The manuscript "Modelling N_2O 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 optimised.

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.

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.

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?).

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?
- 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?
- 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.
- (Maybe I didn't notice) It seems Figure 3 and 5 are not referenced in the texts.
- Is section 2.5 missing? Or should 2.6 be renamed to 2.5?
- Lines 334-336 I think you want to emphasize that fertiliser application does not lead to peak emissions because there are other complex interactions with environmental factors. The sentences read a bit unclear.

- 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?
- 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.