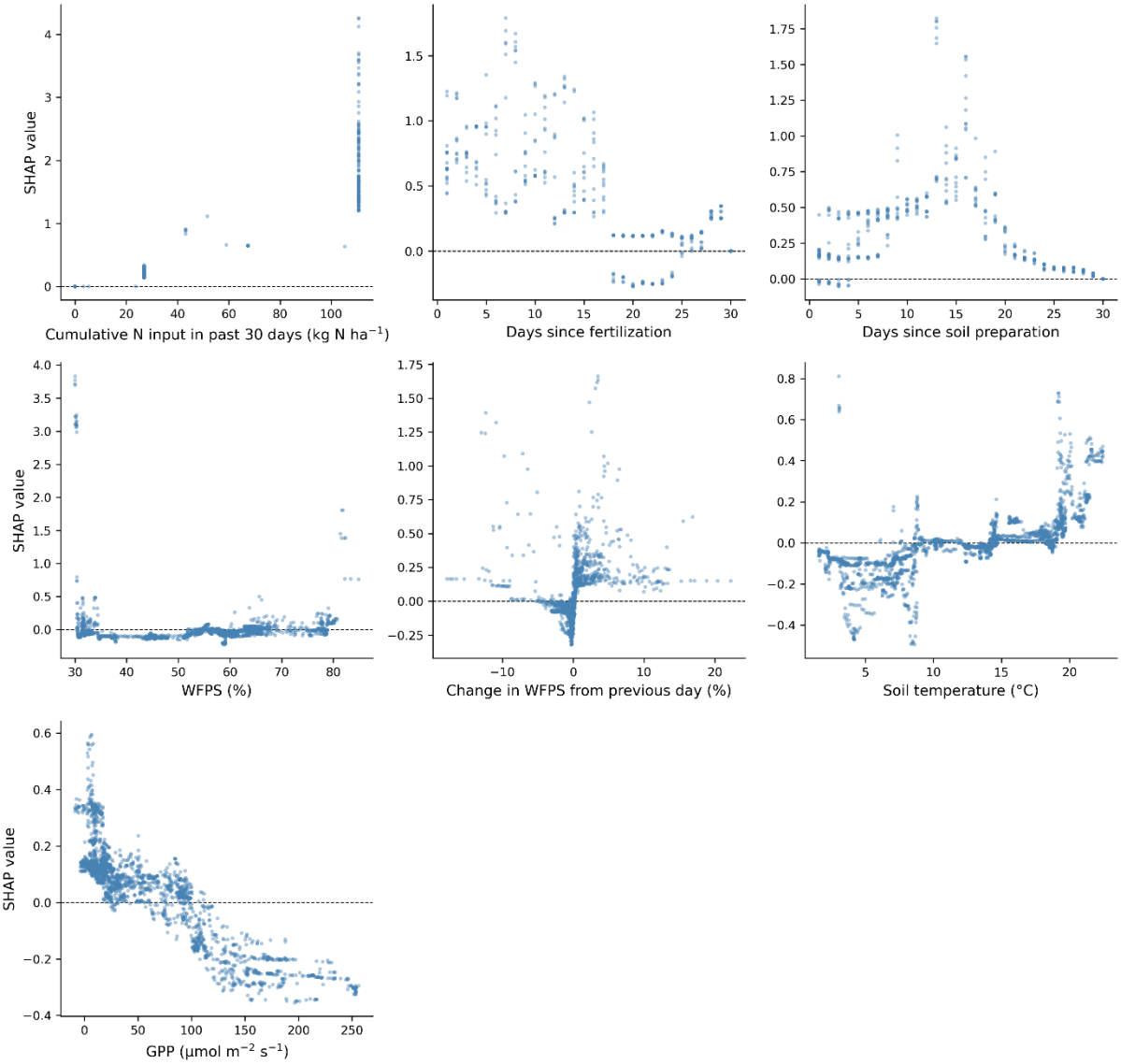

Figure R1.3 SHAP dependence plots for the seven drivers used in the XGBoost model for N_2O fluxes. Each panel shows the SHAP value of a given driver for N_2O fluxes against its observed values across all data, reflecting the driver's marginal contribution to the modelled N_2O flux.

L510-515, Figure 5: The binning strategy used here to quantify the relationships between WFPS, GPP, and flux, is statistically fragile. The bins create artificial step functions in what is likely a linear relationship, and also reduces the number of data points for statistical analysis. It also is able to demonstrate an interaction exists, but does not quantify that relationship in terms of strength or behavior. A better approach might be to model a multiple linear regression with an interaction term. Partial r2 could also be used to more precisely determine the relative explanatory power of each factor and the interaction.

We thank the reviewer for this constructive suggestion. We further explored regression-based alternatives (e.g., multiple linear regression with an interaction term and partial R^2), but the required assumptions were not met in our dataset (non-normal/heteroscedastic residuals and substantial temporal autocorrelation), making such inference unreliable.

We therefore will not rely on the binned approach for interpretation. Instead, we will replace the binned analysis with a SHAP dependence plot for GPP only, including a LOESS smoother (new Fig. R1.4). This retains the full dataset and provides a continuous, model-consistent visualization of the relationship without imposing bin-related step functions or requiring linear-model assumptions. While this plot does not explicitly quantify an interaction term, it more robustly illustrates the pattern of the GPP effect in the fitted model than the current binning approach. The corresponding text describing this relationship will be revised accordingly in the Results/Discussion.

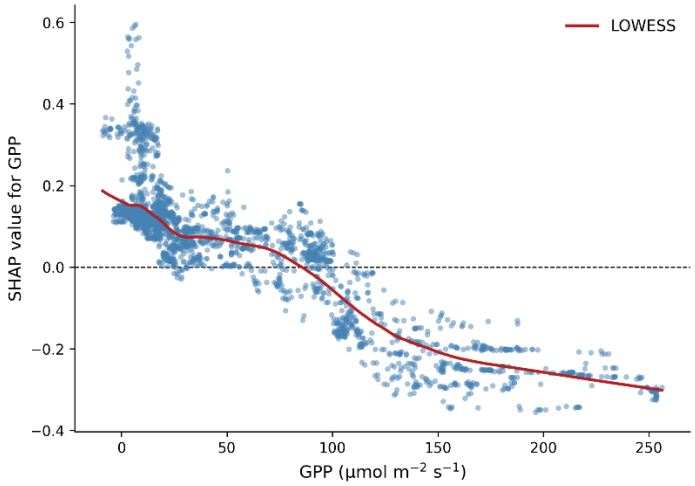

Figure R1.4 SHAP partial dependence plot, showing the SHAP value of GPP against its observed values across all data, reflecting GPP's marginal contribution to the modelled N_2O fluxes. A LOWESS smoother is overlaid to highlight the overall trend.

L559-560: True, but could this also be due to increasing contribution of nitrification to the net N2O emissions? How could you be so sure about Fig 7B secondary vertical axis?

Thank you for the comment. The secondary axis in Fig. 7B represents the reduced fraction (1 − r) derived using the FRAME model based on a closed Rayleigh formulation ($M = M_0 + E \cdot \log(r)$), where r is the residual N_2O fraction remaining after reduction. Thus, 1 − r quantifies the fraction of N_2O that has been reduced to N_2. We will clarify this explicitly in both the main text and the figure caption to avoid ambiguity.

Regarding the isotopic pattern, while an increased contribution of nitrification could indeed explain higher site preference (SP) values, it would not account for the pronounced enrichment in $\delta^{18}O$ observed. Therefore, the simultaneous rise in both $\delta^{18}O$ and SP, particularly given the magnitude of the $\delta^{18}O$ shift, is more consistent with enhanced N_2O reduction rather than a shift toward greater nitrification alone.

This interpretation is further supported by the FRAME model results, which indicate that the relative contributions of the different production pathways remained stable during the last three sampling dates (Fig. R3). To make this clearer to readers, we will include Fig. R1.5 in the Appendix.

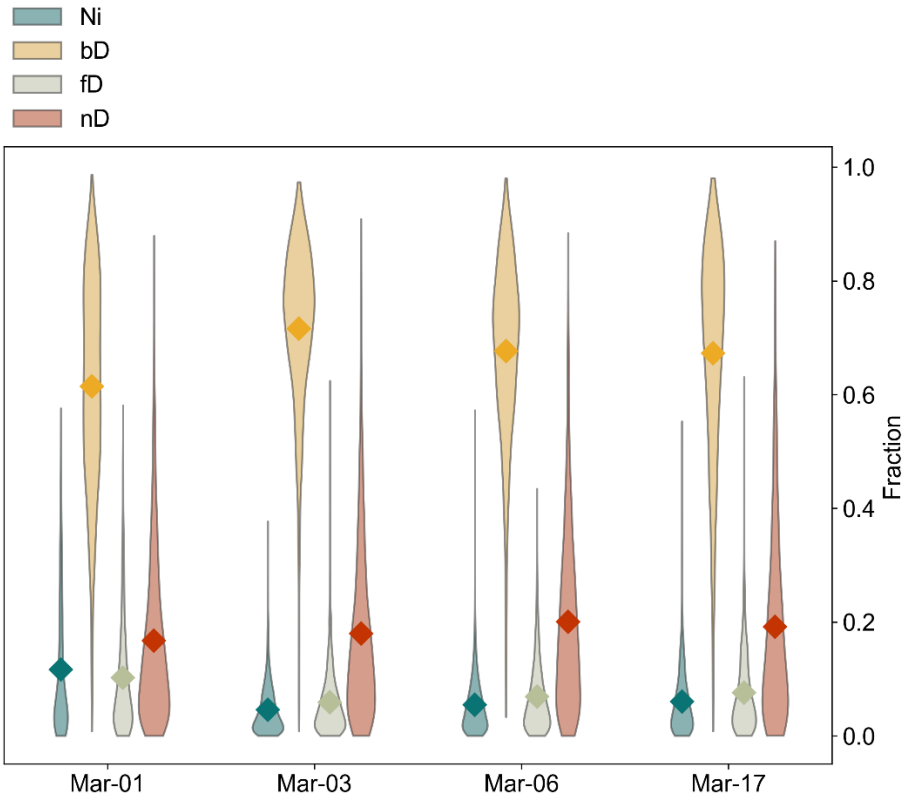

Figure R1.5. FRAME-estimated N_2O source fractions in March 2023 following fertilization (27–28 February), shown as violin plots for denitrification, nitrifier denitrification, nitrification, and fungal denitrification (means indicated by diamonds).

L565-566: Relatedly, could not the increased role of nitrifier denitrification also indicate a decreased rate of denitrifier denitrification overall?

Thank you for this comment. As clarified above, (1−r) represents the fraction of produced N_2O reduced to N_2 before emission. We revised the wording to avoid ambiguity, replacing "more complete denitrification" with "greater N_2O reduction to N_2". In general, nitrifier and bacterial denitrification are very difficult to distinguish, as their source signatures overlap (Figure 7a in the manuscript). We are therefore careful not to overinterpret the FRAME model results on bD and nD contributions.

L646-650: The N2O offset effect on net CO2 is based on NEE and not actually on long-term C sink, which is soil C stabilization. Not all of the NEE sink from a growing season is translated into soil C gain that is stabilized for longer period beyond the growing season. The studies using long-term experiments showed much higher N2O offset effects in fertilized systems in other regions (see Dhaliwal et al., 2025; doi.org/10.1002/jeq2.70046). Therefore, the short-term nature of the study may have underestimated N2O effects.

Thanks for the comment. Please be assured, we are fully aware that a long-term C slink is not equivalent with a net CO2 sink and thus paid extra care not to mislead the reader but only talk about a "net CO2 sink". We thus fully agree that the 70% offset refers to the NEE-based seasonal GHG budget and should not be interpreted as an offset of long-term soil C sequestration. We have therefore revised the wording to avoid linking the observed net CO_2 sink directly to a long-term climate benefit. We also added that seasonal CO_2 uptake does not necessarily translate into

stabilized SOC gains; indeed, previous long-term measurements at the same site reported net SOC losses based on soil C stock changes over more than a decade (Emmel et al., 2018).

L656-661: Mineral N sampling also showed modestly high N availability during this period, likely related to low uptake and mineralization from residue and microbes

Thanks for this comment. We will add this to the revised manuscript.

Section 4.4, L705-708: I will take this line as an opportunity for a discussion on the idea of synchronicity between N supply and demand as a key feature in determining N2O emissions, and as a key idea in this manuscript. The authors place much significance on this idea throughout their introduction, use of methods, and the discussion here. In my view, the evidence for nitrogen synchronicity as the preeminent determinant of N2O emissions, rather than simply N availability, is a bit scattered. For example, Figure 4 excellently illustrates flux and GPP, yet mineral N availability is missing as the third key component needed to make this case. From Fig 4 there certainly does seem to be an inverse relationship between flux and GPP, but I wonder the extent to which GPP is also negatively correlated with N availability. Given the inverse relationship between GPP and mineral N (the nitrogen is moving directly from the soil and into the plant biomass), more evidence is needed to prove that GPP is independently negatively affecting N2O emissions, beyond soil N.

This leads me to my concern about the authors' interpretation of GPP's role in the SHAP analysis. Because soil mineral N concentrations were not included as input features in the Random Forest model, the model likely suffers from omitted variable bias. While 'days since fertilization' is included as a predictor, it is a very rough and linear proxy that fails to capture the dynamic, non-linear depletion of the soil N pool. In the absence of a direct N-substrate variable, the machine learning algorithm quite possibly utilizes GPP as a better mathematical proxy for the diminishing N pool. Consequently, the high importance attributed to GPP in the SHAP analysis may simply reflect N exhaustion rather than a mechanistic suppression of N2O production by plant uptake (synchronicity). To truly substantiate the claim that synchronicity is key to emissions, the authors would ideally show that for a given level of soil mineral N, higher GPP results in lower emissions. Without including measured soil N data in the modeling workflow, the current results show correlation with the crop's growth phase, but do not sufficiently decouple plant demand from simple substrate limitation to make strong claims such as this one.

Thank you for this thoughtful comment. We agree that without continuously measured soil mineral N as a predictor, the XGBoost/SHAP analysis cannot fully disentangle the role of GPP as a proxy for crop N demand from concurrent changes in mineral N availability. Soil mineral N was measured only on a limited number of dates and therefore could not be included in the model without introducing large temporal gaps or substantial interpolation uncertainty. Moreover, to our knowledge, no continuously measuring device exists that reliably measures Nmin in soil solution. Consequently, the SHAP importance of GPP should not be interpreted as evidence for an independent causal effect of GPP on N_2O emissions at a fixed level of soil mineral N, but as a proxy for crop demand based on a continuously available GPP record.

Overall, our aim, was indeed to assess how continuously available variables describing crop activity (GPP), management (fertilization and soil disturbance), and soil–climatic conditions (WFPS and TS) explained temporal variation in N_2O fluxes. While high-frequency mineral N measurements would be great and would have provided a more direct representation of substrate availability, our predictor set allowed us to assess how N inputs, soil disturbance, and

crop activity were associated with N_2O flux dynamics through their links to N supply and demand. We will clarify that GPP is interpreted as an indicator of crop development and N demand within the N synchrony framework, not as an independent mechanistic driver separated from mineral N availability. We will therefore revise the Discussion to state more clearly that the observed negative relationship between GPP and N_2O fluxes likely reflects crop development, during which increasing plant N demand likely coincides with declining mineral N availability.

References

Bouwman, A. F.: Direct emission of nitrous oxide from agricultural soils, Nutr Cycl Agroecosyst, 46, 53–70, https://doi.org/10.1007/BF00210224, 1996.

Cowan, N., Levy, P., Moring, A., Simmons, I., Bache, C., Stephens, A., Marinheiro, J., Brichet, J., Song, L., Pickard, A., McNeill, C., McDonald, R., Maire, J., Loubet, B., Voylokov, P., Sutton, M., and Skiba, U.: Nitrogen use efficiency and N_2O and NH_3 losses attributed to three fertiliser types applied to an intensively managed silage crop, Biogeosciences, 16, 4731–4745, https://doi.org/10.5194/bg-16-4731-2019, 2019.

Cowan, N., Levy, P., Maire, J., Coyle, M., Leeson, S. R., Famulari, D., Carozzi, M., Nemitz, E., and Skiba, U.: An evaluation of four years of nitrous oxide fluxes after application of ammonium nitrate and urea fertilisers measured using the eddy covariance method, Agricultural and Forest Meteorology, 280, 107812, https://doi.org/10.1016/j.agrformet.2019.107812, 2020.

Emmel, C., Winkler, A., Hörtnagl, L., Revill, A., Ammann, C., D'Odorico, P., Buchmann, N., and Eugster, W.: Integrated management of a Swiss cropland is not sufficient to preserve its soil carbon pool in the long term, Biogeosciences, 15, 5377–5393, https://doi.org/10.5194/bg-15-5377-2018, 2018.

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, Science of The Total Environment, 903, 166389, https://doi.org/10.1016/j.scitotenv.2023.166389, 2023.

Lognoul, M., Debacq, A., De Ligne, A., Dumont, B., Manise, T., Bodson, B., Heinesch, B., and Aubinet, M.: N_2O flux short-term response to temperature and topsoil disturbance in a fertilized crop: an eddy covariance campaign, Agricultural and Forest Meteorology, 271, 193–206, https://doi.org/10.1016/j.agrformet.2019.02.033, 2019.

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, Science of The Total Environment, 849, 157541, https://doi.org/10.1016/j.scitotenv.2022.157541, 2022.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A., and Stauch, V. J.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agricultural and Forest Meteorology, 147, 209–232, https://doi.org/10.1016/j.agrformet.2007.08.011, 2007.

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.

Qian, H., Yuan, Z., Chen, N., Zhu, X., Huang, S., Lu, C., Liu, K., Zhou, F., Smith, P., Tian, H., Xu, Q., Zou, J., Liu, S., Song, Z., Zhang, W., Wang, S., Liu, Z., Li, G., Shang, Z., Ding, Y., van Groenigen, K. J., and Jiang, Y.: Legacy effects cause systematic underestimation of N_2O emission factors, Nat Commun, 16, 2775, https://doi.org/10.1038/s41467-025-58090-0, 2025.

Vekuri, H., Tuovinen, J.-P., Kulmala, L., Papale, D., Kolari, P., Aurela, M., Laurila, T., Liski, J., and Lohila, A.: A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates, Sci Rep, 13, 1720, https://doi.org/10.1038/s41598-023-28827-2, 2023.

Zhu, S., McCalmont, J., Cardenas, L. M., Cunliffe, A. M., Olde, L., Signori-Müller, C., Litvak, M. E., and Hill, T.: Gap-filling carbon dioxide, water, energy, and methane fluxes in challenging ecosystems: Comparing between methods, drivers, and gap-lengths, Agricultural and Forest Meteorology, 332, 109365, https://doi.org/10.1016/j.agrformet.2023.109365, 2023.