Response Letter

We are grateful to all comments and suggestions from three reviewers and have carefully addressed their concerns point by point. Major changes include:

(1) We have conducted uncertainty analysis for all pure machine learning models and KGML models presented in out study to include the machine learning model uncertainties;

(2) The uncertainties of process-based model *ecosys* and its performance over various ecosystem for N_2O and CO_2 have been added into the maintext;

(3) We have added LSTM results into the supplement and comparing with all other models for reference;

(4) We have added a new paragraph in discussion to address the concerns of KGML-ag limitations;

(5) We have clarified all the confusing parts which have been pointed out by reviewers, and corrected typo and grammatical errors.

By changing these major concerns and many other minor comments and suggestions, we believe the quality of this manuscript is improved. Below, please find our detailed responses point-by-point.

Please be aware of the formatting of all responses:

1. Reviewer comment in **black**, response in **blue** and quotation from the main text in ***red***;
2. The line number is based on the clean version of the revised manuscript, not the track change version.

To Reviewer 1

Liu et al. presented a promising predictive framework that combined a process-based model (physical knowledge and pre-train dataset) and a machine learning model for agroecosystem N_2O emission estimate. The modeling framework is robust and thoroughly validated. This work will be an important milestone towards a better understanding, monitoring, and predicting agroecosystem greenhouse gas emissions.

The paper is well organized and written. Below are some of my comments that may help elucidate the strength and limitations of the proposed KGML-ag framework.

Response: We really appreciate that the reviewer recognized our efforts in developing the proper knowledge guided machine learning framework for agroecosystem. To improve the quality of this study, we have carefully revised the manuscript based on the reviewer's comments and suggestions shown as below:

1. Robustness of physical (prior) knowledge

ecosys model plays a central role in guiding the ML model in terms of structure and providing a pre-train dataset. It will be important to discuss the structure uncertainty in *ecosys* N_2O module, including e.g., underlying theories, major processes, difference/similarity to the classic leaky pipe type model (Davidson et al., 2000), and so on.

Reference:
Davidson, E. A., Keller, M., Erickson, H. E., Verchot, L. V., & Veldkamp, E. (2000). Testing a conceptual model of soil emissions of nitrous and nitric oxides: using two functions based on soil nitrogen availability and soil water content, the hole-in-the-pipe model characterizes a large fraction of the observed variation of nitric oxide and nitrous oxide emissions from soils. Bioscience, 50(8), 667-680.

Response: Thank you so much for this suggestion. In this revision, we have added a detailed description on the major processes of N_2O production and transfer in *ecosys* model, and on the differences between traditional pipeline N_2O model and *ecosys* model. You can find the description in the manuscript section 2.2.1 (from Line 136 to 147) as:

"It represents N_2O evolution in the microbe-engaged processes of nitrification-denitrification using substrate kinetics that are sensitive to soil nitrogen availability, soil temperature, soil moisture, and soil oxygen status (Grant and Pattey 2008). Two groups of microbial populations, autotrophic nitrifiers and heterotrophic denitrifiers, produce N_2O with specific competitive or cooperative relations in *ecosys* when O_2 availability fails to meet O_2 demand for their respirations and NO_2^- become alternative electron acceptors. N_2O transfer within soil layers and from soil to the atmosphere is driven by concentration gradient using diffusion-convection-dispersion equations, in the forms of gaseous and aqueous N_2O under control of volatilization-dissolution (Grant et al., 2016). Unlike the pipeline model described by Davidson et al. (2000) , which mainly consider the correlations of N_2O production with nitrogen availability and of N_2O emitting with soil water content, *ecosys* enables integrative effects of energy, water, nitrogen availability on N_2O production and N_2O transfer via the microbial population dynamics and their interactions with soil, plant, and atmospheric dynamics, under diverse meteorological and anthropogenic disturbances (e.g. runoff, drainage, tillage, irrigation, soil erosion)."

Again ecosys provides pretrain dataset, which has its own uncertainty and biases. It's worthwhile to at least show some ecosys model performance across various different conditions at agroecosystems. For example, does ecosys pick up the high-frequency signals (fluctuation) of CO2/N2O flux that are observed in the chambers data? If not, is that the reason why PGML-ag could not capture the high fluctuation of CO2/N2O emissions in the field?

Response: We really appreciate this comment which suggests to show the capability of *ecosys* model as the domain knowledge provider. To show the *ecosys* model performance on simulation of CO_2 and N_2O emissions at field, we have added detailed quantitative comparisons between model simulations and observations in the manuscript section 2.2.1 (from line 149 to 154):

"For the agricultural ecosystems in the US Midwest, whose simulations are used for synthetic data in this study, the performance of ecosys on CO_2 and N_2O fluxes have been extensively benchmarked, including CO_2 exchange (NEE, $R^2 = 0.87$) and leaf area index (LAI, $R^2 = 0.78$) from six flux towers, USDA census reported corn yield ($R^2 = 0.83$) and soybean yield ($R^2 = 0.80$), satellite-derived GPP for corn ($R^2 = 0.83$) and soybean ($R^2 = 0.85$) from Illinois, Iowa and Indiana, and cumulative N_2O emissions ($R^2 = 0.36$) across eight Midwestern states (Wang et al., 2021; Yang et al., 2022)."

If you are interested in the more detailed performance of field level N_2O emission simulation using *ecosys* model, you may review 1) the papers of Grant et al (2006, 2008) to find the influences of fertilizer rate and temperature on N_2O emissions in fertilized agriculture soil; 2) the paper of Grant et al (1999) to find the influences of spring thawing; and 3) the papers of Grant et al (2010, 2016) to check the N_2O simulation performances at managed forest and grassland.

2. It's not obvious which variables are used as inputs or intermediate variables and how that relates to the feature importance ranking. It will be better to show each variable in Figure 1. For example, W will be temperature and precipitation. Furthermore, feature importance analysis highlight NH3, H2, N2, O2, CH4, ET, CO2 are important variables that drive N2O emission (~ L230). It's not clear in the main text, how this feature importance ranking helps the design of PGML-ag. What can we get out of this feature importance analysis?

Response: Thanks for pointing out the confusing part of how feature importance related to KGML model development. In this revision, we have extended descriptions in Figure 1 caption to explain W, SCP and IMVs that are used in our study.

"Figure 1: The model structures. a) The *ecosys* model; b) Gated recurrent unit (GRU) model; c) KGML-ag1 model with a hierarchical structure; d) KGML-ag2 model with a hierarchical structure using separated GRU modules for IMV predictions. Specifically, in our KGML model design, weather forcings (W) include temperature (TMAX, TDIF), precipitation (PRECN), radiation (RADN), humidity (HMAX and HDIF) and wind speed (WIND); soil/crop properties (SCP) include bulk density (TBKDS), sand content (TCSAND), silt content (TCSILT), pH (TPH), cation exchange capacity (TCEC), soil organic carbon (TSOC), planting day of the year (PDOY) and crop type (CROPT); IMVs include CO_2 flux, soil NO_3^- concentration, soil NH_4^+concentration, and soil volumetric water content (VWC)."

Feature importance analysis was the first step in our study to learn the knowledge from synthetic data generated by the *ecosys* model and to investigate the correlation between input/intermediate variables and N_2O fluxes. The importance rankings help us to put low/median/high attention to available variables during model development (e.g. CO_2 was tested as a higher ranking variable than others so that we paid high attention to it by testing two different combinations of IMVs w/o CO_2). In addition, the rankings will provide guidance of future N_2O related measurement, which is discussed in section 4.3. We have revised paragraph two in section 2.2.4 to highlight how feature importance rankings help our model development (from line 252 to 258).

"Variables ranked high in feature importance analysis are considered with priority during model development. To develop a functionable KGML-ag, we further investigated the feature importance of four IMVs that are available from mesocosm observations including CO_2, NO_3^-, VWC and NH_4^+, which were ranked 7th, 20th, 58th, 60th respectively in 92 input features of synthetic data (Fig. S2a). We used these four available IMVs to create two input combinations: 1) CO_2 flux, NO_3^-, VWC and NH_4^+ (IMVcb1), and 2) NO_3^-, VWC and NH_4^+ (IMVcb2). The objective of building IMVcb2 was to investigate the importance of highly ranked variable CO_2 flux (by removing it from the inputs), and the impact of mixing-up flux and non-flux variables on model performance. "

3. There is a lack of discussion on uncertainty in PGML-ag, which is fundamentally important for predictive modeling. Also, what about chamber measurements uncertainty?

Response: Thank you for pointing out this concern for predictive modeling. To address the uncertainty of the machine learning models and KGML-ag model, we have conducted 10 ensemble experiments for different model structures (DT, RF, GB, XGB, ANN, GRU, KGML-ag1 and KGML-ag2). Corresponding method part in section 2.1 has been updated (from line 125 to 129).

"We further benchmarked KGML-ag models and uncertainties with other pure ML models without considering temporal dependence, including Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB) from the sklearn package (https://scikit-learn.org/stable/), Extreme Gradient Boosting (XGB) from the XGBoost package (https://xgboost.readthedocs.io/en/latest/) and a 6-linear-layer artificial neural network (ANN) with the mesocosm experiment data by 10 times ensemble experiments (Fig. 4-5; Fig. S6-8);"

The new results have been updated in Figure 4 and Figure 5 (also as Figure R1 and R2 below) in the main text and Figure S6-S7 (also as Figure R3 and R4 below) in the supplementary. We have also updated values in section 3.3 accordingly. For chamber measurement uncertainty, we have cited the original thesis (Miller L., 2021) including the mesocosm experiment settings, instruments and related measurement uncertainties (e.g. Figure 2.2 in the thesis). In our study, we also used a data augmentation method to cover the uncertainties caused by converting hourly observations to daily observations. The data augmentation method has been described in section 2.2.2 paragraph 3.

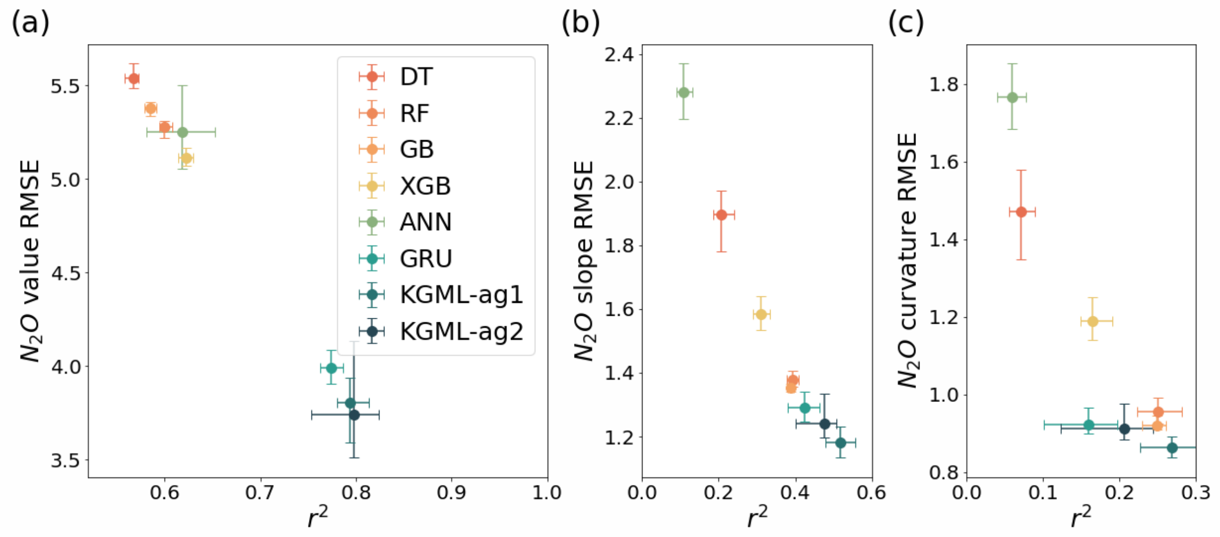


Figure R1: The comparisons of overall prediction accuracy for N_2O value (a), 1st order gradient (slope, b) and 2nd order gradient (curvature, c) between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN, GRU), and KGML-ag models. Different color symbols represent the different models. The x- and y-error bars are coming from the maximum and minimum scores of ensemble experiments. The dot represents the mean score of the ensemble experiments.

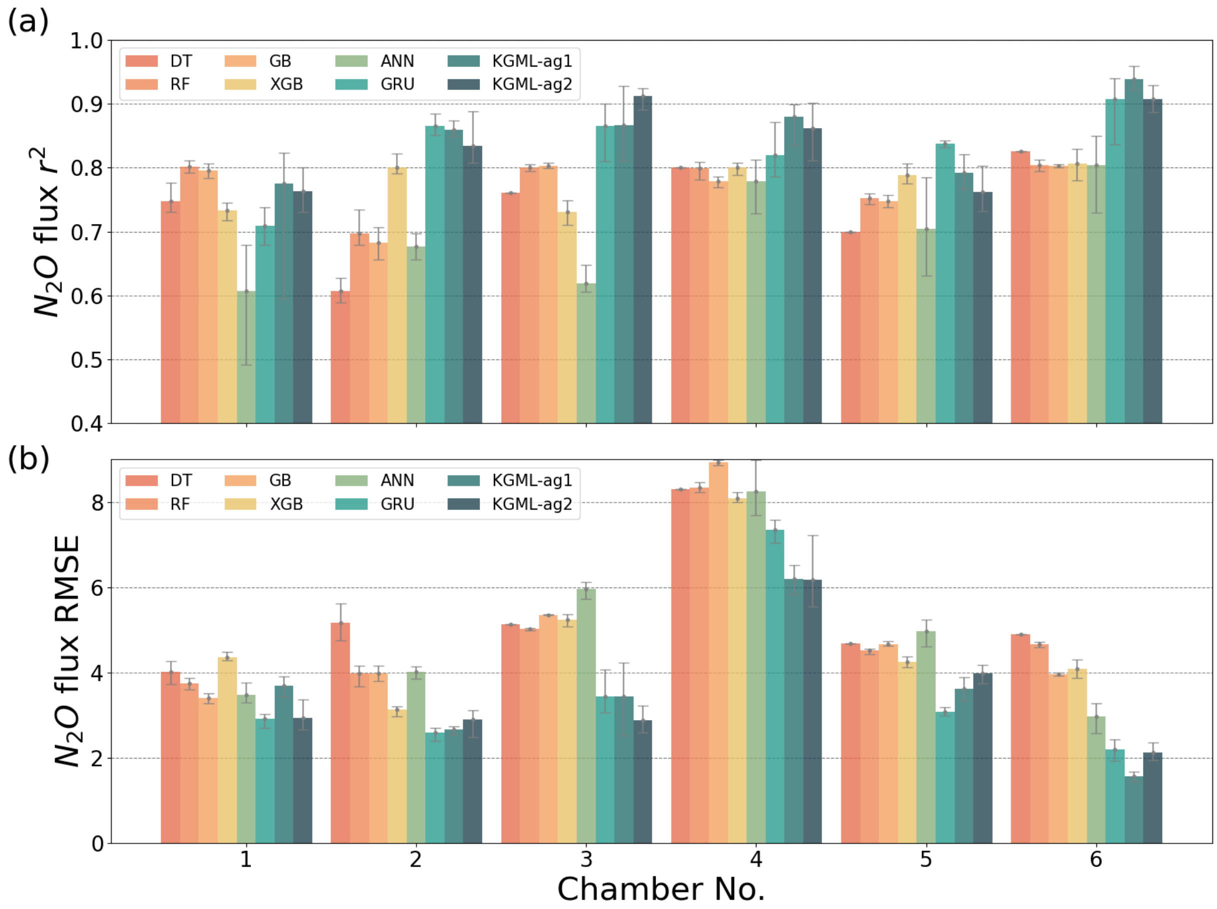


Figure R2: The comparisons of N_2O flux prediction accuracy r^2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU), and KGML-ag models in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

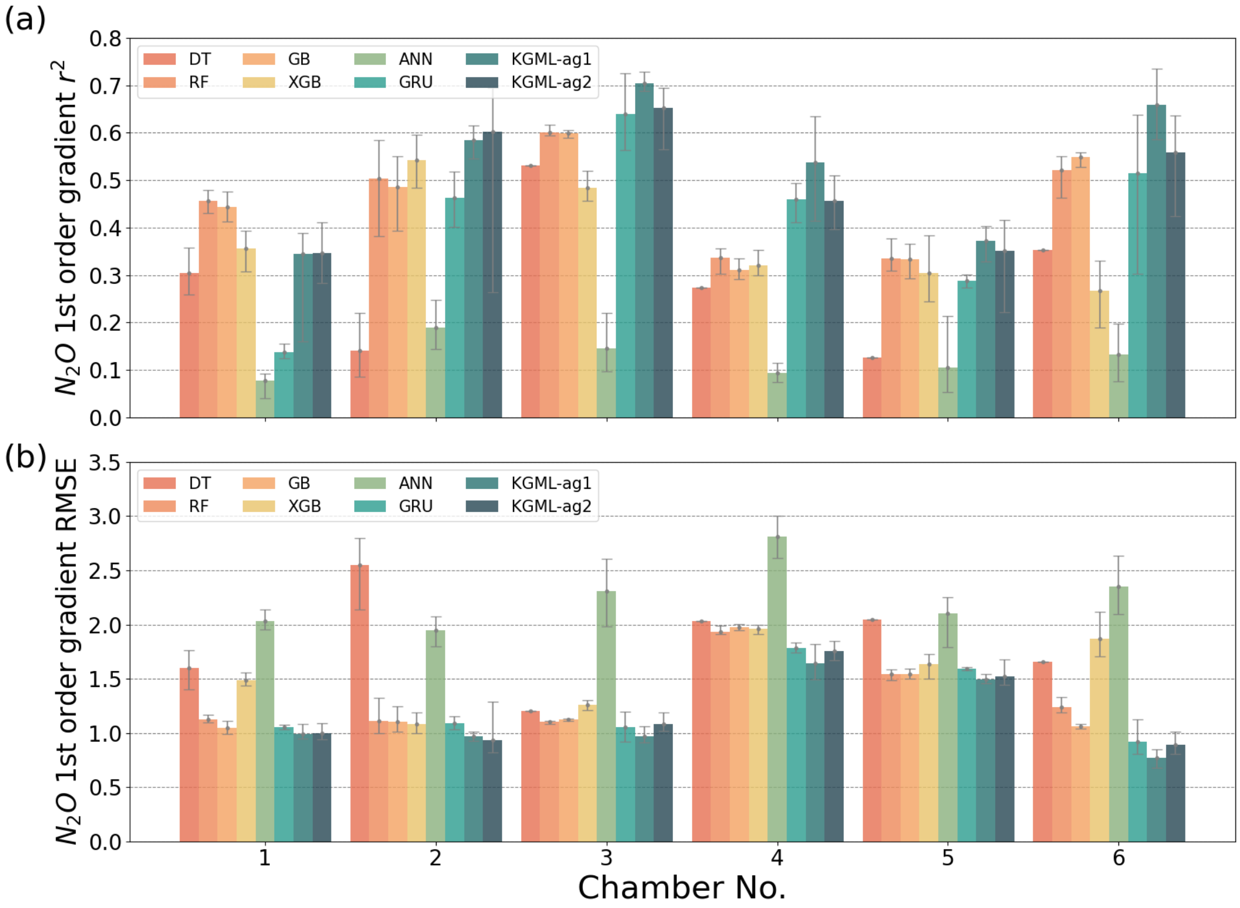


Figure R3: The comparisons of N2O 1st order gradient prediction accuracy r2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag models in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

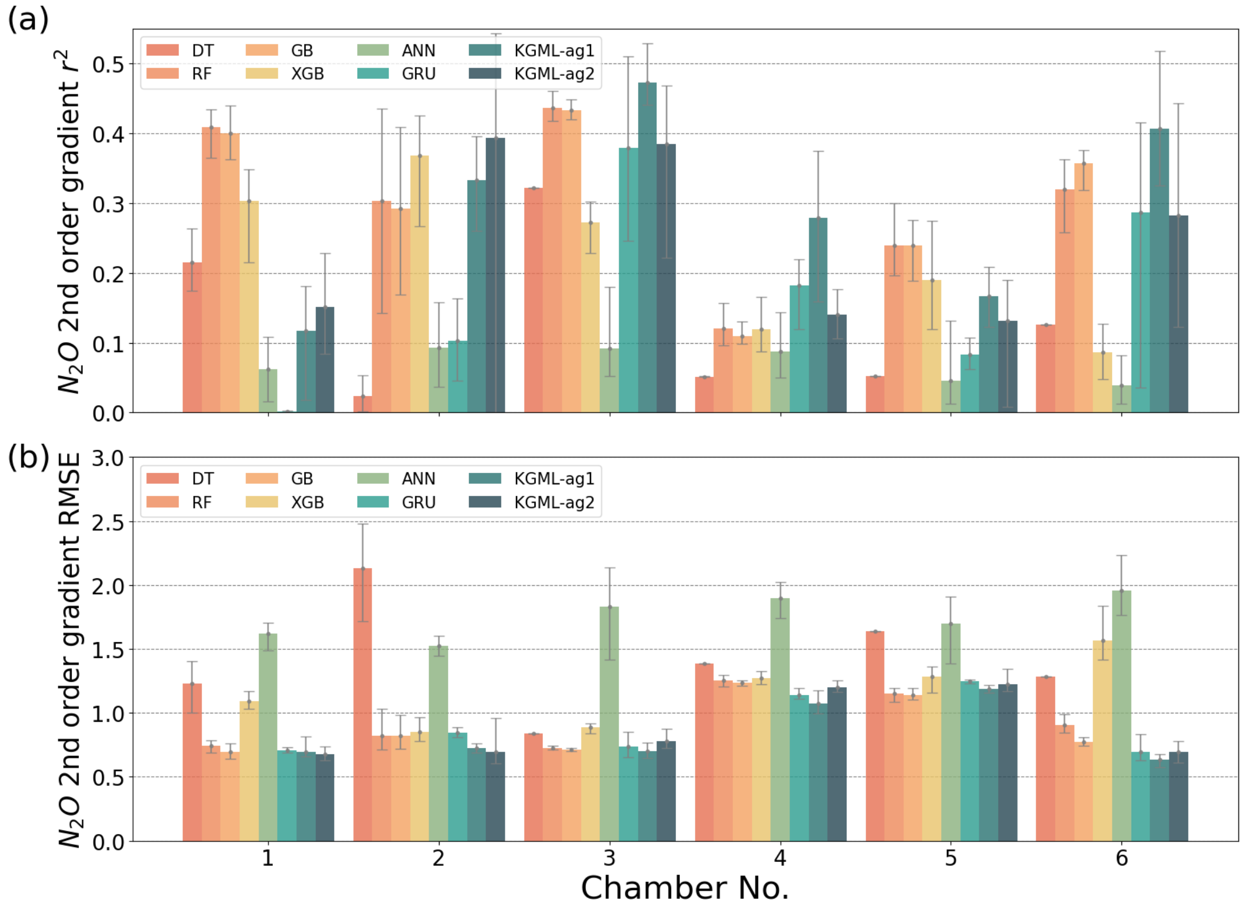


Figure R4: The comparisons of N_2O 2nd order gradient prediction accuracy r^2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag models in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

L254 based on the structure of process representation in ecosys

Response: We have revised the sentence based on your suggestion (Line 276 to 278).

"We built a hierarchical structure based on the structure of process representation in *ecosys* to first predict IMVs and then simulate N_2O with predicted IMVs;"

References:

Grant, R. F., Black, T. A., Jassal, R. S., & Bruemmer, C.: Changes in net ecosystem productivity and greenhouse gas exchange with fertilization of Douglas fir: Mathematical modeling in *ecosys*. Journal of Geophysical Research: Biogeosciences, 115(G4), 2010.

Grant, R. F., & Pattey, E.: Mathematical modeling of nitrous oxide emissions from an agricultural field during spring thaw. Global Biogeochemical Cycles, 13(2), 679-694, 1999.

Grant, R. F., & Pattey, E.: Temperature sensitivity of N_2O emissions from fertilized agricultural soils: Mathematical modeling in ecosys. Global biogeochemical cycles, 22(4), 2008.

Grant, R. F., Neftel, A., & Calanca, P.: Ecological controls on N_2O emission in surface litter and near-surface soil of a managed grassland: modelling and measurements, Biogeosciences, 13(12), 3549-3571, 2016.

Grant, R. F., Pattey, E., Goddard, T. W., Kryzanowski, L. M., & Puurveen, H.: Modeling the effects of fertilizer application rate on nitrous oxide emissions, Soil Science Society of America Journal, 70(1), 235-248, 2006.

To Reviewer 2

General comments

This manuscript presents a new method for estimating N2O flux from cropland. The inputs to the method are known fertilization rate, weather forcings, soil and crop properties. The method also requires initial concentrations of nitrate ions, ammonium ions, and water in the soil, and optionally CO2 flux. The method employs gated recurrent networks organized in a hierarchical structure to mirror the time-dependence and causality present in the process.A process-based model provides pre-training data, and fine-tuning is done using observations from mesocosm experiments. The trained neural network models outperform the process-based model and many basic machine learning approaches.

The methodology employed is both novel and sound. The use of GRUs in hierarchical structures is well-justified and appropriate to the problem. The models have been well-validated, and various alternate choices for model architecture have been explored. I believe this work represent a substantive advance in modelling science. Below I list specific comments which I hope will serve to improve the manuscript.

Response: We really appreciate the reviewer's recognition of our work and all other valuable comments and suggestions mentioned below. Just as the reviewer summarized, we want to incorporate the domain knowledge learned from agroecosystem process-based model *ecosys* to the advanced machine learning models to combine the advantages from both kinds of state-of-art works. This effort is trying to build a new body of research for simulating the agriculture ecosystem and KGML-ag in this study is a demonstration case simulating N_2O flux from mesocosm experiments. To further improve our study, we have carefully revised the manuscript to address all reviewer’s comments. The specific responses can be found in the following letters.

Specific comments

1. The use of the term "initials" confuses me. Upon first reading I thought it referred to the acronyms for various intermediate variables. I think it actually refers to the initial values of a sequence. Is this usage standard? If not, I recommend a different phrase such as "initial values" in place of the word "initials." Alternatively, clarify the meaning of the term in the manuscript.

Response: Thanks so much for pointing out this term which may cause confusion. Just as you said, the term “initials” in the manuscript are most referring to the “initial values”. It indeed will cause some confusion since we also use the term “initial” as a verb for the knowledge guided initialization. Thus we have replaced “initials” to “initial values” throughout the manuscript.

2. Another possible explanation for why KGML-ag2 better predicts IMVs but does not predict N2O as well is that KGML-ag1 may learn to use the IMVs as a kind of extra hidden layer, encoding information relevant to N2O predictions in them.

Response: We really appreciate your interesting explanation about why KGML-ag2 predicts better IMVs but worse N_2O fluxes. In both KGML-ag1 and KGML-ag2, the IMVs were first predicted from KGML-ag-IMV modules and then input into the KGML-ag-N_2O modules. The only difference between

KGML-ag1 and KGML-ag2 is that KGML-ag2 explicitly simulates each IMV by using individual KGML-ag-IMV modules. Thus, using IMVs as a kind of extra hidden layer may happen in both models in KGML-ag-N_2O modules. But since KGML-ag1 has interactions between predicted IMVs and lower complexity, it may be easier for the KGML-ag1-N_2O module to get the useful knowledge from IMVs.

Moreover, your valuable thought draws us to deeply review the model structures and data qualities. The observational data, including the IMVs of CO_2, NH_4^+, NO_3^- and VWC, are not perfect and may have many noises or be lacking some key information. KGML-ag2-IMV module may only follow what we have for IMVs to generate accurate IMV predictions without any extra information, while KGML-ag1-IMV module may perform like an encoding layer to predict IMVs with extra information relevant to N_2O flux, just as you mentioned.

In this revision, we decided to keep our explanation to make our discussion more focused and accessible to a broader audience. But we will find a larger dataset to test both explanations in subsequent ML-oriented technical papers.

3. Why not include KGML-ag2 in Figure 4? I can see simplifying the comparison by choosing only the best-performing model.

Response: The reviewer is right that we excluded KGML-ag2 in the previous Figure 4 to simplify the comparison. To address the reviewer's concern, we have added similar 10 ensemble experiments for KGML-ag2 and updated Figure 4 (also as Figure R5 below). We can see that although KGML-ag2 has similar mean performance as the KGML-ag1 but it has much larger uncertainties. Moreover, the best scores for slope and curvature are all from KGML-ag1.

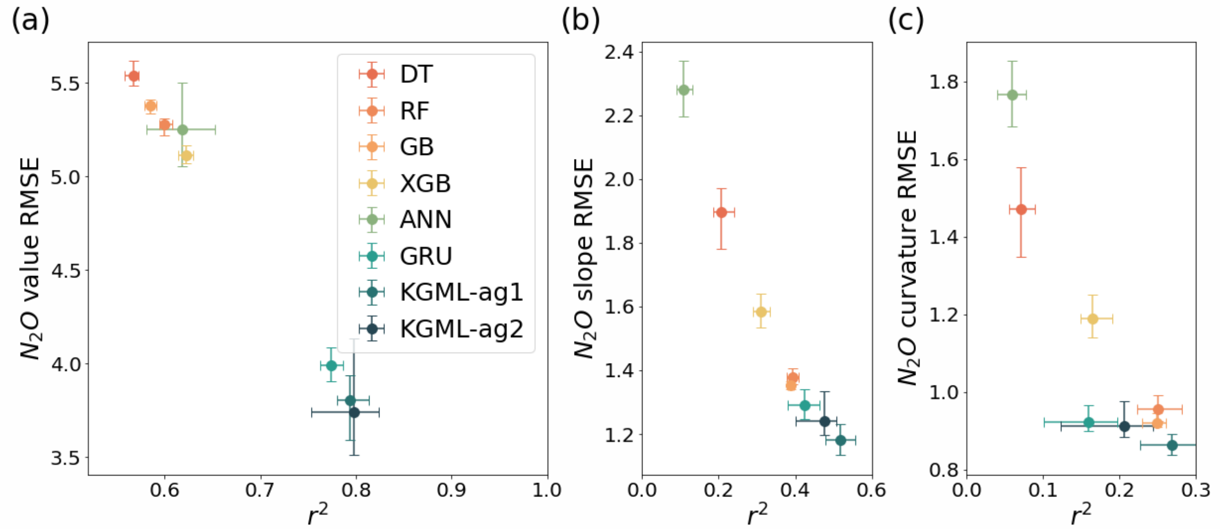


Figure R5: The comparisons of overall prediction accuracy for N_2O value (a), 1st order gradient (slope, b) and 2nd order gradient (curvature, c) between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN, GRU), and KGML-ag models. Different color symbols represent the different models. The x- and y-error bars are coming from the maximum and minimum scores of ensemble experiments. The dot represents the mean score of the ensemble experiments.

We have also updated the corresponding figures including Figure 5, Figure S6-S7, and section 3.3 (From line 399 to 405).

“The results from eight different models showed that KGML-ag1 comparing with other pure ML models consistently provided the lowest RMSE (3.59-3.94 mg N m^{-2} day^{-1}, 1.14-1.23 mg N m^{-2} day^{-2}, and 0.84-0.89 mg N m^{-2} day^{-3}) and highest r^2 (0.78-0.81, 0.48-0.56, and 0.23-0.31) for N_2O fluxes, slope and curvature, respectively (Fig. 4). This indicated that KGML-ag1 outperformed other pure ML models in capturing both the magnitude and dynamics of N_2O flux. KGML-ag2 presented slightly better mean scores for N_2O flux predictions than KGML-ag1, but worse scores for slope and curvature and larger uncertainties. This proved the hypothesis discussed in section 3.2 that KGML-ag2 didn’t benefit the magnitude and dynamics predictions of N_2O flux with its more complex structure and less connections between IMVs”

4. Many standard deep learning models were included for comparison, but an LSTM was not among them. I would expect the LSTM to perform similarly to the GRU. I don't think it is crucial that an LSTM be included in this comparison. However, if the GRU outperforms an LSTM, it could provide further justification for choosing to use a GRU instead of an LSTM. Again, I could understand simplifying the comparison by including only one recurrent neural network.

Response: We fully agree with your comments on LSTM. We have tested both GRU and LSTM as mentioned in section 2.2.3, and preliminary results showed similar performance between the two neural network structures. However, to simplify the comparison and streamline the discussion, we fixed GRU as the basis for pure machine learning models and the KGML models.

To address the reviewer’s concern, we have conducted similar 10 ensemble experiments of LSTM and the comparisons are presented here in Figure R6 and in the supplement Figure S8 (best model in ensemble experiment). From Figure R6 demonstration case, the LSTM with r_L^2 of 0.72 and r_U^2 of 0.73 is better than GRU model (r_L^2 of 0.60 and r_U^2 of 0.57) but worse than KGML-ag1 (r_L^2 of 0.78 and r_U^2 of 0.86). This further proved our conclusion that KGML-ag1 better represents complex dynamics of N_2O flux than other pure machine learning models.

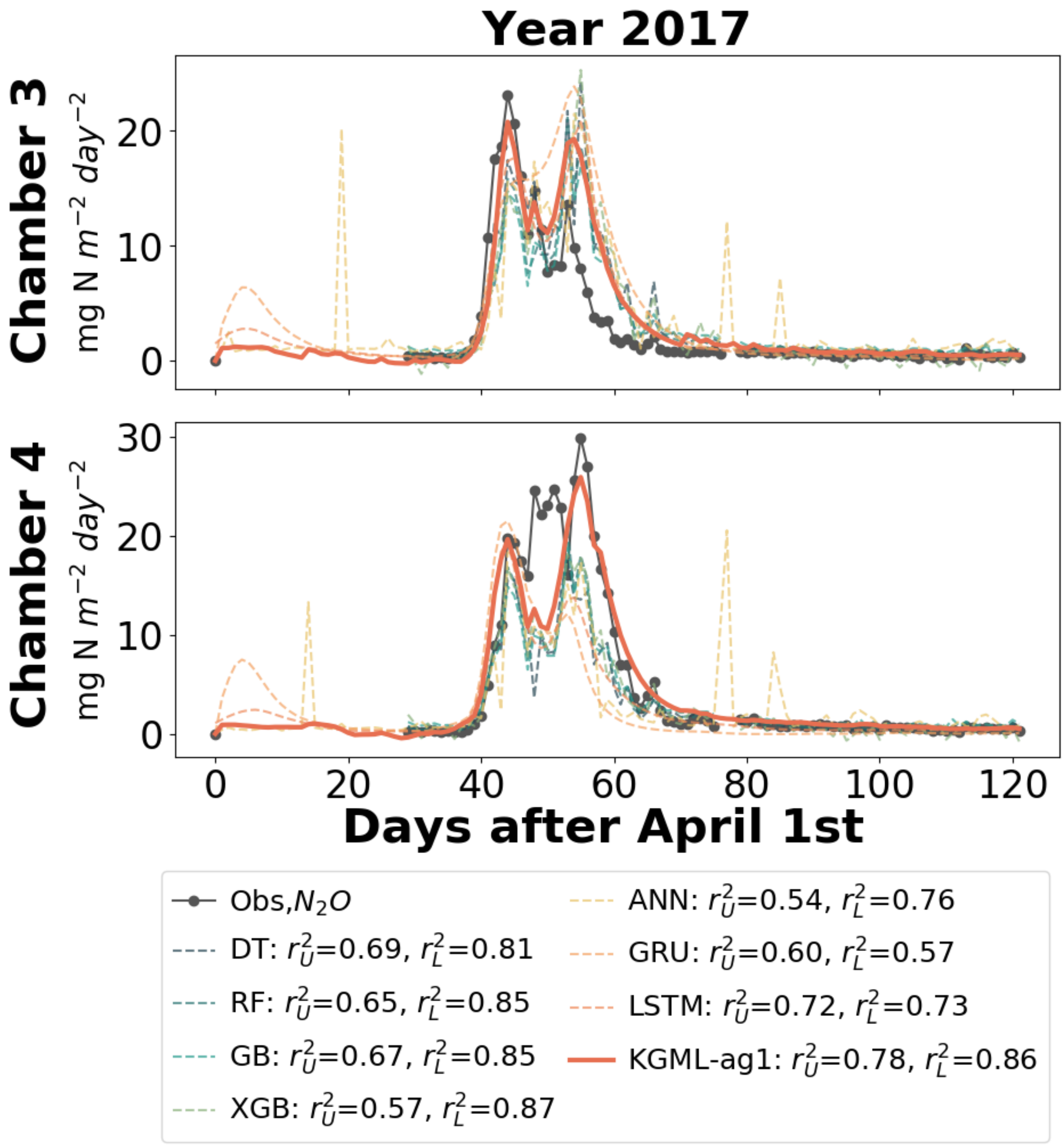


Figure R6: N_2O flux time series comparisons between KGML-ag1 predictions (red solid line), pure ML models (other colored dashed line) and observations (black-dot line) from cross-validation on two representative panels of chamber 3 and 4 in 2016. The r^2 value was calculated between observations and model simulations. r^2_U represents the r^2 value from upper panel (chamber 3) and r^2_L represents the r^2 value from lower panel (chamber 4). The LSTM model has been tested by similar 10 ensemble experiments as GRU. The best LSTM model was chosen to present here compared with other models.

5. You tested two input combinations, IMVcb1 and IMVcb2, but it is not clear how that test informed the model development.

Response: Thank you for finding this unclear part in our manuscript. We have added more descriptions to clarify why we have tested two combinations in section 2.2.4 paragraph 2 (From line 252 to 258).

"Variables ranked high in feature importance analysis should be primarily considered during model development. To develop a functionable KGML-ag in real world, we further investigated the feature importance of four IMVs that are available from mesocosm observations including CO_2, NO_3^-, VWC and NH_4^+, which were ranked 7th, 20th, 58th, 60th respectively in 92 input features of synthetic data (Fig. S2a). We used these four available IMVs to create two input combinations: 1) CO_2 flux, NO_3^-, VWC and NH_4^+ (IMVcb1), and 2) NO_3^-, VWC and NH_4^+ (IMVcb2). The objective of building IMVcb2 was to investigate the importance of highly ranked variable CO_2 flux (by removing it from the inputs), and the impact of mixing-up flux and non-flux variables on model performance. "

Moreover, tests using IMVcb1 (with CO_2) and IMVcb2 (without CO_2) indicate that high ranking variables detected from feature importance analysis based on synthetic data (like CO_2 flux ranks 7th in 92 input features) can also be similarly important in N_2O predictions with real observed data. Therefore the feature importance results could benefit feature selection in real data. We have added the results and discussion in section 3.2 last paragraph (From line 395 to 397).

"In addition, we also found all KGML-ag models would perform better by using IMVcb1 (with CO_2) than using IMVcb2 (without CO_2) in real data tests, indicating feature importance analysis based on synthetic data can be a reasonable substitute for analysis with the often limited real-world data."

6. The reason for evaluating slope and curvature in addition to N2O value could be stated more clearly.

Response: We have added more explanations in section 3.2 paragraph 2 (From line 373 to 375).

"Slope represents the speed of N_2O flux changes through time and curvature represents the acceleration. Assessing prediction performance with these two metrics will reveal the model robustness on capture variable dynamics, which is critical when predicting fast-change variables with hot moments like N_2O."

7. I recommend that the paragraph starting at line 194 be rewritten for clarity. First, data augmentation is a class of methods, not a single method. Second, Meyer et al. use copula-based models in particular to augment datasets. Do you use copula-based methods? The way this reference is cited suggests that you follow their approach. Third, do you randomly sample observed data, or synthetically generated data, or both? Do you randomly sample only the data which are hourly, e.g., air temperature, net radiation, N2O, CO2, and VWC? How is the daily value calculated from the sampled data? I did not find the answers to these questions to be clear from the text.

Response: We really appreciate your detailed comments on the data augmentation method. In this revision, we have deleted the confusing sentence "Data augmentation is a typical practice in ML when training data is limited (Meyer et al., 2021)" because we did not intend to highlight one particular method, but only to explain the data augmentation concept using one recent citation. To your second question, the augmentation method is only used on observed data and corresponding weather forcings. To your third question, we only randomly sample the data which are hourly. Lastly, we used the average of the 16 hours (or maximum valid hours) of data to represent the daily values. We have addressed all those questions in the new paragraph in section 2.2.2 (from line 213 to 223):

“To reduce overfitting and increase the generalization of the trained model based on the small amount of mesocosm data, we applied the following method to augment the experimental measurements and weather forcings to 1000 times larger by sampling hourly data and averaging them to daily scale. In this method, 16 hours (or maximum valid hours) of data are randomly selected from 24 hours of data to compute their mean as the daily value. Since 3/4 of the day are covered by the selected data (16 hours /24 hours), the augmented daily values should be representative enough for the source day and meanwhile present slight variations. Furthermore, the observation ratio, (24 hours - missing hours) / 24 hours, can be used as the weights in loss function to inject the data quality information in model optimization. If the day has more than 16 hours missing values, we consider the observations in that day as not trustworthy and drop the day by setting the weight to 0. This method can not only augment the data to 1000 times larger but also deal with the missing values in observed data inherently. The total amount of observed mesocosm data and related weather forcings are augmented to 122 days x 3 years x 6 chambers x 1000 data samples in this study.”

8. How well does the model perform out-of-sample? Out-of-sample performance is mentioned in the introduction, but the discussion does not address it.

Response: We totally agree with the reviewer that out-of-sample performance would be critical for predictive models. Thus we have mentioned in the introduction that out-of-scenario ability is the limitation of machine learning models. In our study, we have compared the out-of-sample performance between different models using the period without any observation data in section 3.2 paragraph 1 (from line 363 to 366):
“For the region without observation data (normally before day 25), KGML-ag1 predicted stable N_2O fluxes close to 0 mg N m^{-2} day^{-1} (which is close to the reality in the experiment setting) while GRU caused anomalous peaks of fluxes. This is because KGML-ag1 has learned knowledge for the whole period from the pretraining process with *ecosys* model generated synthetic data, but GRU model has no prior knowledge for the period without any data in observations;”
and section 3.3 last paragraph (from 424 to 429):
“From these comparisons, we infer that without considering temporal dependence and pretraining process, the tree-based model including DT, RF, GB and XGB and deep learning model ANN predicted erratic peaks in almost every missing data point, while GRU model was stable in small gaps and only presented poor performance in long missing period (before 25 day). This improvement by GRU model can be attributed to the structure of GRU that naturally keeps the historical information using hidden states, which enables GRU to consider the temporal dependence and make consistent predictions over time.”
Moreover, the objective for this study is to explore ways to incorporate knowledge into ML models for improving agriculture ecosystem simulation. The mesocosm experiment measured many inputs and intermediate variables in addition to the output of N_2O fluxes, thus serving as a unique testbed. Continuous N_2O flux data with a comprehensive set of input and intermediate variables, especially those

at hourly or daily scales, are very limited. Some recent projects funded by the US Department of Energy have started to collect such datasets in real-world fields, but the data has not been released. While we fully understand the importance of out-of-sample testing, working with another dataset is beyond the scope of this manuscript.

Technical corrections

- At line 239, Sec. 4.4 does not exist.

Response: We have corrected the sentence by replacing 4.4 to existing 4.3 (Line 262).
"... and would guide future N2O related measurements and KGML model development (discussed in Sec. 4.3)."

- At line 240, I believe this should refer to Fig. 1c and 1d, not 1b and 1c.

Response: We have corrected this mistake (line 264).
"Next we used the knowledge learned from synthetic data to develop the structure of KGML-ag (Fig. 1c-d)."

- Tables 1 and 2 have identical captions but different contents.

Response: We have corrected this by replacing the right caption.
"Table 2: Prediction accuracy comparisons between non-pretrained GRU model and KGML-ag1."

- Sections 4.1 and 4.2 are both entitled "Interpretability of KGML-ag."

Response: We have replaced the section 4.2 title to "Lessons for KGML-ag development"

To Reviewer 3

The authors are proposing the development of a new approach KGML-ag to machine learning in estimating N2O emissions from fertilized agricultural fields. This approach involves using data generated from a process model and a mesocosm experiment to tune the relationships and their parameters among input and intermediate variables by which N2O emissions are thought to be governed. The advantages of this approach over process models are simplified input data requirements, more rapid model execution, and possibly more accurate simulation of N2O fluxes measured in experiments for which the model is tuned.

Response: We really appreciate the reviewer correctly recognizing our efforts and achievements. We want to incorporate the domain knowledge learned from agroecosystem process-based model *ecosys* to the advanced machine learning models to combine the advantages from both. Developing KGML-ag is one of the very first few attempts to realize the concept of hybrid modeling (Reichstein et al. 2019 Nature) in simulating agroecosystem biogeochemistry. To further improve our manuscript, we have carefully revised the content based on all reviewers' comments and suggestions.

The ability of this approach to simulate N2O emission events under controlled laboratory conditions is impressive. It should be noted that the N2O emissions in Fig. 2 and the soil NO3 contents in Fig. 3 are much larger than those commonly encountered in field conditions. However the relationships and their parameters upon which this approach is based are not disclosed to the reader, and so remain a 'black box'. For example, in section 4.1 the processes governing the time course of N2O emissions following a urea application are described, but the method by which these processes were represented in KGML is not.

Response: We have double checked the N_2O emission and NO_3^- concentration magnitude from mesocosm and comparing with other field studies under similar conditions (Fassbinder et al., 2013; Grant et al., 1999, 2006, 2008; Hamrani et al., 2020; Venterea et al., 2011). It turned out that our magnitude for N_2O (peak value around 20 mg N m^{-2} day^{-1}) and NO_3^- (peak value around 50 g N m^{-2}) are within the field observed ranges for managed crop soils. The reviewer's impression that these values being "too large" is likely because of the different units we used. Here all units are converted to daily scale as a default setting in *ecosys*, while other studies often report N fluxes using mg N m^{-2} h^{-1} for N_2O flux and mg N kg^{-1} for NO_3^- concentration (in this case, peak values in our experiment are 1 mg N m^{-2} h^{-1} and 40 mg N kg^{-1}). To avoid future misunderstandings of the data, we first add a sentence in data description section 2.2.2 to include the comparisons with other studies (From line 198 to 201) and then add units in Figure 2 and Figure 3 caption to notify readers about the different units being used.

"The magnitude of N_2O flux and NO_3^- soil concentration and their responses following fertilizer application from this mesocosm experiment are consistent with several field studies of agricultural soils (Fassbinder et al., 2013; Grant et al., 1999, 2006, 2008; Hamrani et al., 2020; Venterea et al., 2011)."

"Figure 2: N_2O flux time series comparisons among pure non-pretrained GRU predictions (blue line), KGML-ag1 predictions (red line) and observations (black line-dot) from cross-validation. The N_2O flux unit is mg N m^{-2} day^{-1}."

"Figure 3: IMVs prediction from KGML-ag1. The black-dot line represents observations and the red line represents the results from KGML-ag1. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber. The CO_2 flux and soil NO_3- concentration units are g C m^{-2} day^{-1} and g N m^{-2}, respectively."

"Figure 3 Contd.: IMVs prediction from KGML-ag1. The black-dot line represents observations and the red line represents the results from KGML-ag1. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber. The soil NH_4^+ concentration and soil VWC units are g N m^{-2} and m^3 m^{-3}, respectively."

We would like to note that this study is one significant step towards none-black box use of machine learning, but fully opening the black box is one of the frontiers in ML research that still has a long way to go. We partially opened the black box by incorporating domain knowledge into a completely black box ML model via three efforts: 1) building a hierarchical structure (with black-box GRU model as basis) to simulate the important intermediate variables (IMVs) first; then the predicted IMVs are used as the additional inputs in target variable simulation (e.g. N_2O), which will provide an opportunity to track those IMVs during the simulation period; 2) pretraining the KGML model with a process-based model so that the KGML model can perform as a surrogate model of the process-based model; 3) other techniques like using initial values to preserve state, feature importance analysis and stepwise training and fine tuning etc. With these implementations, our KGML model not only outperformed pure ML models but also was more interpretable. The ability to predict IMVs also shed light on model improvement, which is not possible or much more complicated with pure ML models.

Regarding the relationships and parameters, we will make the KGML-ag code and neural network weights open through Github once the review process is done. But explicitly describing these like what is often done for process-based models is not practical because KGML-ag is essentially a neural network model, and readers may not be able to infer much directly from layers, nodes and weights.

Finally, we agree with the reviewer that in some cases why KGML performed so well needs to be explained, but this would not deny our contribution towards opening the "black box". To reflect the reviewer's concern, we have added in the discussion section 4.3 last paragraph (from line 558 to 562) that:

"Finally, at the current stage we can not claim to have completely opened the black box of KGML-ag, but this framework is a significant step towards this goal. For example, some ideas implemented in our study, such as using pretraining to transfer knowledge from PB model to ML model, incorporating causal relations by hierarchical structure, predicting IMVs for tracking middle changes and using initial values as input to reduce data demand, would shed light on the future KGML-ag model improvement."

As for all black box approaches to modelling, it is vitally important that KGML be subjected to tests with truly independent datasets, i.e. datasets that are completely separate, and preferably very different, from those used in model calibration. Impressive results can always be achieved by calibrating enough parameters, but are these parameters robust? The extent to which such testing of KGML was conducted in this paper is not clear. At the very least, for this paper to be publishable, calibration and validation of KGML must be clearly distinguished, and clear evidence of independent testing must be provided. Further description of the key relationships and their parameters that govern N2O emissions in the model should also be provided so as to improve confidence in its robustness.

Response: We agree with the reviewer that out-of-sample testing is critical for model development. In this work all results reported in Figure 4 and Figure 5 are from leave-one-out experiment. For example, we trained KGML with data from chamber 1-5 and tested it against the left out chamber 6 as the model performance. Another out-of-sample test is by comparing the prediction performance during the periods without any chamber observation data (i.e. before April 25th of each year). Results show that KGML-ag1 predicted stable N_2O fluxes close to 0 mg N m^{-2} day^{-1} (which is close to the reality in the experiment setting) while GRU caused anomalous peaks of fluxes. This highlighted the power of KGML because KGML-ag1 has learned "knowledge" for the whole period from the pretraining process using *ecosys* model generated synthetic data. Relevant text can be found in 363-366:

"For the region without observation data (normally before day 25), KGML-ag1 predicted stable N2O fluxes close to 0 mg N m-2 day-1 (which is close to the reality in the experiment setting) while GRU caused anomalous peaks of fluxes. This is because KGML-ag1 has learned knowledge for the whole period from the pretraining process with *ecosys* model generated synthetic data, but GRU model has no prior knowledge for the period without any data in observations;"

and in lines 424-429:

"From these comparisons, we infer that without considering temporal dependence and pretraining process, the tree-based model including DT, RF, GB and XGB and deep learning model ANN predicted erratic peaks in almost every missing data point, while GRU model was stable in small gaps and only presented poor performance in long missing period (before 25 day). This improvement by GRU model can be attributed to the structure of GRU that naturally keeps the historical information using hidden states, which enables GRU to consider the temporal dependence and make consistent predictions over time."

We understand these two out-of-sample tests are not in the sense of being "very different" from what the KGML model was developed. However, this is so far the best data we can access. The mesocosm experiment data we used in this study has provided a comprehensive set of inputs and intermediate variables in addition to the output of N_2O fluxes, thus serving as a unique testbed. Continuous N_2O flux measurements along with a comprehensive set of input and intermediate variables, especially those at hourly or daily scales, almost do not exist or are not publicly accessible. Some recent projects funded by the US Department of Energy have started to collect such gold standard dataset under field conditions, but the data needs to be accumulated for another one or two years before release. We anticipate that gold standard data will significantly benefit the development of the KGML-ag model.

Finally, we argue that the novelty and robustness of our study can be justified in a different perspective. Our results show that a well-calibrated *ecosys* is not able to reproduce many dynamics of observed N_2O fluxes (Fig. S9) regardless how we tune *ecosys* parameters. A pure ML model can better reproduce the time series, but still has missed several key peaks in growing season while falsely predicted spring peak emissions even though fertilizers were not applied until several days later (Fig. 2). The KGML-ag1 leveraged the advantage of *ecosys* and the pure ML model, and outperformed both (Fig. 2). These nested comparisons clearly demonstrate the power of KGML as a framework. While we do not argue that KGML-ag is a perfect model that would be directly applicable to other places, sharing our approach will provide food-for-thought to the community on how to build a hybrid biogeochemical model that is computationally more efficient and more robust than both process-based and ML-based models. We have added new discussions about this concern in the last paragraph of section 4.3 (from line 562 to 566).

“Besides, we acknowledge the importance of further testing the KGML-ag over completely independent datasets, but results presented in this manuscript are sufficient to justify the power of KGML as a framework. The mesocosm experiment data we used in this study has provided a comprehensive set of inputs and intermediate variables in addition to the output of N_2O fluxes, thus serving as a unique testbed. We expect our validation results will be more solid once more gold standard data of N_2O fluxes along with other relevant inputs and intermediate variables become publicly available.”

In the Discussion, the authors rightfully address some of the factors that may limit the robustness of KGML. These limitations will likely become more apparent when the authors conduct tests of KGML under field conditions. Addressing these factors, as described by the authors, appears to require that KGML more closely resemble process-based models, and may reduce the computational advantages claimed for the KGML approach.

Response: The reviewer’s concern on decreased performance in field application is legit, and is a good hypothesis to test when more dataset become available. At this stage, we do not know whether or not these limitations will become more apparent under field conditions. But we are currently collecting new gold standard data of inputs, intermediate and N_2O fluxes from both field and lab experiments, which will be used to test the reviewer’s hypothesis. We would also like to acknowledge that KGML-ag’s limitations apply to both pure ML model and process-based models under field conditions, so it is very likely KGML-ag will continue to outperform both.

Another concern by the reviewer is that further development of KGML will make it resemble process-based models, thereby reducing the computational advantages. We argue this is unlikely because the application of neural networks is faster than process-based models by multiple orders. To surrogate as many components of process-based models as possible is one research frontier in hybrid modeling for earth system science (Reichstein et al. 2019 Nature; Irrgang et al. 2021 Nature Machine Intelligence), with latest advances occurred in weather forecast (Bauer et al. 2021 Nature Computational Science). By using a hybrid model, computationally inefficient components of PB can be identified one by one, and be replaced with more efficient ML-based surrogates to eventually obtain the most efficient model, thereby resolving the concern raised by the reviewer. We have added the new discussion at the end of section 4.3 to address the reviewer’s concern (from line 566 to 573).

“Moreover, incorporating more and more domain knowledge into KGML-ag will be inevitable in further improvement, but we don’t think KGML-ag will become inefficient as it becomes more like the PB model. In fact, to efficiently surrogate components of PB models has been proposed as a research frontier in hybrid modeling for earth system science (Reichstein et al., 2019; Irrgang et al., 2021), with latest advances occurring in weather forecasts (Bauer et al., 2021). By using a hybrid model, computationally inefficient components of PB can be identified one by one, and be replaced with more efficient ML-based surrogates to eventually obtain the most efficient model. Further KGML-ag model development will also need to balance efficiency, accuracy and interpretability.”

Reference:

Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C., & Wedi, N. P.: The digital revolution of Earth-system science. Nature Computational Science, 1(2), 104-113, 2021.

Fassbinder, J. J, Schultz, N. M, Baker, J. M, & Griffis, T. J.: Automated, Low-Power Chamber System for Measuring Nitrous Oxide Emissions, Journal of environmental quality, 42, 606. doi: 10.2134/jeq2012.0283, 2013.

Grant, R. F., & Pattey, E.: Mathematical modeling of nitrous oxide emissions from an agricultural field during spring thaw. Global Biogeochemical Cycles, 13(2), 679-694, 1999.

Grant, R. F., & Pattey, E.: Temperature sensitivity of N_2O emissions from fertilized agricultural soils: Mathematical modeling in *ecosys*. Global biogeochemical cycles, 22(4), 2008.

Grant, R. F., Pattey, E., Goddard, T. W., Kryzanowski, L. M., & Puurveen, H.: Modeling the effects of fertilizer application rate on nitrous oxide emissions, Soil Science Society of America Journal, 70(1), 235-248, 2006.

Hamrani, A., Akbarzadeh, A., & Madramootoo, C. A.: Machine learning for predicting greenhouse gas emissions from agricultural soils, Science of The Total Environment, 741, 140338, 2020.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. Nature Machine Intelligence, 3(8), 667-674, 2021.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204, 2019.

Venterea, R. T., Maharjan, B., & Dolan, M. S.: Fertilizer source and tillage effects on yield‐scaled nitrous oxide emissions in a corn cropping system. Journal of Environmental Quality, 40(5), 1521-1531, 2011.

KGML-ag: A Modeling Framework of Knowledge-Guided Machine Learning to Simulate Agroecosystems: A Case Study of Estimating N_2O Emission using Data from Mesocosm Experiments

Licheng Liu[1], Shaoming Xu[2], Jinyun Tang[3~~4~~], Kaiyu Guan[4~~5~~,5~~6~~,6~~7~~], Timothy J. Griffis[7~~8~~], Matthew D. Erickson[7~~8~~], Alexander L. Frie[7~~8~~], Xiaowei Jia[8~~9~~], Taegon Kim[1, 9], Lee T. Miller[7~~8~~], Bin Peng[4~~5~~,5~~6~~,6~~7~~], Shaowei Wu[10], Yufeng Yang[1], Wang Zhou[4~~5~~,5~~6~~], Vipin Kumar[2], Zhenong Jin[1,11~~3~~]*

[1]Department of Bioproducts and Biosystems Engineering, University of Minnesota, Saint Paul, MN, 55108, USA
[2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455, USA
~~[3]Institute on the Environment, University of Minnesota, Saint Paul, MN, 55108, USA~~
[3~~4~~]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4~~5~~]Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[5~~6~~]Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[6~~7~~]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[7~~8~~]Department of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN 55108, USA
[8~~9~~]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, 15260, USA
[9]Department of Smart Farm, Jeonbuk National University, Jeonju, Jeollabuk-do, 54896, Republic of Korea
[10]School of Physics and Astronomy, University of Minnesota, Minneapolis, MN, 55455, USA
[11]Institute on the Environment, University of Minnesota, Saint Paul, MN, 55108, USA

Correspondence to: Zhenong Jin (jinzn@umn.edu)

Abstract.

Agricultural nitrous oxide (N_2O) emission accounts for a non-trivial fraction of global greenhouse gases (GHGs) budget. To date, estimating N_2O fluxes from cropland remains a challenging task because the related microbial processes (e.g., nitrification and denitrification) are controlled by complex interactions among climate, soil, plant and human activities. Existing approaches such as process-based (PB) models have well-known limitations due to insufficient representations of the processes or uncertainties~~constraints~~ of model parameters, and to leverage recent advances in machine learning (ML) a new method is needed to unlock the "black box" to overcome its limitations such as~~due to~~ low interpretability, out-of-sample failure and massive data demand. In this study, we developed a first-~~ ~~of-~~ ~~the~~its ~~kind knowledge-guided machine learning model for agroecosystems (KGML-ag), by incorporating biogeophysical/chemical domain knowledge from an advanced PB model, *ecosys*, and tested it by comparing simulating daily N_2O fluxes with real observed data from mesocosm experiments. The Gated Recurrent Unit (GRU) was used as the basis to build the model structure. To optimize the model performance, we have investigated a range of ideas, including: 1) Using initial values of intermediate variables (IMVs) instead of time series as model input to reduce data demand; 2) Building hierarchical structures to explicitly estimate IMVs for further N_2O prediction; 3) Using multitask learning to balance the simultaneous training on multiple variables; and 4) Pretraining with millions of synthetic data generated from *ecosys* and fine tuning with mesocosm observations. Six other pure ML models were developed

Formatted: Numbering: Continuous

Formatted: Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font color: Auto

Formatted: Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font color: Auto

using the same mesocosm data to serve as the benchmark for the KGML-ag model. Results show that KGML-ag did an excellent job in reproducing the mesocosm N_2O fluxes (overall r^2 = 0.81, and RMSE = 3.6 mg N m^{-2} day^{-1} from cross-validation). Importantly KGML-ag always outperforms the PB model and ML models in predicting N_2O fluxes, especially for complex temporal dynamics and emission peaks. Besides, KGML-ag goes beyond the pure ML models by providing more interpretable predictions as well as pinpointing desired new knowledge and data to further empower the current KGML-ag. We believe the KGML-ag development in this study will stimulate a new body of research on interpretable ML for biogeochemistry and other related geoscience processes.

1 Introduction

Nitrous oxide (N_2O), with its global warming potential 273 ± 118 times greater than that of carbon dioxide (CO_2) for a 100-year time horizon, is one of the major~~important~~ greenhouse gases (IPCC6; Forster et al., 2021). The increasing rate of atmospheric N_2O concentration during the period 2010-2015 is 44% higher than during 2000-2005, mainly driven by increased anthropogenic sources that have increased total global N_2O emissions to ~17 Tg N yr^{-1} (Syakila and Kroeze, 2011; Thompson et al., 2019). It is estimated that approximately 60% of the contemporary N_2O emission increases are from agriculture management at global scale (Pachauri et al., 2014; Robertson et al., 2014; Tian et al., 2020), but the estimation uncertainty can exceed 300% (Barton et al., 2015; Solazzo et al., 2021). Quantifying N_2O emissions from agricultural soils is extremely challenging, partly because the related microbial processes, mainly about incomplete denitrification and nitrification, are controlled by many environment and management factors such as temperature/water conditions, soil/crop properties, and N fertilization rate, all of which together have collectively led to large temporal and spatial variabilities of N_2O emissions (Butterbach-Bahl et al., 2013; Grant et al., 2016).

Process-based (PB) models are often used for simulating N_2O fluxes from ~~the~~ agroecosystems, but they have some inherent limitations, including incomplete knowledge of the processes, low accuracy due to the under-constrained parameters, expensive computing cost, and rigid structure for further improvements, that we could not resolve by using PB model itself. For example, an advanced agroecosystem model, *ecosys* (Grant et al., 2003, 2006, 2016), simulates N_2O production rates through nitrification and denitrification processes when oxygen (O_2) is limited, with equations considering the influence from related substrate concentrations (e.g., NO_2^-, N_2O, and CO_2), nitrifier and denitrifier populations, and soil thermal, hydrological physical and chemical conditions. The produced N_2O accumulates, transfers in gaseous phase, aqueous phase, over different soil layers, and eventually exchanges with atmosphere at the soil surface. Other PB models, including DNDC (Zhang et al., 2002; Zhang and Niu, 2016), DAYCENT (Del Grosso et al., 2000; Necpálová et al., 2015), and APSIM (Keating et al., 2003; Holzworth et al., 2014), have also included processes to simulate N_2O production, but adopt different parameterizations using static partition parameters to estimate N_2O emission from nitrification, and other empirical parameters to control the influence on nitrification from soil water content, pH, temperature and substrate concentrations. Besides, N_2O is intimately connected with the soil organic carbon (SOC) dynamics, because soil nitrifiers and denitrifiers interact strongly with aerobic and anaerobic heterotrophs that process SOC evolution, and all of these microbes are driven by shared environmental variables including soil temperature, moisture, redox status, and physical and chemical properties (Thornley et al., 2007). As expected, these connections make it difficult for PB models, even the most advanced ones like *ecosys*, to find sufficient representations of the physical and biogeochemical processes or obtain enough data to calibrate a large number of model parameters with strong spatio-temporal variations. Thus, novel approaches are needed for addressing the big challenge of agricultural N_2O flux simulations.

Machine learning (ML) models can automatically learn patterns and relationships from data. Recent studies have investigated the potential to predict agricultural N_2O emission with ML models, including random forest (RF, Saha et al., 2021), metamodelling with extreme gradient boosting (XGBoost) (Kim et al., 2021), and deep learning neural network (DNN) (Hamrani et al., 2020). Notably, Hamrani et al. (2020) compared nine widely used ML models for predicting agricultural N_2O. That study pointed out that the long short term memory (LSTM) model with recurrent networks containing memory cells as building blocks will be most suitable for N_2O predictions, but the challenge remains with respect to the ability of capturing the sharp peak of N_2O fluxes and lag time between N fertilizer application and the emission peak. Although there is an increasing interest in leveraging recent advances in machine learning, capturing this opportunity requires going beyond the ML limitations, including limited generalizability to out-of-sample scenarios, demand for massive training data, and low interpretability due to the "black-box" use of ML (Karpatne et al., 2017). PB models with their transparent structures built by representations of physical and biogeochemical processes, seem to be exact complementary to ML models. Thus, combining the power of ML model and PB model understanding innovatively is likely a path forward.

The above need to integrate ML and PB models can be potentially~~ssibly~~ addressed by the newly proposed framework of Knowledge-guided Machine Learning (KGML) models. In the review by Willard et al. (2021), five research frontiers have been identified regarding the development of KGML for diverse disciplines including earth system science, they are: 1) Loss function design according to physical or chemical laws (Jia et al., 2019, 2021; Read et al., 2019); 2) Knowledge-guided initialization through pretraining ML models with synthetic data generated from PB models (Jia et al., 2019, 2021; Read et al., 2019); 3) Architecture design according to causal relations or adding dense layers containing domain knowledge (Khandelwal et al., 2020; Beucler et al., 2019, 2021); 4) Residual modeling with ML models to reduce the bias between PB model outputs and observations (Hanson et al., 2020); and 5) Other hybrid modeling approaches combining PB and ML models (Kraft et al., 2021). These recent advances in KGML pave the pathway to a more efficient, accurate and interpretable solution for estimating N_2O fluxes from the agroecosystem.

In this study, we present a~~the~~ first-of-its-kind attempt of developing a~~the~~ KGML for agricultural GHG fluxes prediction (KGML-ag) with knowledge-guided initialization and architecture design, and demonstrate the potential of KGML-ag with a case study on quantifying N_2O flux observed by a multi-year mesocosm experiments. We designed the KGML-ag structure based on the causal relations of related N_2O processes informed by an advanced agroecosystem model, *ecosys* (Grant et al., 2003, 2006, 2016). We used the synthetic data generated from *ecosys* to design the KGML-ag input/output, and to pre-train the KGML-ag model to learn the basic patterns of each variable. Observations from multi-season controlled-environment mesocosm chambers (Miller, 2021, thesis; Miller et al., 2021, in review) were used to refine the pretrained KGML-ag and evaluate the model performance. Since there is limited literature that guides the development of KGML-ag and not a one that directly addressed GHG fluxes, we investigated a range of ideas to optimize the model performance, including: 1) Using initial values of intermediate variables (IMVs) instead of sequences as model input to reduce data demand; 2) Building hierarchical structures to explicitly estimate IMVs for further N_2O prediction; 3) Using multitask learning to balance the simultaneous training on multiple variables; and 4) Pretraining with millions of synthetic data generated from *ecosys* and fine tuning with mesocosm observations. Although we evaluated the KGML-ag models with real measurements only from a mesocosm experiment, the lessons learned from the development process and various KGML-ag structures can be transferred to other data, other variables and large scale simulations, therefore have broader implications on further KGML related research in agriculture. We believe this study will stimulate a new body of research on interpretable machine learning for biogeochemistry and other related topics in geoscience.

2 Methods

2.1 Experimental design overview

To develop and evaluate the KGML-ag models and compare their performance with pure ML models, we designed the following experiments:

1) With the synthetic data, we developed and pretrained multiple KGML-ag models to learn general patterns and interactions among variables, and evaluated their model performance (Fig. S2, Table 1);
2) With the observed data, we finetuned multiple KGML-ag models to adapt real-world situations, and evaluated their model performance (Fig. 2-3; Fig. S3-5; Table 2-3);
3) We further benchmarked KGML-ag models and uncertainties with other pure ML models without considering temporal dependence, including Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB) from the sklearn package (https://scikit-learn.org/stable/), Extreme Gradient Boosting (XGB) from the XGBoost package (https://xgboost.readthedocs.io/en/latest/) and a 6-linear-layer artificial neural network (ANN) with the mesocosm experiment data by 10 times ensemble experiments (Fig. 4-5; Fig. S6-8);
4) We conducted a few small experiments to further investigate how various model configurations, such as the pretraining process, data augmentation and IMV initial values would influence KGML-ag model performance (Table 3).

2.2 KGML-ag structure development

2.2.1 Generating synthetic data with *ecosys*

We generated synthetic data using a PB model, *ecosys*. The *ecosys* model is an advanced agroecosystem model constructed from detailed biophysical and biogeochemical rules instead of using empirical relations (Grant et al., 2001). It represents N_2O evolution in the microbe-engaged processes of nitrification-denitrification using substrate kinetics that are sensitive to soil nitrogen availability, soil temperature, soil moisture, and soil oxygen status (Grant and Pattey 2008). Two groups of microbial populations, autotrophic nitrifiers and heterotrophic denitrifiers, produce N_2O with specific competitive or cooperative relations in ecosys when O_2 availability fails to meet O_2 demand for their respirations and NO_2^- become alternative electron acceptors. N_2O transfer within soil layers and from soil to the atmosphere is driven by concentration gradient using diffusion-convection-dispersion equations, in the forms of gaseous and aqueous N_2O under control of volatilization-dissolution (Grant et al., 2016). Unlike the pipeline model described by Davidson et al. (2000) , which mainly consider the correlations of N_2O production with nitrogen availability and of N_2O emissions~~emisiontsting~~ with soil water content, ecosys enables integrative effects of energy, water, nitrogen availability on N_2O production and N_2O transfer via the microbial population dynamics and their interactions with soil, plant, and atmospheric dynamics, under diverse meteorological and anthropogenic disturbances (e.g. runoff, drainage, tillage, irrigation, soil erosion). Many previous studies have demonstrated its robustness in simulating agricultural carbon and nitrogen cyclings at different spatial/temporal scales, and under different management practices (Grant et al., 2003, 2006, 2016; Metivier et al., 2009; Zhou et al., 2021). For the agricultural ecosystems in the US Midwest, whose simulations are used for synthetic data in this study, the performance of *ecosys* on CO_2 and N_2O fluxes have been extensively benchmarked, including CO_2 exchange (NEE, $R^2 = 0.87$) and leaf area index (LAI, $R^2 = 0.78$) from six flux towers, USDA census reported corn yield ($R^2 = 0.83$) and soybean yield ($R^2 = 0.80$), satellite-derived GPP for corn ($R^2 = 0.83$) and soybean ($R^2 = 0.85$) from Illinois, Iowa and Indiana, and cumulative N_2O emissions ($R^2 = 0.36$) across eight Midwestern states (Wang et al., 2021; Yang et al., 2022). Therefore, *ecosys* is an appropriate choice of domain knowledge provider and synthetic data generator in the development of KGML models. We generated daily synthetic data including N_2O flux and 76 IMVs (e.g. CO_2 flux from soil, layerwise soil NO_3^- concentration, layerwise soil temperature, and layerwise soil moisture; detailed in Table S1) from *ecosys* simulations for 2000-2018 over 99 randomly selected counties in Iowa, Illinois, and Indiana, USA. We used hourly meteorological inputs (downward shortwave radiation, air temperature, precipitation, relative humidity, and wind speed) from the phase 2 of North American Land Data Assimilation System (NLDAS-2, Xia et al., 2012) and layerwise soil properties (e.g. bulk density, texture, pH, SOC concentration) from the SSURGO database (Soil Survey Staff, 2020) as inputs to *ecosys*. Crop management except N fertilization rates were configured to the same settings as mesocosm experiments (described in Sec 2.2.2). To increase the variability in synthetic data, we implemented 20 different N fertilization rates ranging from 0 to 33.6 g N m^{-2} (i.e. 0 to 300 lb N ac^{-1}) in each simulation of 99 counties, and more detailed information for model setup refers to Zhou et al. (2021).

Commented [1]: Grant, R. F., & Pattey, E. (2008). Temperature sensitivity of N2O emissions from fertilized agricultural soils: Mathematical modeling in ecosys. Global Biogeochemical Cycles, 22(4). https://doi.org/10.1029/2008gb003273
Grant, R. F., Neftel, A., & Calanca, P. Ecological Controls on N2O Emission in Surface Litter and Near-surface Soil of a Managed Pasture: Modelling and Measurements. Biogeosciences 13 3549–71

Commented [2R1]: Hi Yufeng, we expect a little more details of the similarity and differences between ecosys model and the pipe line model (like what mentioned in figure 2, Davidson et al., 2000, https://academic.oup.com/bioscience/article/50/8/667/243260) of N2O simulation.
I remember your previous version of manuscript (https://docs.google.com/document/d/1KbQnglvd_pFfPccIl5eq592fKzbSVybL5mOVciSwmq8/edit?usp=sharing) section 2.1 para2 has very detailed description of ecosys model N2O part.

Here could you help us 1) briefly check Davidson et al., 2000 N2O pipe line model and 2) compare with ecosys model to find the similarity and differences? (like something I intuitively think, a) ecosys comprehensively calculated some parts ...

Commented [3R1]: I added those general descriptions on ecosys N2O emissions. I've provided more details as needed and comparisons with the pipe model

Commented [4]: @yang6956@umn.edu Hi Yufeng, I may need your help here adding few sentences to 1) Explain the ecosys model structure of simulating N2O and the differences between simple pipeline model ; 2) ecosys model performance on various ecosystem for CO2/N2O. You should have pretty enough materials before from your paper for thi ...

Commented [5R4]: Original related review comments:

Robustness of physical (prior) knowledge

ecosys model plays a central role in guiding the ML model i ...

Commented [6R4]: Moreover, Wang has summarized few previous model simulation comparisons with field level observations here, in case you need (https://docs.google.com/document/d/1tRTQi7R-mCwMN_uqT-5klskl8cNad-Rs0Umeffk0u0U/edit?usp=sharing). In the maintext we may ...

Commented [7]: A problem here with ecosys N2O studies is that R2 or other quantitatives are missing in daily N2O flux comparison

Commented [8R7]: Hi Yufeng, is there a way to cite your paper in current stage? if yes please provide the detailed citation, thanks @yang6956@umn.edu

Commented [9R7]: Yang, Y., Liu, L., Zhou, W., Guan, K., Kim, T., Tang, J., Peng, B., Zhu, P., Grant, R. F., Griffis, T. J., Jin, Z., (2022). Distinct driving mechanisms of non-growing season N2O emissions call for spatial-specific mitigation strategies in the US Midwest. One Earth. Submitted.
Is it proper to provide the journal name directly as just ...

The generated synthetic data were then processed for further use by KGML-ag development. Meanwhile, the hourly weather forcings were converted to seven daily variables, including the maximum air temperature (TMAX_AIR, ºC), difference between the maximum and the minimum air temperature (TDIF_AIR, ºC), the maximum humidity (HMAX_AIR, fraction), difference between the maximum and the minimum humidity (HDIF_AIR, fraction), surface downward shortwave radiation (RADN, W m^{-2}), precipitation (PREC, mm day^{-1}), and wind speed (WIND, m s^{-1}). Six soil properties were retrieved from the SSURGO database, including total averaged (depth weighted averaged for all layers) bulk density (TBKDS, Mg m^{-3}), sand content (TCSAND, g kg^{-1}), silt content (TCSILT, g kg^{-1}), pH (TPH), cation exchange capacity (TCEC, $cmol^{+}$ kg^{-1}) and soil organic carbon (TSOC, g C kg^{-1}); and two crop properties were retrieved, including planting day of the year (PDOY) and crop type (CROPT, 1 for corn and 0 for soybean). Finally, each synthetic data sample has daily N_2O flux, 76 selected IMVs, 7 weather forcings (W), 1 N fertilization rate (FN, g N m^{-2}) and 8 soil/crop properties (SCP) (Fig. 1.a; Table S1). The periods from April 1st to July 31st (122 days) were selected to cover the mesocosm observations (around 30 days before and 90 days after N fertilizer date). The total amount of synthetic data sample is 122 days x 18 years x 99 counties x 20 N fertilizer rates (about 4.3 million data points). We randomly selected the samples from 70 counties for training, 10 counties for validation, and 19 counties for testing.

2.2.2 Mesocosm experiments for KGML-ag model fine-tuning and evaluation

Observations were acquired from a controlled-environment mesocosm facility on the St. Paul campus of the University of Minnesota. Soil samples were sourced in 2015 from a farm in Goodhue County, MN (44.2339º N and 92.8976º W), which had been under corn-soybean rotation for 25 years. Six chambers with a soil surface area of 2 m^2 and column depth of 1.1 m were used to plant continuous corn during 2015-2018 and monitor the N_2O flux response to different precipitation treatments. The experiment also measured other environmental variables including air temperature and photosynthetically active radiation (PAR), which were controlled to mimic the outdoor ambient environment. Granular urea fertilizer was hand broadcasted and incorporated to a depth of 0.05 m to each chamber at a rate of 22.4 g N m^{-2} (200 lb N ac^{-1}) on May 1st of 2015, May 4th of 2016 and May 3rd of 2017, and 10.3 g N m^{-2} (92 lb N ac^{-1}) on May 8th of 2018. Corn hybrid (DKC-53-56RIB) were hand planted to a depth of 0.05 m in two rows spaced 0.76 m apart 3-5 days after fertilizer application, at a seeding rate of 35,000 seeds ac^{-1} in 2015 to 2017, and 70,000 seeds ac^{-1} in 2018 but thinned upon emergence to ensure 100 percent emergence at 35,000 seeds ac^{-1}. Crops were harvested at the end of September by cutting the stover five inches above the soil. Hourly N_2O fluxes (mg N m^{-2} h^{-1}) and CO_2 fluxes (g C m^{-2} h^{-1}) were measured using non-steady-state flux chambers with a CO_2 analyzer (LI-10820 for 2016 and LI-7000 for 2017 and 2018, LI-COR Biosciences, Lincoln, NE) and a N_2O analyzer (Teledyne M320EU, Teledyne Technologies International Corp, Thousand Oaks, CA) (Detail method can be retrieved from Fassbinder et al., 2012, 2013). We also collected soil moisture at 15 cm depth (VWC as abbreviation of volumetric water content, m^3 m^{-3}), weekly 0-15 cm depth soil NO_3^- + NO_2^- concentration (NO_3^- for short in the following text, g N Mg^{-1}), soil NH_4^+ concentration (NH_4^+, g N Mg^{-1}), and related environment variables including air temperature, radiation, humidity and soil/crop properties from three growing seasons during 2016-2018 and six mesocosm chambers (Fig. S1). The magnitude of N_2O flux and NO_3^- soil concentration and their responses following fertilizer application from this mesocosm experiment are consistent with several field studies of agricultural soils (Fassbinder et al., 2013; Grant et al., 1999, 2006, 2008; Hamrani et al., 2020; Venterea et al., 2011). More details about the mesocosm facility and experimental design can be found in the thesis of Miller L. (2021).

The observed data were then processed to fine-tune and evaluate the KGML-ag models. The N_2O flux and four IMVs and weather variables were collected from the measurements in the selected period (i.e., April 1st to July 31st). Weekly NO_3^- (short for soil NO_3^- within 0-15 cm depth), and NH_4^+ (short for soil NH_4^+ within 0-15 cm) were linearly interpolated to the daily time scale on days containing VWC (short for soil VWC in 15 cm) data. Hourly air temperature, net radiation, N_2O (short for N_2O fluxes from soil), CO_2 (short for CO_2 fluxes from soil) and VWC were resampled to daily scale. All SCP were derived from mesocosm measurements except that TCEC was derived from the SSURGO database according to the soil origin. We used the leave-one-out cross-validation (LOOCV) method for the finetuning and evaluation process. Each time we used one chamber data for validation and another five chambers' data for model finetuning.

~~To increase the model generalization and avoid overfitting, we used the data augmentation method to enrich the finetuning data set to be 1000 times larger. Data augmentation is a typical practice in ML when training data is limited (Meyer et al., 2021). In particular, we randomly sampled 16 hours of data from a 24-hours period in each day and chamber, and then used the sampled data to calculate the daily value. If less than 16 missing values existed in 24 hours, we used the above method to sample the data and calculated a fraction number (24-missing value number)/24 to record valid data fraction in the mask matrix. If more than 16 missing values were found, we dropped this point and recorded 0 in the mask matrix. The final sample has daily N_2O flux, 4 IMVs, 7 weather forcing variables and 8 static soil/crop properties (similar to synthetic data). The total amount of augmented observed data sample is 122 days x 3 years x 6 chambers x 1000 data augmentations. The mask matrix is of the same size as the observed data sample but its elements range from 0 to 1.~~

To reduce overfitting and increase the generalization of the trained model based on the small amount of mesocosm data, we applied the following method to augment the experimental measurements and weather forcings to 1000 times larger by sampling hourly data and averaging them to daily scale. In this method, 16 hours (or maximum valid hours) of data are randomly selected from 24 hours of data to compute their mean as the daily value. Since 3/4 of the day are covered by the selected data (16 hours /24 hours), the augmented daily values should be representative enough for the source day and meanwhile present slight variations. Furthermore, the observation ratio, (24 hours - missing hours) / 24 hours, can be used as the weights in loss function to inject the data quality information in model optimization. If the day has more than 16 hours missing values, we consider the observations

Commented [10]: @xu000114@umn.edu Hi Shaoming, I may need your help here to clarify the data augmentation method. You may modify this part as many as you want but with a track. The basic strategy is to answer the three questions from Referee#2. The related comments list below:

**recommend that the paragraph starting at line 194 be rewritten for clarity. First, data augmentation is a class of methods, not a single method. Second, Meyer et al. use copula-based models in particular to augment datasets. Do you use copula-based methods? The way this reference is cited suggests that you follow their approach. Third, do you randomly sample observed data, or synthetically generated data, or both? Do you randomly sample only the data which are hourly, e.g., air temperature, net radiation, N2O, CO2, and VWC? How is the daily value calculated from the sampled data? I did not find the answers to these questions to be clear from the text.
Assigned to Shaoming Xu

in that day as not trustworthy and drop the day by setting the weight to 0. This method can not only augment the data to 1000 times larger but also deal with the missing values in observed data inherently. The total amount of observed mesocosm data and related weather forcings are augmented to 122 days x 3 years x 6 chambers x 1000 data samples in this study.

Commented [11]: Please check if this paragrah is well fit your needs to replace the previous paragraph. @lichengl@umn.edu

2.2.3 Gated Recurrent Unit (GRU) as the basis of KGML-ag

Hamrani et al. (2020) compared different models and reported that LSTM provided the highest accuracy in predicting N_2O fluxes, because N_2O flux is time dependent by its production/consumption nature and LSTM simulates target variables~~variable~~ by considering both current and historical states. The LSTM model, proposed by Hochreiter and Schmidhuber (1997), uses a cell state as an internal memory to preserve the historical information. At each time step, it creates a set of gating variables to filter the input and historical information and then uses the processed data to update the cell state. Similar to LSTM, GRU is a gated recurrent neural network but only keeps one hidden state (Cho et al., 2014). Though simpler than LSTM, GRU is proved to have similar performance (Chung et al., 2014). Our preliminary test on synthetic data for N_2O prediction showed that GRU indeed provided similar or higher accuracy and model efficiency under different model settings than LSTM (Table S2). This is possible~~likely~~ because simpler models with fewer weights and hyperparameters are more robust in combating the overfitting problem. Therefore, we choose GRU as the basis of KGML-ag development.

2.2.4 Incorporating domain knowledge to the development of KGML-ag

To quantitatively reveal the correlations between N_2O fluxes and IMVs and guide the KGML-ag development, we conducted ~~the~~ feature importance analysis by a customized 4-layer GRU ML model (Fig. 1b). Each layer of the model has a GRU cell with 64 hidden units. The 4-layer structure makes the model deeper and capable of capturing complex interactions. Between each GRU cell, 20% of the output hidden states are randomly dropped by replacing them with zero values (so called 20% dropout) to avoid overfitting. A linear dense layer is used to map the final output to N_2O. We first trained GRU models using~~by~~ synthetic data with different combinations of IMVs as inputs to predict the N_2O fluxes (original-test, Table S2). The feature importance analysis of well-trained models was then implemented by replacing one input feature with a Gaussian noise with mean $\mu=0$ and standard deviation $\sigma=0.01$, while keeping others untouched (new-test). The importance score was calculated by the new-test's root mean square error (RMSE) (replacing one feature) minus the original-test's RMSE (no replacing). RMSE was calculated by $\frac{\sqrt{\sum_1^N (y_i - y_i')^2}}{N}$ where N is the total number of observations across time and space, y_i is i-th measurement from synthetic data or observed data and y_i' is its corresponding prediction.

To find important variables for N_2O flux prediction in an ideal situation where~~that~~ all variables are available, we conducted a feature importance analysis for GRU models with all IMVs and basic inputs including FN, 7 W and 8 SCP (Fig. S2a). Results indicated that flux variables including NH_3, H_2, N_2, O_2, CH_4, evapotranspiration (ET) and CO_2 had significant influence on the model performance. Variables ranked high in feature importance analysis are~~should be~~ ~~primarily~~ considered with priority during model development. To develop a functionable KGML-ag~~ in real world~~, we further investigated the feature importance of four IMVs that are available from mesocosm observations including CO_2, NO_3^-, VWC and NH_4^+, which were ranked 7th, 20th, 58th, 60th respectively in 92 input features of synthetic data (Fig. S2a). We used these four available IMVs to create two input combinations: 1) CO_2 flux, NO_3^-, VWC and NH_4^+ (IMVcb1), and 2) NO_3^-, VWC and NH_4^+ (IMVcb2). The objective of building IMVcb2 was to investigate the importance of highly ranked variable CO_2 flux (by removing it from the inputs), and the impact of mixing-up flux and non-flux variables on model performance. We tested the feature importance of the GRU models built with IMVcb1 and IMVcb2 to check whether they would help in N_2O prediction (Fig. S2b-c). All the feature importance results above indicated the correlation intensity between N_2O and many other variables, which would help the KGML-ag model development and interpretation in this study (rest of this section and Sec. 3.1), and would guide future N_2O related measurements and KGML model development (discussed in Sec. 4.3~~4~~).

Next we used the knowledge learned from synthetic data to develop the structure of KGML-ag (Fig. 1c~~b~~-d~~c~~). Previous studies for KGML models have used physical laws, e.g., conservation of mass or energy, to design the loss function for constraining the ML model to produce physically consistent results (Read et al., 2019; Khandelwal et al., 2020). However, for complex systems like agroecosystems, it is challenging to incorporate physical laws, such as mass balance for N_2O, into the loss function due to the incomplete understanding of the processes and the lack of mass balance related data for validation. An alternative solution is to incorporate such information in the design of the neural network (Willard et al., 2021). Effectiveness of such an approach was demonstrated by Khandelwal et al. (2020) in the context of modeling stream flow in a river basin using Soil & Water Assessment Tool (SWAT). They used a hierarchical neural network to explicitly model IMVs (e.g., soil moisture, snow cover) and their relationships with the target variable (streamflow) and showed that this model is much more effective than a neural network that attempts to directly learn the relationship between input drivers and the target variables. Following this idea, we identified four desired features of an effective KGML-ag model, including: 1) We used initial values instead of sequence of the IMVs from synthetic data or observed data to provide a solid starting state for the ML system and reduce the IMV data demand, and then used the rest of the data to further constrain the prediction of IMVs; 2) We built a hierarchical structure based on the structure of process representation in~~causal relations derived from~~ *ecosys* to first predict IMVs and then simulate N_2O with predicted IMVs; 3) We trained all variables together using multitask learning to reach the best prediction scores, which generalized the model and incorporated interactions between IMVs and N_2O; 4) We initialized the KGML-ag model by pretraining with~~using~~ synthetic data before using real observed data to transfer physical knowledge, which further reduced the demand on large training samples and aided in faster convergence for fine-tuning.

To meet these desired features, we proposed two KGML-ag models (Fig. 1c-d). The first model, KGML-ag1, is a hierarchical structure containing two modules to simulate IMVs and N_2O sequentially. Each module is a 2-layer 64 units GRU ML model. The inputs to the module of the KGML-ag1 model for IMV predictions (KGML-ag1-IMV module) are FN, 7W and 8SCP together with the initial values of IMVs, and the outputs are IMV predictions. The inputs to the module of the KGML-ag1 model for N_2O predictions (KGML-ag1-N_2O module) are FN, 7W, 8SCP and predicted IMVs from KGML-ag1-IMV, and the output is the target variable N_2O. Linear dense layers were coded for both modules to map output states to IMVs or N_2O. The dropout method was applied to drop 20% of the state output between GRU cells and dense layers. The second model, KGML-ag2, is also a hierarchical structure similar to KGML-ag1, but has multiple KGML-ag2-IMV modules to explicitly simulate IMVs by tuning them separately in the fine-tuning process (discussed in Sec. 2.2.5). Each KGML-ag2-IMV module in KGML-ag2 is a 2-layer 64 units GRU cell with the inputs of FN+7W+8SCP and one IMV initial value, and the output of one IMV prediction. The KGML-ag2-N_2O module collects the IMV predictions from KGML-ag2-IMV modules and predicts the N_2O with inputs of FN+7W+8SCP and predicted IMVs.

2.2.5 Strategies for pretraining and fine-tuning processes

To increase the efficiency of the training process, we used the Z-normalization ($\frac{(X-\mu)}{\sigma}$, where X is the vector of a particular variable over all the data samples in the data set; μ is the mean value of X; σ is the standard deviation of X) method to normalize each variable separately on synthetic data. Then the scaling factors (μ, σ) derived from *ecosys* synthetic data for each variable were used to Z-normalize observed data into the same ranges as synthetic data. As mentioned in Sec. 2.2.1, the TDIF_AIR, HDIF_AIR were used instead of absolute min temperature (TMIN_AIR) and humidity (HMIN_AIR). This is done because TMIN_AIR and HMIN_AIR follow similar trends as TMAX_AIR and HMAX_AIR, making Z-normalization numerically poorly defined. Using the difference between maximum and minimum can provide a clearer information of daily air temperature/humidity variation.

During the pretraining process, we initialized the IMV of KGML-ag using the first day value of synthetic IMV time series. Adam optimizer with a start learning rate of 0.0001 was used for the training process. The learning rate would decay by 0.5 times after every 600 training epochs. At each epoch, synthetic data samples were randomly shuffled before being input to the model to predict N_2O (and IMVs if any). The mean square error (MSE) loss (calculation was equal to the square of RMSE) or sum of MSE loss (if multitask learning) between predictions and *ecosys* synthetic observations were calculated to optimize the weights of GRU cells. After the training process updated the model's weights, the validation process was performed to evaluate the model performance based on untouched samples with RMSE and the square of Pearson correlation coefficient (r^2). r^2 was calculated as $\frac{\left(\sum_i (y_i'-\underline{y_i'})(y_i-\underline{y_i})\right)^2}{\sum_i (y_i'-\underline{y_i'})^2(y_i-\underline{y_i})^2}$, where y_i is the i-th measurement from synthetic data or observed data, y_i' is its corresponding prediction, $\underline{y_i}$ is the mean of the measurement y in diagnosing space and $\underline{y_i'}$ is the mean of the predicted y' in diagnosing space. If both validated r^2 and RMSE were better than the best values in previous epochs, the updated model in this epoch would be saved. Normalized RMSE (NRMSE, calculated by RMSE/(max-min) of each variable observation) was introduced to evaluate IMV predictions between variables with different value ranges.

During the fine-tuning process, we used estimated IMV initial values of 1.0 g C m^{-2}, 0.2 m^3 m^{-3}, 0.0 g N Mg^{-1}, and 20.0 g N Mg^{-1} for CO_2, VWC, NH_4^+, and NO_3^- respectively, from starting day (April 1st) to the day before the first day of real observations, as input to KGML-ag models. Then the first-day values of observed IMVs were input into KGML-ag during the rest days of the period as IMV initial values. In addition, as described in Sec. 2.2.2, we used a data augmentation method to augment the total amount of data 1000 times larger for the fine-tuning process. The purpose of this data augmentation method was to increase the generalization of the fine-tuned model and to overcome the overfitting due to small sample size. The mask matrix was elementarily multiplied to the output matrix to calculate the MSE, r^2 and RMSE only for days with observations. The similar optimizer was used with an initial learning rate of 0.00005 and decay fraction of 0.5 per 200 epochs. Other training/validation methods in each epoch were similar to the pretraining process. Specifically, in the KGML-ag1 model finetuning process, we first froze the KGML-ag1-N_2O module and only trained the KGML-ag1-IMV module for IMVs. After finishing the KGML-ag1-IMV module training, we froze the KGML-ag1-IMV module and trained the KGML-ag1-N_2O module for N_2O. In the KGML-ag2 fine-tuning process, the similar freezing method was used but different KGML-ag2-IMV modules were trained separately one by one.

2.3 Development environment description

We used the Pytorch 1.6.0 (https://pytorch.org/get-started/previous-versions/) and python 3.7.9 (https://www.python.org/downloads/release/python-379/) as the programing environment for the model development. In order to use the GPU to speed-up the training process, we installed cudatoolkit 10.2.89 (https://developer.nvidia.com/cuda-toolkit). A desktop with Nvidia 2080 super GPU was used for code development and testing. The Mangi cluster (https://www.msi.umn.edu/mangi) from High Performance Computing of Minnesota Supercomputing Institute (HPC-MSI, https://www.msi.umn.edu/content/hpc) with 2-way Nvidia Tesla V100 GPU was used in training processes which consumed longer time and bigger memories.

3 Results

3.1 Pretraining experiments using synthetic data from *ecosys*

In the pretraining stage, the GRU model with 76 IMVs achieved the best performance in predicting N_2O fluxes (r^2=0.98, RMSE =0.54 mg N m^{-2} day^{-1} and normalized RMSE (NRMSE) = 0.01) on the test set of synthetic data generated from *ecosys* (Table

1). The high performance was due to some flux IMVs such as NH_3, H_2, O_2, CO_2 and ET, which are highly correlated to N_2O (Fig. S2a), were used as input to the model. The good performance of GRU with all IMVs indicates that ML models are able to perfectly mimic *ecosys* when sufficient information about IMVs is available. The GRU model with only basic input of N fertilizer rate, 7 weather forcings, and 8 soil/crop properties (FN+7W+8SCP) had the accuracy of r^2=0.89 and RMSE = 1.37 mg N m^{-2} day^{-1} (Table 1). The relatively low performance is likely because this model failed to capture several highly nonlinear pathways that are employed by *ecosys* to predict N_2O (e.g., one influence pathway from precipitation to N_2O can be: Precipitation → soil moisture → N components solubility/concentration → nitrification/denitrification rate/amount → soil N_2O concentration → gas N_2O flux). When adding sequences of IMV combinations (i.e., IMVcb1 of CO_2 flux, NO_3^-, NH_4^+ and VWC, and IMVcb2 of NO_3^-, NH_4^+ and VWC), the GRU models performed slightly better than the GRU model using only basic inputs, achieving r^2 of 0.92 and 0.90, respectively (Table 1). The KGML-ag1 with IMVcb1 and IMVcb2 initial values provided better performance (both r^2 = 0.90) than GRU with basic input and comparable performance to the GRU with inputs of IMVcb1 and IMVcb2 sequence. Besides, KGML-ag1 provided predicted IMVs of CO_2, NO_3^-, NH_4^+, and VWC with r^2 over 0.91, and NRMSE below 0.06 (Table 1). KGML-ag2 also provided comparable N_2O performance but relatively better IMVs performance of r^2 over 0.92 and NRMSE below 0.05. Results indicated that KGML-ag models with IMV initial values as extra input performed similar or better than pure ML models in synthetic data.

Formatted: Font: Italic

3.2 KGML-ag evaluation using observed data from mesocosm

After being fine-tuned with observed data, KGML-ag1 had N_2O prediction overall accuracy of r^2=0.81 and RMSE=3.6 mg N m^{-2} day^{-1}, while non-pretrained GRU model provided r^2=0.78 and RMSE=4.0 mg N m^{-2} day^{-1}, and pretrained GRU model provided r^2=0.80 and RMSE=3.77 mg N m^{-2} day^{-1} (Table 3). The time series of N_2O predictions from KGML-ag1 and the non-pretrained GRU model were further compared (Fig. 2), from which we found at least two advantages of using KGML-ag1 for N_2O predictions: 1) For the region without observation data (normally before day 25), KGML-ag1 predicted stable N_2O fluxes close to 0 mg N m^{-2} day^{-1} (which is close to the reality in the experiment setting) while GRU caused anomalous peaks of fluxes. This is because KGML-ag1 has learned knowledge ~~“common sense”~~ for the whole period from the pretraining process with *ecosys* model generated synthetic data, but GRU model has no prior knowledge for the period without any data in observations; 2) Although KGML-ag1 had a lower accuracy than GRU in some chambers, KGML-ag1 can better capture the temporal dynamics of N_2O fluxes compare to GRU, especially when the fluxes are highly variable (e.g. Fig 2 chamber 2).

To validate KGML-ag1 robustness, we further investigated the KGML-ag1 and GRU model performance in different temporal windows, shrinking from the whole period to the N_2O peak occurrence time (days 1-122, day 30-80, day 40-65 and day 45-60 for year 2016-2018), and performance in N_2O flux, first order gradient of N_2O (slope) and second order gradient of the N_2O (curvature) (Table 2). Slope represents the speed of N_2O flux changes through time and curvature represents the acceleration. Assessing prediction performance with~~on~~ these two metrics will reveal the model robustness on capture variable dynamics, which is critical when predicting fast-change variables with hot moments like N_2O. First of all, the overall r^2 and RMSE of

KGML-ag1 for values, slope and curvature were always better than GRU. In particular, KGML-ag1 captured the peak region (e.g., days 45-60) much better than GRU in both magnitude and dynamics (Table 2, Fig 2). Even for chamber 2 and 5 in which KGML-ag1 made worse N_2O predictions than GRU (Δr^2 ranging from -0.07 to -0.03), it better captured temporal dynamics than GRU in terms of slope (Δr^2 ranging from 0.08 to 0.16) and curvature (Δr^2 from 011 to 0.23) (Table 2). For other chambers, KGML-ag1 outperformed GRU consistently. For chamber 1, KGML-ag1 had worse N_2O predictions RMSE than GRU but the Δr^2 increased as the window shrinks to the peak emission time (0.07 → 0.13). The slope and curvature for chamber 1 also indicated that KGML-ag1 captured the dynamics much better than GRU. For chamber 3, KGML-ag1 predicted better N_2O but presented worse slope and curvature RMSE than GRU (Table 2). However, when explicitly investigating the time series of N_2O flux, slope and curvature in each year, KGML-ag1 outperformed GRU more significantly in 2017, the year with more complex temporal dynamics of N_2O fluxes, than in 2016 and 2018, especially for chamber 3 (Fig. 2; Fig. S3-4). This investigation supported that KGML-ag1 was more capable for complex dynamics predictions.

Interestingly, the fine-tuned KGML-ag1 model predicted reasonable IMVs including CO_2, NO_3^-, NH_4^+, and VWC with overall r^2 of 0.37, 0.39, 0.60, and 0.33 and NRMSE of 0.14, 0.21, 0.09 and 0.18, respectively (Table 3). The time series comparisons between IMV predictions and observations further indicated that KGML-ag1 could reasonably capture both magnitude and dynamics (Fig. 3). KGML-ag2 presented better IMVs predictions than KGML-ag1, with overall r^2 of CO_2, NO_3^-, NH_4^+, and VWC increasing by 0.37, 0.17, 0.06 and 0.51, and NRMSE decreasing by 0.05, 0.03, 0.01 and 0.10, respectively, but a slightly lower r^2 (decreasing 0.02) of N_2O (Table 3; Fig. S5). This indicated that explicitly simulating each IMV with separated KGML-ag2-IMV modules did not benefit the N_2O flux prediction accuracy, likely due to increasing model complexity which resulted in reduced~~ing~~ stability and ignoring the IMV interactions. In addition, we also found all KGML-ag models would perform better by using IMVcb1 (with CO_2) than using IMVcb2 (without CO_2) in real data tests, indicating feature importance analysis based on synthetic data can be a reasonable substitute for analysis with the often limited real-world data.

Commented [12]: model stability?

Commented [13R12]: Yes the model stability will be reduced due to more parameters in ML models to be determined in KGML-ag2.

3.3 KGML-ag comparing with other pure ML models

The results from eight~~seven~~ different models showed that KGML-ag1 comparing with other pure ML models consistently provided the lowest RMSE (3.59-3.94~~60~~ mg N m^{-2} day^{-1}, 1.14-1.23~~20~~ mg N m^{-2} day^{-2}, and 0.84-0.89~~7~~ mg N m^{-2} day^{-3}) and highest r^2 (0.78-0.81, 0.48-0.56~~51~~, and 0.23-0.31~~8~~) for N_2O fluxes, slope and curvature, respectively (Fig. 4). This indicated that KGML-ag1 outperformed other pure ML models in ~~both~~ capturing both the magnitude and dynamics of N_2O flux. KGML-ag2 presented slightly better mean scores for N_2O flux predictions than KGML-ag1, but worse scores for slope and curvature and larger uncertainties. This proved the hypothesis discussed in section 3.2 that KGML-ag2 didn't benefit the magnitude and dynamics predictions of N_2O flux with its more complex structure and less connections between IMVs.

Within the tree-based models (DT, RF, GB and XGB), the simplest model DT provided the worst predictions for N_2O flux, slope and curvature. The XGB model provided the highest N_2O flux accuracy with r^2 of 0.61-0.63~~2~~ and RMSE of 5.07-5.17~~11~~

mg N m^{-2} day^{-1}, while the GB model provided best slope and curvature predictions with r^2 of 0.38-0.40~~42~~ and 0.23-0.26~~8~~, and RMSE of 1.34-1.37~~1~~ mg N m^{-2} day^{-2} and 0.91-0.95~~88~~ mg N m^{-2} day^{-3}, respectively. The highest N_2O flux accuracy and relatively low slope and curvature accuracy of the XGB model implied that there is a trade-off between the abilities of capturing dynamics and magnitude.

In the group of deep learning models including ANN, GRU and KGML-ag1, ANN provided the worst predictions. Even with the better N_2O flux predictions than most tree-based models (except XGB), the slope and curvature predictions of ANN were the worst among all eight~~seven~~ models. This implied that the trade-off between accurately capturing N_2O dynamics to magnitude in ANN was significant. But when considering the temporal dependence, deep learning model GRU and KGML-ag1 outperformed all other models in flux, slope and curvature predictions. This indicated that without considering temporal dependence the improvement in N_2O flux prediction accuracy could be risky by causing the performance drop in capturing dynamics.

The detailed model comparisons in each chamber are shown in Fig. 5 (N_2O flux) and Fig. S6-7 (N_2O slope and curvature), where the results are found to follow the same pattern as described above. In addition, time series comparisons of chamber 3 and 4 in 2017 between different models are presented in Fig. S8 as two examples. From these comparisons, we infer that without considering temporal dependence and pretraining process, the tree-based model including DT, RF, GB and XGB and deep learning model ANN predicted erratic peaks in almost every missing data point, while GRU model was stable in small gaps and only presented poor performance in long missing period (before 25 day). This improvement by GRU model can be attributed to the structure of GRU that naturally keeps the historical information using hidden states, which enables GRU to consider the temporal dependence and make consistent predictions over time.

3.4 Influence of pretraining process, data augmentation and using IMV initial values as input feature

After we pretrained the GRU model with synthetic data, the overall r^2 of N_2O flux predictions in observed data increased by 0.02, 0.12 and 0.14, and RMSE decreased by 0.23 mg N m^{-2} day^{-1}, 0.15 mg N m^{-2} day^{-2} and 0.02 mg N m^{-2} day^{-3} for flux, slope and curvature predictions, respectively, compared to non-pretrained GRU (Table 3 gray region). The gap between the GRU model with pretrain and KGML-ag1 in N_2O value prediction shows the improvement resulting from architecture change (r^2 increases by 0.01 and RMSE decreases by 0.17 mg N m^{-2} day^{-1}). Although pretrained GRU had higher slope and curvature prediction accuracy than KGML-ag models, it still couldn't achieve the current N_2O value prediction accuracy of KGML-ag1. Besides, the KGML-ag models had relatively shallow N_2O prediction modules (2-layer GRU KGML-ag-N_2O module of KGML-ag models vs 4-layer GRU) but included modules for IMV predictions, which therefore increased the model interpretability.

It's worth noting that prediction accuracy of all KGML-ag models dropped without augmenting the training dataset in the fine-tuning process (Table 3 blue region). Moreover, the maximum training epochs increased from 800 to 20000, which resulted in overfitting on the small data set. This indicated that the data augmentation indeed helped the models become more generalizable and gain better accuracy.

Experiments using zero initial values presented a significant drop in every variable's prediction accuracy (Table 3 yellow region). This indicated that the IMV initial values input into the KGML-ag-IMV modules of KGML-ag models influenced not only the IMV prediction but also the N_2O prediction of the KGML-ag-N_2O module. This shows that there is useful information transferred from IMVs in the KGML-ag-IMV module to the KGML-ag-N_2O module.

4 Discussion

In the previous section, we showed that KGML-ag models can outperform ML models, by invoking architectural constraints and PB model synthetic data initialization. Compared to traditional PB models such as *ecosys*, KGML-ag models provide computationally more accurate and efficient predictions (KGML-ag few seconds vs *ecosys* half hour), which is similar to traditional ML surrogate models (Fig. S9). But KGML-ag goes beyond that by providing more interpretable predictions than pure ML models.

4.1 Interpretability of KGML-ag

The proposed KGML-ag models incorporate causal relations among N_2O related variables/processes as shown in Fig. S10. Managements, weather forcings and initial values of IMVs influence soil water, soil temperature and soil properties, which influence the availability of O_2 and N as well as the microbe populations in soil, and further influence the nitrification and denitrification rates. N_2O is produced during both nitrification and denitrification when soil O_2 concentration is limited. Our KGML-ag follows this hierarchical structure by designing KGML-ag-IMV modules representing the soil processes for IMVs predictions (Fig. 1c-d).

To better explain the time series predictions of N_2O flux (Fig. S1; Fig. 2-3), we separated the observations of each year into three periods: leading period (before N_2O increasing), increasing period (increasing to the peak) and decreasing period (peak decreasing to near zero). During the leading period, both NH_4^+ and CO_2 were increasing immediately in the following few days following urea N fertilizer application, indicating that urea was decomposing into NH_4^+ and CO_2 in soil water. With accumulating NH_4^+ in soil, nitrification started producing NO_3^- and consuming O_2. N_2O didn't respond to the fertilizer immediately due to enough O_2 in soil. Then when the soil became sufficiently hypoxic, N_2O fluxes entered an increasing period with N_2O being produced by nitrification and denitrification processes. CO_2 fluxes were relatively low and NH_4^+ kept decreasing during this period. Finally, when soil NH_4^+ was exhausted and NO_3^- started decreasing due to denitrification, N_2O

fluxes then entered the decreasing period. CO_2 flux was related to urea decomposition during the leading period, and was more closely related to O_2 demand in other periods. The KGML-ag predictions of N_2O and IMV captured the three periods and transition points, demonstrating the connections between those variables following the description as above (Fig. 3; Fig. S5). Although KGML-ag1 obtained lower IMVs prediction accuracy compared to KGML-ag2, it captured the general trends and was doing better for transitions, especially in NH_4^+ predictions. KGML-ag2 overfitted on the observations and ignored the correlations between IMVs, which resulted in loss in pretrain knowledge, poorer performance in the leading period, and erratic predictions in the period with missing observations (before day 25).

4.2 Lessons for KGML-ag development~~Interpretability of KGML-ag~~

The development of KGML-ag in our study is suitable to predict not only N_2O but also other variables, such as CO_2, CH_4 and ET, with complicated generation processes relying on the historical states. To develop a capable KGML model, we need to carefully address three questions:

What kind of ML model is suitable for developing KGML? The answer could be determined by the dominant variation type of the target variable in the data. If the dominant type is temporal variance, like flux variables in high temporal resolution (e.g., daily, or hourly), we should consider ML models with temporal dependency. RNN models such as GRU used in this study, and CNN models such as casual CNN (Oord et al., 2016) can be good starting ML models. If the dominant type is spatial variation, like variables in coarse temporal resolution (e.g., monthly or annually) but with high diversity due to soil property, land cover and climate, we should consider ML models with the ability to deal with edges, hotpoints and categories, such as CNN;

What physical/chemical constraints can be used to build KGML models? Although physical rules such as mass balance or energy balance are conceptually straightforward and were proved capable of constraining KGML in predicting lake phosphorus and temperature dynamics (Hanson et al., 2020; Read et al., 2019), they were excluded in this study according to our preliminary analysis. The reason is that the mass balance equation of N in the agriculture ecosystem includes too many unknown and unobservable components such as N_2 flux, NH_3 flux, N leaching, microbial N, plant N and soil/plant exchange, which collectively introduce large uncertainties in balance equations and make them hard to be directly applied in the KGML-ag framework. Other related physical (e.g., diffusion, solution) or chemical (e.g., nitrification, denitrification) processes cannot be easily added into the KGML-ag structure as rules due to lack of understanding of the process. Instead, as mentioned in Sect. 2.2.4, we used hierarchical structure to enforce an architectural constraint and causal relations among variables, and pretraining processes to infuse knowledge from *ecosys* to KGML-ag models.

How to involve PB models in the KGML development? An advanced PB model like *ecosys* built upon biophysical and biochemical rules instead of empirical relations will be a good basis to learn the process, guide the structure and provide the

Commented [14]: as a personal side comment: I think part of the difficulty is the involved many processes operate on different time scales, making the conservation constraint muc harder to impose.

Commented [15R14]: This is a great point! We kind of discussed this a little bit before, on modules transition, static variables/slow change variables/fast change variables, and boundary conditions, which may be all related to your point. I will keep thinking and investigating from this point with experiments.

constraints for KGML. The generated synthetic data in this study helped us get some knowledge about variables such as their general trends, dynamics and correlations. Such knowledge can be transferred to KGML models from synthetic data in the pretraining process, which can reduce the efforts to collect large numbers of real-world observation data. Moreover, while KGML shows great potential beyond PB models, we reckon that equally important for improving N_2O modeling is to continue improving our understanding of the related processes and mechanisms. Novel data collection and incorporating new understanding into PB models (e.g., *ecosys*) could provide foundation to further empower KGML (see further discussion in Sect. 4.3).

4.3 Limitation and possible improvement

First, the KGML-ag models in this study are limited by the available observed data. Some IMVs with high feature importance scores (e.g., O_2 flux, N_2 flux) or at different depths (e.g., soil NO_3^- at 5 cm depth, VWC at 5 cm depth), and data out of growing seasons are not included. The direct consequences are that some important processes cannot be well represented by the current KGML-ag (e.g., O_2 demand and N availability for nitrification and denitrification). Further improvement of KGML should consider three categories of data: target variable N_2O flux, IMVs and basic inputs (Fig. 1a). For N_2O flux observation, we lack sub-hourly to sub-daily observations to capture the hot moment of emission during 0-30 days after N fertilizer applications. Besides, the non-growing season can provide 35-65% of the annual direct N_2O emissions from seasonally frozen croplands and lead to a 17–28 % underestimate of the global agricultural N_2O budget if ignoring its contribution (Wagner-Riddle et al., 2017), but we can barely find observations from non-growing seasons. For IMVs, we found oxygen demand indicator (e.g., O_2 concentration or flux, CO_2 flux, CH_4 flux), N mass balance related variables (e.g., N_2 flux, soil NO_3^-, soil NH_4^+, N leaching) and soil water and temperature, can be used to better constrain the processes and therefore improve the KGML performance. Rohe et al. (2021) also indicated the importance of O_2, CO_2 and N_2 soil fluxes for N_2O predictions. In addition, the layerwise soil observations (e.g., soil NO_3^-, soil VWC) at 0-30 cm depth can be used to significantly improve the KGML model quality, according to our feature importance analysis (Fig. S2a). Moreover, continuous monitoring on these variables during the whole year is preferred rather than only during the growing season, since N_2O flux is largely influenced by previous states. To apply the KGML-ag to large scale, other observational data including basic inputs of soil/crop properties (e.g., soil bulk density, pH, crop type), management information (e.g., fertilizer, irrigation, tillage) and weather forcings along with N_2O flux observations are critical for fine-tuning and validating the developed KGML-ag and therefore explicitly simulating the N_2O or IMVs dynamics under specific conditions. Recent advances in remote sensing and machine learning have enabled estimating these variables with high-resolution at a large scale (Peng et al., 2020)

Second, the physical/chemical constraints can be more comprehensive in KGML-ag models. Although current KGML-ag models are well-initialized with *ecosys* synthetic data and constrained by causal relations of processes with hierarchical structure, the predicted N_2O flux and IMVs can still violate some basic physical rules like mass balance. As we discussed in

Sec. 4.2, it will be challenging to add physical rules like mass balance equation for N in a complicated agriculture ecosystem due to data limitations such as missing observations on certain key variables. Using inequalities instead of equations for mass balance may be one alternative solution. For example, we could use ReLU to add in a limitation for N mass balance residues which are calculated from known terms not larger than an empirical static value. Besides, better understanding of processes in the N cycle from fieldworks and lab experiments can also help us design new constraints. This limitation is also partially related to the data limitation and can be overcomed by involving more complete N_2O data to introduce more powerful constraints to KGML-ag.

Third, the KGML-ag currently are suffering from dealing with physical/chemical~~chamical~~ boundary transitions. Boundary transitions are common in the real world, such as phase change, volume solubility, and soil porosity etc. A detailed PB model generally coded plenty of "if/else/switch" statements inside to deal with the boundaries. But KGML-ag models based on the GRU are better at capturing continuous changes, rather than discrete changes. One solution is to include data with boundary information. In this study, involving IMVs like O_2, CO_2 and N_2, which already have boundary information like water freezing point, N pool volumes and other complicated boundaries related to soil/crop properties, can significantly improve the model performance. The data with boundary information could be continuous observation or estimated value from existing data. By using initial values to predict IMVs, KGML-ag in this study can partially solve the boundary transition problem when observation data is limited. Another solution is designing new structures of KGML-ag, such as combining ReLU function or including CNN model which are robust for discrete situations to the RNN models, or designing new constraints to limit the model working within the thresholds.

Finally, at the current stage we can not claim to have completely opened the black box of KGML-ag, but this framework is a significant step towards this goal. For example, some ideas implemented in our study, such as using pretraining to transfer knowledge from PB model to ML model, incorporating causal relations by hierarchical structure, predicting IMVs for tracking middle changes and using initial values as input to reduce data demand, would shed light on the future KGML-ag framework improvement. Besides, we acknowledge the importance of further testing the KGML-ag over completely independent datasets, but results presented in this manuscript are sufficient to justify the power of KGML as a framework. The mesocosm experiment data we used in this study has provided a comprehensive set of inputs and intermediate variables in addition to the output of N_2O fluxes, thus serving as a unique testbed. We expect our validation results will be more solid once more gold standard data of N_2O fluxes along with other relevant inputs and intermediate variables become publicly available. Moreover, incorporating more and more domain knowledge into KGML-ag will be inevitable in further improvement, but we don't think KGML-ag will become inefficient as it becomes more like the PB model. In fact, to efficiently surrogate components of PB models has been proposed as a research frontier in hybrid modeling for earth system science (Reichstein et al., 2019; Irrgang et al., 2021), with latest advances occurring in weather forecasts (Bauer et al., 2021). By using a hybrid model, computationally inefficient components of PB can be identified one by one, and be replaced with more efficient ML-based surrogates to eventually obtain the most efficient model. Further KGML-ag model development will also need to balance efficiency, accuracy and interpretability.

5 Conclusions

In this study, two KGML-ag models have been developed, validated, and tested for agricultural soil N_2O flux prediction using synthetic data generated by the PB model *ecosys* and observational data from a mesocosm facility. The results show that KGML-ag models can outperform PB and pure ML models in N_2O prediction in not only magnitude (KGML-ag1 $r^2 = 0.81$ vs best ML model GRU $r^2 = 0.78$) but also dynamics (KGML-ag1 accuracy minus GRU accuracy, slope $\Delta r^2 = 0.06$ and curvature $\Delta r^2 = 0.08$). KGML-ag can also defeat the PB model *ecosys* in efficiency by completing *ecosys*'s half-hour job within a few seconds. Compared to ML models, KGML-ag models can better represent complex dynamics and high peaks of N_2O flux. Moreover, with IMV predictions and hierarchical structures, KGML-ag models can provide biogeophysical/chemical information about key processes controlling N_2O fluxes, which will be useful for interpretable forecasting and developing mitigation strategies. Data demand for the KGML-ag models is significantly reduced due to involving IMV initial values and pretrain processes with synthetic data. This study demonstrated that the potential of KGML-ag application in the complex agriculture ecosystem is high and illustrates possible pathways of KGML-ag development for similar tasks. Further improvement of our KGML-ag models can involve general principles to further constrain the predictions through loss functions or architectures, but call for more detailed, high temporal resolution N_2O observation data from field measurements.

Code and Data Availability

The code and data used in this study can be found at https://doi.org/10.5281/zenodo.5504533.

Author contributions

LL and~~,~~ ZJ~~, JT KG and VK~~ conceived the study. ~~TJG, MDE, ALF and LTM conducted mesocosm experiments and provided observed data. KG,~~ WZ and YY conducted *ecosys* simulations and provided synthetic data. LL and SX processed the data and ~~wrote~~ developed the KGML-ag model ~~code~~. LL, SX and SW carried the experiments out with supervisions from~~.~~ ZJ, JT, KG and VK. TJG, MDE, ALF and LTM shared mesocosm observations and interpreted the data.~~, BP and WZ supervised the experiments and advised on analysis from agricultural domain science perspective. VK, XJ and SX advised on the code and analysis from computer science perspective.~~ LL wrote the ~~original~~ first draft of manuscript with further editing from TK on figure and tables. ZJ, SX, JT, KG, XJ, BP, YY, ~~and~~ WZ and VK further edited the manuscript ~~and ZJ, KG and VK provided supervision~~.

Competing interests

The authors declare that they have no conflict of interest.

References

Barton, L., Wolf, B., Rowlings, D., Scheer, C., Kiese, R., Grace, P., ... & Butterbach-Bahl, K.: Sampling frequency affects estimates of annual nitrous oxide fluxes, Scientific reports, 5(1), 1-9, 2015.

Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C., & Wedi, N. P.: The digital revolution of Earth-system science. Nature Computational Science, 1(2), 104-113, 2021.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P.: Enforcing analytic constraints in neural networks emulating physical systems, Physical Review Letters, 126(9), 098302, 2021.

Beucler, T., Rasp, S., Pritchard, M., & Gentine, P.: Achieving conservation of energy in neural network emulators for climate modeling, arXiv preprint arXiv:1906.06622, 2019.

Butterbach-Bahl, K., Baggs, E. M., Dannenmann, M., Kiese, R., & Zechmeister-Boltenstern, S.: Nitrous oxide emissions from soils: how well do we understand the processes and their controls? Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1621), 20130122, 2013.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259, 2014.

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio.: Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555, 2014.

Daw, A., Thomas, R. Q., Carey, C. C., Read, J. S., Appling, A. P., & Karpatne, A.: Physics-guided architecture (pga) of neural networks for quantifying uncertainty in lake temperature modeling, In Proceedings of the 2020 siam international conference on data mining (pp. 532-540), Society for Industrial and Applied Mathematics, 2020.

Del Grosso, S. J., Parton, W. J., Mosier, A. R., Ojima, D. S., Kulmala, A. E., & Phongpan, S.: General model for N_2O and N2 gas emissions from soils due to dentrification, Global biogeochemical cycles, 14(4), 1045-1060, 2020.

Fassbinder, J. J, Schultz, N. M, Baker, J. M, & Griffis, T. J.: Automated, Low-Power Chamber System for Measuring Nitrous Oxide Emissions, Journal of environmental quality, 42, 606. doi: 10.2134/jeq2012.0283, 2013.

Fassbinder, J. J., Griffis, T. J., & Baker, J. M.: Evaluation of carbon isotope flux partitioning theory under simplified and controlled environmental conditions, Agricultural and forest meteorology, 153, 154-164, 2012.

Forster, P., Storelvmo, T., Armour, K. , Collins, W., … & Zhang, H.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press. In Press, 2021.

Gilhespy, S. L., Anthony, S., Cardenas, L., Chadwick, D., del Prado, A., Li, C., ... & Yeluripati, J. B.: First 20 years of DNDC (DeNitrification DeComposition): model evolution, Ecological modelling, 292, 51-62, 2014.

Grant, R. F.: Modeling Carbon and Nitrogen Dynamics for Soil Management, (Boca Raton, FL: CRC Press) A review of the Canadian ecosystem model ecosys 173–264, 2021.

Grant, R. F., & Pattey, E.: Modelling variability in N_2O emissions from fertilized agricultural fields, Soil Biology and Biochemistry, 35(2), 225-243, 2003.

Grant, R. F., & Pattey, E.: Temperature sensitivity of N_2O emissions from fertilized agricultural soils: Mathematical modeling in ecosys. Global biogeochemical cycles, 22(4), 2008.

~~Grant, R. F., Neftel, A., & Calanca, P.: Ecological controls on N_2O emission in surface litter and near-surface soil of a managed grassland: modelling and measurements, Biogeosciences, 13(12), 3549-3571, 2016.~~

Grant, R. F., Neftel, A., & Calanca, P.: Ecological controls on N_2O emission in surface litter and near-surface soil of a managed grassland: modelling and measurements, Biogeosciences, 13(12), 3549-3571, 2016.

Grant, R. F., Pattey, E., Goddard, T. W., Kryzanowski, L. M., & Puurveen, H.: Modeling the effects of fertilizer application rate on nitrous oxide emissions, Soil Science Society of America Journal, 70(1), 235-248, 2006.

Hamrani, A., Akbarzadeh, A., & Madramootoo, C. A.: Machine learning for predicting greenhouse gas emissions from agricultural soils, Science of The Total Environment, 741, 140338, 2020.

Hanson, P. C., Stillman, A. B., Jia, X., Karpatne, A., Dugan, H. A., Carey, C. C., ... & Kumar, V.: Predicting lake surface water phosphorus dynamics using process-guided machine learning, Ecological Modelling, 430, 109136, 2020.

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ... & Keating, B. A.: APSIM–evolution towards a new generation of agricultural systems simulation, Environmental Modelling & Software, 62, 327-350, 2014.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. Nature Machine Intelligence, 3(8), 667-674, 2021.

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., & Kumar, V.: Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles, ACM/IMS Transactions on Data Science, 2(3), 1-26, 2021.

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V.: Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles, In Proceedings of the 2019 SIAM International Conference on Data Mining (pp. 558-566), Society for Industrial and Applied Mathematics, 2019.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... & Kumar, V.: Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Transactions on knowledge and data engineering, 29(10), 2318-2331, 2017.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., ... & Smith, C. J.: An overview of APSIM, a model designed for farming systems simulation, European journal of agronomy, 18(3-4), 267-288, 2003.

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., ... & Kumar, V., Physics guided machine learning methods for hydrology, arXiv preprint arXiv:2012.02854, 2020.

Kim, T., Jin, Z., Smith, T., Liu, L., Yang, Y., Yang, Y., ... & Zhou, W.: Quantifying nitrogen loss hotspots and mitigation potential for individual fields in the US Corn Belt with a metamodeling approach, Environmental Research Letters, 2021.

Kraft, B., Jung, M., Körner, M., Koirala, S., & Reichstein, M.: Towards hybrid modeling of the global hydrological cycle, Hydrology and Earth System Sciences Discussions, 1-40, 2021.

Meyer, D., Nagler, T., & Hogan, R. J.: Copula-based synthetic data augmentation for machine-learning emulators. Geoscientific Model Development, 14(8), 5205-5215, 2021.

Miller, L. T. , Griffis, T. J., Erickson, M. D., Turner, P. A., Deventer, M. J., Chen, Z., Yu, Z., Venterea, R.T., Baker, J. M., and Frie, A. L.: Response of nitrous oxide emissions to future changes in precipitation and individual rain events, Journal of Environmental Quality, In review, 2021

Miller, L. T., Assessing Agricultural Nitrous Oxide Emissions and Hot Moments Using Mesocosm Simulations, (Master Thesis, University of Minnesota) Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/219276, 2021

Necpálová, M., Anex, R. P., Fienen, M. N., Del Grosso, S. J., Castellano, M. J., Sawyer, J. E., ... & Barker, D. W.: Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling, Environmental Modelling & Software, 66, 110-130, 2015.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K.: Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499, 2016.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., ... & van Ypserle, J. P.: Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change (p. 151). Ipcc, 2014.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., ... & Kumar, V.: Process-guided deep learning predictions of lake water temperature, Water Resources Research, 55(11), 9173-9190, 2019.

Peng, B., Guan, K., Tang, J., Ainsworth, E. A., Asseng, S., Bernacchi, C. J., ... & Zhou, W.: Towards a multiscale crop modelling framework for climate change adaptation assessment, Nature plants, 6(4), 338-348, 2020.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204, 2019.

Robertson, M., BenDor, T. K., Lave, R., Riggsbee, A., Ruhl, J. B., & Doyle, M.: Stacking ecosystem services, Frontiers in Ecology and the Environment, 12(3), 186-193, 2014.

Rohe, L., Apelt, B., Vogel, H. J., Well, R., Wu, G. M., & Schlüter, S.: Denitrification in soil as a function of oxygen availability at the microscale, Biogeosciences, 18(3), 1185-1201, 2021.

Saha, D., Basso, B., & Robertson, G. P.: Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems, Environmental Research Letters, 16(2), 024004, 2021.

Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., & Janssens-Maenhout, G.: Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases, Atmospheric Chemistry and Physics, 21(7), 5655-5683, 2021.

Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., & Janssens-Maenhout, G.: Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases, Atmospheric Chemistry and Physics, 21(7), 5655-5683, 2021.

Syakila, A., & Kroeze, C.: The global nitrous oxide budget revisited, Greenhouse gas measurement and management, 1(1), 17-26, 2011.

Thompson, R. L., Lassaletta, L., Patra, P. K., Wilson, C., Wells, K. C., Gressent, A., ... & Canadell, J. G.: Acceleration of global N_2O emissions seen from two decades of atmospheric inversion, Nature Climate Change, 9(12), 993-998, 2019.

Thornley, J. H., & France, J.: Mathematical models in agriculture: quantitative methods for the plant, animal and ecological sciences, Cabi, 2007.

Tian, H., Xu, R., Canadell, J. G., Thompson, R. L., Winiwarter, W., Suntharalingam, P., ... & Yao, Y.: A comprehensive quantification of global nitrous oxide sources and sinks, Nature, 586(7828), 248-256, 2020.

Venterea, R. T., Maharjan, B., & Dolan, M. S.: Fertilizer source and tillage effects on yield-scaled nitrous oxide emissions in a corn cropping system. Journal of Environmental Quality, 40(5), 1521-1531, 2011.

Wagner-Riddle, C., Congreves, K. A., Abalos, D., Berg, A. A., Brown, S. E., Ambadan, J. T., ... & Tenuta, M.: Globally important nitrous oxide emissions from croplands induced by freeze–thaw cycles, Nature Geoscience, 10(4), 279-283, 2017.

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, arXiv preprint arXiv:2003.04919, 2020.

Yang, Y., Liu, L., Zhou, W., Guan, K., Kim, T., Tang, J., Peng, B., Zhu, P., Grant, R. F., Griffis, T. J., Jin, Z.: Distinct driving mechanisms of non-growing season N_2O emissions call for spatial-specific mitigation strategies in the US Midwest. Agriculture and Forest Meteorology~~One Earth~~. Submitted, 2022.

Zhang, Y., & Niu, H.: The development of the DNDC plant growth sub-model and the application of DNDC in agriculture: a review, Agriculture, Ecosystems & Environment, 230, 271-282, 2016.

Zhang, Y., Li, C., Zhou, X., & Moore III, B.: A simulation model linking crop growth and soil biogeochemistry for sustainable agriculture, Ecological modelling, 151(1), 75-108, 2002.

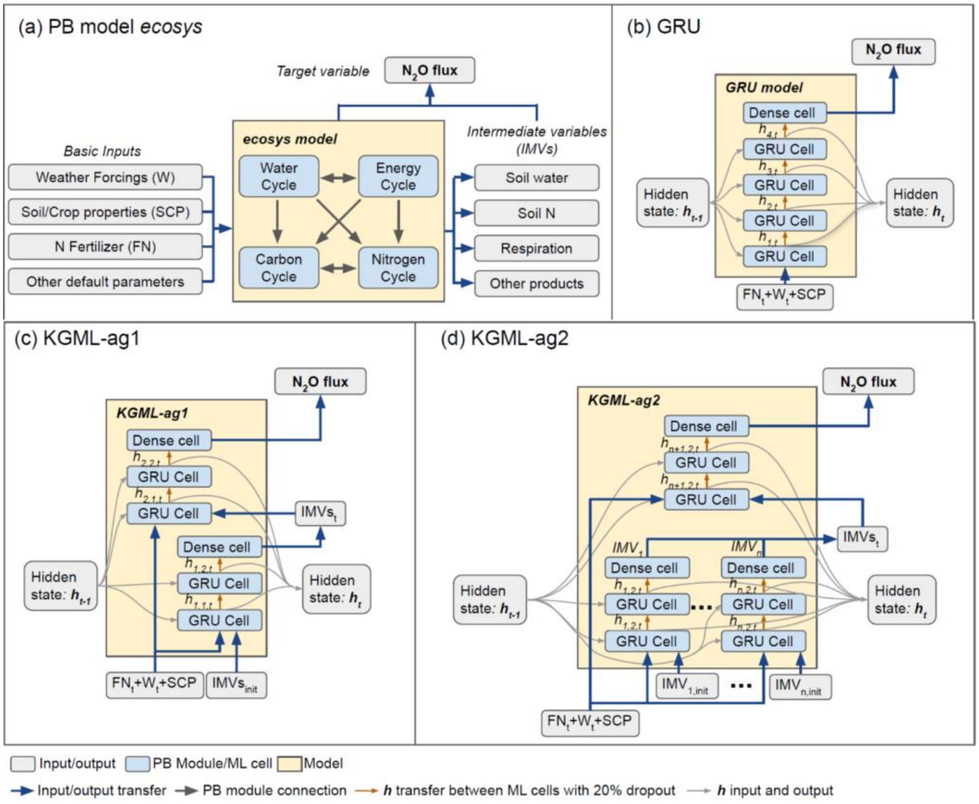


Figure 1: The model structures~~frame~~s. a) The *ecosys* model~~ frame~~; b) Gated recurrent unit (GRU) model~~ frame~~; c) KGML-ag1 model with a ~~frame of~~ hierarchical structure; d) KGML-ag2 model with a ~~frame of~~ hierarchical structure with separated GRU modules for IMV predictions. Specifically, in our KGML model design, weather forcings (W) include temperature (TMAX, TDIF), precipitation (PRECN), radiation (RADN), humidity (HMAX and HDIF) and wind speed (WIND); soil/crop properties (SCP) include bulk density (TBKDS), sand content (TCSAND), silt content (TCSILT), pH (TPH), cation exchange capacity (TCEC), soil organic carbon (TSOC), planting day of the year (PDOY) and crop type (CROPT); IMVs include CO_2 flux, soil NO_3^- concentration, soil NH4+concentration, and soil volumetric water content (VWC).

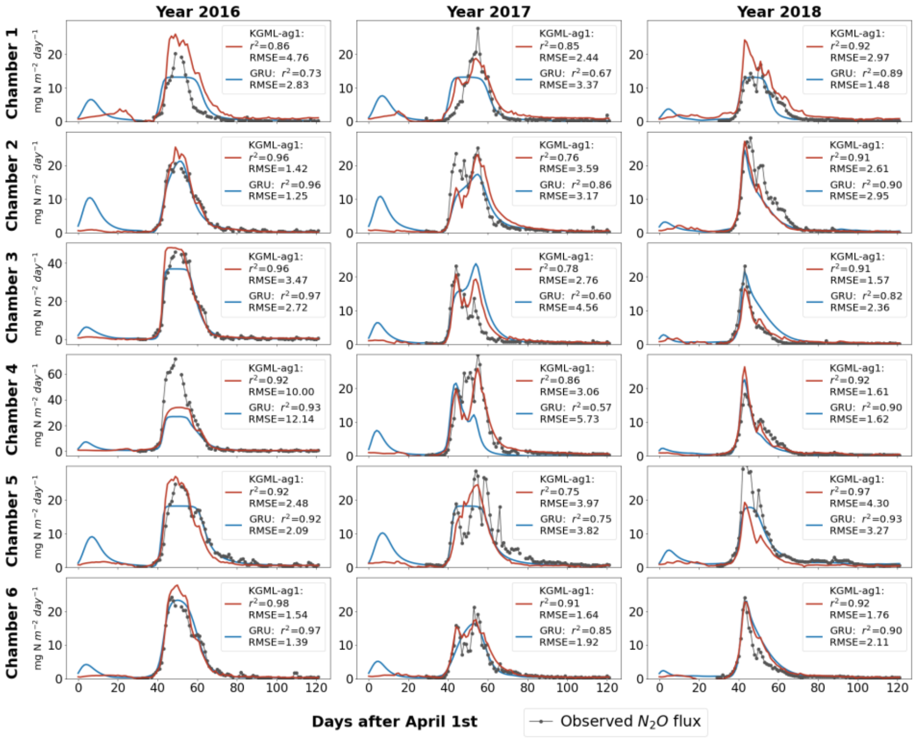


Figure 2: N_2O flux time series comparisons among pure non-pretrained GRU predictions (blue line), KGML-ag1 predictions (red line) and observations (black line-dot) from cross-validation. The N_2O flux unit is mg N m^{-2} day^{-1}.

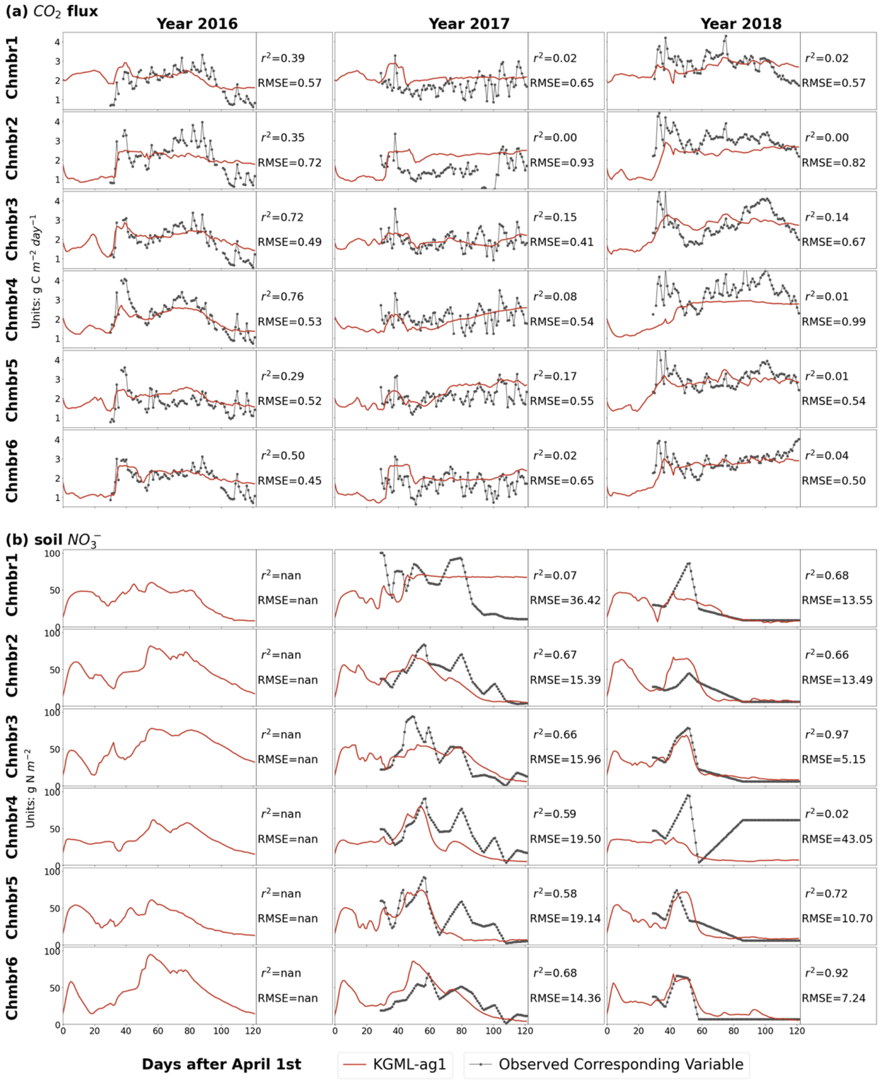
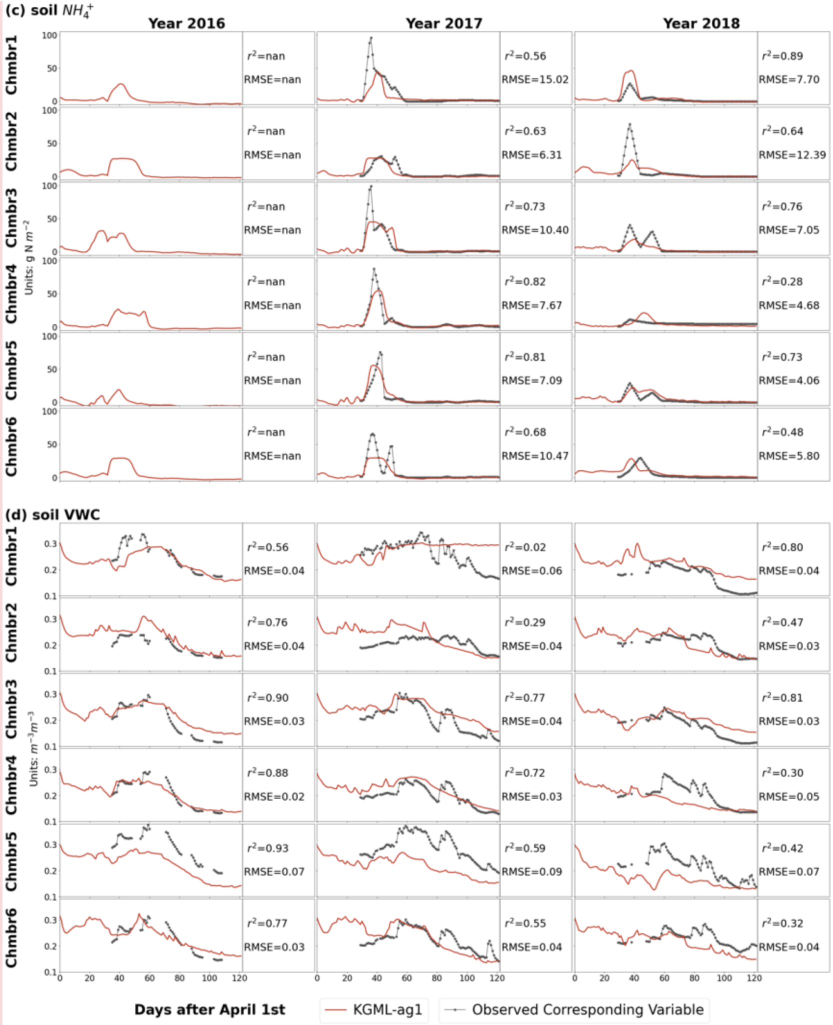

(a) CO_2 flux
Year 2016
Year 2017
Year 2018
Units: g C m^{-2} day^{-1}
Chmbr1
r^2=0.39 RMSE=0.57
r^2=0.02 RMSE=0.65
r^2=0.02 RMSE=0.57
Chmbr2
r^2=0.35 RMSE=0.72
r^2=0.00 RMSE=0.93
r^2=0.00 RMSE=0.82
Chmbr3
r^2=0.72 RMSE=0.49
r^2=0.15 RMSE=0.41
r^2=0.14 RMSE=0.67
Chmbr4
r^2=0.76 RMSE=0.53
r^2=0.08 RMSE=0.54
r^2=0.01 RMSE=0.99
Chmbr5
r^2=0.29 RMSE=0.52
r^2=0.17 RMSE=0.55
r^2=0.01 RMSE=0.54
Chmbr6
r^2=0.50 RMSE=0.45
r^2=0.02 RMSE=0.65
r^2=0.04 RMSE=0.50
(b) soil NO_3^-
Units: g N m^{-2}
Chmbr1
r^2=nan RMSE=nan
r^2=0.07 RMSE=36.42
r^2=0.68 RMSE=13.55
Chmbr2
r^2=nan RMSE=nan
r^2=0.67 RMSE=15.39
r^2=0.66 RMSE=13.49
Chmbr3
r^2=nan RMSE=nan
r^2=0.66 RMSE=15.96
r^2=0.97 RMSE=5.15
Chmbr4
r^2=nan RMSE=nan
r^2=0.59 RMSE=19.50
r^2=0.02 RMSE=43.05
Chmbr5
r^2=nan RMSE=nan
r^2=0.58 RMSE=19.14
r^2=0.72 RMSE=10.70
Chmbr6
r^2=nan RMSE=nan
r^2=0.68 RMSE=14.36
r^2=0.92 RMSE=7.24
0 20 40 60 80 100 120
Days after April 1st
KGML-ag1
Observed Corresponding Variable

Figure 3: IMVs prediction from KGML-ag1. The black-dot line represents observations and the red line represents the results from KGML-ag1. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber. The CO_2 flux and soil NO_3^- concentration units are g C m^{-2} day^{-1} and g N m^{-2}, respectively.


(c) soil NH_4^+
Year 2016
Year 2017
Year 2018
Chmbr1
Chmbr2
Chmbr3
Chmbr4
Chmbr5
Chmbr6
Units: g N m^{-2}
0
50
100
r^2=nan
RMSE=nan
r^2=0.56
RMSE=15.02
r^2=0.89
RMSE=7.70
r^2=nan
RMSE=nan
r^2=0.63
RMSE=6.31
r^2=0.64
RMSE=12.39
r^2=nan
RMSE=nan
r^2=0.73
RMSE=10.40
r^2=0.76
RMSE=7.05
r^2=nan
RMSE=nan
r^2=0.82
RMSE=7.67
r^2=0.28
RMSE=4.68
r^2=nan
RMSE=nan
r^2=0.81
RMSE=7.09
r^2=0.73
RMSE=4.06
r^2=nan
RMSE=nan
r^2=0.68
RMSE=10.47
r^2=0.48
RMSE=5.80
(d) soil VWC
Units: $m^{-3}m^{-3}$
0.1
0.2
0.3
r^2=0.56
RMSE=0.04
r^2=0.02
RMSE=0.06
r^2=0.80
RMSE=0.04
r^2=0.76
RMSE=0.04
r^2=0.29
RMSE=0.04
r^2=0.47
RMSE=0.03
r^2=0.90
RMSE=0.03
r^2=0.77
RMSE=0.04
r^2=0.81
RMSE=0.03
r^2=0.88
RMSE=0.02
r^2=0.72
RMSE=0.03
r^2=0.30
RMSE=0.05
r^2=0.93
RMSE=0.07
r^2=0.59
RMSE=0.09
r^2=0.42
RMSE=0.07
r^2=0.77
RMSE=0.03
r^2=0.55
RMSE=0.04
r^2=0.32
RMSE=0.04
0
20
40
60
80
100
120
Days after April 1st
KGML-ag1
Observed Corresponding Variable

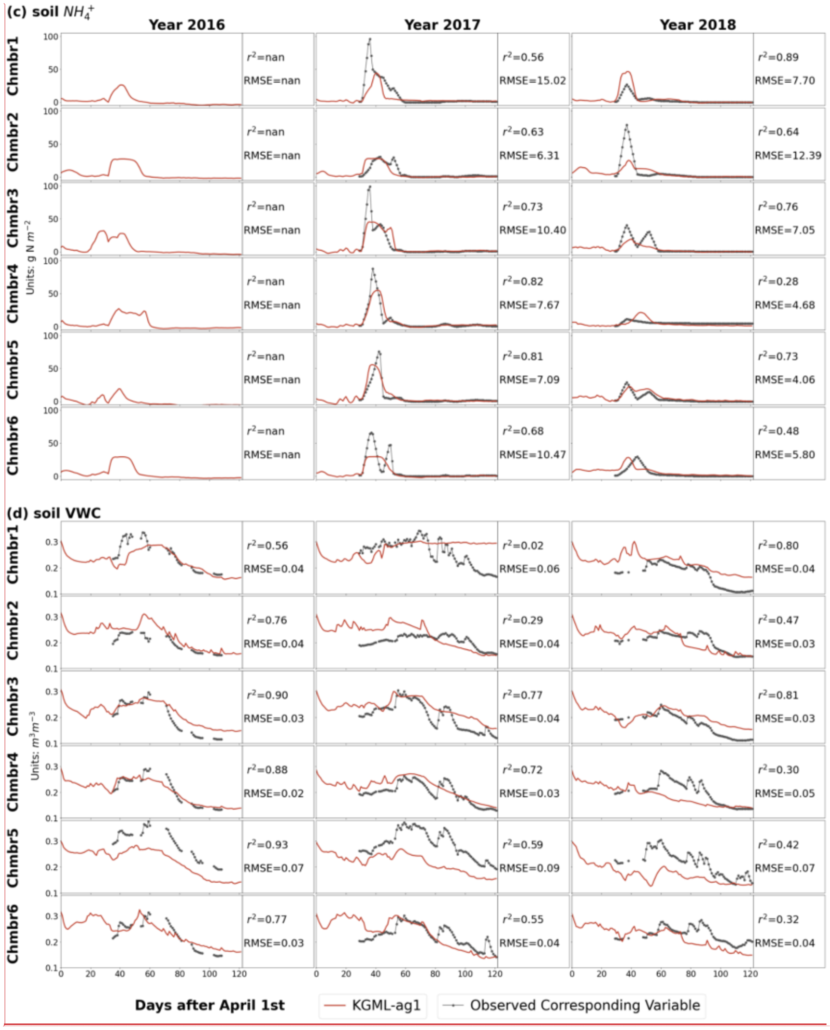
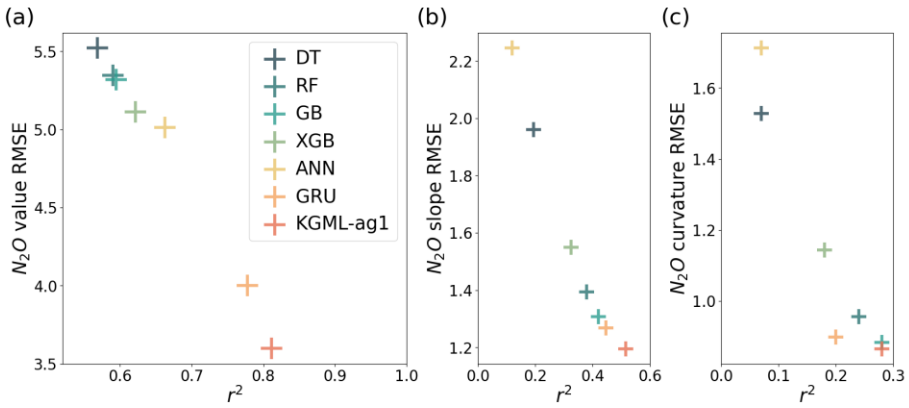
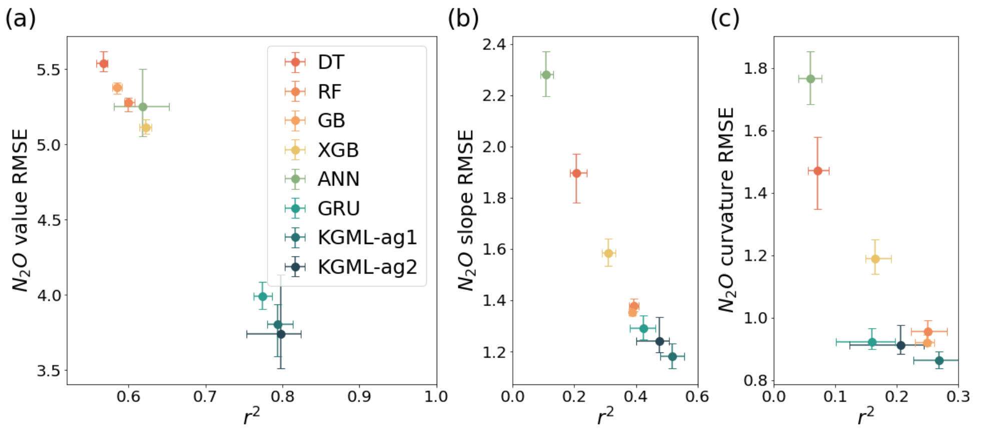


Figure 3 Contd.: IMVs prediction from KGML-ag1. The black-dot line represents observations and the red line represents the results from KGML-ag1. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber. The soil NH_4^+ concentration and soil VWC units are g N m^{-2} and m^3 m^{-3}, respectively.

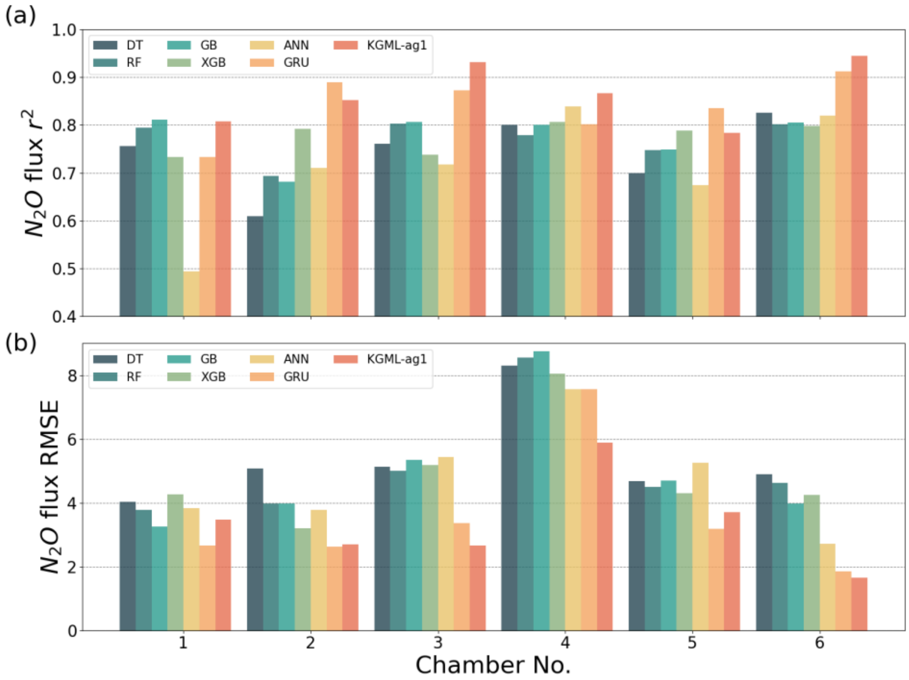
Commented [16]: A typo of VWC units has been fixed.





Formatted: Font: (Default) Arial, (Asian) Arial, 11 pt

Figure 4: The comparisons of overall prediction accuracy for N_2O value (a), 1st order gradient (slope, b) and 2nd order gradient (curvature, c) between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag~~1~~ models. Different color symbols represent the different models. The x- and y-error bars are coming from the maximum and minimum scores of ensemble experiments. The dot represents the mean score of the ensemble experiments.


(a)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
N_2O flux r^2
DT
RF
GB
XGB
ANN
GRU
KGML-ag1
(b)
8
6
4
2
0
N_2O flux RMSE
1
2
3
4
5
6
Chamber No.

771

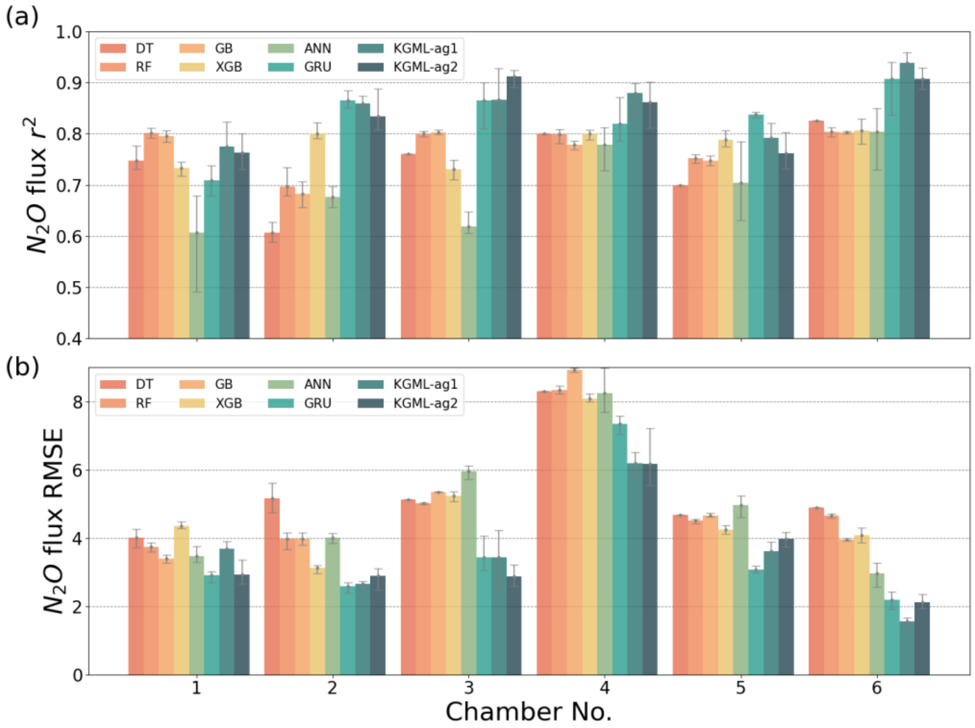


Figure 5: The comparisons of N_2O flux prediction accuracy r^2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag~~1~~ models in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

Formatted: Font: (Default) Arial, (Asian) Arial, 11 pt

Table 1: Pretrain results for different model and IMV combinations using *ecosys* synthetic data.

			N_2O		CO_2		NO_3^-		NH_4^+		VWC	
No.	Pretrain Model	Input Feature N	r^2	RMSE	r^2	NRMSE	r^2	NRMSE	r^2	NRMSE	r^2	NRMSE
1	GRU+76IMVs	76 IMVs+FN+7Ws+8SCP	0.98	0.54	--[a]	--	--	--	--	--	--	--
2	GRU+IMVcb1	4 IMVs+FN+7Ws+8SCP	0.92	1.15	--	--	--	--	--	--	--	--
3	GRU+IMVcb2	3 IMVs+FN+7Ws+8SCP	0.90	1.26	--	--	--	--	--	--	--	--
4	GRU	FN+7Ws+8SCP	0.89	1.37	--	--	--	--	--	--	--	--
5	KGML-ag1+IMVcb1_ini	FN+7Ws+8SCP+4IMV_ini	0.90	1.24	0.91	0.06	0.95	0.03	0.98	0.03	0.95	0.04
6	KGML-ag1+IMVcb2_ini	FN+7Ws+8SCP+3IMV_ini	0.90	1.26	--	--	0.94	0.03	0.97	0.03	0.95	0.04
7	KGML-ag2+IMVcb1_ini	FN+7Ws+8SCP+4IMV_ini	0.90	1.27	0.92	0.05	0.95	0.02	0.98	0.03	0.96	0.04
8	KGML-ag2+IMVcb2_ini	FN+7Ws+8SCP+3IMV_ini	0.91	1.19	--	--	0.95	0.00	0.99	0.02	0.95	0.04

[a]The empty slot indicates that the model does not predict that variable.

Table 2: Prediction accuracy comparisons between non-pretrained GRU model and KGML-ag1.~~Pretrain results for different model and IMV combinations using *ecosys* synthetic data.~~

		N_2O, KGML-ag1 minus GRU				N_2O 1st order gradient, KGML-ag1 minus GRU				N_2O 2nd order gradient, KGML-ag1 minus GRU			
	No.	All time[b]	Day 30-80	Day 40-65	Day 45-60	All time	Day 30-80	Day 40-65	Day 45-60	All time	Day 30-80	Day 40-65	Day 45-60
Δr^2 [a]	All data	0.03[c]	0.04	0.07	0.10	0.07	0.07	0.07	0.15	0.08	0.08	0.09	0.11
	Chamber1	0.07	0.10	0.20	0.13	0.18	0.18	0.19	0.14	0.08	0.09	0.09	0.02
	Chamber2	-0.04	-0.05	-0.07	-0.05	0.08	0.09	0.09	0.16	0.20	0.20	0.20	0.23
	Chamber3	0.06	0.06	0.08	0.06	0.04	0.04	0.04	0.13	-0.01	-0.01	-0.01	0.07
	Chamber4	0.06	0.08	0.12	0.07	0.05	0.05	0.05	0.14	0.07	0.07	0.08	0.12
	Chamber5	-0.05	-0.06	-0.07	-0.03	0.09	0.09	0.10	0.16	0.13	0.13	0.15	0.11
	Chamber6	0.03	0.04	0.08	0.17	0.14	0.14	0.15	0.22	0.12	0.13	0.14	0.23
ΔRMSE[a]	All data	-0.41	-0.56	-0.84	-1.19	-0.07	-0.10	-0.14	-0.20	-0.03	-0.05	-0.07	-0.08
	Chamber1	0.80	1.06	1.21	1.70	0.00	0.00	-0.02	0.00	0.05	0.07	0.10	0.18
	Chamber2	0.08	0.11	0.07	-0.04	-0.10	-0.13	-0.18	-0.14	-0.10	-0.14	-0.19	-0.22
	Chamber3	-0.71	-0.96	-1.30	-2.09	0.03	0.04	0.07	-0.25	0.09	0.13	0.17	0.08
	Chamber4	-1.68	-2.27	-3.09	-3.81	-0.11	-0.15	-0.21	-0.26	-0.05	-0.07	-0.09	-0.16
	Chamber5	0.53	0.69	0.86	0.99	-0.10	-0.14	-0.20	-0.23	-0.09	-0.12	-0.18	-0.14
	Chamber6	-0.20	-0.27	-0.37	-0.61	-0.14	-0.20	-0.29	-0.33	-0.07	-0.10	-0.15	-0.19

[a]The difference of r^2 (Δr^2), and difference of RMSE (ΔRMSE, units are mg N m^{-2} day^{-1}, mg N m^{-2} day^{-2}, mg N m^{-2} day^{-3} for N_2O value, 1st order gradient and 2nd order gradient, respectively) were calculated by values from KGML-ag1 minus values from GRU.

[b]Results from different time windows of different chambers during the period of April 1st-July31st (Days1-122) were detected.

[c]Blue cells mean KGML-ag1 outperforms GRU, while yellow cells mean the opposite.

Table 3: Experiments for measuring GRU and KGML-ag models performance, and influence of pretraining process, training data augmentation and IMV initial values.

			N_2O		N_2O 1st order gradient		N_2O 2nd order gradient		CO_2		NO_3^-		NH_4^+		VWC	
No.	**Retrain Model**	**Experiment**	**r^2**	**RMSE**	**r^2**	**RMSE**	**r^2**	**RMSE**	**r^2**	**NRMSE**	**r^2**	**NRMSE**	**r^2**	**NRMSE**	**r^2**	**NRMSE**
1	GRU, baseline[a]	No Pretrain	0.78	4.00	0.45	1.27	0.20	0.90	--[b]	--	--	--	--	--	--	--
2	GRU	Pretrain	0.80	3.77	0.57	1.12	0.34	0.82	--	--	--	--	--	--	--	--
3	KGML-ag1+ IMVcb1_ini	Original setting	0.81	3.60	0.51	1.20	0.28	0.87	0.37	0.14	0.39	0.21	0.60	0.09	0.33	0.18
4	KGML-ag1+ IMVcb2_ini	Original setting	0.80	3.71	0.49	1.22	0.21	0.91	--	--	0.37	0.22	0.53	0.10	0.33	0.19
5	KGML-ag2+ IMVcb1_ini	Original setting	0.79	3.77	0.48	1.23	0.22	0.90	0.74	0.09	0.46	0.18	0.66	0.08	0.84	0.08
6	KGML-ag2+ IMVcb2_ini	Original setting	0.78	3.91	0.47	1.24	0.20	0.91	--	--	0.49	0.18	0.69	0.08	0.84	0.08
7	KGML-ag1+ IMVcb1_ini	No augmentation	0.80	3.73	0.49	1.22	0.22	0.90	0.38	0.14	0.38	0.21	0.61	0.09	0.37	0.17
8	KGML-ag1+ IMVcb2_ini	No augmentation	0.77	4.04	0.41	1.31	0.13	0.95	--	--	0.38	0.21	0.53	0.10	0.35	0.18
9	KGML-ag2+ IMVcb1_ini	No augmentation	0.76	4.06	0.45	1.27	0.16	0.95	0.69	0.10	0.21	0.25	0.60	0.09	0.80	0.09
10	KGML-ag2+ IMVcb2_ini	No augmentation	0.74	4.27	0.48	1.23	0.21	0.90	--	--	0.40	0.21	0.60	0.09	0.81	0.09
11	KGML-ag1+ IMVcb1_ini	Zero initial values	0.48	6.27	0.26	1.49	0.08	1.00	0.19	0.16	0.25	0.25	0.47	0.12	0.14	0.25
12	KGML-ag1+ IMVcb2_ini	Zero initial values	0.49	5.94	0.31	1.41	0.13	0.95	--	--	0.31	0.25	0.38	0.13	0.24	0.25
13	KGML-ag2+ IMVcb1_ini	Zero initial values	0.48	6.05	0.12	1.66	0.01	1.09	0.58	0.12	0.34	0.25	0.21	0.13	0.56	0.31
14	KGML-ag2+ IMVcb2_ini	Zero initial values	0.39	6.60	0.15	1.59	0.04	1.01	--	--	0.16	0.27	0.27	0.12	0.53	0.31

[a] Gray region includes the experiments with original simulation settings as described in Sec. 2 and dark gray refers to the baseline GRU simulation; Blue region includes the experiments without data augmentation during the finetuning process; And yellow region includes the experiments of replacing original IMV initial values with zeros.

[b] The empty slot indicates that the model does not predict that variable.