#Reviewer 1

General comments:

This manuscript takes on the important task of improving N2O emissions predictions from grasslands using machine learning, crucially through the use of a large multi-site dataset. This is significant development from the earlier work on this topic. However, several aspects of the configuration of LSTMs and the preprocessing of data require clarification or improvement in order to provide a fair and transparent evaluation of the model's performance. Specifically, I have concerns regarding potential data leakage in the train/test splitting strategy, and the ambiguity regarding the use of past target variable values as inputs.

These details may seem esoteric, but without them, it is impossible to discern whether the RMSE achieved here is impressive or disappointing. That being said, I feel that these changes can be implemented—or at least clarified—relatively easily, and are necessary to validate the study's conclusions. I also strongly encourage the authors to make their model code and data available via a public repository. This would resolve the interpretability issues raised in this review and ensure the reproducibility of the study.

I believe that with these methodological issues resolved, this study could make valuable contribution to the field of environmental modeling, with broad potential impact.

AC: Thank you for acknowledging the study as significant development and for your constructive comments and suggestions. We addressed your concerns regarding potential data leakage in the train/test splitting strategy. Your concern that any past target variable values would have been used as inputs are a misunderstanding, this was and is not the case which we clarify below. In order to ensure transparency, we now make the model and specifications available via a public repository given in Code and data availability section.

Specific comments:

L105-114: Equation 1 suggests that you have manually included four additional time-lagged features of temporal drivers for each of the prior four days. If I have correctly interpreted this, it is unclear to me why you have manually created these time-lagged features given that you are using an LSTM model. The LSTM model already contains memory cells, forget-gates, an attention layer, and other architectural features which remember past events and learn which are important. As an LSTM model already has access to previous time-steps, forcing the model to use manually-created ones may interfere with the model's attempt to learn and forget appropriately. Please clarify whether the 't-1' etc indicate manually included time-lagged features, and if so, justify your use of them. It would be helpful to explicitly state the shape of the 3D tensor used in the LSTM (N_Samples, Time_Steps, N_Features).

AC: We did not manually include time-lagged features to interfere with the LSTMs internal operations, indeed we had to do preprocessing. The Equation 1 is only explaining how the LSTM walks on the data sequences and what is needed to predict one days of N2O emission: When the sequence length is set to 5, the LSTM processes the input time series using sliding windows of five consecutive time steps, provided that the total sequence length exceeds this value. For each input window, the LSTM processes one shot of five time steps **independently** and updates its hidden state and cell state through the recurrent computations of the network. The input, forget, and output gates also function within a window of length 5 and regulate how information from the current input and the previous hidden state is incorporated, retained, or discarded in the cell state. The model could also be trained in batches. In this case, the model receives several input sequences (each of length five) in **parallel** during each training step. The forward pass is performed **independently** for each sequence within the batch, producing a set of predictions. The loss is then computed across all sequences in the batch and averaged (or aggregated) to obtain the batch loss. So, if you want to evaluate the model with new data for 1 day, you will need to enter the values mentioned in Eq. 1.

(N_Samples, Time_Steps, N_Features) in our data: (14500,5,4)→ in which 4 is the number of temporal features --> How LSTM looks into this regardless of the N-Samples: a length 5 time shot at each round.

we improve the description L109:L110 to provide more clarity: „To estimate the daily N_2O-N emission, the sequential LSTM model walks through the data sequences and in each round processes a subsequence of length 5 which include the information as Eq. 1 for estimating average daily fluxes: "

On the other hand, detailed preprocessing had to be applied on the daily data of each site to ensure we have continuous sequences of given daily value, which we explain further.

Firstly, we had to preprocess the large sequence of daily data for each site to remove the days including gaps in measurements of input variables such as soil moisture. This leads to smaller uncontentious subsequences. For instance, only 300 days of data rows for one year might be have measurements for all the input features, causing gap days in between.

Secondly, we had to consider the fact that there are also many days with gaps in N_2O emissions value. For instance, only for 100 days in one year the N_2O observation were available for a site. For each of these days, we needed the 5 subsequent days to be available after gap removal in first stage. Therefore, the sequences had to be preprocessed again to extract the sequences of length 5 for which we have one N_2O emission measurement and the input drivers for 5 subsequent days. Please see a part of the data in Fig. 2 below as an example of how the gaps in N2O emission dataset looked like:

date	Precipitation	Soil moist_10cm	soil Temp_10cm	Fertilization	n2o_ugram
2019-02-11 00:00:00	8.6	37.75759837	2.839139329		17.14227734
2019-02-12 00:00:00	0	36.94179706	2.937299357		
2019-02-13 00:00:00	0	36.33231574	2.831384935		
2019-02-14 00:00:00	0	35.9283916	2.731060942		
2019-02-15 00:00:00	0	35.66553495	2.836581746		
2019-02-16 00:00:00	0	35.45681966	2.904428867		
2019-02-17 00:00:00	0	35.28278148	3.043306248		
2019-02-18 00:00:00	0	35.12288128	3.136523853		
2019-02-19 00:00:00	0	34.9851447	3.355586143		90.06761842

Figure 2. a snapshot of dataset showing the gaps in the data. The green part of the sequence could be considered and used for LSTM, while the red subsequence is useless due to unavailability of the N_2O value.

Finally, after gap removal the smaller subsequences created for all sites had to be shuffled together, to ensure at each batch the LSTM sees the data from different sites. If you don't do this shuffle and just stack the sequences of all sites together, you'll see a much lower performance. This was a very important aspect that we discovered after investigations.

In conclusion, we didn't create individual time lags, we just had to shorten the big sequences of data to smaller length of 5 because of removing gaps and shuffling all sites.

L120-130: After a careful read, I have interpreted the authors' data splitting as follows. There are two subsets, one training and one testing subset. All of the Rottenbuch-in-Fendt data was siloed to the testing dataset, while all Neustift-in-Neustift data was used for training. All data from a single year of the Chamau site was withheld for the testing dataset, while all previous years were included in training. The rest of the data appears to be split at the treatment level, with all three replications and experiment years of a given treatment either sent to training or testing. This is somewhat different from the description given in the abstract, as the test set included not just a holdout site and withheld years as described, but also parallel years of separate treatments: "The trained LSTM model showed strong predictive performance (RMSE of 18 $\mu gm{-}2\ h\ {-}1$, and Relative RMSE of 270%) when evaluated on a test set that included both data from an independent soil and withheld years from training procedures." Importantly, withheld years and withheld sites make up only 320 and 259 data points of the total 14591 measurements in the testing set (~2% each).

I firstly want to commend the authors for clearly taking data handling seriously and devising a more rigorous method of splitting than a simple randomized split. I also appreciate that authors have provided a loose description of criteria and thought that went into dividing datasets into testing and training and their desire to represent a variety of soil, environmental, and management properties in both groups. That being said, I do still have concerns about the potential for data leakage given the methods described. Most crucially, I believe it generally best practice when working with sequential data to avoid having data from the same locations and time periods represented in both testing and training pools, even if there was a difference in management technique (e.g. extensive/intensive). As most experiments use one-factor-at-a-time, separate treatments typically experience mostly the same conditions at the same time, including weather and other factors which may affect biogeochemical systems and which may not be captured by model inputs (Zhu et al., 2023). Thus there is often a similarity in flux patterns over time, where factor differences may lead to a difference in peak height etc, but the overall shape of flux over time is similar. Thus when split across testing and training datasets, information about flux patterns and their drivers may leak from a parallel treatment, giving the model an unfair advantage during predictions.

In my view, how the authors have handled data from the Chamau site and Rottenbuch-in-Fendt are both strong examples of data splitting. From the Chamau site, the model had the opportunity to learn from previous years' data during training but was then evaluated on a separate year, with no opportunity to 'cheat' and gain information about events and effects that occurred during that year from other treatments. The Rottenbuch-in-Fendt is the strongest example of a true holdout testing the model's generalizability, as the authors point out. Arguments about the model's generalizability thus would be stronger based upon the model's performance on these two subsets. Aggregate RMSE and other metrics could be reported independently for withheld sites and withheld years, without the overwhelming influence of additional data included in the test set where parallel treatments are represented in training data. Authors may also wish to consider adding additional sites/years to these pools, given the small quantity of data handled in these manners.

AC: Considering your details discussion, also a similar concern about data leakage from other reviewers, we changed the strategy to apply your feedback. The train/test split is now ensuring to assess temporal generalizability using a hold-out set from Fendt-in-Fendt and Graswang-in-Fendt, in which all intensive and extensive treatments of them in 2020 are hold out from training, similar to Chamau. Cross-site test set to assess the spatial generalizability is Rottenbuch in Fendt as it was before. We corrected this in the paper in L126:L139

We now report error metrics separately on both the temporal and the spatial test sets in section 3.2 and table 5, and report it like this also in the abstract.

L140-142, L179-185: The use of log-transform and the loss function selected is an interesting combination of model techniques. I think it would be useful to provide a greater discussion on the effect of this combination on model behavior, given a non-statistical audience and a highly non-traditional loss term. The log transformation compresses the data distribution, forcing the model to minimize relative prediction errors on a relative, rather than absolute, scale. Then, the loss function penalizes prediction errors *exponentially,* placing a massive (approaching infinite) influence on any large errors. Thus the cost of a single large error will easily outweigh a thousand small errors. The result is a highly risk-averse model (avoids large errors at all cost), and which prioritizes consistency (always predicting within the same order of magnitude), in contrast to precision and accuracy on the bulk of data with poor predictions on outliers. It also likely forces the model to pay close attention to hot moments. It's an interesting and potentially effective strategy depending on the goals, but I feel the biogeochemist audience would benefit from more explicit discussion of these effects.

AC:

We chose this approach due to the following reasons:

(1) Applying the Log-transformation on highly right-skewed data is a crucial step before applying a model: The log transformation reduces skewness and stabilizes variance, resulting in better-behaved gradients, faster convergence, and more stable and effective training (West 2021; Bishop, 2006, Ch. 5; Hastie et al., 2009, Sec. 2.4; Box & Cox, 1964). In highly skewed data, rare extreme values dominate the gradient in gradient-based models such as neural networks, which is known to impair optimization and slow convergence in deep neural networks (Goodfellow et al., 2016, Ch. 8). Variance-stabilizing transformations such as logarithmic scaling are commonly used in recurrent neural forecasting models to address strong skewness and scale heterogeneity (Salinas et al., 2020; Smyl 2020). Without log-transformation, the extreme values create high gradients, dominate the training, and cause the model even not learn the relationship in the data but focus on the rare 5% of the extreme events only. Our goals is not to only predict large extremes, but learn the N2O emission pattern, shown in both low and high fluxes. We briefly mentioned the importance of log-transform in L150:L153

(2) Loss function: The log-transform helps training stability which is a must in this problem as discussed, but we still needed a proper loss function for the specific problem. Our goal is to have a model optimized to learn both high and low emissions, rather than only extremely high values. A weighted loss function ensures a trade-off between the target transformation for better convergence, and high performance of the model on high emissions. In addition, the exponential loss function is executed on log-transformed data, making it not only penalize large errors on high values, but also on low fluxes, which ensures the model learns both.

To ensure more details about this we add extra description in methods section as well a discussion: L339:L342

L164-175: LSTM models can be configured in three ways: (1) using ground-truth target values from previous time-steps as inputs (teacher-forcing or closed-loop), (2) using the model's own past predictions as inputs (autoregressive or open-loop), or (3) relying solely on environmental drivers without including any version of the past target variable as an input feature in the 3D tensor. I could not find any mention of whether some version of previous flux time-steps were included in the 3D tensor. This is a crucial detail, as flux is often highly autocorrelated, yet in practical scenarios, actual past flux values will not be known. Using teacher-forcing is thus appropriate during training to accelerate convergence, but inappropriate during model evaluation on testing datasets, as this would provide highly useful but practically impossible information. A discussion of these distinctions and how precisely the LSTM model was set up in both training and testing phases is immensely needed in order to assess the model's true predictive power.

AC: Thanks for spotting the lack of clarity in our description. Option 3 was used for this paper, relying solely on the environmental drivers in both training and test phases. We add this in lines L118-L119 „The implemented sequential LSTM models in this study rely solely on the temporal and static input features shown in Eq. 1 in both training and testing phases, and no information from the N_2O emissions on previous days is used for modelling the N_2O emission of the current day."

L222-224: I interpret this as you have evaluated two different hyperparameters for the n_steps/timesteps parameter of input_shape argument of the LSTM layer, those being 5 and 10. This sets a hard limit on the number of days your model can "look backward" at previous events. I am curious why such short timeframes were considered, especially given that a fertilization input may have an impact on soil nitrogen content and thus denitrification rates for many weeks, particularly if the fertilization happens to occur during a dry period. I note that you explain in the discussion that emissions nearly always peak shortly after fertilization, thus reducing the benefit of longer sequence inputs. However, I also notice from Figure 4 that moderate peaks often occur long after fertilization- much longer than the 10 days maximum evaluated. I am curious whether longer input sequences may improve predictions on such moderate peaks, for example if soil nitrogen is still moderately elevated a month after fertilization and a significant precipitation event occurs.

This also compounds my confusion regarding the aforementioned time-lagged features. The input_shape argument of an LSTM layer already defines the ceiling of how many previous time-steps the model is capable of considering, and attention and forget gates are used to determine which data among this pool to actually focus on or forget.

Thank you for the valid point. Initially, we started with sequence length 5, then we checked if longer sequences had an effect on model performance. Several points led to the decision to remain with shorter sequences in the publication:

(1) Changing the sequence length to 10 (and hyperparameter tuning) or 20 didn't bring significant improvements to the model so we didn't investigat longer sequences. We added one line to Table 4, so that the audience knows sequence length 20 was also tested. Please look at the Table 1 below showing how the model acts on samples of test set when we include a longer sequence of length 20:

```
=== LOG SCALE ===
                      R2  Spearman Spearman_CI 95%   RMSE    MAE   Bias
Dataset
Train              0.536     0.587    [0.58, 0.60]  0.733  0.604  0.063
Test               0.380     0.570    [0.55, 0.59]  0.693  0.542  0.352
Independent test -0.328     0.361    [0.34, 0.38]  0.715  0.607  0.401

=== REAL SCALE ===
                      R2  Spearman Spearman_CI 95%    RMSE     MAE   Bias
Dataset
Train              0.346     0.587    [0.58, 0.60]  43.570  12.469 -6.292
Test               0.417     0.570    [0.55, 0.59]  26.806   8.501  1.740
Independent test -0.154     0.361    [0.34, 0.38]  11.677   4.925  1.947
```

Table 1. Statistical metrics on a model trained on sequence length 20.

(2) Longer input sequences (length **X**) increase the amount of data required for the model to function effectively, which makes evaluation at a new site more challenging. This is because the LSTM requires **X consecutive days of input** to estimate the **N_2O flux for a given day**. As the required sequence length increases, the amount of usable data decreases due to gaps in input variables such as soil moisture and temperature. Consequently, we chose to keep the sequence length—and thus the model's data requirements—as small as possible to ensure robust performance. The potential benefits of longer sequences (e.g., **30 days or more**) will be investigated in future work using an expanded dataset.

we also added a statement in line L401:L403 in the Implications and limitations section to ensure using a longer sequence is known as a potential future work: „Firstly, expanding the dataset to encompass more geographically and climatically diverse sites, along with a wider range of management practices and legacy effects on N_2O emissions, would allow for the development of more generalized N_2O emission models for grasslands across different agroecosystems".

L235-236: I see that RMSE for the holdout set are given individually for timeseries in Fig 4, but I would appreciate an aggregate performance evaluation for the Rottenbuch-in-Fendt holdout listed here for comparison's sake.

AC: We separated now according to spatial and temporal test set and provide the performance evaluation in our updated table 5.

L269-271: Clarification is again needed here as to whether time-lagged features were directly represented as model features. If the importance of previous time-steps of features were determined by using standard PFI, which then varied the values of e.g. feature "soil_temperature_t-1" and ultimately assigned it an importance value, then this method would be invalid. In the case of using explicit time-lagged features, previous values of features are represented in the model in two places: the engineered feature, as well as the internal Hidden State of the LSTM. Even if the time-lagged feature were altered by PMI, the model would still have access to the true value in its hidden state, and thus the model could simply ignore the shuffling by PFI.

If this is a misinterpretation and time-lagged features are not explicitly represented, then I would like to see more detail on how PFI was adapted to assess the importance of previous time-steps given the LSTM architecture.

AC: We addressed this misinterpretation in the reply to your other comment above, time-lagged features are not explicitly represented to LSTM.

To calculate PFI for individual time steps: we got each sequence of 5 days generated with preprocessing, and shuffled the features only in that sequences. For instance, in each sequence of length 5 we only shuffled the precipitation within that single sequence, which reflects shuffling the time steps. So, if you had precipitation only in first day for 18mm and not in other days, then this value will move to another day randomly, for instance day 3 and in this way we could see how the models estimation on this sequence change and get the relative importance from PFI.

After getting comments about the PFI for time lags, and considering different opinions on this topic between data scientists which doubts whether this is a proper approach or might cause misunderstanding about model's results, we decided to remove this approach and solely rely on Fig. 6 as an standard way which the PFI is calculated in other scientific publications. We believe that such a method needs very wide discussion about all the aspects and the implementation method, which its out of the scope of this paper.

L349-353: Given that no process-based model was tested against the LSTM model, I question the need for this whole section of comparing with process model, which is not the scope of this study. I am skeptical that any meaningful comparison of predictive performance metrics can be made across different datasets, especially given the extreme temporal and spatial variability of N2O which translates to major variability in data distributions. Some datasets are simply easier to predict due to lesser influence of hot moments, and thus comparisons of models evaluated on different datasets should not be made without at least acknowledging their tenuousness.

AC: We removed the section 4.4 according to your feedback and the feedback from other reviewers.

L361-365: Again, considering that data exists for process-based model predictions at the same site (and year?) as one of the modeled sites in this study (Chamau), I would prefer that authors limit their comparisons of LSTM vs process-based simulations where these comparisons can be made fairly.

AC: We removed the section 4.4 according to your feedback and the feedback from other reviewers.

Technical corrections:

L24. Spell out full word 'approximately' Done.

L30. '2' subscript Done.

L237: "Having relative RMSE higher than 1 is a known effect of the imbalanced dataset." Citation needed.

AC: We removed RRMSE, as we got many feedbacks and investigated further that its not a proper metric for this problem with highly skewed data. We instead use a set of metrics in Table 5 to better interpreted the performance: section 3.2.

Other minor comments:

Table 1 and 2: Dataset size - clarify if this is daily flux observation numbers

Done.

L81: Better to say "flux" and not "flux rate" Done.

Fig 1: It would be helpful to include flux distribution for the test data dataset too. Done, please see Fig 1.

Table 5 and associated discussion: The table does not need to be in the main manuscript and can be presented as supplementary data.

AC: We actually provide the details to those related literature there to avoid huge amount of text. One would need the paper to follow the studies.

L297-303 sounding like the authors are criticizing the previous studies and can be shortened. These are some of the initial studies on this topic and should be acknowledged accordingly with limitations that the current study is addressing.

We improved this section, L316:L325

Fig 6: While feature importance was assessed separately for each site in Fig 6, the discussion on any variation in variable importance across sites is weak and contributes little to the understanding of differential N2O dynamics. Also, why time-lagged features were not included in Fig 6 analysis.

Please look into section 4.3 feature importance analysis, we added few more lines on this. Time lagged features were not included in Fig. 6 as this is the standard and general approach for PFI to be implemented on the features, regardless of time-lags. Fig. 7 which used to include time-lags due to the reasons mentioned in previous comments.

#Reviewer 2

This study evaluates the application of LSTM networks to model spatial and temporal N2O emission dynamics across montane grasslands in Central Europe. By training on multi-site data including meteorological and soil drivers, the authors assess the model's ability to generalize to independent sites, concluding that the LSTM approach offers predictive performance comparable to established process-based biogeochemical models. My overall feeling about this manuscript is "negative", but still feel it could make some contributions to the body of literature. As such, I recommend a major revision.

AC: Thank you for your time to review this manuscript. We adjusted our manuscript regarding your comments and suggestions. The main updates based on your concerns are changing the train and test split, rephrasing the model from generic to regional and restructuring the literature review paragraph. Please find the detailed answers below.

Major comments:

[1] Inadequate Contextualization and Benchmarking. The authors haven't acknowledged the state-of-the-art in GHG modeling, such as various advanced pure DL approaches, or those combining process-based model with DL typically known as Physics-Informed Neural Networks (PINN) and Knowledge-Guided Machine Learning (KGML). By omitting these latest advancements, the introduction falsely framed standard LSTM as a novel frontier. Furthermore, the benchmarks in Table 5 appear to be low-performance baselines. To avoid the appearance of cherry-picking, the authors need to validate their model against advanced contemporary literature rather than outdated or weak standards.

AC:

The novelty of the study lies mainly in considering real data of several sites at different elevations and kilometers apart in one model. The use of LSTM is valid because it is a deep learning model with the ability to learn complex sequential data and we had enough data to train

it. We agree that we could provide a broader introduction to the advances in deep learning models that are applicable to this question. We have thus now updated our manuscript to include KGML example in lines L316:L325 as well as table 6 and L418. To facilitate a constructive review process, we kindly ask the reviewer to specify the study or studies they are referring to. To the best of our knowledge, no comparable multi-site application for N_2O has been published to date. We are, of course, happy to include additional studies for comparison if appropriate references are provided.

[2] Over-Ambitious Scope and Insufficient Performance. The goal of developing a "generic model" using only 122 site-year combinations is over-ambitious and unsupported by the results. The reported Relative RMSE of 270% suggests the model error is nearly three times the average observed emission, contradicting the claim of "strong predictive performance." This poor performance likely stems from insufficient training strategies or architectural limitations along with data noise. Realistically, the authors should reframe this as a regional case study and objectively analyze why the model failed to generalize better, rather than over-claiming its predictive utility. In that case, I would still treat this as a valuable study that documents "what works well or not, by how much" in the surge of AI applications.

AC:

According to your feedback, we worked on the topic of model performance metrics further and decided to remove the RRMSE due to existance of a highly right skewed dataset and the feedbacks from reviewers. RRMSE or R^2 are not enough metrics for performance in a highly skewed dataset, since division by a low observed mean value penalizes performance heavily so we should combine several metrics to create a proper interpretation about the performance. we have added new statistical metrics, as shown in Table 5, to improve the analysis of the results. Using the Spearman correlation metrics, we demonstrate that the model captures the pattern in daily fluctuations well and is not overfitted. Please refer to section 3.2, 'Model performance'.

Additionally, we discuss the model's weaknesses and the reasoning behind them in more detail in Section 4.2 in the discussion:L349:L360

„Although the model captures daily fluctuations in many cases, it shows error in other cases such as estimating the magnitude of very low fluxes at Rottenbuch in Fendt extensive in June 2016 (Fig. 4 f). In general, the model captures ranking structure well as shown by the spearman metrics in Table 5 but in some cases it shows overestimation of low fluxes or underestimation of high fluxes as reflected by the R^2 score for the Cross-site test set.“[3] Lack of Technical Reproducibility. The methodology consumes excessive space describing rudimentary ML concepts ("ML 101") while omitting essential technical specifications. Critical details regarding feature engineering and model architecture—specifically how the attention mechanism was implemented within the LSTM—are missing. The authors need to replace generic descriptions with the precise technical details required for reproducibility.

AC: We shortened the section about the model architecture B, instead explain the attention further and improved Fig 2 including more details: section 2.4.1.

We explained feature engineering in L140-L154, mentioning that the decision about the correct scaling method is made after investigating several other feature scaling such as standard scaler and min-max scaler considering the characteristics of different input features. To stay within the scope of the paper, however, we did not go into detail about different experiments with other scaling methods, instead providing the same level of detail as in other publications.

Specific comments:

L15, "withheld years from training procedures": why? likely a data leakage issue.

AC: The goal is here to assess the temporal generalizability. Please find our new validation strategy and train-test split which ensures avoiding all potential data leakage issues in the manuscript as well as in the answers to reviewer 1: In the updated version we have a clear definition of a Spatial Generalization Test Set (Rottenbuch, intensive as well as extensive lysimeters) and a Temporal Generalization Test Set (Chamau hold-out, an additional Fendt and Graswang hold-out year): L126:L140

L35: need to update the 2013 reference because there're already several recent ML/DL advances for N2O simulations. I know Klaus Butterbach-Bahl is aware of those, but not sure if he had a recent review paper reflecting the state-of-the-art.

AC: We also used Wang et al., 2021, nevertheless we keep the Butterbach-Bahl Reference since it provides a good basis. We hope this addresses your concern.

L43: the introduction should review existing ML-based literature and disclose strength and weakness, rather than discussing the opportunities

AC: We compare our work to the existing ML literature in the discussion section, and think that the introduction will get lengthy if we already discuss different ML approaches there.

L65: plz add information about when the data was collected in Table 1

AC: To keep the table short we provide the references to detailed information on sampling in the Table caption. (Kiese et al., 2018; Soltani et al., 2018; Gilgen & Buchmann, 2009; Roth, 2006; Hörtnagl et al., 2014)

L100: the data split for Chamau is problematic. 2015 is between training years which makes "test set" not a real out-of-sample test.

AC: In case of this problem, the N2O emission values are only predicted based on short-term climate and soil properties and not the previous N_2O values in past periods, therefore all years are independent and choosing a in-between-year as a hold-out test set wouldn’t make any issue.

The reason we didn’t select 2019 and 2020 as test samples for chamau, is that these years had dryer climates comparing to other years, causing specific patterns in the emissions which was important for the model to learn them. For instance, these years showed peaks after fertilization which happened not at the few days after fertilization but later. Such a pattern is rarely visible in German sites and older years of chamau, which avoids the model to receive such info during the training. In time series analysis, maintaining the temporal order of data is crucial. However, if the final year includes atypical events (e.g., extreme weather conditions due to climate change), using it as a test set might not provide a realistic assessment of the model's performance under normal conditions. In such cases, selecting an earlier year for testing can offer a more representative evaluation. In this way models learns the climate change affects.

L117: sorry, you can't claim generic model here. 6 sites only occupies a very small fraction of the high-dimensional feature space that jointly determines N2O production.

AC: Well, our intention was to set the contrast to site-specific models. You are right that generic model is a too broad term, thus we changed the term to ‘regional model’.

L123-125: It's hard for readers to follow this narrative. consider visualizing by a decision tree or a separate table apart from Table 2 that mingled soil information with data split

AC: We improved the train test split, also we rephrased the text for better understanding: please see L126:L140

L128: personally I feel it's meaningless to compare LSTM for gap-filling, which may be workable but not the best choice; instead, you could use "Bi-LSTM" that captures the contextual information of a whole time series

AC: We Also tested Bi-LSTM, but the performance was worse compared the current LSTM model.

L163: section 2.4.1 need to be significantly expanded to allow a full evaluation, such as the dimension of the LSTM and Dense layers and what was the output of the 1st "Dense layers", and how attention was implemented in the pipeline

AC: We expanded the section with more details. Also removed architecture B according to your comment.

L185: in eq.(4), with "loss" term also appear in the exponential power, this weighted loss can be quite unstable.

AC: Exponential weighting functions may introduce numerical instability if poorly constrained. In our study, we ensured training stability: model inputs were scaled, the peak-weight parameter was conservatively bounded (1-1.5), and training convergence was carefully monitored. No gradient divergence or optimization instability was observed. The exponential weighting therefore functioned as a controlled mechanism to reduce bias toward the mean and improve the representation of extreme events, which constitute a primary objective of the study. Please see this plot showing the stable training in progress until 100 epochs of training.

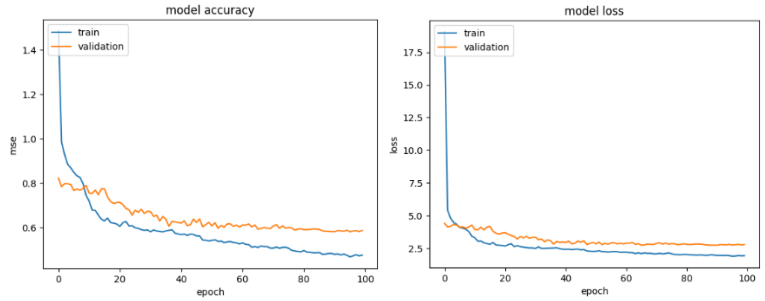


Figure 1. Training and validation progress during 100 epochs of LSTM training

L200, Table 3: what's the point to test "0.01" and "0.05" for dropout?

AC: The changes of dropout from 0.01 to 0.05 changes the models ability in capturing very low fluxes depending on other hyperparameters such as batch size. One leads to a flat line threshold on the low-value fluxes and won’t learn them, and the other learns low fluxes better. The value of dropout and is influence had to be configured with other hyperparameters.

L218: If this is true, you don't need to report architecture B, which is not a normal structure that one would like to test.

AC: We reduced the details about architecture B and added more details about the main architecture: please see Section 2.4.1

L233: in this section, RMSE alone is insufficient to describe the model performance, at least add R2

AC: We added the metrics R2, Spearman’s correlation, RMSE, MAE and bias in Table 5, and described them in section 3.2

L255: Figure 3 suggests the model has a low performance, but given the limited information given, it's hard to say whether it is because of low training skills or missing key features or the model simply didn't work for this particular dataset

AC: We expanded the performance metrics to describe the performance in more details. Please see Table 5 as well as section 3.2

L274: time series plot for Chamau and Fendt is impressive, but the authors need to convince people they were not cherry-picked and Chamau is not because of data leakage.

AC: We changed the train test split based on the comments we got from all reviewers in L126:L140. From each Fendt, Graswang the last year is hold out from all of their treatments. The fact that you consider the timeseries plots 'impressive' while error measures are not fully reflecting this impression is now better addressed using a combination of performance metrics which could better interprete the results in section 3.2 which supports our argumentation that the model performance is quite promising.

L295: this study used a relatively larger dataset, which also means more noise. As a result, the current LSTM structure is insufficient to fit a "generic model" that works for multiple sites, soil and managements. This problem is much more challenging than the authors written or expected, so I won't blame the low model performance, but claiming this study has solved the problem or even reached an acceptable milestone is misleading.

AC: We reframed our phrasing in the introduction, and in other section: now we refer to a 'regional model'.

L346: section 4.4 is redundant in this manuscript

AC: We removed section 4.4 according to your feedback.

Reviewer 3

General comments

RC: This is a very interesting study that tackles the difficult task of simulating N2O emissions using LSTM models across European grasslands. I think that the dataset compilation is a valuable contribution to the community. To help this manuscript reach its full potential, I have identified a few key areas where the approach should be strengthened regarding data aggregation, validation strategy and data transformation. In addition, the Discussion would benefit from a stronger connection between findings in this study with existing literature. I believe this extra effort is necessary to prove the robustness of the presented approach and will help this paper to make a significant contribution to this field. For this reason, I recommend major revisions.

AC: Thank you for your detailed and constructive review. We appreciate your feedback and implement your suggestions. Specifically, regarding the validation strategy (data split): We improved the train/ test split to ensure there is no doubt of data leakage. Regardig data transformation: we replied to your comments about log-transformation and loss function to clarify themotivation and correct implementation. Regarding model performance analysis, we added new metrics regarding the model performance and made sure we report specifically. We refined the discussion by strengthening the context of our findings with the results of existing publications. We hope that these improvements and revisions will meet the high standards of Biogeosciences and provide an interesting, publication-ready manuscript.

Major comments

Aggregation to daily values

L70-97: The manuscript states that sub-daily flux measurements (from both automatic chambers and eddy covariance) were "aggregated into daily values". However, the specific method of aggregation and the quality control criteria for this aggregation are not defined here. In addition, I am missing info about which data were aggregated (gap-filled or non-gap-filled). Especially for the eddy covariance (EC) sites Chamau and Neustift this would be crucial information for reproducibility, since half-hourly EC data are frequently rejected during quality control which often leaves days with very few valid data points. Then, calculating a daily value from only a few sparse measurements (which may be biased towards specific times of day) can introduce significant errors. It is therefore currently unclear if the daily training targets represent true daily averages, or if they are biased artifacts of incomplete data.

I therefore suggest that the authors please specify the aggregation method and give more detailed info about the used time series (gap-filled or non-gap-filled) that were aggregated. In case non-gap-filled time series were used, the minimum data availability threshold required to accept a day as valid should be given. If days with low coverage were included, please discuss the potential bias this introduces to the model. If gap-filled time series were used for the aggregation, the used gap-filling models should be given.

AC: For N2O, due to its sporadic nature gap-filling using temporal interpolation or look-up tables can be misleading, since preceeding fertilizer events are usually not considered and the period before a fertilizer event can be errorously filled with too high values. We wanted to avoid comparing ML model results to modelled data, which seems to be the case if larger gaps in the sub-daily timeseries were filled with a more sophisticated empirical model, we prefer to compare the ML model only with direct measurements. Since the investigated N2O fluxes showed no diurnall cycle we consider this approach justified and implemented it for the Tereno data which we directly processed. We added the dra QA/QC was performed by the experimentalists publishing the original data. Due to differences in how we received the data from experimantalists, the approach of aggredating sub-daily to daily data differed across sites.

1. For Chamau, Switzerland: EC post-processing was done using EddyPro, and only fluxes labelled with quality flags 0 and 1 were included (See Fuchs et al. 2018; For details section on QA/QC and post-processing). Days with at least

30% mesurements were used. Diurnal cycles were absent, thus an averaging to daily values without previous gap-filling was justifed.

2. For TERENO, Germany: Experimentalists went through data collection: The robot places the chamber every 15 minutes on a lysimeter. However, if the wind is too strong, the robot cannot lower the chamber correctly, so that it may go into park position at high wind speeds. There is a higher absolute measurement frequency in Graswang compared to Fendt.

   We received the data which included up to 6 measurements per day and we included all days with at least 2 out of 6 possible measurements on that day.

3. For Neustift, Austria: EC post-processing and gap-filling of sub-daily data was performed by Hörtnagl, L., & Wohlfahrt, G. (2014): „Gaps less than or equal to 2 h were filled by linear interpolation. For the filling of larger gaps a lookup table was generated, using flux data in a time window of 14 days around the missing flux value and Tsoil bin widths of 1 ◦C.“ We used this data that was gap-filled and then aggregated to daily values.

discussion section 4.5 Implications and limitations : L398-L400„Additionally, the number of available sub-daily flux measurements varied across days in the TERENO and Chamau non-gap-filled datasets due to gaps resulting from quality control procedures and instrumental failures (Kiese et al., 2018; Feigenwinter et al., 2023). Such uneven data availability can introduce bias into flux estimation modelling **(Lucas-Moffat et al., 2022;** Teng et al., 2020).

Input variables

L103-114 and Table 4: The model looking back at drivers of the past 5 days is likely too short to capture the causal drivers of emission peaks, because the soil is potentially primed by an event long before the emissions actually happen (legacy effect). For example, if fertilization happened 6 days ago then the model cannot see it and it assumes e.g. "rain and no fertilizer" in its time window. Therefore, a longer sequence would be more appropriate (e.g. 30 days or longer). One architecture was tested using 10 days, but this might also be too short. Maybe the authors have already tested longer windows (10+) and can extend Table 4 with this additional information. It is still possible that the shortest window works best, but it is important to see this in comparison to other tests.

Thank you for the valid point. Initially, we started with sequence length 5, then we checked if longer sequences had an effect on model performance. Several points led to the decision to remain with shorter sequences in the publication:

(1) Changing the sequence length to 10 (and hyperparameter tuning) or 20 didn’t bring significant improvements to the model so we didn’t investigat longer sequences. Please look at the Table 1 belowshowing how the model acts on samples of test set when we include a longer sequence of length 20:

```
=== LOG SCALE ===
                      R2  Spearman Spearman_CI 95%   RMSE    MAE   Bias
Dataset
Train              0.536     0.587    [0.58, 0.60]  0.733  0.604  0.063
Test               0.380     0.570    [0.55, 0.59]  0.693  0.542  0.352
Independent test -0.328     0.361    [0.34, 0.38]  0.715  0.607  0.401

=== REAL SCALE ===
                      R2  Spearman Spearman_CI 95%    RMSE     MAE   Bias
Dataset
Train              0.346     0.587    [0.58, 0.60]  43.570  12.469 -6.292
Test               0.417     0.570    [0.55, 0.59]  26.806   8.501  1.740
Independent test -0.154     0.361    [0.34, 0.38]  11.677   4.925  1.947
```

Table 1. Statistical metrics on a model trained on sequence length 20.

(2) Longer input sequences (length **X**) increase the amount of data required for the model to function effectively, which makes evaluation at a new site more challenging. This is because the LSTM requires **X consecutive days of input** to estimate the **N_2O flux for a given day**. As the required sequence length increases, the amount of usable data decreases due to gaps in input variables such as soil moisture and temperature. Consequently, we chose to keep the sequence length—and thus the model’s data requirements—as small as possible to ensure robust performance. The potential benefits of longer sequences (e.g., **30 days or more**) will be investigated in future work using an expanded dataset.

we also added a statement in line L401:L403 in the Implications and limitations section to ensure using a longer sequence is known as a potential future work: „Firstly, expanding the dataset to encompass more geographically and climatically diverse sites, along with a wider range of management practices and legacy effects on N_2O emissions, would allow for the development of more generalized N_2O emission models for grasslands across different agroecosystems“.

Dataset splitting

L120-130, Data leakage: The manuscript describes a data splitting strategy where parallel treatments (e.g., intensive vs. extensive management) from the same site and year are separated into training and testing sets. However, this approach introduces significant data leakage because the parallel treatmens experience identical daily meteorological drivers. Basically, the model is exposed to the (specific) weather sequences of the test period during training. I would assume that performance metrics on the test set likely reflect the ability of the model to memorize these weather patterns rather than the true ability of the model to generalize. A test for generalization would require the test data to be independent in time (unseen years) or space (unseen sites).

I therefore suggest that the authors restructure their validation strategy. This could be done by withholding entire years across all treatments or by withholding entire sites (as was done with Rottenbuch) from the training set.

RC: Thank you for your feedback on our validation strategy and train-test split. We restructured the concept and addressed your concern about data leakage. In the updated version we have a clear definition of a cross-site Test Set for independence in space (Rottenbuch, intensive as well as extensive lysimeters) and a Temporal Test Set for independence in timeWe specify this in the paper in L126-L140.

L125/L130, Train/test splitting:

In Section 2.3, the reported train/test split (13000 training vs. 15000 testing) is highly unconventional for machine learning and risks underfitting. It is unusual for a test set to be larger than the training set, I think I have not seen this in other studies yet. Unless there is a very specific requirement for these unusual data sizes that I am not aware of (and a proper justification), this needs to be addressed in an updated version of the manuscript.

RC: The reason for a huge test set is cosndiering a cross-site test that has many years of data in total 8690 data rows. Our new train-test split results now in a more conventional split (14500 for training vs. 8690 for cross-site testing and 3160 for temporal testing).

Log-transformation of N2O data and loss function

L140-157, L178-189: The text describes that daily N2O fluxes were log-transformed, meaning that the model ouputs predictions on a log scale which are then back-transformed. This transformation seems to be somewhat in conflict with L180 that emphasizes the "importance of catching high peaks". To me it seems counterintuitive to first compress extreme values/peaks (log transformation), but then applying a custom exponential loss function that basically attempts to re-weight the peaks. By compressing the target variable, it incentivizes the model to focus on fitting the "background" noise of low magnitude emissions rather than learning the rare, high magnitude events that drive the N2O budget. Therefore, the ability of the model to predict timing and magnitude of emission peaks is limited. Regarding the loss function, as pointed out the penalty grows exponentially as the error increases. If used in a dataset with rare high peaks this makes the model highly unstable during training.

I am sure that there were good reasons why the authors choose this approach, and I therefore suggest that the authors add their reasoning to the text. If the authors want to reconsider their approach and aim for optimizing the predictive power for peak events, I recommend to train the model on the linear N2O data in combination with a loss function that is robust to outliers (to improve stability).

AC: We chose this approach due to the following reasons:

(1) Applying the Log-transformation on highly right-skewed data is a crucial step before applying a model: The log transformation reduces skewness and stabilizes variance, resulting in better-behaved gradients, faster convergence, and more stable and effective training (West 2021; Bishop, 2006, Ch. 5; Hastie et al., 2009, Sec. 2.4; Box & Cox, 1964). In highly skewed data, rare extreme values dominate the gradient in gradient-based models such as neural networks, which is known to impair optimization and slow convergence in deep neural networks (Goodfellow et al., 2016, Ch. 8). Variance-stabilizing transformations such as logarithmic scaling are commonly used in recurrent neural forecasting models to address strong skewness and scale heterogeneity (Salinas et al., 2020; Smyl 2020). Without log-transformation, the extereme values create high gradients, dominate the training, and cause the model even not learn the relationship in the data but focus on the rare 5% of the exterem events only. Our goals is not to only predict large extremes, but learn the N2O emission pattern, shown in both low and high fluxes. We briefly mentioned the importance of log-transform in L150:L153

(2) Loss function: The log-transform helps training stability which is a must in this problem as discussed, but we still needed a proper loss function for the specific problem. Our goal is to have a model optimized to learn both high and low emissions, rather than only extremely high values. A weighted loss function ensures a trade-off between the target transformation for better convergence, and high performance of the model on high emissions. In addition, the exponential loss function is executed on log-transformed data, making it not only penalyze large errors on high values, but also on low fluxes, which ensures the model learns both.

(3) Training stability: Exponential weighting functions may introduce numerical instability if poorly constrained. But, to ensure stability, model inputs were scaled, the peak-weight parameter was conservatively bounded (1-1.5), and training convergence was carefully monitored. No gradient divergence or optimization instability was observed. The exponential weighting therefore functioned as a controlled mechanism to reduce bias toward the mean and improve the representation of extreme events, which constitute a primary objective of the study.

(4) The loss function in our study was designed to emphasize rare high-magnitude flux events, which are systematically underweighted by conventional losses such as MSE or MAE in highly right-skewed datasets. MSE, MAE, Huber loss, weighted Huber and weighted MSE were also tested, which all understimate high peaks and/or overstimated low fluxes with a flat line threshold.

(5) The use of standardization and normalization in combination with different loss functions such as MAE, wighted MSE, Huber loss as well as the custom loss was tested. Indeed huge amount of experiments using approaches which theoritically seem to be a proper fit for this problem where tested. Finally, the investigations showed that log-transform was the best approach, also underpinned by all the aspects we discussed above in 1).

An example of another study which used log-transformation and a weighted loss function on highly skewed data was Ghosh et al., 2025: "Short-term precipitation prediction in India using ConvLSTM", which used weighted mean squanred error loss function.

LSTM

L163-175: In its current state, Section 2.4.1 lacks technical specifications that would enable others to reproduce the model. The parts in the text that mention "a set of LSTM layers" and "several dense layers" are missing details, e.g. the specific layer dimensions (units per layer) used in the final model. Similarly, the "attention mechanism" should be explained, at least briefly, even if more details are given in the provided reference. Also, attention can be applied in different places in the pipeline and there are multiple ways for its implementation. These details, among others, are currently missing, but required for reproducibility. I also suggest to add some of these details to Fig. 2.

AC: In order to increase clarity, we now put the technical specifications of the LSTM also in the method section 2.4.1. In the previous version we had put it only in the result section: section 2.4.1 LSTM model architecture

. To ensure reproducibility and full documentation, we put the code to Github, and the link could be found in the Data and code availability section. This code includes the loss function, and the LSTM model and specific values of hyperparameters which are finally selected for the model.

Hyperparameter tuning

L190-200: I appreciate showing the ranges used for hyperparameter tuning, but this section also needs to show the optimized values that the authors decided to use in the final model. In addition, in this section I have another concern regarding validation. From the description it is not clear which data were used for validation, so I assume the test set was used instead of a validation set. Section 2.3 only mentions a test set. A typical split for train/validation/test split would be e.g. 60/20/20. To my knowledge, optimization algorithms like Hyperband require a validation metric to compare the differrent configurations. If the test set was used for optimization, this would be a "training on the test set" error (it would mean the model has seen the "answers" to the tests it is taking), a form of data leakage. If the training set was used for optimization, the hyperparameters are likely overfitted to the training data. And finally, the "manual adjustments" are in my view difficult to justify after the optimization procedure, because it means that the model was optimized, but optimal parameters were then not applied. Generally, manual tuning based on observing model output introduces researcher bias and implies that the model was tweaked until it *looked* right on the evaluation data. Regarding the ranges for tuning: it is important that the authors justify the specific ranges chosen and describe e.g. why the search space for neuron count was so constrained (it is very narrow and specific).

AC: (1) After the train/test split which is mentioned in section 2.3, the test set is completly hold out of the training, the training set is partitioned further to train/validation with 80/20 ratio. So, no info from the cross-site or temporal test gets into training According to your feedback, we add a short and clear explanation about it in L 130-132: „During the model training process, the training dataset partitions into training and validation parts following an 80:20 split as a requirement for machine learning model training. The training partition is used to fit the model, while the validation partition facilitates hyperparameter optimization—such as learning rate decay—and helps monitoring model's progress during training."

(2) Regarding manual hyperparameter tuning: The hyperparameter search space and manual adjustments are required to ensure reproducibility. Importantly, adjustments were performed exclusively using the training and validation datasets to avoid information leakage, and final model performance was assessed on an independent test set. HPOs provide a great opportunity toward finding a good model, but none of them are perfect (Raiaan et al., 2024). Manual adjustments, guided by domain expertise and iterative experimentation, can further refine model performance by incorporating nuanced understanding of the problem context and model behavior. Manual adjustments included small changes in dropout, learning rate, loss function coefficient and batch size. For instance, we changed dropout from 0.05 to 0.03 or batch size from 120 to 100 to see the potential improvements. The value of 0.03 was not in the HPO search space, but as 0.05 was in one of the best configurations returned by HPO, so we tested the values around it like 0.03 and 0.07. The number of neurons and layer regularization were retained as determined by Hyperband. Without these adjustments, the model tended to generate predictions around center of distribution that yielded lower statistical error metrics but failed to accurately represent abrupt increases or decreases in emissions, which constituted an important objective of N2O emission estimation. Ultimately, model development required careful trade-offs among the loss function design and weighting, batch size, learning rate, and dropout. Compared with typical modeling tasks, this process demanded substantially more experimentation, including the evaluation of alternative data-scaling strategies.

(3) Regarding narrow neurons size: the ranges of hyperparameters to go through hyperpatameter tuning is a decision based on expertise. We also tested a higher number of neurons in few experiments, but that didn't increase the performance and the search space for the number of neurons was then narrowed for HPO algorithm and higher number of neurons were excluded from the rest of experiments. For other hyperparameters, the ranges are best practice values which are always tested for LSTMS.

Feature importance

L202-209, Fig.6, Fig. 7: PFI is used to assess the importance of drivers at individual time steps. However, this is methodologically flawed for highly autocorrelated time series such as soil moisture since the value at time t is nearly identical to t-1. So by shuffling only the value at time t the model can recover the information from the unshuffled values at t-1 (or t-2 ...) stored in the memory of the LSTM. Consequently, the importance score of the respective feature is reduced and continuous drivers might be undervalued in comparison to e.g. fertilization. Therefore, I am not sure Fig. 7 can work because the lags are not independent. For Fig. 6, it is not immediately clear to me why the lagged variables are missing. I assume that maybe all time steps of a single variable were shuffled together, and then an overall importance is given (sort of a "grouped" PFI based on grouped shuffling)?After getting comments about the PFI for time lags, and considering different opinions on this topic between data scientists which doubts whether this is a proper approach or might cause misunderstanding about model's results, we decided to remove Fig. 7 and solely rely on Fig. 6 as an standard way which the PFI is calculated in other scientific publications. We believe that such a method needs very wide discussion about all the aspects and the implementation method, which its out of the scope of this paper.

In Figure 6, shows if we initially shuffle all the values of a variable, regardless of lags and then execute the model on it, so what you mentioned is correct, exactly the overall importance of that variable is considered.

Model performance

L230 Table 4: The relative RMSE of 270% is very high, and one of the reasons could be that the log transformation distorts the optimization. The log transformation minimizes the (logarithmic) error, but after the results are back-transformed to the linear scale the bias on low-flux days becomes disproportionately large (relative to the mean). Also, I assume the errors on peak events increase dramatically. The authors should compare this performance against e.g. the predicted mean of the training set for every day and then check whether the LSTM can outperform this naive baseline.--> we added R^2 for this case.

In light of the high RMSE, it is important that the authors discuss why the error is so high, especially if it is driven by a few very high outlier values, or if the model is systematically biased. A scatter plot of predicted vs. observed fluxes on the linear scale (not log-log) can help here to diagnose this. Currently, I cannot see how the model is "capable to transfer results across sites".

AC: The RRMSE for N2O predictions usually largely depends on the dataset. Since the RRMSE divides by the site mean, a low divisor creates a high RRMSE. Due to the sporadic nature of N2O which creates a highly skewed distribution of N2O values, the reported RMSEs and also higher values than usually reported for predictions of other greenhouse gases like CO2 (and CH4). Only few extreme errors combined with very low mean of a highly right skewed dataset highly increase RRMSE and R2 scores, thus the model's performance can not be judged only by these metrics. Therefore, also by considering your later comment on the fact that relative RMSE is probably not a good metric for this dataset, and the comments of other reviewers in requiring more metrics, we improve our model performance anaylsis by reporting several metrics which could help better analysing the results. Please see Table 5 in section 3.2 model performance for detailed result analysis on the train set, temporal hold out test set and cross-site test set.

Regarding log-transformation and specifically its importance on highly right-skewed datasets to stabilize training, we refer to the answer earlier in this document.

Specific comments

L10/L61: The site Neustift is described as "North Austria". Please correct this to "Western Austria". Corrected.

L25: Change "greenhouse gas" to "anthropogenic greenhouse gas". Done.

L30: The sentence "The Pearson correlation..." presents specific results in the introduction ("between -0.11 to 0.16 for our dataset"), which should not be done. I suggest the authors rephrase this sentence and remove these specific calculation outcomes, they have their place in the Results section.
We removed this from the introduction and added to the results, L223-L225

L33: Change "long periods" to "long time periods"
Done!

L56: I would not go so far to mention "general applicability". One independent grassland is not enough for this claim. Changed to „regional applicability "

L62: "no grazing": Some of these sites have grazing. If the authors are refering to specific time periods used in their study, please add this info.
Thanks a lot for mentioning this- We'll correct this in the paper. Chamau had some days grazing in 2014-few days in 2015 and some days in 2016. We remove phrase „no grazing" from L62. Also in section 4.5, L434-435 ist mentioned that: „For instance, different grazing practices could affect the N2O emission patterns, which were not considered in this study due to limited data (only in Chamau there were 3 years including grazing practices and other sites in this study were non-grazed)."

L65: I suggest to give the FLUXNET identifiers for all sites, which in case of Chamau is "CH-Cha" (instead of "CH-CHA", note the lowercase letters) and in case of Neustift "AT-Neu" (instead of "AT-NEU"). Done!

L65 table caption, and other places throughout the manuscript: "Hörtnagel et al., 2014": the correct name is "Hörtnagl"
thank you, corrected!
L65 table header: I think there is a typo in the header and "(kg h-1 y-1)" should be "(kg ha-1 y-1)". I would interpret "h-1" as "per hour".
You are definitly right, thanks for noticing this, corrected!

L65 Table 1: Please give full units for "Annual N2O", i.e., is this "kg N ha-1 yr-1"? The header indicates that it is "kg N2O ha-1 yr-1", because the header mentions specifically N2O, but from the magnitude of the values I would assume N2O-N. Its is indeed N2O-N, corrected!

L65/L100 What is "Dataset size"? I am also not sure how this dataset size matches up with the dataset sizes given in Table 2, e.g., Graswang has dataset size 8708 in Table 1, but numbers in Table 2 add up to 1670. Please check and add missing info.

The dataset size refers to the number of daily values used, thanks for spotting the inconsistency here. As the size of data after pre-processing is given in table 2, we remove repetetive info from table 1.
It was a mistake in table 1. Initially we calculated all the data received for each grassland before pre-processing to report available data. In Graswang the records of soil moisture and temperature and climate are there for 2014-2020, while N2O is only given in 2020 (as mentioned in Table 2). This was mistakenly not consiered in the initial counting.

L71: For the three German sites, please add "respectively" after giving the altitudes.Done!
L71: "asl": I suggest changing to "a.s.l.", because this journal generally uses periods for text abbreviations. Done!
L81/82: The given references point to subpanels a and c in Fig. A1, but this figure does not show any subpanel letters. Solved!

L85: "is part of the FLUXNET initiative": I suggest to change to "is part of Swiss FluxNet". Done!

L88: "by separating two parcels at north and south of the site" change to "by establishing separate parcels in the northern and southern parts of the site", if I understand this correctly.Done!

L89: "in south" should be "in the south"Done!

L100 Table 2: Great to see (and have available) the different soil properties for the different locations. Would it make sense to include the soil properties as static drivers in the model? It definitly could make sense, mainly for an expanded dataset and future works. For this dataset, due to high correlation between some of the soil properties only pH and BD are used. For instance, SOC and BD are highly correlated in the current dataset. Also,clay content and BD showed a strong negative correlation.

L116: "among German grassland" should be "across German grasslands". Done!

L117: "capable to transfer" should be "capable of transferring" Done!

L125: "13000 data sequences": to me it is not immediately clear what "data sequence" means. I assume one data sequence is what is shown in Eq. 1, i.e., a single input block that is fed into the LSTM (correct?). I suggest to add more info here since the term "data sequence" is mentioned for the first time.

That correct, we meant the available number of sequences that the model looks at. Because, due of gaps of input parameters in some days, the time series were not continous. But, you are correct, this makes a confusion. We use the number of data rows after removing gaps instead to avoid any confusion. Change to „data rows“

L130: "sequence" should be "sequences".Done!

L161: "should be considered" is prescriptive language and OK in other places. In the Methods section I suggest to focus on reporting what "was" done to clearly describe the specific actions taken in this study, not what "should" be done. Done! We changed it to „were considered“

L213: Please add this reference for Matplotlib: https://doi.org/10.1109/MCSE.2007.55 Done!

L235: Given the low flux values (in combination with sometimes emission peaks), I do not think that the *relative* RMSE is a good indicator

We agree. We changed this to other indicators, please look into section 3.2 Table 5.

L235: From the figures I take that flux values are expressed using N2O-N, but in the text there are several places where this is not clear. I suggest to add a sentence in the Methods that fluxes are always expressed using N2O-N. In this regard, the header in Table 1 is slightly misleading because it mentions "Annual N2O" with incomplete units, but the numbers seem to give cumulative N2O-N.

Corrected the table 1 as well as anywhere a flux amoint was mentioned. Also added this to L65: The amount of fluxes is always expressed as N_2O-N amount.

L236: Please explain "total bias"

I think you mean L246. we meant on the whole test set, not individual sites. We now report metrics differenly for each part of test set as in table 5 and section 3.2

L260: Is there a reason why the time lagged variables are shown in Fig. 7, but not in Fig. 6? Maybe I have missed this somewhere.

Two different approaches of studying feature importance were taken to show feature importance from different perspectives. No other reason behind.

L275 Fig. 4: Typo "occurance" --> "occurrence"

Done!

L290-296: This could be shortened, most was already mentioned before.

We shortened this section, please look into lines L316:L325

L297-306: This currently reads more like a literature review than a discussion. The limitations of previous studies are listed, but the direct connection to findings in this study are missing. I suggest the authors revise this section to synthesize these citations with their results. For example, how their study addresses (or contrasts with) methodological gaps they identified in prior work (e.g., lack of independent test sites).

We improved this section mentioning the methodological gaps that we are filling, please look into lines L316:L325

L308-326: Although this is part of the Discussion, there are no references to existing literature and many of the sentences belong to the Results section. Please revise this section and contextualize findings by citing previous work that supports or contradicts results from this study. For example, is the observed moisture dependency consistent with established theory?

In the revised version section 4.2

L330: "Soil temperature, ..." This sentence seems to be missing words, please fix. Done!

L331: I am critical of the claim that maximum air temperature can substitute for soil temperature with "minimal impact". Air and soil temperatures often decouple due to factors like snow cover (insulation) and vegetation shading. And of course, there are different soil depths and different thermal lags. In case a sensitivity analysis was performed to demonstrate that the performance drop is indeed negligible, it could be mentioned solely in the context of this study rather than implying it is a viable strategy for sites globally.

Thats totally a valid point, but in our case we actually had a high correlation between soil_tempretaure and tmax, and we tested creation of the model with either of them and didn't see any significant performance drop when using tmax instead of soil temperature. Therefore, in the paper we also emphasized that in case of a high correlation they could be replaced: „but in case of high Pearson correlation with maximum air temperature (0.86 in our dataset)", the two variables could be replaced.

L334-337: Context to literature is missing Done!

L338-345: Here, some interesting points are raised from the literature (pH, BD), but they are not discussed in relation to findings in this study. Instead, the final two sentences jump back to soil moisture but without giving references.

we expanded this section with adding a conclusion sentence and more references please see L381:L387 . In general for each input variable one short paraghraph is discussed.

L346-369: I suggest to limit this section to sites that were also used in this study. A comparison with sites from other locations is difficult, especially for N2O that is characterized by high spatial variability.

considering the fact that we got critics about this section from all reviewers, who also questioned if we even need this section, we remove it from this publication!

References:

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Ghosh, T., Anand, S., Nannewar, R. G., & Nagaraj, N. (2025). Deep Learning for Short-Term Precipitation Prediction in Four Major Indian Cities: A ConvLSTM Approach with Explainable AI. *arXiv preprint arXiv:2511.11152.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, **26**(2), 211–243.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Raiaan, M. A. K., Sakib, S., Fahad, N. M., Mamun, A. A., Rahman, M. A., Shatabda, S., & Mukta, M. S. H. (2024). A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks. *Decision Analytics Journal, 11*, 100470. https://doi.org/10.1016/j.dajour.2024.100470

West, R.M. (2021) Best Practice in Statistics: The Use of Log Transformation. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, 59, 162-165. DOI: 10.1177/00045632211050531.