Report #2:

Thanks a lot for following up on the review process. The input is highly appreciated.

1. In the manuscript, you wrote: "Long short-term memory network or LSTM (Hochreiter and Schmidhuber, 1997) is a machine learning model that detects the dynamic relationships in sequential data. It has previously been used for different applications such as flood forecasting (Le et al., 2019), stock market prediction (Chhajer et al., 2022), and predicting water table depth in agricultural areas (Zhang et al., 2018), and therefore is a promising method for modelling soil N2O emissions."
which indicates LSTM has not been used for modeling N2O, and your study is the first time to do so, which is not true. I think the authors should acknowledge RNN models have been tested to simulate N2O emissions and results are mixed. And your contribution is to adding a more cases to let us know when it works well vs not. That's good enough. But certainly there's no meaning to emphasize how LSTM is a novel and strong methods for predicting N2O emissions.

   Thank you for this comment. Implying that this is the first application of LSTM to model soil N_2O emissions was not our intention. Therefore, to avoid any misunderstanding, we changed the statement to:
   "It has previously been used for different applications such as flood forecasting (Le et al., 2019), predicting water table depth in agricultural areas (Zhang et al., 2018), as well as modelling soil N_2O and CO_2 emissions (Hamrani et al., 2020)."

   Our intention in the original paragraph was simply to highlight that LSTM is well suited to modelling processes with temporal dependencies and sequential patterns, such as soil N_2O emissions. We did not intend to claim novelty in the application of LSTM to N_2O emission modelling. Rather, the novelty of our study lies in evaluating the performance of an LSTM model using a unique, high-frequency, multi-year field dataset from an intensively managed grassland, thereby providing further evidence of the conditions under which LSTM models can successfully simulate soil N_2O emissions.

2. I'm not convinced by their explanations about low performance in certain site-year. The current revision is more like "why some low performing cases are not too terrible", and "we have the excuse". It's not about why the model couldn't perform well in some extreme cases, and here're a few possible "whys". For example, the explanation, even hypothetically, should clarify why your LSTM structure could capture fertilization in several cases (Fig 4) even though the inputs explicitly include~~s~~ fertilization.

   Thank you for your comment. We have substantially expanded the Discussion to better address the potential causes of these prediction errors, L58-L68:

"Within the Cross-site test set, several site-year evaluations showed good predictive ability with R^2 values ranging from 0.24 to 0.64. In contrast, some site-years remained challenging, with R^2 ranging from −25 to 0.14. At the challenging site-years, substantial overestimation of N2O emissions particularly following fertilization events was noticed. These overestimations likely occurred because the model learned strong associations between fertilization events, favorable soil temperature, and soil moisture conditions from the training dataset. However, the observed emissions after such conditions remained unusually low in several instances (Fig. 4 i, l, n). This suggests that, additional controlling factors which were not represented in the model inputs have likely influenced the response. For instance, differences in substrate availability, microbial community composition, soil physical properties, or site-specific conditions also determine whether fertilization leads to substantial N_2O production. Moreover, low-emission site-years represent relatively rare conditions within the available dataset. Therefore, an expanded training set with broader range of environmental conditions and emission responses could improve the model's ability to recognize such atypical responses. Nevertheless, despite the poor R^2 values in these challenging site-years, Spearman correlation coefficients remained relatively high (0.42–0.61), indicating that the model captured the temporal dynamics of N_2O emissions even when the magnitude of individual peaks was overestimated."

3. I am not satisfied with how the authors addressed the request for a transparent and professional description of the LSTM architecture. Standard practice for explaining deep learning models—and justifying specific architectural modifications—relies on a concise set of mathematical equations. The current qualitative framing fails to achieve the "technical completeness" the authors claim to target. To ensure both clarity and rigor, the authors must provide the formal equations defining their modified LSTM model, and then explain why certain architecture changes are meaningfully make a difference.

   We agree that a transparent description of deep learning architectures is essential for reproducibility and that mathematical formulations are valuable for explaining modified model structures. However, we would like to clarify that the LSTM model architecture itself was not modified in this study; we employed the well-established standard LSTM units. Based on the research problem, model performance depends on how the input data are structured and whether additional components, such as an attention mechanism, are incorporated. In our study, we did not alter the internal structure, equations, or operations of either the LSTM units or the attention mechanism. The main methodological choices in our study were therefore not modifications of the LSTM equation itself, but rather the optimization of the model configuration by selecting appropriate layers, stacking strategies, and model settings through extensive experimentation to identify a suitable architecture for this specific application.

   To avoid ambiguity, we have added the following statement to direct the reader to the original formulation of the LSTM and attention layers:

"The mathematical formulation of the LSTM architecture and the attention mechanism are described in detail in the original publications (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2015)."

In addition, we have expanded the description of the model architecture and implementation details in the Methods section previously as requested. The layer configuration and overall model structure are provided in Fig. 2, and the complete implementation is available in the Code Availability section. Since we use standard LSTM, attention, dropout, and dense layers, the model can be reproduced using the corresponding established implementations available in common deep learning frameworks such as TensorFlow and considering the stacking method and hyperparameters provided in Code availability section.

Report #1

I find that the manuscript has been improved, and I am generally happy that the majority of my comments are addressed.One question that the authors should still think about is the negative R2 in cross-site evaluation (L297-300). Current explanations are mathematically possible, but is there a way to really prove it, e.g., any other useful indicators?

Thank you for this valuable comment. We agree that the explanation of negative R^2 values in the cross-site evaluation should be supported by additional analyses. To show that even few extreme errors can ruin R^2 significantly we revised following paragraph accordingly, L298-L304:

"Substantial overestimations were observed at Rottenbuch in Fendt 2017 and 2020 intensive management campaigns following fertilization (Fig. 4 i, l, n). In 2015, 2016, and 2019 fertilization events occurring during warm periods were associated with pronounced emission peaks even exceeding 150 µg m^{-2} h^{-1} (Fig. 4 h, j, k). Despite similar temperature and moisture, the 2017 and 2020 campaigns did exceptionally not exhibit elevated N_2O fluxes after fertilization. The model nevertheless predicted several large emission peaks after fertilization. This results in the extreme prediction errors at these site-years contributing to the negative R^2 (Fig. 4 i, l, n). For instance, excluding the three extreme outliers from Rottenbuch in Fendt 2017 intensive in Fig. 4 l, resulted in R^2 increasing from −15.36 to −0.51. However, the bias value remains high, equal to 4.10, indicating that the model shows levels of overestimations also for other days at this challenging site-year, contributing further to the negative R^2 (Fig. 4 l). "

1. L23: please remove "and production".
   Done!
2. L65: Do you mean "The magnitude" or "The intensity" of daily fluxes?
   We corrected the sentence: *"Daily N_2O fluxes always were always expressed as N_2O-N in this paper."*
   This sentence was mainly added as a request from another reviewer to ensure its explicitly mentioned in the paper that the N part of the N_2O fluxes is reported in our dataset.

3. L101: Table 2. I suggest that the authors double check the fertiliser rates at each site. For example, the avg fertiliser of 99 kg N/ha/yr for CHAMAU seems quite low. Did you take the mean value of two parcels?
   Thanks a lot for spotting this. We double checked and corrected the value for Chamau. The reason of a misinterpretation of the value was that in the Chamau dataset, from which we extracted the fertilizer N, had a column called N.Fertilizer_kg.ha, which seemed to indicate the fertilizer N amount but was not the case. We have taken a deeper look into the meta-data according to your feedback and found the correct column of fertilizer N amount, while the former wrong column was indicating the biomass related N.
   In addition, the fertilizer N value for Tereno dataset is adjusted, as the values for one year were missing previously and are completed now.