Response Letter to Editor

We really appreciate all comments and suggestions from the editor and have carefully addressed all concerns and comments point by point. Main changes include:

1) Added more descriptions on *ecosys* model evaluation on N_2O flux;
2) Added discussion on comparisons between mesocosm data uncertainty with model simulation uncertainty;
3) Added more descriptions of leave-one-out cross validation repeatedly to remind reader we only presented validation results;
4) Added discussion on mesocosm data slightly higher than field studies;
5) Removed coloured cell/values as recommended from file validation and corrected other minor mistakes.

We believe the quality of this manuscript has improved after the revision. 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

Detailed responses begin:

Post-review decision: gmd-2021-317

The authors have done a good job responding to reviewer comments and concerns. There are some remaining improvements to be made, I think, but these are relatively minor. Thus, I recommend publication with minor revisions (with a final review by me).

Response:

We are grateful for the editor's recommendation and recognition of our efforts. We have carefully revised the manuscript to address all his comments and suggestions.

Primary comments:

Reviewer 1 comment 2

Reviewer 1 asked specifically about ecosys performance with regard to high-frequency fluctuations. It's unclear whether that's included in the R^2 value for NEE listed on line 151— what frequency was that analysis at?

Response:

Thanks so much for pointing these out. To address this concern, we have revised the related test and added in the frequency information for flux tower NEE and Reco data used in *ecosys* validation (line 148 to 152).

"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 have been extensively benchmarked, including CO_2 exchange (daily Reco R^2 = 0.80-0.86; daily NEE, R^2 = 0.75-0.89) 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) in the US Midwest (Zhou et al., 2021)."

There's nothing in the new text about N_2O performance; please add some text pointing to Grant et al. papers as you did for in the response to the reviewer (although that level of detail isn't necessary): "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."

Response: We really appreciate the comments and suggestions from the editor. A new sentence has been added to describe the detailed validation scores for *ecosys* N_2O flux simulation in hourly frequency and to include those Grant et al. papers listed for validation in various ecosystems (line 152 to 155).

"In addition, *ecosys* model can capture the dynamics and magnitude of N_2O flux in hourly frequency (R^2 = 0.2-0.4 and RMSE = 0.1-0.2 mg N m^{-2} h^{-1} in Grant et al., 2008; R^2 = 0.28-0.37 and RMSE = 0.2-0.28 mg N m^{-2} h^{-1} in Grant et al., 2003), and in various ecosystems (e.g. agriculture soil in Grant et al., 2006, 2008; forest in Grant et al., 2010; and grassland in Grant et al., 2016)."

The Wang et al. (2021) reference is missing from the References, and I can't check Yang et al. (2022) because it's only been submitted.

Response:

Thanks so much for pointing out these. The Yang et al. (2022) paper and related accumulated N_2O flux validation results have been removed from text since the formal citation is not available at this stage. Besides, the actual first author for "Wang et al. (2021)" is Zhou, Wang. Thus we have corrected the citation to "Zhou et al. (2021)" and added the reference to the list (line 732 to 734).

"Zhou, W., Guan, K., Peng, B., Tang, J., Jin, Z., Jiang, C., ... & Mezbahuddin, S.: Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for US Midwestern agroecosystems. Agricultural and Forest Meteorology, 307, 108521, 2021."

Reviewer 1 comment 3

Yes, you cited the Miller (2021) thesis, which has chamber measurement uncertainty. But I think the reviewer was saying it would be good to explicitly compare the uncertainty in the simulations to the uncertainty in the chamber observations. Please add some discussion of This.

Response:

Thanks much for your suggestions. Beside previous added ensemble experiments (presented in section 3.3 and figure 4 and 5), we have added the comparisons of uncertainty distribution ranges and standard deviations between mesocosm observation and KGML-ag1 model simulations. The details can be found in results section 3.3 first paragraph (line 405 to 409):

"Meanwhile, we have calculated the uncertainty of mesocosm measurement due to converting hourly data to daily data during 30-80 days by using augmented value minus mean of the augmented values (-10.2 to 10.4 mg N m^{-2} day^{-1}, and standard deviation =1.4 mg N m^{-2} day^{-1}). KGML-ag1 during the same period has comparable uncertainties based on ensemble simulations (calculated by ensemble value minus mean of ensemble values; -14.4 to 15.2 mg N m^{-2} day^{-1}, with standard deviation = 1.3 mg N m^{-2} day^{-1})."

Reviewers 2 and 3: Out-of-sample performance

The reviewers seem to have missed that Chamber 6 served as an out-of-sample evaluation. This should be made clearer throughout the manuscript.

• Figs. 2 and 3 should indicate (graphically and in caption) which chamber was out-of-sample

• Figs. 4 and 5 should only include the out-of-sample chamber as the observation. (It's unclear whether this is already the case.) This should be mentioned in their captions.

• Same for Tables 1–3.

Response:

We really appreciate your suggestions. Indeed we should repeatedly remind reviewers and readers that model performance is evaluated based on a leave-one-out cross validation (LOOCV) method. Each time when we trained and evaluated a model, the six chambers' data were split into five chambers for training and the remaining one as validation, thus serving as out-of-sample evaluation. Following the editor's suggestion, we have first added a sentence in method section 2.2.2 paragraph 2 to explain the validation process (line 210 to 213):

"We used the leave-one-out cross-validation (LOOCV) method for the evaluation process. Each time we used five chambers' data for model finetuning and another one chamber data for validation. For example, if we used chamber 1-5 to train the model, then chamber 6 would serve as the out-of-sample data to validate the results. Only the validation results would be presented in our study."

Moreover, we have added descriptions about LOOCV being used and all results are from validation data sets to 2-5 figure captions:

"Figure 2: Leave-one-out cross validation of time series of N_2O flux (mg N m^{-2} day^{-}) predicted by the pure non-pretrained GRU model (blue line) and KGML-ag1 model (red line). Observations are shown as black line-dots. Validation results for each chamber were based on out-of-sample predictions by models trained by other five chambers."

"Figure 3: Leave-one-out cross validation of time series of IMVs predicted by KGML-ag1 model (red line). Observations are shown as black line-dots. Validation results for each chamber were based on out-of-sample predictions by models trained by other five chambers. 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 Mg^{-1}, respectively. "

"Figure 3 Contd.: Leave-one-out cross validation of time series of IMVs predicted by KGML-ag1 model (red line). Observations are shown as black line-dots. Validation results for each chamber were based on out-of-sample predictions by models trained by other five chambers.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 Mg^{-1} and m^3 m^{-3}, respectively."

"Figure 4: The comparisons of overall prediction accuracy from leave-one-out cross validation 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 models. The overall performances were calculated by comparing out-of-sample predictions (each chamber's predictions were from models trained by other five chambers) from all validated chambers with observations. 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 5: The comparisons of N_2O flux prediction accuracy r^2 (a) and (b) RMSE from leave-one-out cross validation, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models

(ANN and GRU) and KGML-ag models in six chambers. Validation results for each chamber were based on out-of-sample predictions by models trained by other five chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments."

and table 2 footnote a:

"[a]Leave-one-out cross validation results for each chamber were based on out-of-sample predictions by models trained by other five chambers. The "All data" performances were calculated by comparing out-of-sample predictions from all validated chambers with observations. "

and table 3 footnote c:

""The leave-one-out cross validation overall performances were calculated by comparing out-of-sample predictions (each chamber's predictions were from models trained by other five chambers) from all validated chambers with observations."

Table 1 presented the pretrained results from synthetic data and it used ecosys model generated data to train and evaluate the model. Only the results from testing data (synthetic testing data from 19 counties) were presented. We have also added this information into the table 1 title:

"Table 1: Pretrain results for different model and IMV combinations using *ecosys* synthetic data. Only performances from testing data sets (synthetic data from 19 counties) were presented. "

You also point to how the model behaves when there is no chamber observation data as an additional out-of-sample test. While you may suspect it does well, without any measurement data, it's not justifiable to use this as a certain measure of performance. Please revise lines 424–429 to reflect that (e.g., "poor assumed performance," "assumed improvement).

Response:

Thanks so much for the suggestion. We have added a new sentence at the beginning to declare this assumption and revised the part in section 3.3 last paragraph to reflect the *assumed* performance increase (line 431 to 438):

"For periods without any observed data, we assumed that the good model predictions should be stable, consistent with the nearest period and close to the reality in the experiment setting (e.g. no erratic peak and N_2O flux near 0 mg N m^{-2} day^{-1} before day 25). 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 the GRU model was stable in short missing period (1-2 days of missing data) and only presented poor performance in long missing period (before day 25). This improvement by the GRU model may 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."

Reviewer 3: N_2O fluxes and NO_3 concentrations higher than normal

In your reply to Reviewer 3, you posited that you saw peaks of N_2O around 20 mgN m^{-2} day^{-1} and NO_3 around 40 mgN kg^{-1}. However, it appears from Figs. 2 and 3 that those are actually about 60 mgN m^{-2} day^{-1} and 95 gN m^{-2}, respectively. Compare to the papers you cited (units converted as necessary to match yours):

N_2O emissions (mgN m^{-2} day^{-1})	Reference
4.8	Venterea et al. (2011) Figs 3–4
8.2	Fassbinder et al. (2013), p. 612
19.2	Grant & Pattey (1999) abstract
18	Grant et al. (2006) Fig. 2
34	Grant & Pattey (2008) Fig. 3
38	Grant & Pattey (2008) Fig. 4
50	Hamrani et al. (2020)

NO_3 concentration (mgN kg^{-1})	Reference
7.1	Grant & Pattey (1999) Table 3
80	Venterea et al. (2011) Fig. 8

It's not a problem that your peaks are higher than seen in other studies, but as Reviewer 3 suggested, this should be disclosed (perhaps in the Discussion).

Response:

We really appreciate the editor's efforts on checking the details of the N_2O flux and NO_3^- soil concentration in the cited references. We have also double-checked all numbers in the references and our mesocosm data set. The N_2O fluxes from all six mesocosm chambers during the peak regions (45-60 days) are 16.9±11.7 mg N m^{-2} day^{-1}, except for abnormal values (> 40 and up to 71 mg N m^{-2} day^{-1}) in 2016 of chamber 3 and 4. The NO_3^- soil concentrations from all six mesocosm chambers during the peak regions (45-60 days) are 59.3±20.7 g N Mg^{-1}, with highest values up to 95.2 mg N m^{-2} day^{-1}. NO_3^- data from our mesocosm experiments are from 0-15cm but values in Grant et al. (1999) Table 3 are from all layers ranging from 1cm to 115cm. Grant et al. (1999) indicated that the surface layer of the soil would normally have a higher NO_3^- concentration level (31.36 g N Mg^{-1} for 0-15cm). We admit that field level data as listed in references can be different from our mesocosm data (e.g. drainage condition, wind), and our data are in the high end of the ranges summarized from previous field studies.

Following the editor's suggestion, we have revised the description of the data in method section 2.2.2 paragraph 1 and added another field N_2O study (Grant et al., 2016, Fig.4, N_2O peak around 100 mg N m^{-2} day^{-1}) to the reference list (line 199 to 202):

"The magnitude of N_2O flux and NO_3^- soil concentration and their responses following fertilizer application from this mesocosm experiment are slightly higher than several field studies of agricultural soils (Fassbinder et al., 2013; Grant et al., 1999, 2006, 2008, 2016; Hamrani et al., 2020; Venterea et al., 2011)."

Besides, we have added discussion in the discussion section 4.3 first paragraph (line 523 to 527):

"The mesocosm measurements of N_2O fluxes (16.9±11.7 mg N m^{-2} day^{-1} during days of 45-60; Highest value is 71 mg N m^{-2} day^{-1}) and NO_3^- soil concentrations (59.3±20.7 g N Mg^{-1} during days of 45-60; Highest value is 95.2 g N Mg^{-1}) are at the high end of the range that has been observed by field studies (Fassbinder et al., 2013; Grant et al., 1999, 2006, 2008, 2016; Hamrani et al., 2020; Venterea et al., 2011)."

Moreover, we have found that the unit of NO_3^- and NH_4^+ soil concentrations (both should be g Mg^{-1} throughout the manuscript) in some places of the main text is not consistent. This is a writing mistake but will not affect any results interpretation. We have fixed them in Figure 3, Figure S1 and Figure S5.

Minor corrections

• L140-1: Should be "respiration, and NO_2^- becomes an alternative electron acceptor" (note not "respirations")

Response: We have corrected this in line 139 to 140:

"... when O_2 availability fails to meet O_2 demand for their respiration, and NO_2^- become alternative electron acceptors."

• L143: Should be "considers"

Response: We have corrected this in line 142 to 143:

"Unlike the pipeline model described by Davidson et al. (2000) , which mainly considers the correlations of N_2O production with nitrogen availability and of N_2O emissions with soil water content, ..."

• L216-8:

- o Should be "Since up to"
- o 16/24 is ⅔, not ¾
- o Should be "of the day is"
- o "and meanwhile present slight variations"—it's unclear what this means

Response: We have corrected the mistakes and revised the sentence to make it more clear in line 218 to 220:

"Since up to 2/3 of the day is covered by the selected data (16 hours /24 hours), the augmented daily values should be representative enough for the source day and with slight variations from each other. "

• L257: "the highly ranked"

Response: We have corrected this in line 258 to 260:

"The objective of building IMVcb2 was to investigate the importance of the 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. "

• L375: What is a "hot moment"? Please define for less-technical readers.

Response: We have added the definition of "hot moment" in line 376 to 378:

"... which is critical when predicting fast-change variables with hot moments (a short period of time with rare events like flux increasing quickly) like N_2O."

• L560: "from a PB model to an ML model"

Response: We have corrected this in line 572 to 573:

"... such as using pretraining to transfer knowledge from a PB model to a ML model ..."

• L565: "We expect our validation results will be more solid" is a little too casual and vague. Maybe something like, "We expect to further validate and refine our model"

Response: We have revised the sentence based on the editor's advice in line 578 to 580:

"We expect to further validate and refine our KGML-ag model once more gold standard data of N_2O fluxes along with other relevant inputs and intermediate variables become publicly available."

• L567: “Will be inevitable” why?

Response: We agree with the editor that here “inevitable” is too strong and ignore other possibilities. Thus we have replaced it with “possible” and revised the sentence in line 580 to 581:

“Moreover, incorporating more and more domain knowledge into KGML-ag will be possible for further improvement, ...”

• L568: “surrogate” isn’t a verb. Maybe replace “to efficiently surrogate” with “efficiently emulating”

Response: We agree with the editor about this replacement and have revised the sentence in line 581 to 582:

“In fact, to efficiently emulate components of PB models has been proposed as a research frontier in hybrid modeling for earth system science ...”

• Line 723: “structuress”

Response: We have corrected this in figure 1 caption:

“Figure 1: The model structures. ...”