Supplement of Geosci. Model Dev., 15, 2839–2858, 2022
https://doi.org/10.5194/gmd-15-2839-2022-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

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 et al.

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

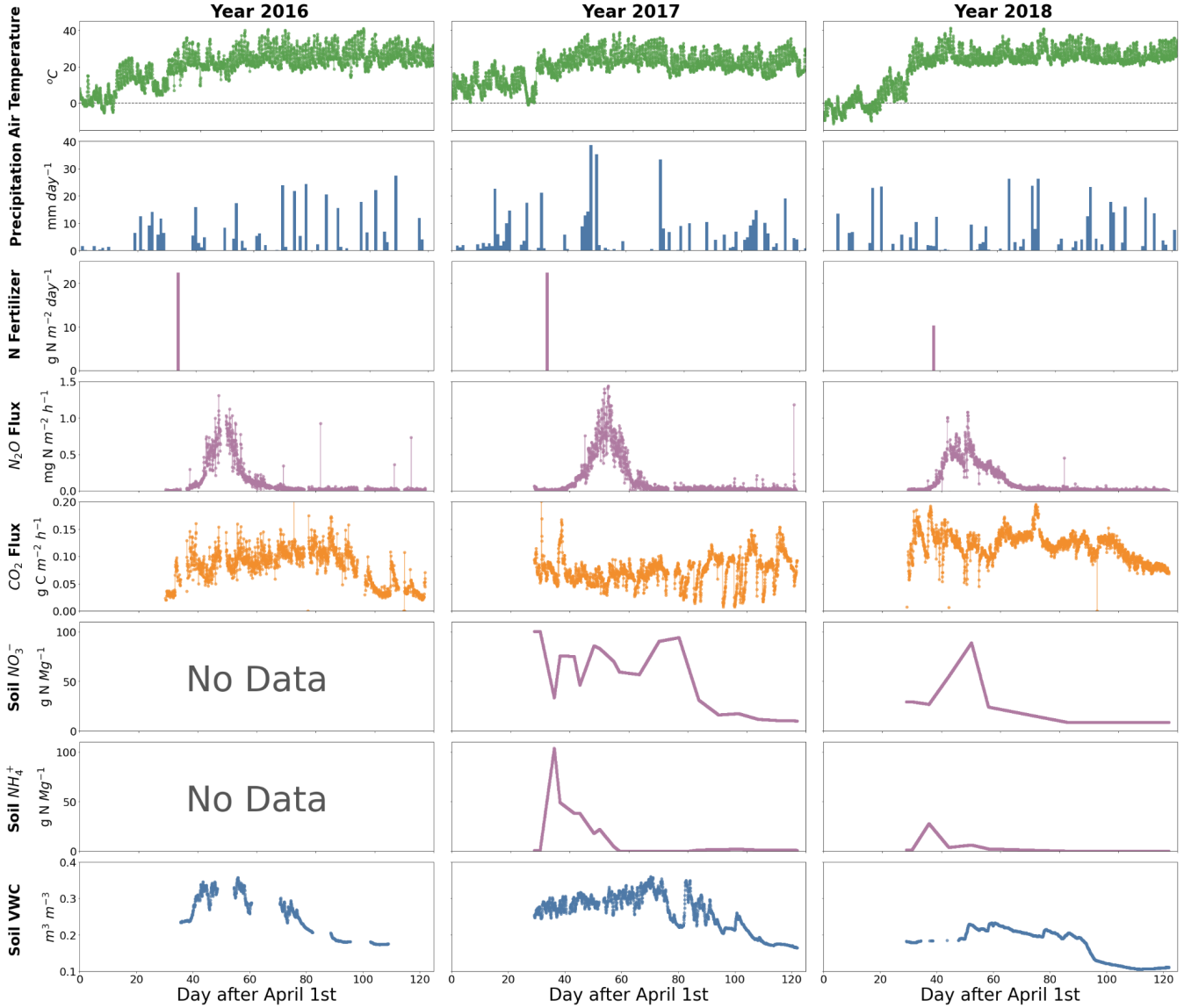

Figure S1: Time series example of observation data collected from mesocosm chamber 1. The precipitation, N fertilizer, Soil NO_3^- and NH_4^+ data are in the daily time scale, while other data are in the hourly time scale. Temperature presents in green; water related variables (precipitation and soil VWC) are in blue; N related variables (N fertilizer, N_2O flux, Soil NO_3^- and NH_4^+) are in purple; and CO_2 is in orange. Anomaly points will be down-weighted by daily averaging method with quality check in later processes, which is mentioned in section 2.2.2 last paragraph.

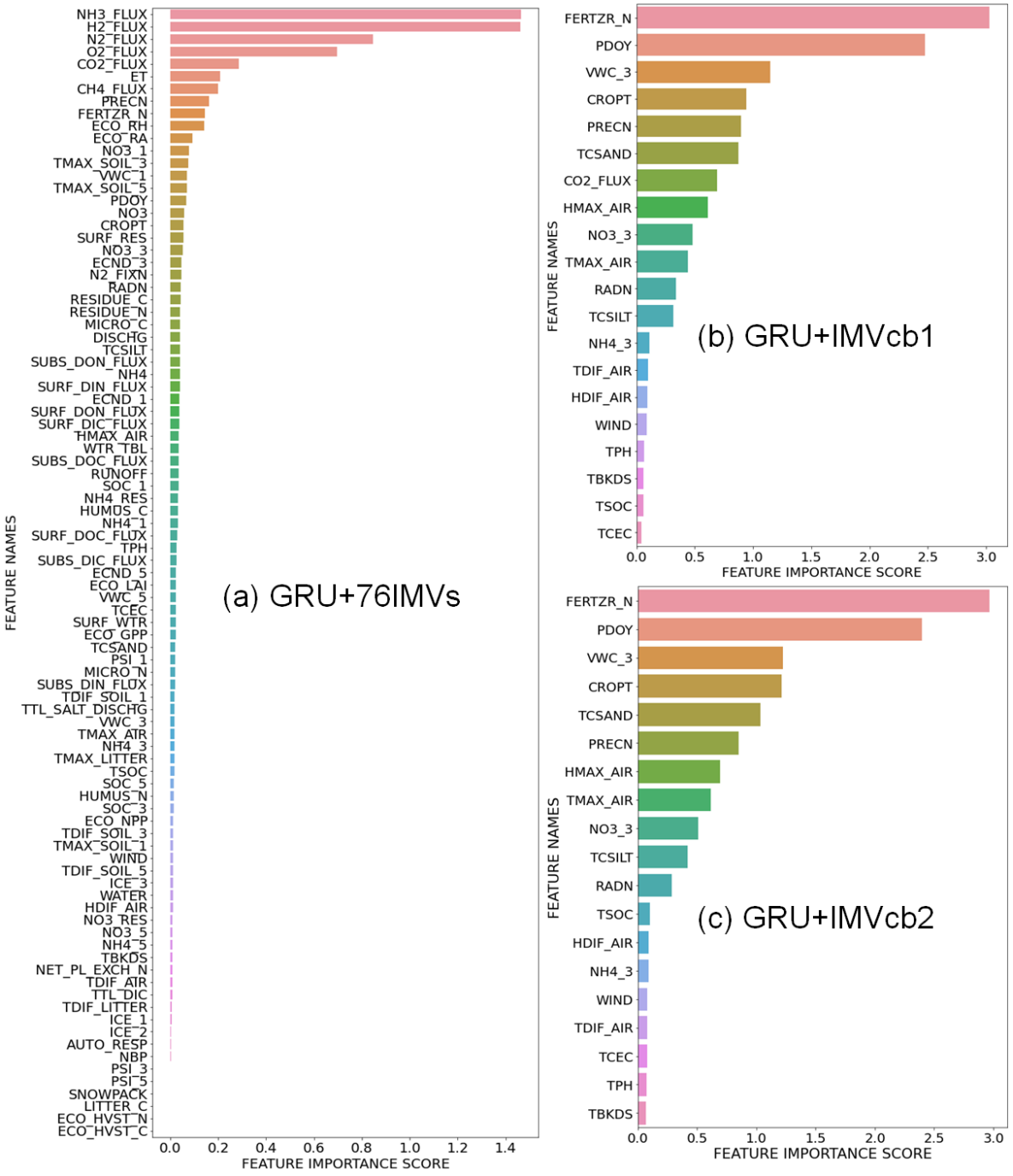

Figure S2: Feature importance test for intermediate variables (IMVs) with GRU models. To be noticed, the VWC, NO_3^-, and NH_4^+ from third layer soil, which are presented in the main text are abbreviated here as VWC_3, NO_3_3, and NH_4_3 to be distinguished from the same variables from 1st and 5th layers. Details of the variables can be found in Table S1.

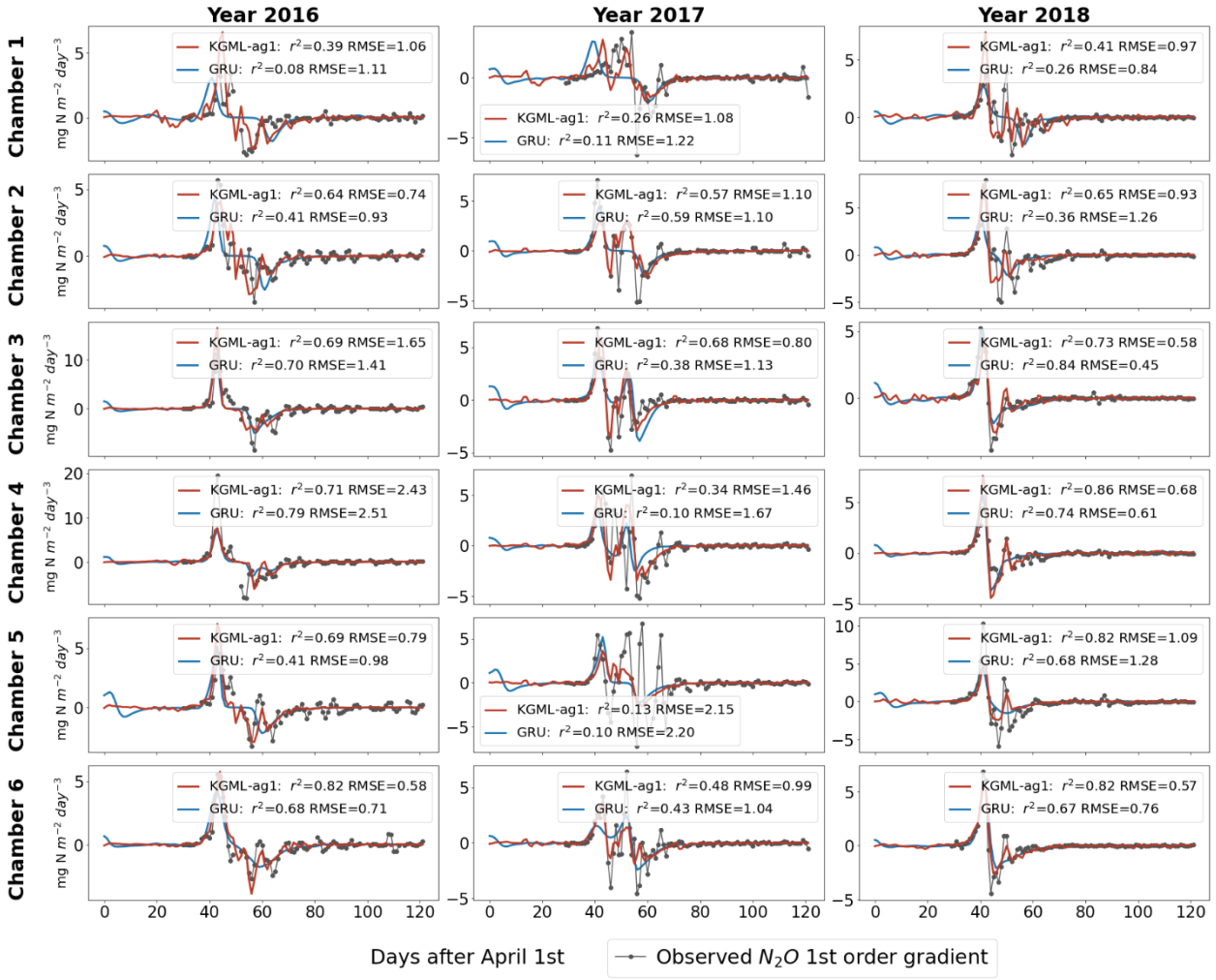

Figure S3: N_2O flux 1st order gradient time series comparisons between non-pretrained GRU model and KGML-ag1. The black-dot line represents the observation, while blue represents GRU and red represents KGML-ag1.

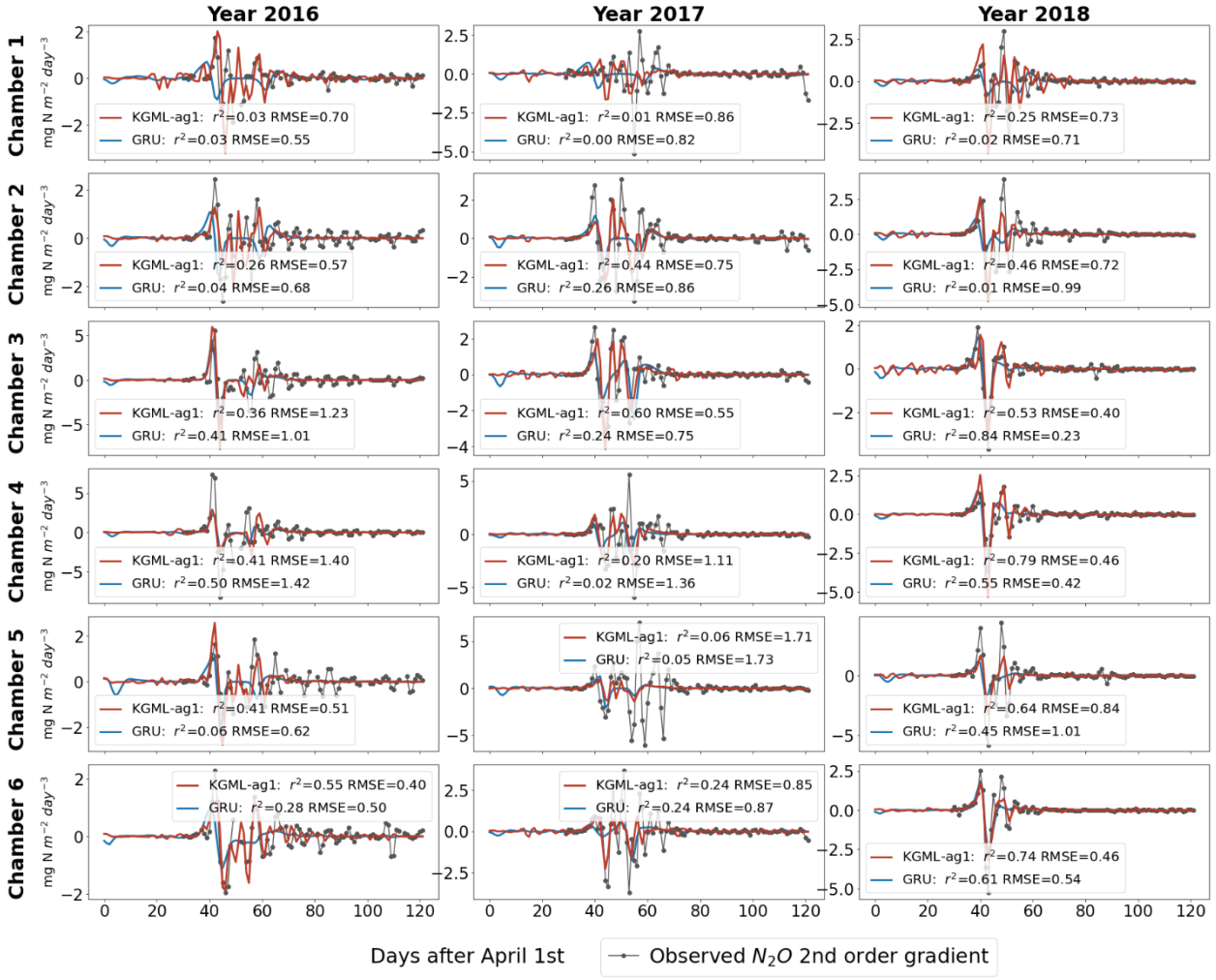

Figure S4: N_2O flux 2nd order gradient time series comparisons between non-pretrained GRU model and KGML-ag1. The black-dot line represents the observation, while blue represents GRU and red represents KGML-ag1.

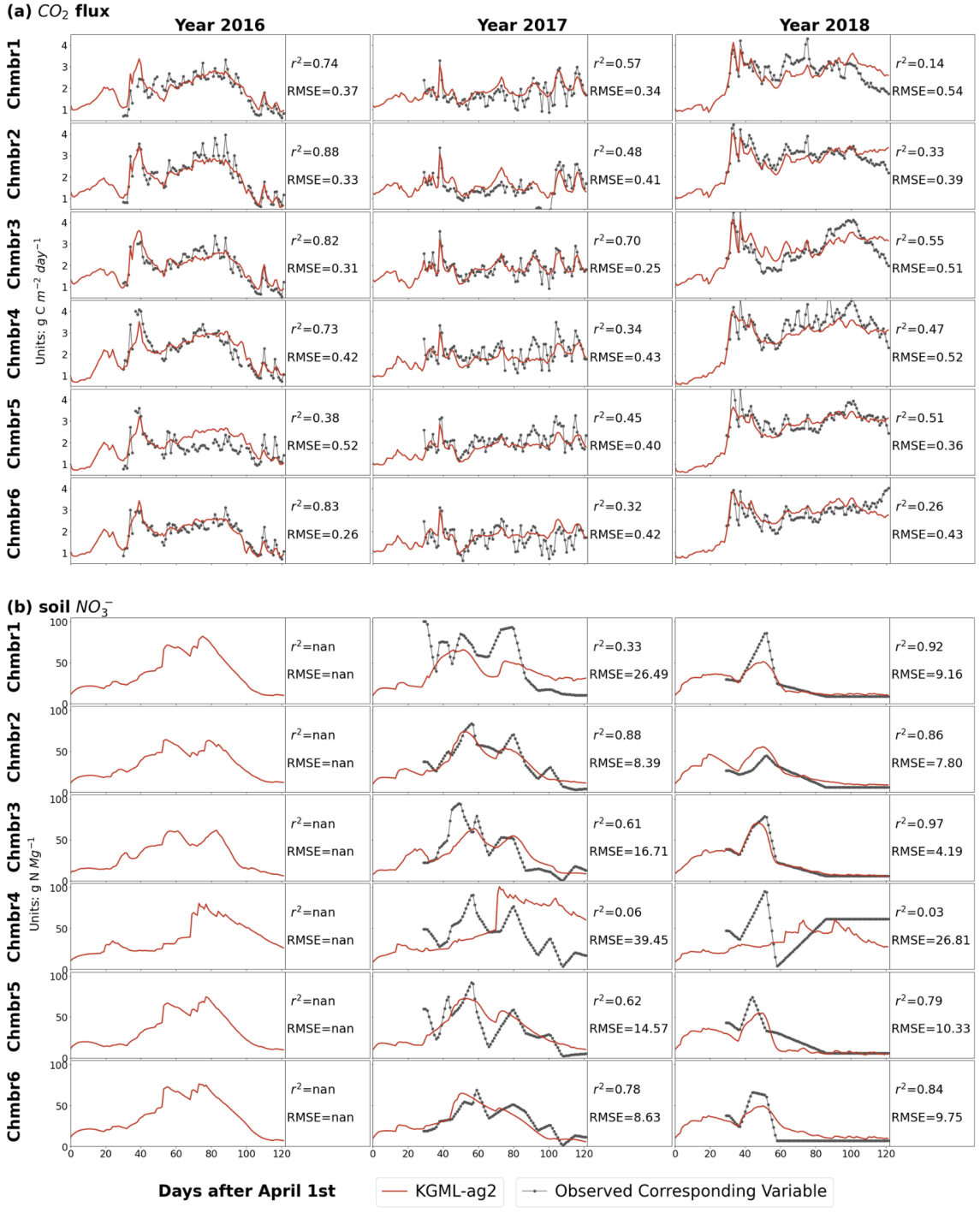

Figure S5: IMVs prediction from KGML-ag2. The black-dot line represents observations and the red line represents the results from KGML-ag2. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber.

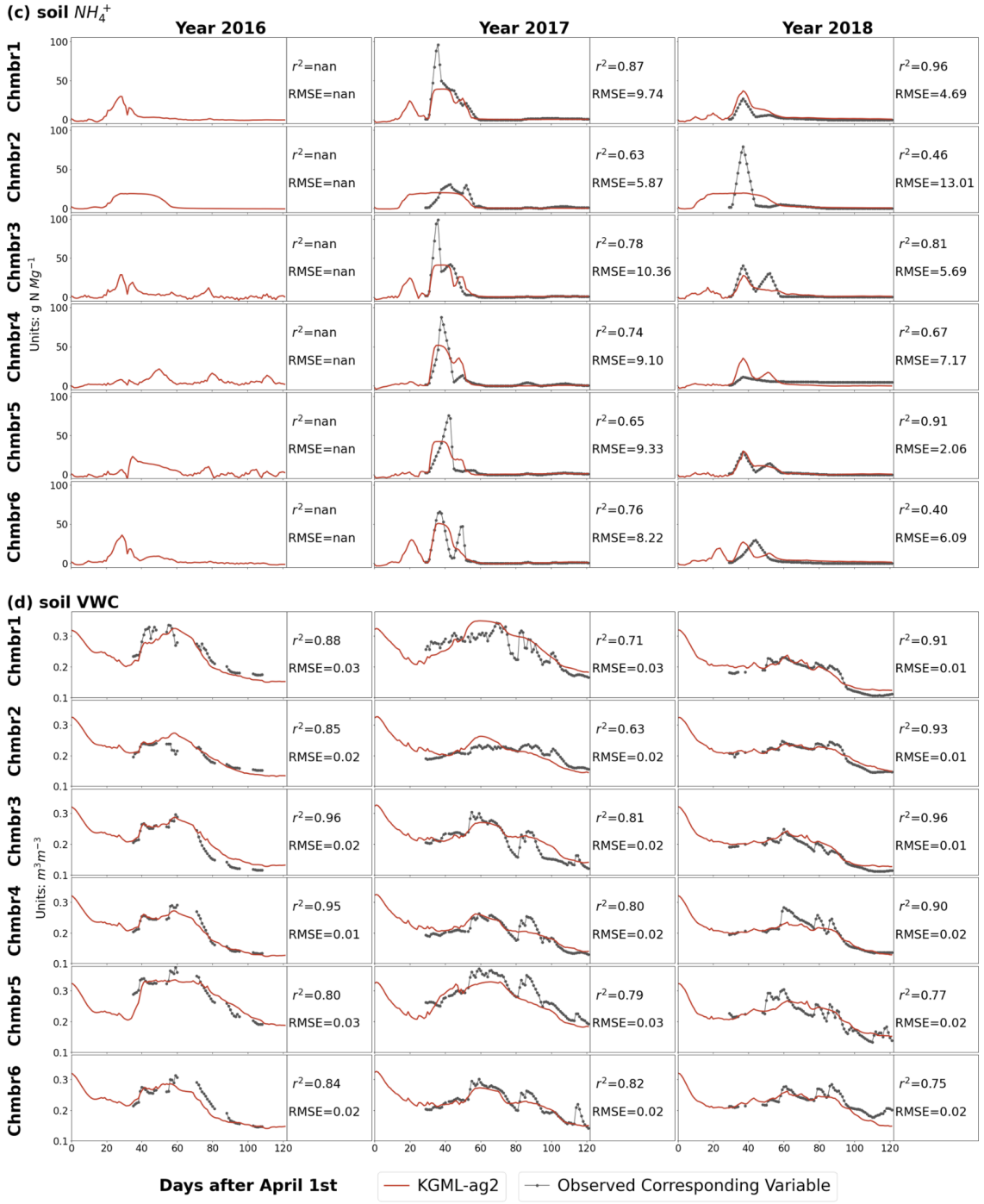

Figure S5 contd.: IMVs prediction from KGML-ag2. The black-dot line represents observations and the red line represents the results from KGML-ag2. Chmb is the abbreviation for chamber. r^2 and RMSE are calculated and present in each year and chamber.

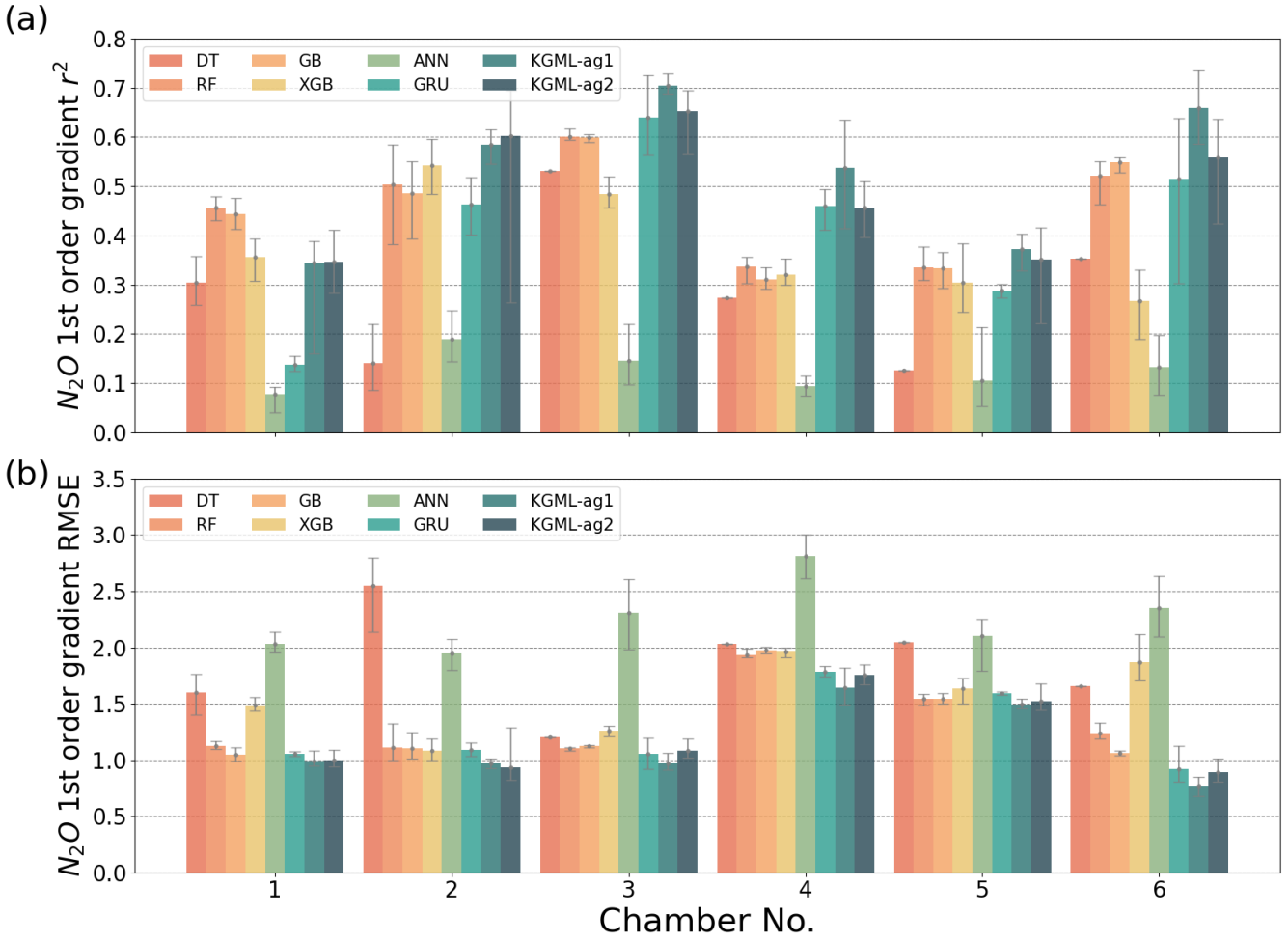

Figure S6: The comparisons of N_2O 1st order gradient prediction accuracy r^2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag models in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

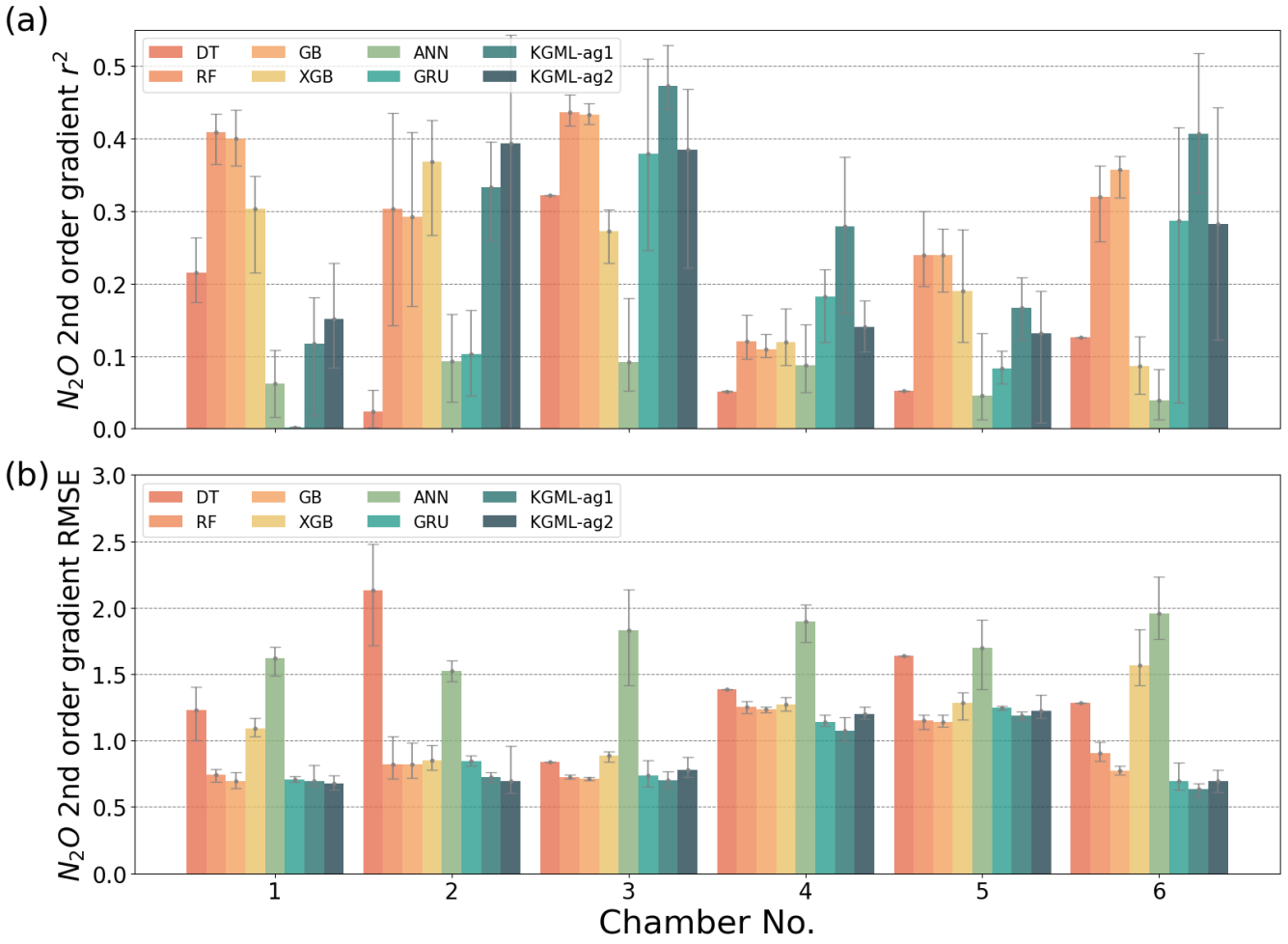

Figure S7: The comparisons of N_2O 2nd order gradient prediction accuracy r^2 (a) and (b) RMSE, between four tree-based ML models (DT, RF, GB and XGB), two deep learning models (ANN and GRU) and KGML-ag1 model in 6 chambers. The gray error bars are coming from the maximum and minimum scores of ensemble experiments.

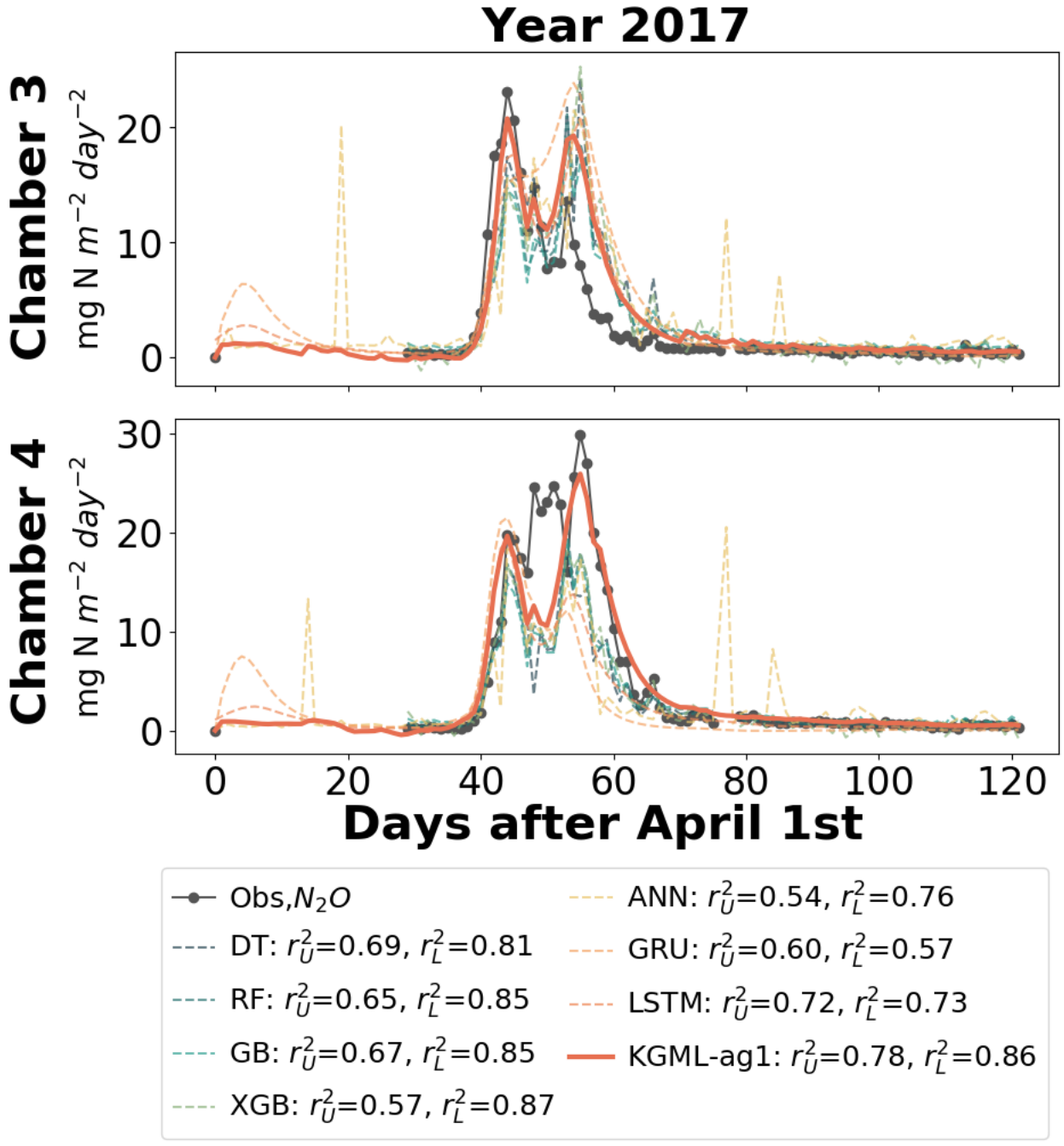

Figure S8: N_2O flux time series comparisons between KGML-ag1 predictions (red solid line), pure ML models (other colored dashed line) and observations (black-dot line) from cross-validation on two representative panels of chamber 3 and 4 in 2016. The r^2 value was calculated between observations and 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.

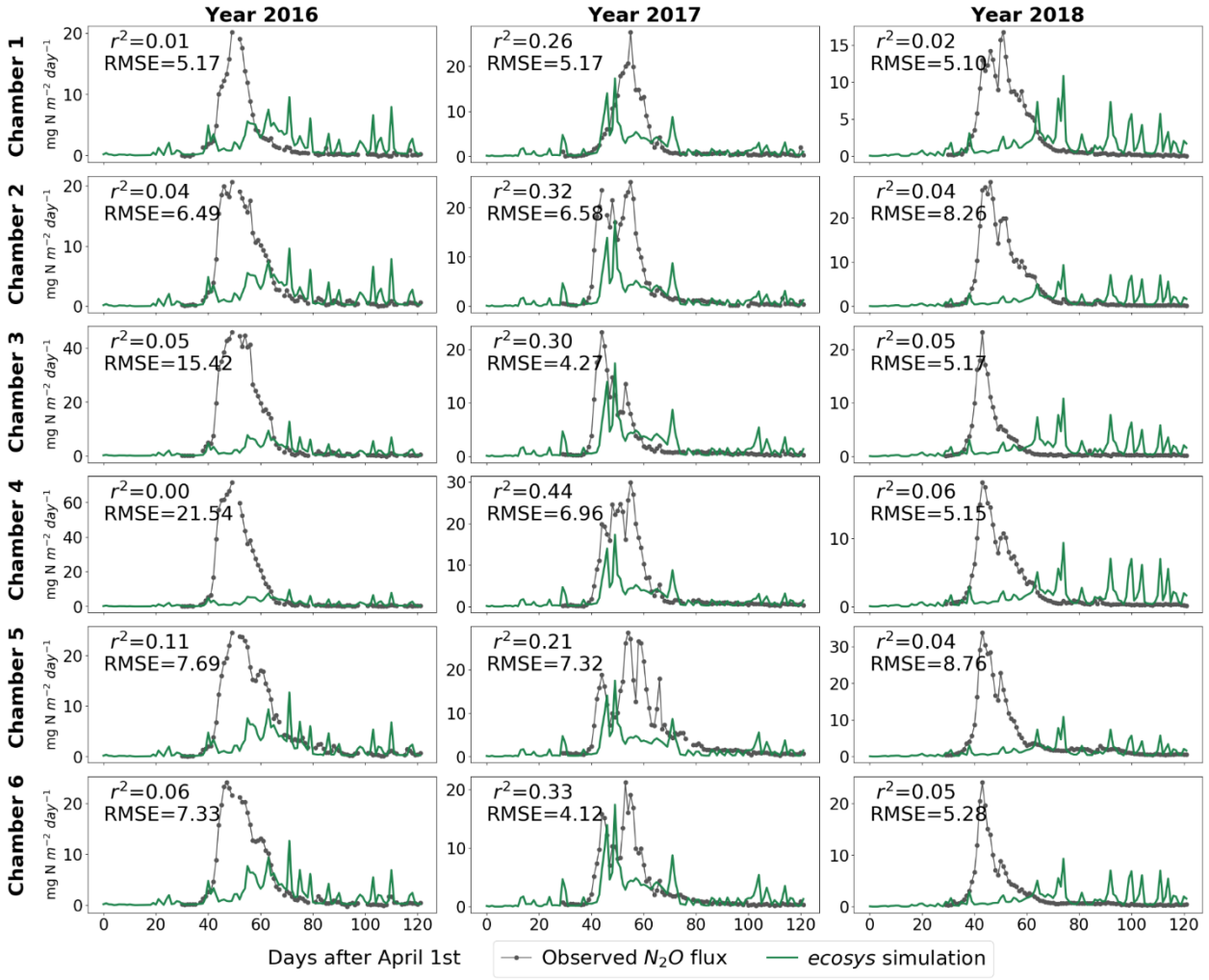

Figure S9: N_2O flux time series comparisons between *ecosys* simulations (green line) and observations (black-dot line).

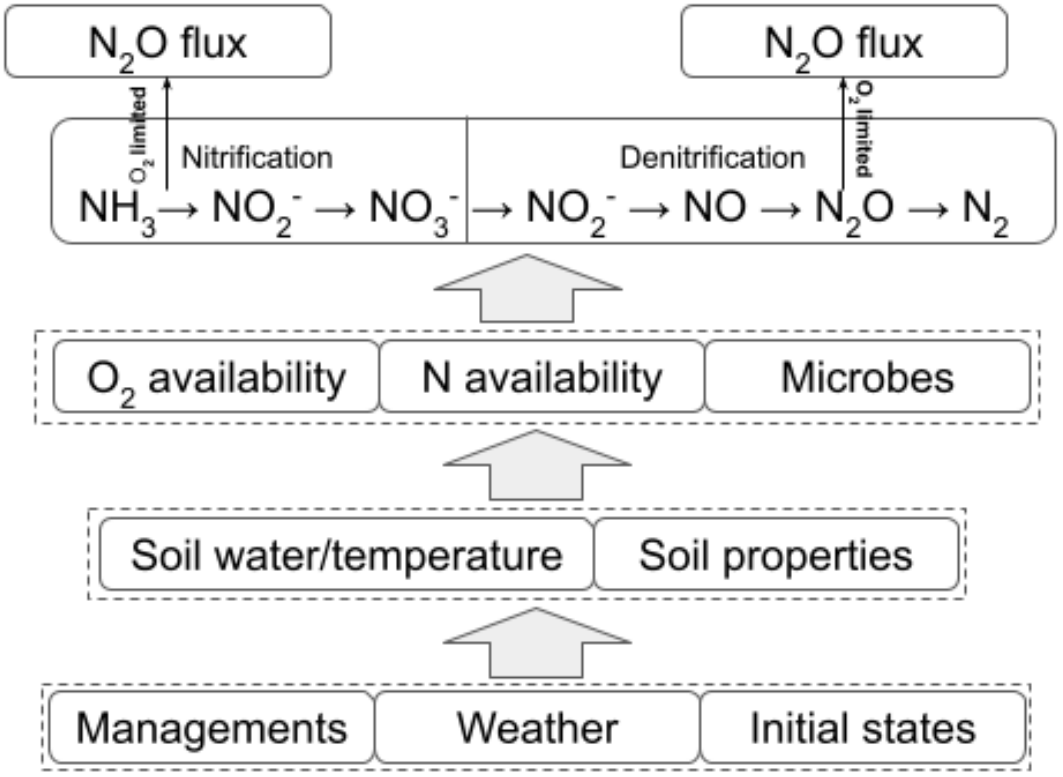

Figure S10: The simplified schema of N_2O flux related variables and processes.

Table S1: Variable short abbreviation, category (IMV represents intermediate variable, W represents weather forcing, FN represents the N fertilizer rate, SCP represents soil/crop property and target variable), description and units.

No.	Abbreviation	Variable categary	Descriptions	Units
1	RESIDUE_C	IMV	Total residue C on soil surface and in soil profile	g C m^{-2}
2	HUMUS_C	IMV	Total particulate + non-particulate C in soil profile	g C m^{-2}
3	LITTER_C	IMV	C in above + below-ground litterfall	g C m^{-2}
4	CO2_FLUX	IMV	CO_2 flux at the soil surface	g C m^{-2} day^{-1}
5	O2_FLUX	IMV	O_2 flux at the soil surface	g O_2 m^{-2} day^{-1}
6	AUTO_RESP	IMV	Below-ground autotrophic (root) respiration	g C m^{-2} day^{-1}
7	MICRO_C	IMV	Microbial C in all residue and humus complexes	g C m^{-2}
8	SURF_RES	IMV	Residue C on soil surface and in soil profile	g C m^{-2}
9	CH4_FLUX	IMV	CH_4 flux at the soil surface	g C m^{-2} day^{-1}
10	SURF_DOC_FLUX	IMV	Flux of organic C across all external surface boundaries in runoff and sediment	g C m^{-2} day^{-1}
11	SUBS_DOC_FLUX	IMV	Flux of organic C across all external subsurface boundaries in water dischage	g C m^{-2} day^{-1}
12	SURF_DIC_FLUX	IMV	Flux of inorganic C across all external surface boundaries in runoff and sediment	g C m^{-2} day^{-1}
13	SUBS_DIC_FLUX	IMV	Flux of inorganic C across all external subsurface boundaries in water dischage	g C m^{-2} day^{-1}
14	NBP	IMV	Net biome productivity	g C m^{-2} day^{-1}
15	SOC_1	IMV	Residue + humus C in soil layer 1, 5cm depth	g C m^{-2}
16	SOC_3	IMV	Residue + humus C in soil layer 3, 15cm depth	g C m^{-2}
17	SOC_5	IMV	Residue + humus C in soil layer 5, 28cm depth	g C m^{-2}
18	H2_FLUX	IMV	H_2 flux at the soil surface	g H_2 m^{-2} day^{-1}
19	ECO_HVST_C	IMV	C removed in harvest	g C m^{-2}
20	ECO_LAI	IMV	Leaf area index	m^2 m^{-2}
21	ECO_GPP	IMV	Gross primary productivity	g C m^{-2} day^{-1}
22	ECO_RA	IMV	Autotrophic respiration	g C m^{-2} day^{-1}
23	ECO_NPP	IMV	Net primary productivity	g C m^{-2} day^{-1}
24	ECO_RH	IMV	Heterotrophic respiration	g C m^{-2} day^{-1}
25	TTL_DIC	IMV	Total stocks of dissolved inorganic C	g C m^{-2}
26	ET	IMV	Evapotranspiration rate	mm day^{-1}
27	RUNOFF	IMV	Overland surface flow	mm day^{-1}
28	WATER	IMV	The total amount of water in the rooting zone of the soil profile	mm day^{-1}
29	DISCHG	IMV	Water discharge flux through all subsurface boundaries	mm
30	SNOWPACK	IMV	The equivalent water content of snow + ice + water in the snowpack	mm
31	VWC_1	IMV	The volumetric water content in soil layer 1, 5cm depth	m^3 m^{-3}
32	VWC_3	IMV	The volumetric water content in soil layer 3, 15cm depth	m^3 m^{-3}
33	VWC_5	IMV	The volumetric water content in soil layer 5, 28cm depth	m^3 m^{-3}

34	SURF_WTR	IMV	Near surface volumetric water content	$m^3\ m^{-3}$
35	ICE_1	IMV	The volumetric ice content in soil layer 1, 5cm depth	$m^3\ m^{-3}$
36	ICE_2	IMV	The volumetric ice content in soil layer 3, 15cm depth	$m^3\ m^{-3}$
37	ICE_3	IMV	The volumetric icecontent in soil layer 5, 28cm depth	$m^3\ m^{-3}$
38	PSI_1	IMV	The matric water potential in soil layer 1, 5cm depth	Mpa
39	PSI_3	IMV	The matric water potential in soil layer 3, 15cm depth	Mpa
40	PSI_5	IMV	The matric water potential in soil layer 5, 28cm depth	Mpa
41	WTR_TBL	IMV	Depth of the water table from the surface	m
42	RESIDUE_N	IMV	Total residue N on soil surface and in soil profile	$g\ N\ m^{-2}$
43	HUMUS_N	IMV	Total particulate + non-particulate N in soil profile	$g\ N\ m^{-2}$
44	FERTZR_N	FN	N fertilizer applied	$g\ N\ m^{-2}$
45	NET_PL_EXCH_N	IMV	Net N exchange between soil and plants	$g\ N\ m^{-2}\ day^{-1}$
46	NH4	IMV	Total $NH_4^+ + NH_3$ in the soil profile	$g\ N\ m^{-2}$
47	NO3	IMV	Total NO_3^- in soil profile	$g\ N\ m^{-2}$
48	SURF_DON_FLUX	IMV	Flux of organic N across all external surface boundaries in runoff and sediment	$g\ N\ m^{-2}\ day^{-1}$
49	SUBS_DON_FLUX	IMV	Flux of organic N across all external subsurface boundaries in water dischage	$g\ N\ m^{-2}\ day^{-1}$
50	SURF_DIN_FLUX	IMV	Flux of inorganic N across all external surface boundaries in runoff and sediment	$g\ N\ m^{-2}\ day^{-1}$
51	SUBS_DIN_FLUX	IMV	Flux of inorganic N across all external subsurface boundaries in water dischage	$g\ N\ m^{-2}\ day^{-1}$
52	N2O_FLUX	Target variable	N_2O flux at the soil surface	$g\ N\ m^{-2}\ day^{-1}$
53	NH3_FLUX	IMV	NH_3 flux at soil and plant surfaces	$g\ N\ m^{-2}\ day^{-1}$
54	N2_FIXN	IMV	Aerobic + anaerobic non-symbiotic N_2 fixation + symbiotic N_2 fixation	$g\ N\ m^{-2}\ day^{-1}$
55	MICRO_N	IMV	Total microbial N in all residue and humus complexes	$g\ N\ m^{-2}$
56	NH4_1	IMV	Total $NH_4^+ + NH_3$ concentration in soil layer 1, 5cm depth	$g\ N\ m^{-2}$
57	NH4_3	IMV	Total $NH_4^+ + NH_3$ concentration in soil layer 3, 15cm depth	$g\ N\ m^{-2}$
58	NH4_5	IMV	Total $NH_4^+ + NH_3$ concentration in soil layer 5, 28cm depth	$g\ N\ m^{-2}$
59	NO3_1	IMV	Total $NO_3^- + NO_2^-$ concentration in soil layer 1, 5cm depth	$g\ N\ m^{-2}$
60	NO3_3	IMV	Total $NO_3^- + NO_2^-$ concentration in soil layer 3, 15cm depth	$g\ N\ m^{-2}$
61	NO3_5	IMV	Total $NO_3^- + NO_2^-$ concentration in soil layer 5, 28cm depth	$g\ N\ m^{-2}$
62	NH4_RES	IMV	Residue $NH_4^+ + NH_3$ on soil surface and in soil profile	$g\ N\ m^{-2}$
63	NO3_RES	IMV	Residue $NO_3^- + NO_2^-$ on soil surface and in soil profile	$g\ N\ m^{-2}$
64	ECO_HVST_N	IMV	N removed in harvest	$g\ N\ m^{-2}\ day^{-1}$
65	N2_FLUX	IMV	N_2 flux at the soil surface	$g\ N\ m^{-2}\ day^{-1}$
66	RADN	W	Solar Radiation	$W\ m^{-2}$
67	TMAX_AIR	W	Max air temperature	^oC
68	TDIF_AIR	W	Difference between max and min air temperature	^oC
69	HMAX_AIR	W	Max humidity	fraction
70	HDIF_AIR	W	Difference between max and min humidity	fraction

71	WIND	W	Wind speed	m s^{-1}
72	PRECN	W	Precipitation	mm day^{-1}
73	TMAX_SOIL_1	IMV	The maximum temperature in soil layer 1, 5cm depth	ºC
74	TDIF_SOIL_1	IMV	The difference between max and min temperature temperature in soil layer 1 , 5cm depth	ºC
75	TMAX_SOIL_3	IMV	The maximum temperature in soil layer 3, 15cm depth	ºC
76	TDIF_SOIL_3	IMV	The difference between max and min temperature temperature in soil layer 3, 15cm depth	ºC
77	TMAX_SOIL_5	IMV	The maximum temperature in soil layer 5, 28cm depth	ºC
78	TDIF_SOIL_5	IMV	The difference between max and min temperature temperature in soil layer 5, 28cm depth	ºC
79	TMAX_LITTER	IMV	The maximum temperature in litter	ºC
80	TDIF_LITTER	IMV	The difference between max and min temperature temperature in litter	ºC
81	ECND_1	IMV	Electrical conductivity in soil layer 1, 5cm depth	dS m^{-1}
82	ECND_3	IMV	Electrical conductivity in soil layer 3, 15cm depth	dS m^{-1}
83	ECND_5	IMV	Electrical conductivity in soil layer 5, 28cm depth	dS m^{-1}
84	TTL_SALT_DISCHG	IMV	Total salt discharge through water through all subsurface boundaries	g Mg^{-1} day^{-1}
85	PDOY	SCP	Plant day of the year	day
86	CROPT	SCP	Crop type, 1 for corn and 0 for soybean	unitless
87	TBKDS	SCP	Depth weighted averaged bulk density in soil profile	Mg m^{-3}
88	TCSAND	SCP	Depth weighted averaged sand content in soil profile	g kg^{-1}
89	TCSILT	SCP	Depth weighted averaged silt content in soil profile	g kg^{-1}
90	TPH	SCP	Depth weighted averagedpH in soil profile	unitless
91	TCEC	SCP	Depth weighted averaged cmol+ kg-1 in soil profile	$cmol^{-1}$ kg^{-1}
92	TSOC	SCP	Depth weighted averaged soil organic carbon in soil profile	g C kg^{-1}

Table S2: N_2O prediction accuracy comparisons between LSTM and GRU models on synthetic data, with different combinations of IMVs (+9 or +58IMVs) and different sliding sliding window settings during training (e.g. 2y1y represent window size is 2 years and the window move 1 year after 1 iteration). Training Efficiency is also compared between LSTM and GRU models for the first two experiments, with changing the training counties = 3, 10, 30, 70, validation counties = 1, 2, 5, 10, and batch size (county numbers input in each iteration) = 1, 5, 5, 5, 5.

	N_2O prediction accuracy		Training effeciency			
Experiment settings	Test r^2	Test RMSE	Train=3, val=1, batch =1	Train=10, val=2, batch =5	Train=30, val=5, batch =5	Train=70, val=10, batch =5
LSTM+9IMVs+1y1y	0.74	1.32	3.8s	3.3s	9.2s	22s
GRU+9IMVs+1y1y	0.81	1.08	3.5s	2.7s	7.2s	17s
LSTM+58IMVs+1y1y	0.91	0.6				
GRU+58IMVs+1y1y	0.92	0.59				
LSTM+58IMVs+2y2y	0.86	0.76				
GRU+58IMVs+2y2y	0.9	0.66				
LSTM+58IMVs+2y1y	0.89	0.67				
GRU+58IMVs+2y1y	0.91	0.6				