Suggestions for revision or reasons for rejection

(visible to the public if the article is accepted and published)

The authors have addressed most of my minor comments, but for each of the major comments, I still have concerns and hopefully the authors can understand my constructiveness to the whole modeling community.

[1] Regarding appropriate baselines, I believe my previous comment may have been misunderstood. I am not questioning whether LSTM is a valid approach for this problem. Rather, my concern is that the manuscript's current framing is misleading in presenting this study as the "first" deep learning methodological advance for predicting N2O. To clarify, I do value the contribution of the study in showing what does and does not work, and by how much, when applying LSTM to model N2O fluxes across multiple sites. That is a meaningful contribution in its own right. However, the current novelty statement overstates this contribution and should be revised to more accurately reflect the scope of the advance. I encourage the authors to present the work in a more precise and balanced manner, clear and restrained framing would better serve both the paper and the field.
btw, L316:L325 is not about KGML, are you referring to somewhere else?
A few baselines, for methodology: Liu et al. KGML-ag, https://gmd.copernicus.org/articles/15/2839/2022/
for multi-sites: Sharma et al. https://www.pnas.org/doi/10.1073/pnas.2524808123

Thank you very much for following up on the review process.

- We share the same perspective. We did not include any statement claiming that this study represents the first deep learning approach for this application. The paragraph discussing the research gaps does not present the study as the first of its kind (L350:L357), but rather highlights aspects such as the use of a relatively rich dataset, sequential modelling, and the inclusion of an independent test site within a unified framework.
  Could you please indicate which specific paragraph or statement you are referring to, so that we can revise it accordingly?

- Several statements along the manuscript also in the limitations and implications section indicate the shortcomings of the current model as well as improvement potentials. For instance:
  1. In section 4.2: "Further improvements, particularly through expansion of the training dataset, could enhance the model's ability to capture such episodic events more robustly.
     Although the model captures daily fluctuations in many cases, inaccuracies remain in periods of very low fluxes, such as Rottenbuch in Fendt extensive in October 2017 (Fig. 4 g). During this period, changes in soil moisture, temperature and precipitation were minimal, resulting in limited variability in the driving variables and consequently reduced predicted dynamics."

- For the KGML-ag more specifically lines L347:L349 provide a brief overview in the newly updated manuscript. We improved the statement to be more precise, also added the second reference that you mentioned regarding the 2026 multi-site publication.

[2] The cross-site test performance is markedly lower than the training and temporal test performance, and this important result is not sufficiently discussed. For this reason, I do not find it appropriate to describe the model as demonstrating "promising predictive performance." Figure 3c suggests that the model predicts a number of observations near the mean, despite the much larger variability in the measurements. This should be recognized as a highly challenging test case. I encourage the authors to frame these results more carefully by identifying what the model is able to capture, what it fails to capture, and what this implies for the current limits of cross-site prediction, rather than suggesting that the LSTM approach has already solved the problem.

- While there is a drop in performance in noticed comparing the whole cross-site test with temporal test set, the model performs well as some of the cross-site site-years and not well at others. For a better analysis, we added site-year-specific performance metrics to the headings of the panels in Fig. 4. This allows readers to directly assess model performance for each individual site-year and provides a more detailed view than the aggregated statistics presented in Table 5. The updated figure demonstrates that the model performs fine or very well for 6 site-years, while mainly 4 of cases, exhibit relatively low values for certain metrics. Please see newly added lines with details about this, L295:L307: " For a more detailed assessment of model performance on the cross-site test set, performance metrics were calculated separately for each site-year and are reported in the titles of the individual panels in Fig. 4. The R^2 values ranged from 0.24 to 0.64 for six site-years (Fig. 4 e, f, g, h, j, and k), indicating a moderate to good

agreement between observed and predicted emissions. In contrast, four site-years (Fig. 4 i, l, m, and n) exhibited substantially lower R^2 values, ranging from −25 to 0.14. These results demonstrate that the overall R^2 value of −0.12 reported for the cross-site test set in Table 5 does not reflect a uniformly poor model performance or a systematic failure within the whole cross-site test set. Rather, it is strongly influenced by a small number of challenging site-years that disproportionately affect the aggregated metric. In addition, a proportion of the reduced R^2 at the challenging sites is caused by the few extreme prediction errors at fertilization events. For example, substantial overestimations were observed at Rottenbuch in Fendt 2017 and 2020 intensive management campaigns following fertilization (Fig. 4 i, l, n). In 2015 and 2016, fertilization events occurring during warm periods were associated with pronounced emission peaks exceeding 150 μg m^{-2} h^{-1} for this soil (Fig. 4 j, k). However, despite similar temperature and fertilization conditions, the 2017 and 2020 campaigns did not exhibit elevated N_2O fluxes. The model nevertheless predicted several large emission peaks, resulting in the extreme prediction errors at these site-years leading to negative R^2 (Fig. 4 i, l, n). Even for site-years with low R^2 values, the relatively high Spearman correlations, 0.42-0.61, indicate that the model was generally able to reproduce the temporal dynamics of N_2O emissions."

- To ensure conveying the right message, In the abstract we replace the phrase promising predictive performance with a new statement L14:L15: " The trained LSTM model showed varying predictive performance when evaluated on test datasets, with higher performance at some site-years and lower at others. "
- This issue about figure 4 c, was discussed in L282, to ensure properly applying your feedback we expanded it by mentioning that in case the performance is lower and the bias is higher comparing to other sites: "At Graswang, fluxes were more frequently overestimated (Fig. 4 c, d), showing a significantly lower performance and higher bias comparing to other soils."

[3] The newly added sentences on the model structure and attention mechanism are still overly superficial. They do not provide sufficient quantitative or symbolic detail to make clear how the LSTM and attention mechanism are actually formulated and applied in this study. I recommend that the authors revise this section substantially by including the relevant equations, definitions of variables and inputs, and a precise explanation of how attention is implemented and used. In its current form, the methodological description is not adequate for careful evaluation or reproducibility.

- We appreciate the reviewer's emphasis on clarity regarding the model architecture. To balance the paper's focus for a broader, non-machine-learning audience while ensuring technical completeness, we have briefly expanded the description of the attention layer of our model in the manuscript (L173:L174). Additionally, we have updated Figure 2 to explicitly include the concatenation and dropout layers. To maintain readability, Figure 2 remains conceptual; however, readers interested in the exact architectural details can access the complete LSTM implementation via the GitHub link in the Data and Code Availability section.

#Reviewer 4

The manuscript "Modelling N2O emissions in fertilized grassland with machine learning: a multi-site LSTM approach" by Payrosangari et al. presents the development and application of a machine learning approach that predicts temporal dynamics of agricultural N2O emissions using observational data from grassland across several European sites. While the study is interesting and offers a potentially useful tool for improving N2O estimates in agricultural soils, it does not necessarily advance understanding of the mechanisms and processes that govern N2O emissions. Therefore, it is important for such a machine learning study to 1) demonstrate robust results, i.e., convincingly good model performance, and 2) critically discuss the choice of data, results and limitations. The major concerns with the current manuscript revolve around these points. I find that it sometimes needs more detailed description of the model development and an explanation of why the authors claim their approach is "good or robust". I have compiled several points that may hopefully help improve the manuscript. 1. The authors present several measures and indicators (e,g., R2, RMSE, Bias etc) for evaluating model performance. These are important and useful. However, given the current results, I am not sure whether readers will be convinced that the model is trained "well" and that its parameters have been optimized. Since data from multiple sites were used, it would also be useful to report normalized RMSE (NRMSE) and normalized model bias, which provide a quantitative evaluation of the magnitude of RMSE for each site, or how accurate the model is.

- Thank you for this comment. To provide a more transparent and informative assessment of model performance, we implemented two changes in the revised manuscript.
    - First, we revised the abstract to better reflect the variability in model performance across sites and years. The updated text states in L14: "The trained LSTM model showed varying predictive performance when evaluated on two test datasets, with higher performance at some site-years and lower at others."

- Second, we added site-year-specific performance metrics to the titles of the panels in Fig. 4, along with an explanatory paragraph in the manuscript which also is a reply to your comment→ Section 3.2, L295:L307: " For a more detailed assessment of model performance on the cross-site test set, performance metrics were calculated separately for each site-year and are reported in the titles of the individual panels in Fig. 4. The R^2 values ranged from 0.24 to 0.64 for six site-years (Fig. 4 e, f, g, h, j, and k), indicating a moderate to good agreement between observed and predicted emissions. In contrast, four site-years (Fig. 4 i, l, m, and n) exhibited substantially lower R^2 values, ranging from −25 to 0.14. These results demonstrate that the overall R^2 value of −0.12 reported for the cross-site test set in Table 5 does not reflect a uniformly poor model performance or a systematic failure within the whole cross-site test set. Rather, it is strongly influenced by a small number of challenging site-years that disproportionately affect the aggregated metric. In addition, a proportion of the reduced R^2 at the challenging sites is caused by the few extreme prediction errors at fertilization events. For example, substantial overestimations were observed at Rottenbuch in Fendt 2017 and 2020 intensive management campaigns following fertilization (Fig. 4 i, l, n). In 2015 and 2016, fertilization events occurring during warm periods were associated with pronounced emission peaks exceeding 150 $\mu g\ m^{-2}\ h^{-1}$ for this soil (Fig. 4 j, k). However, despite similar temperature and fertilization conditions, the 2017 and 2020 campaigns did not exhibit elevated N_2O fluxes. The model nevertheless predicted several large emission peaks, resulting in the extreme prediction errors at these site-years leading to negative R^2 (Fig. 4 i, l, n). Even for site-years with low R^2 values, the relatively high Spearman correlations, 0.42-0.61, indicate that the model was generally able to reproduce the temporal dynamics of N_2O emissions."

- NRMSE was included in an earlier version of the manuscript. However, following reviewers' request, we decided to remove this metric because it is not well suited to the characteristics of our dataset. The N_2O flux data are highly right-skewed, with many low values and a small number of extreme emission events. Since NRMSE is calculated as RMSE divided by the mean observed value, its magnitude becomes highly sensitive to the combination of a low mean and a few large prediction errors associated with extreme fluxes. Consequently, NRMSE can become disproportionately large and may not accurately reflect the model's overall predictive performance. As an example, for the Rottenbuch site in Fendt (2015, intensive management; Fig. 4j), the calculated NRMSE is approximately 160%. At first glance, this suggests very poor model performance. However, the model captures the temporal dynamics reasonably well, as indicated by a Spearman correlation of 0.61 (95% CI: 0.53–0.68), an R^2 of 0.42, and a mean bias of −0.54. The inflated NRMSE results primarily from the low mean flux value at this site (approximately 8 units) combined with a relatively large RMSE (12.74), which is driven by a small number of high-emission events. In highly skewed datasets, the mean can be strongly influenced by the distribution of observations and may not provide a robust normalization factor for model error metrics. As a result, NRMSE can overemphasize errors associated with rare extreme values and obscure the model's ability to reproduce the overall temporal behaviour of the fluxes. For these reasons, and in agreement with previous reviewer recommendations, we removed NRMSE from the analysis and instead focused on complementary metrics such as Spearman correlation, R^2, RMSE, MAE, and mean bias, which provide a more balanced assessment of model performance.

2. It is worth explaining (at least briefly) the criteria for data selection in the Methods. Having measurement data from multiple sites is certainly an advantage, but does the model prefer quantity over quality? What is your standards for including a dataset in the training set? A specific case is [Neustift Austria]. You mentioned that this site has a limited sample size, yet it is included because it sits at high elevation and shows low emissions (as written in L129-130). I wonder whether it is necessary to include this site, as too little data may introduce bias. Any expert judgement on why this site has low emissions? Is it simply due to relatively low fertiliser rates? How would the model perform if this site were removed? Maybe you can also report fertiliser application rates (e.g, xxx kg N/ha/yr) in Table 1, which would give some hints on how much applied N is roughly lost through N2O emissions (since no fertiliser controls may not be available). Meanwhile, there were both chamber measurements and eddy covariance measurements. How do the different measurement methods affect your modelling approach? It would be helpful to discuss.

- The quality control procedures were performed by the organizations responsible for data collection, providing confidence in the use of the complete dataset. We also have mentioned in the paper that non gap-filled data was used. Furthermore, the lower N_2O emissions observed at Neustift do not represent anomalous or noisy behavior. Rather, they can be explained by the site's environmental conditions, including low average soil moisture (approximately 20%) and low average soil temperature (9.5°C), as shown in Tables A1a and A1b. In addition, the site receives only a single fertilization event per year, typically in October. The site was retained in the training dataset despite the absence of similar conditions in the test set because the primary objective of this study was to develop a model with regional applicability rather than one optimized for a limited number of grasslands. Including sites with diverse environmental conditions increases the variability represented in the training data and may improve the model's ability to generalize to new locations. A model trained on a broader range of conditions is expected to have a greater likelihood of successfully transferring to previously unseen sites than a model trained on a more homogeneous dataset.
- Removing Neustift didn't improve the results on the other soils in the test set.
- Good point. We have added fertilizer application rates in Kg N $ha^{-1}\ y^{-1}$ to table 2
- Chamber and EC measurements are associated with different sources of uncertainty and may differ in their ability to capture short-term emission peaks and spatial variability. Despite these differences, both measurement types provide robust estimates of surface-atmosphere N_2O exchange and were used as observations for model training. The inclusion of data from both methods increased the

range of environmental conditions and management practices represented in the dataset, supporting our goal of developing a model with broader applicability across sites. Furthermore, site-specific characteristics were provided through the static input variables, allowing the model to learn relationships between environmental drivers and measured fluxes across different measurement settings. To ensure explaining this in the paper we add to L423:425: "Finally, the dataset combined chamber and eddy covariance measurements, which differ in spatial representativeness and associated uncertainties. While this increased the diversity of conditions represented in the training data and supported model generalizability, it may also have contributed to unexplained prediction errors. Future work could investigate the influence of specific measurement methodologies more explicitly."

3. I think Section 2.4.3 Hyperparameter tuning can be improved. The current description is still too general. To improve transparency, the Hyperband approach was used for tuning and the following manual adjustments need to be described in sufficient detail. I see that the authors claimed that the best hyperparameters were only provided in an associated repository. I would suggest adding a new column to explicitly list the chosen values (why not?).

- Instead of explaining the Hyperband in details, we decided to refer to the paper which defined the approach to keep the focus of the paper on the complete picture of the framework and avoiding details about single steps. To apply your feedback we add a short explanation about it to provide a first impression to the audience at L198:L199"Hyperband improves upon conventional random search by efficiently balancing exploration and exploitation, initially evaluating a broad range of hyperparameter configurations and progressively allocating more computational resources to the most promising trials (Li et al., 2018)."
- The reason that the best hyperparameters are mentioned only in Github is that even small changes in the data leads to changes in the best hyperparameters, being changing data volume or adding new sites. So, for any new dataset the tuning process has to be done by the researcher and the best hyperparameter of this specific dataset won't be the optimum for the new one. Therefore, also to stay aligned with the framework of other publications which report ML models but not the details about the model hyperparameters, the very low-level details are avoided to keep the flow.

4. Some descriptions and interpretation of the results are not very clear.

• In 257-262, I find it difficult to state that "…a substantial fraction of observed variability is retained under temporal extrapolation". If R2 is 0.35, it means that most variability (65%) is still unexplained. I would suggest rephrasing the sentence Additionally, you could consider explaining why R2 dropped. Does it indicate some level of overfitting in the training data or missing drivers that vary with time? In the following texts, while the authors claimed that the "model successfully capture the variability in many cases", it is also important to mention (and explain, discuss here or later on in Discussion) that Fig. 4c [Graswang in Fendt 2020 extensive] shows the model consistently overestimating N2O flux and generating a few tiny peaks (e.g., 2020-07 and 2020-09) that do not exist in observations. Moreover, I wonder whether L260-261 make sense, ie., why an overall moderately good representation of flux variability by the model would indicate higher peaks in Chamau than in other sites?

- We removed the word "substantial".
- The lower performance observed on the test set compared with the training set is expected, as the model parameters are optimized using the training data and therefore capture patterns in this subset more effectively. The decrease in R^2 from 0.51 to 0.30 indicates a moderate reduction in predictive performance on unseen data, but does not suggest an overfitting. This is particularly relevant for daily N_2O emissions, which are characterized by high temporal variability and episodic peak events that are difficult to predict. In addition, the model was intentionally developed using a limited set of widely available input variables. Several important drivers of N_2O emissions, such as soil mineral nitrogen content (NH_4^+ and NO_3^-), are not available across all sites and therefore could not be included. As a result, some loss of performance when applied to unseen observations is unavoidable.
  We corrected line 273-275: "The R^2 value representing the proportion of explained variance, was 0.30 for the temporal test set, lower than 0.51 of the training data (Table 5). This indicates that a fraction of observed variability is retained under temporal extrapolation. The reduction in R^2 indicates a moderate loss of predictive skill on test data, which is expected given the highly episodic nature of N_2O emissions and the intentionally limited set of input variables used in this study. "

- We have addressed the model's lower performance at Fig. 4c in the paper L270-L272: "Despite some performance degradation comparing to the training set, the preserved rank correlation shows that the model reliably captures the temporal ordering of emission events across years within the same soils in many cases, while it shows a lower performance at others such as Fig 4 c."

  Also 282:L283 "At Graswang, fluxes were more frequently overestimated (Fig. 4 c, d), showing a significantly lower performance and higher bias comparing to other soils."
- The higher emission peaks predicted for Chamau during the test period likely reflect the model's ability to learn the site-specific characteristics represented in the training data. Within the Chamau training subset, N_2O emissions were generally higher than those observed at many of the other sites. The model's ability to reproduce these differences in emission magnitude across sites suggests

that the selected input variables effectively captured key environmental and management factors controlling site-to-site variability in N_2O fluxes.

In L263 and Table 5 bias shifts from negative in training set to positive in temporal and cross-site validation. Is this common in ML, or does it indeed indicate a systematic shift? In the following sentence, why is bias only discussed in comparison with observed peaks? What is the annual mean flux?

- There is indeed a systematic shift in the data of Tereno, Germany in year 2020 used for testing comparing to previous years as we compared the mean median min and max values of emissions of this year with previous years. The reason is that from the last month of 2019, a more sensitive analyzer was used in Tereno which is more sensitive to lower fluxes. Therefore, the range of lower fluxes is lower compared to what the model has learned during training, causing a positive bias. We have corrected the sentence mentioned by the reviewer to ensure not to only concern bias of large fluxes, rather the general bias. See line 276-278.

• L275-277 I wondered if a negative R2 indicates something just as simple as what is explained. Does it indicate a systematic failure rather than a few outliers? Does your prediction consistently overestimate the flux? Now it is difficult to tell.

- Please refer the reply to your first comment, in which we explained that we add a more detailed analysis of R^2 values to the paper to better prove that the negative R^2 is not a result of a uniformly bad model, but only low performance at some site-years. We also mention that the very low value of R^2 is affected by the few extreme errors, but avoided to mention that they are the only reason for low R^2 value. Please see Section 3.2, L295:L307.

• (Maybe I didn't notice) It seems Figure 3 and 5 are not referenced in the texts.

That is correct! Thanks for picking that up. Unfortunately, they have been unintentionally deleted after the first review changes. We copied them back from the initial uploaded version.

- For Fig. 3 see L259
- For Fig. 5, Please see L314:L320: "In addition to analysis of model performance for daily emissions, the cumulative emissions were assessed. Figure 5 shows the comparison between observed and modelled annual cumulative emissions achieved by aggregating the available daily fluxes. The results show that the model creates an accurate estimation of annual emissions for some soils and in different years with a slight error rate, such as Rottenbuch in Fendt 2015 8 intensive and Rottenbuch in Fendt 2015, 2016, 2017 extensive. But, higher overestimation is evident for Chamau 2015 and Graswang in Fendt 2020 and extreme overestimation is evident for soils such as Rottenbuch in Fendt 2017, 2020 intensive (Fig. 5 a and b)."

• Is section 2.5 missing? Or should 2.6 be renamed to 2.5?

- This section was removed after the first review. We just corrected it.

• Lines 334-336 I think you want to emphasize that fertilizer application does not lead to peak emissions because there are other complex interactions with environmental factors. The sentences read a bit unclear.

- Thank you for this suggestion. We agree that the original phrasing was unclear. Our intention was precisely to highlight that fertilizer application requires specific environmental conditions to trigger peak emissions. We have revised the text (Lines 364–366) as follows: "For instance, fertilization alone did not consistently trigger immediate N2O peaks in this dataset. This aligns with the understanding that nitrogen availability must coincide with favorable environmental conditions—such as optimal soil temperature and moisture—to stimulate the microbial processes responsible for elevated fluxes (Butterbach-Bahl et al., 2013; Wang et al., 2021; Cui et al., 2024)."

• L339-342 These explanations are helpful. However, you only referenced figures that show fluxes without showing precipitation. Is it possible to include weather data (such as temperature and precipitation) and perhaps soil moisture in a composite figure in Appendix/Supplementary materials?

- That's a nice suggestion. We added precipitation in the new version so that the discussion about precipitation gets clearer. However, adding more details such as temperature and soil moisture makes the Figure overly congested and reduces the readability. Please refer to the Table A1 in appendix to get a better perspective about soil temperature and moisture at different soils. In addition, Table 1 shows precipitation and average air temperature for different grasslands.

• L353-355 There are both intensive and extensive grasslands, in which main differences are fertiliser application rates and cuttings. If the fertiliser rate has little variability and does not improve model performance, should cutting be included as a variable in your model? At least it should be discussed why it is not included/considered.

- Fertilizer rate has low variability within the sites in this dataset, but the occurrence of fertilization has a huge effect and causes elevated fluxes. We have added cutting events once, but it didn't improve the results. The reason is potentially the fact the in this dataset the cutting and fertilization events occur mostly at the same time. In lines L429:L430 we mention that further management practices could be added to the data in future work if possible.