Supporting Information for
Quantifying the impact of input data-induced dataset shift on machine learning model applications: A case study of regional reactive nitrogen wet deposition

Yan Zhang[1], Jiani Tan[1,2#], Qing Mu[3], Joshua Fu[4], Li Li[1#]

[1]School of Environmental and Chemical Engineering, Shanghai University, Shanghai, 200444, China
[2]Key Laboratory of Formation and Prevention of Urban Air Pollution Complex, Ministry of Ecology and Environment, Shanghai Academy of Environmental Sciences, Shanghai, 200233, P.R. China
[3]Department of Health and Environmental Sciences, School of Science, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, 215123, China
[4]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, 37996, USA

Correspondence to: Jiani Tan(Jiani-tan@shu.edu.cn) and Li Li (lily@shu.edu.cn)

This file contains:
Figures S1-S14
Table S1-S4

Figure S1

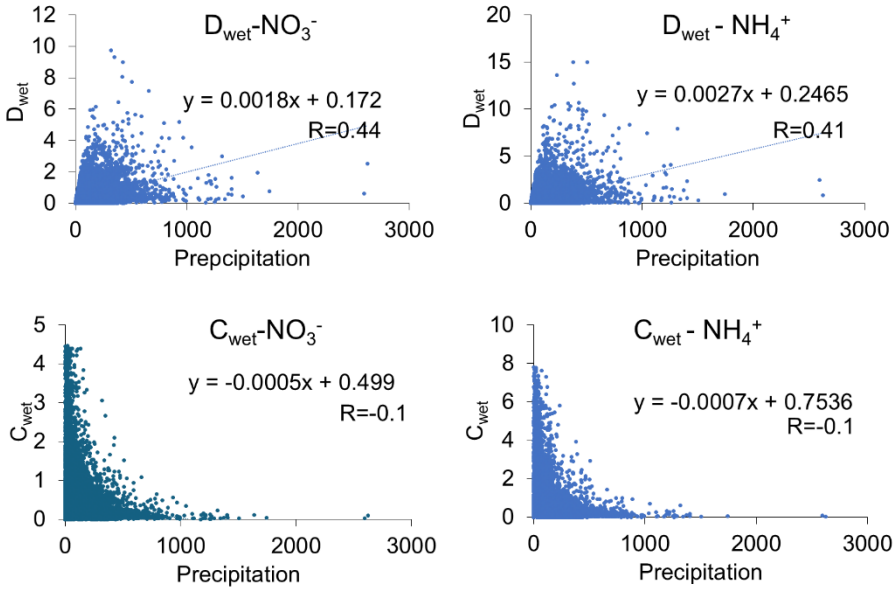

Figure S1. Correlations between precipitation (unit: mm mon^{-1}) and D_{wet} (unit: kg N hec^{-1} mon^{-1}) and C_{wet}. (unit: mg N L^{-1}). Data came from observational datasets.

Figure S2

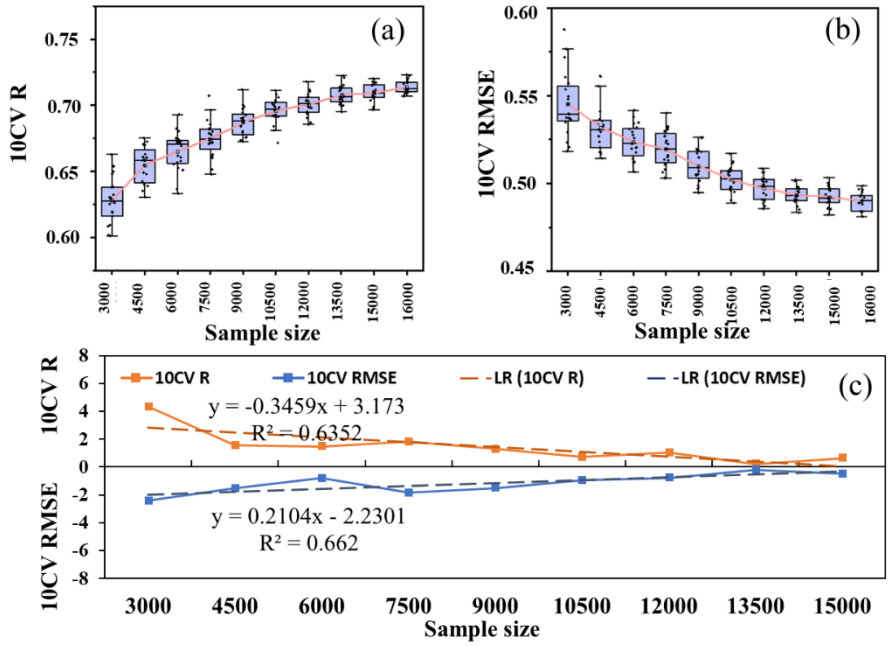

Figure S2 Model performance (R and RMSE) (a-b) and percentage changes (c) with increasing sample sizes for predicting C_{wet}-NO_3^-. The boxplots represent the results from 20 iterations with random data selections.

Figure S3

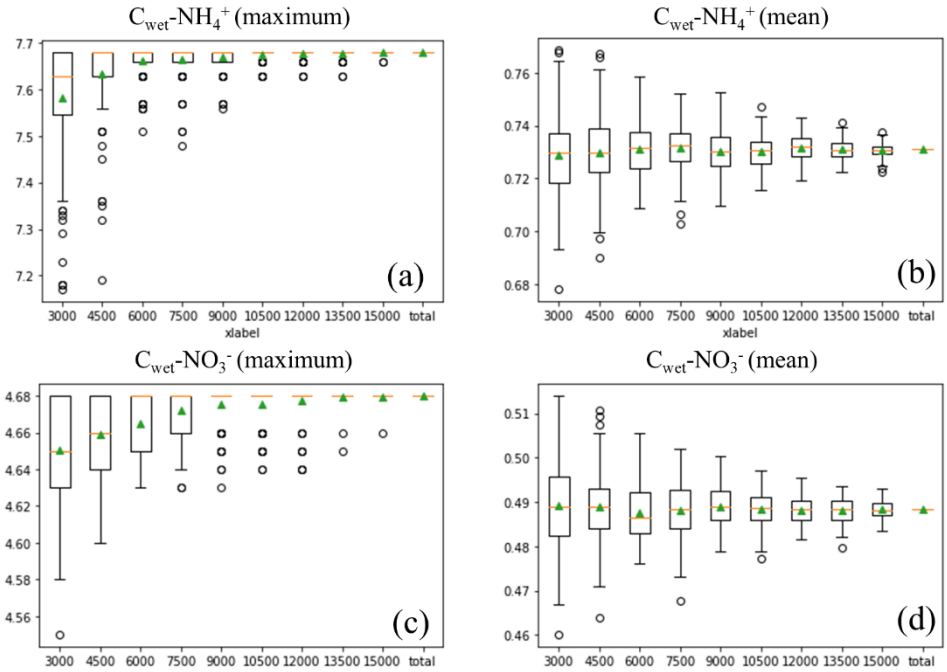

Figure S3. The maximum and average values of C_{wet}-NO_3^- and C_{wet}-NH_4^+ in difference sampling sizes (unit: mg N L^{-1}). The boxplots represent data distributions of 100-time random data selection.

Figure S4

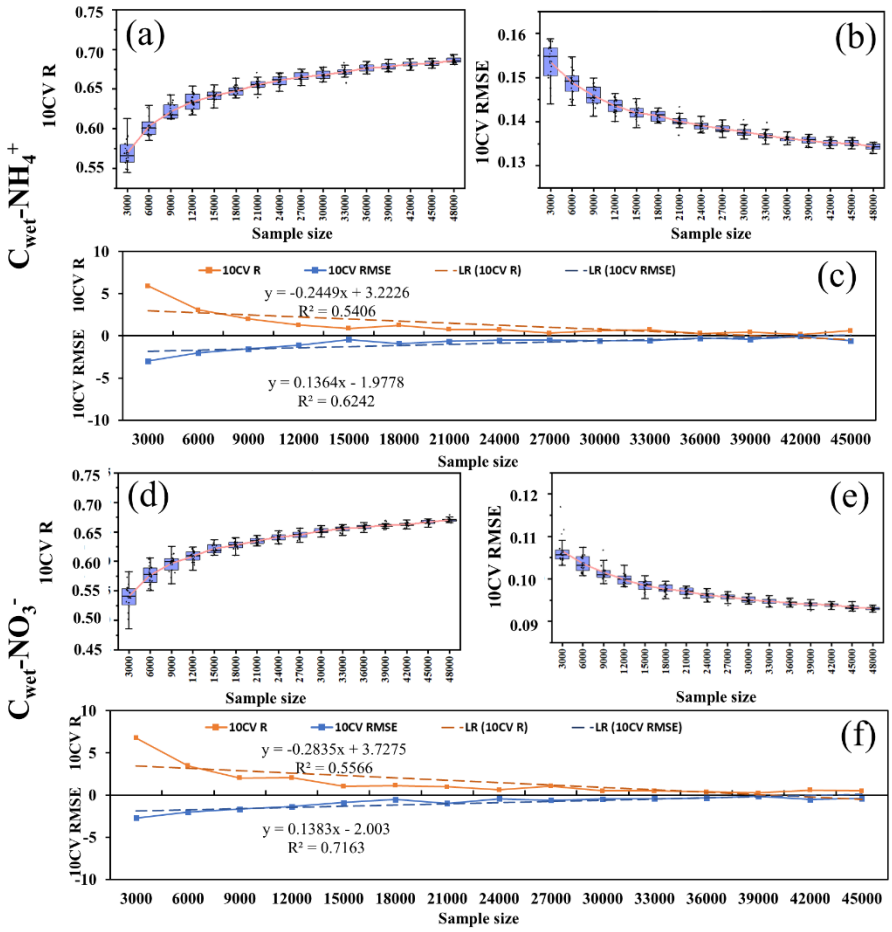

Figure S4. Model performance (R and RMSE) and their percentage changes with increasing sampling sizes for predicting C_{wet}-NH_4^+ (a-c) and C_{wet}-NO_3^- (d-e) (unit: mg N L^{-1}). The boxplots represent the results from 20 iterations with random data selections. We used the same method as Base case (as shown in Fig.4) but using an extensive dataset from NADP (Table 1).

Figure S5

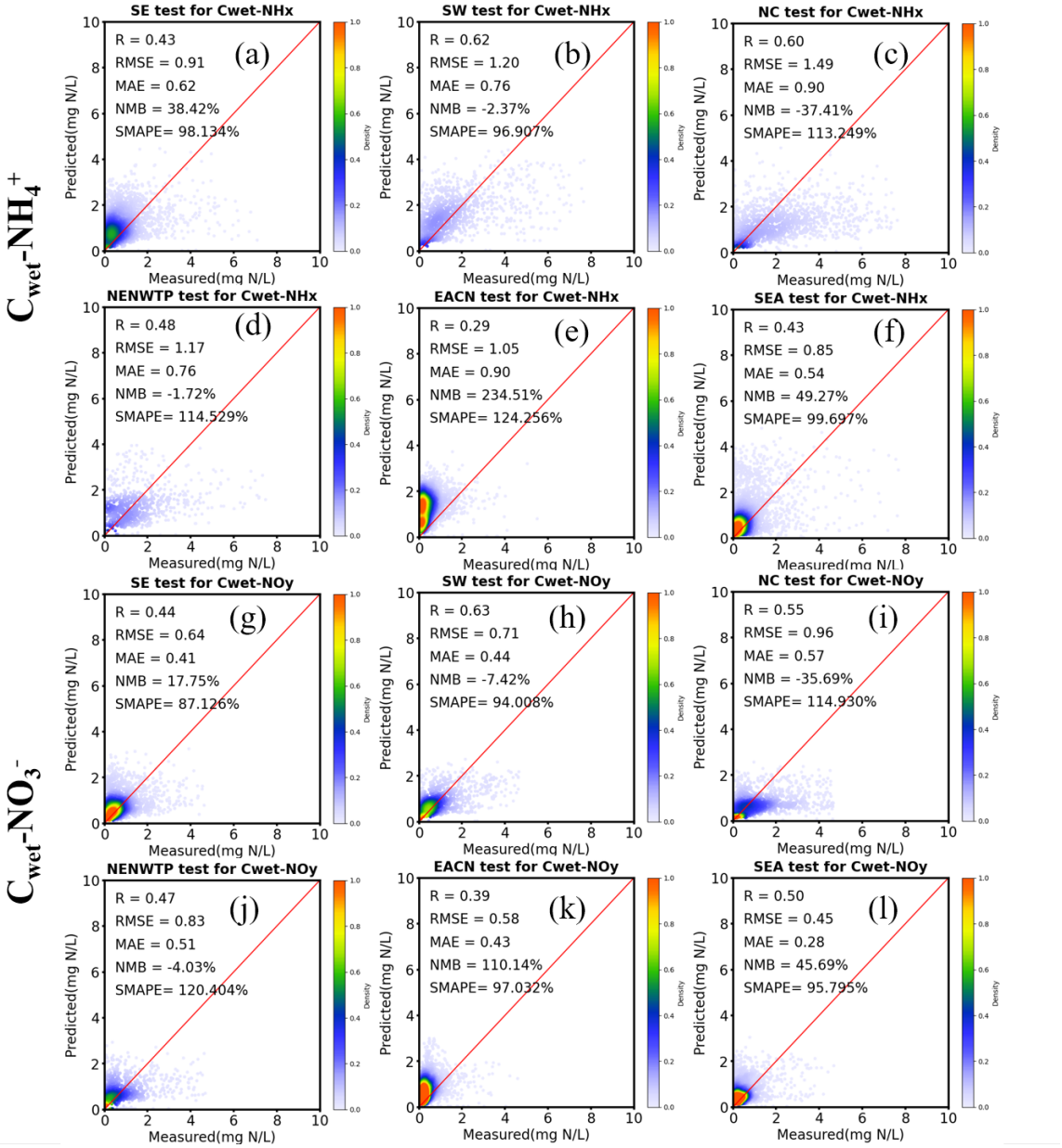

Figure S5. Scatter plots for model performances on predicting C_{wet}-NH_4^+ and C_{wet}-NO_3^- under Case S2 scenarios (unit: mg N L^{-1}).

Figure S6

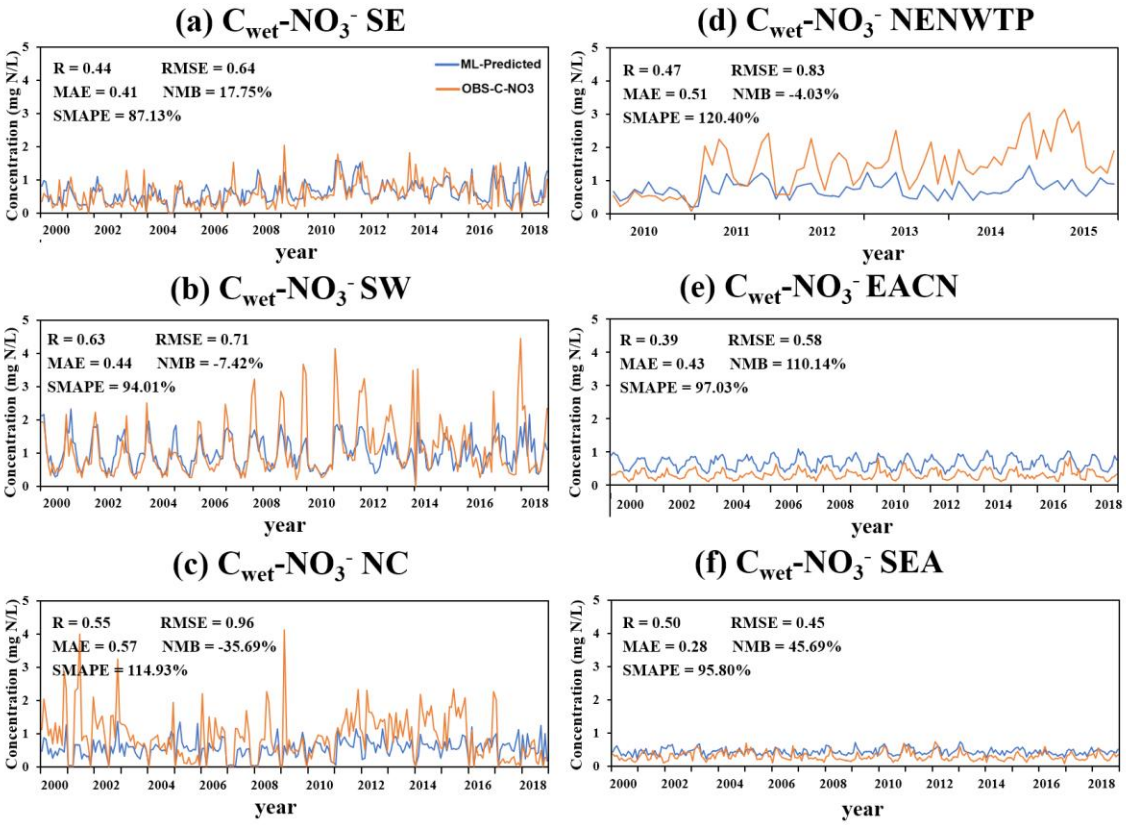

Figure S6. ML model performance on C_{wet}-NO_3^- under Case S2 scenarios for 2000-2018 (unit: mg N L^{-1}). The evaluation for the NENWTP region was conducted for 2010-2015 due to lack of observational data.

Figure S7

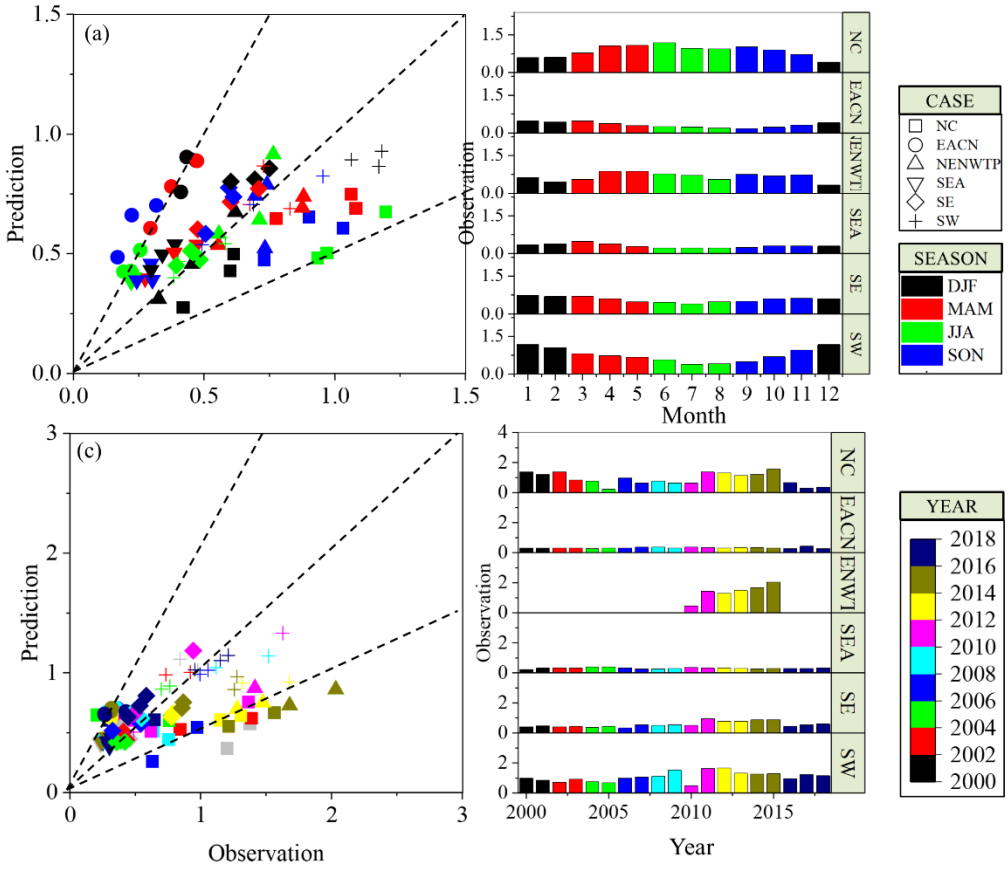

Figure S7. Comparison of ML prediction with observations under Case S2 scenarios for C_{wet}-NO_3^- (unit: mg N L^{-1}). (a-b) Scatter plots of model performance on seasonal variations (a) and inter-annual variations (b). (c-d) Distribution of observed values in four seasons (c) and during 2000-2018 (d).

Figure S8

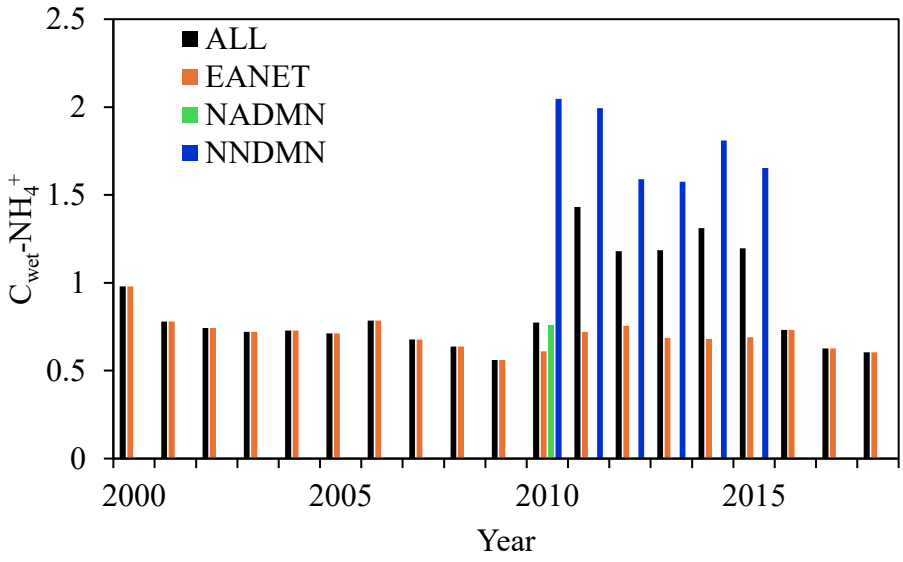

Figure S8. Observed C_{wet}-NH_4^+ values from different data sources from 2000-2018 (unit: mg N L^{-1}). The values are annual average of all sites.

Figure S9

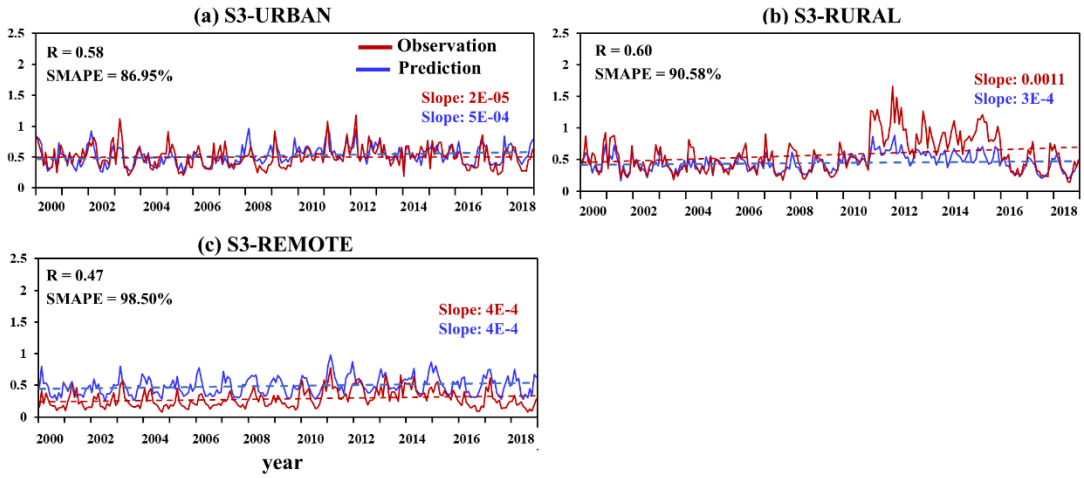

Figure S9. ML model performance on C_{wet}-NO_3^- under Case S3 scenarios for 2000-2018 (unit: mg N L^{-1}).

Figure S10

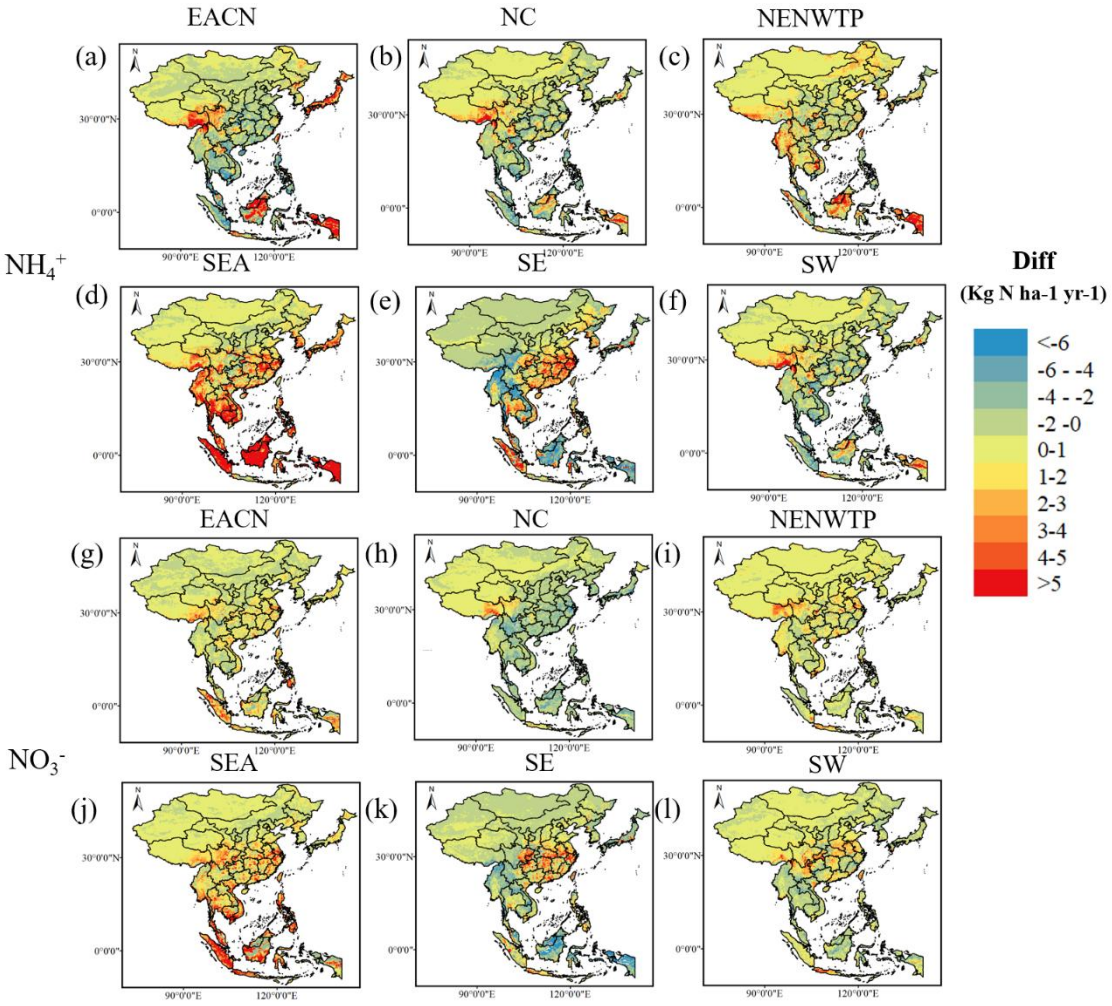

Figure S10. Spatial distribution of difference between Case2 scenarios and Base case for annual accumulated D_{wet}-NH_4^+ and D_{wet}-NO_3^- (unit: kg N hec^{-1} yr^{-1})

Figure S11

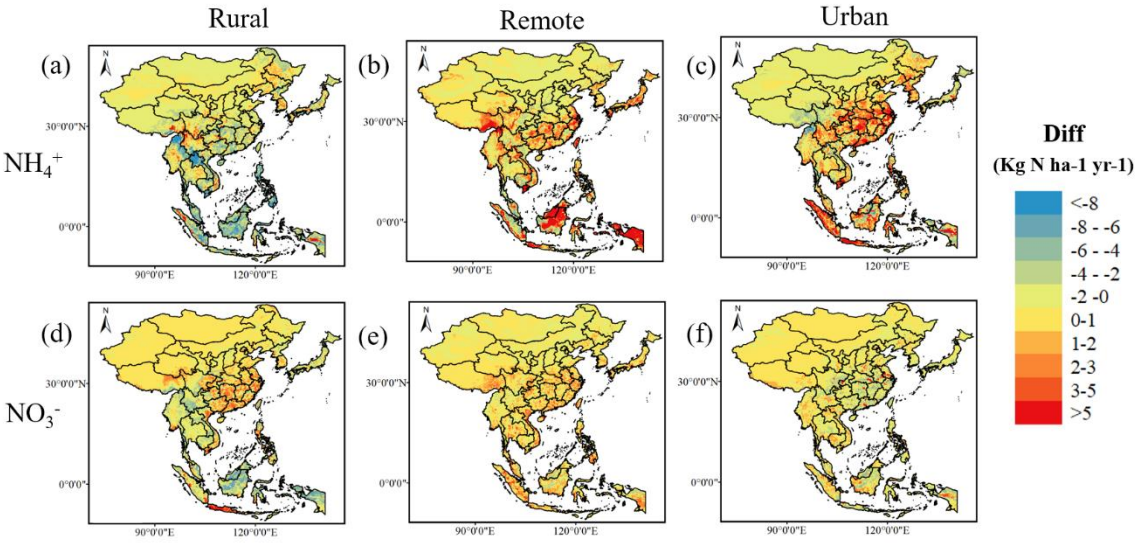

Figure S11. Spatial distribution of difference between Case3 scenarios and Base case for annual accumulated D_{wet}-NH_4^+ and D_{wet}-NO_3^- (unit: kg N hec^{-1} yr^{-1})

Figure S12

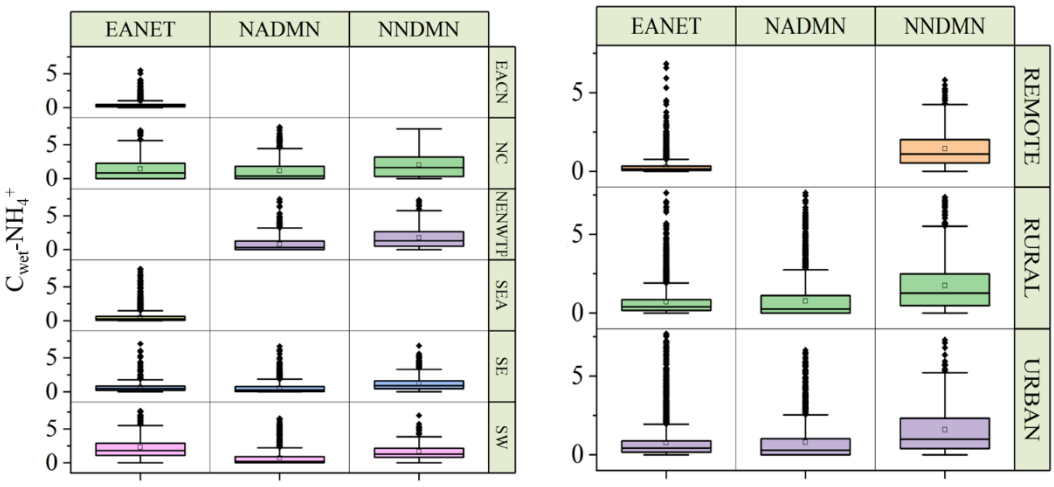

Figure S12. Ranges of C_{wet}-NH_4^+ reported by EANET, NADMN and NNDMN networks in six regions (left panel) and at three site types (right panel).

Figure S13

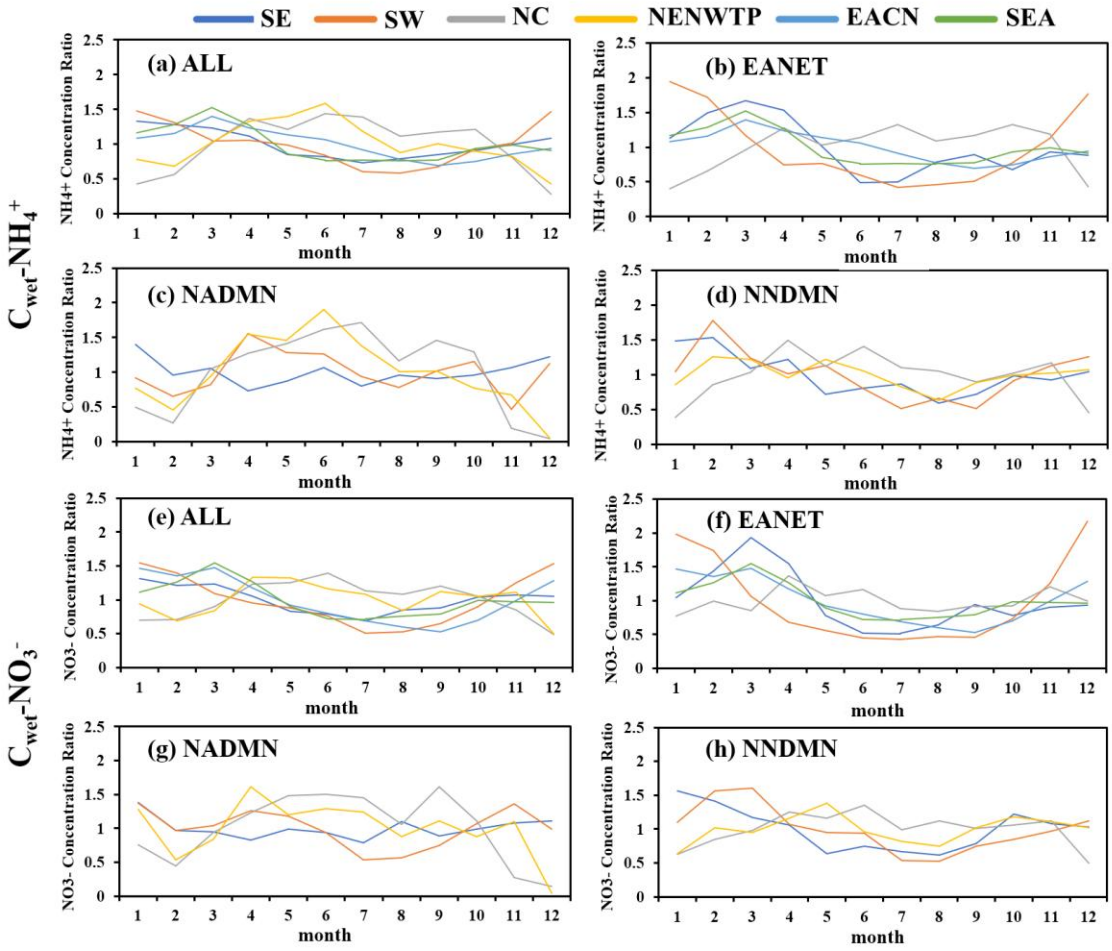

Figure S13. Monthly variations of C_{wet}-NH_4^+ and C_{wet}-NO_3^- in six regions from different observational datasets. The ratios are calculated by comparing monthly average values with annual averages.

Figure S14

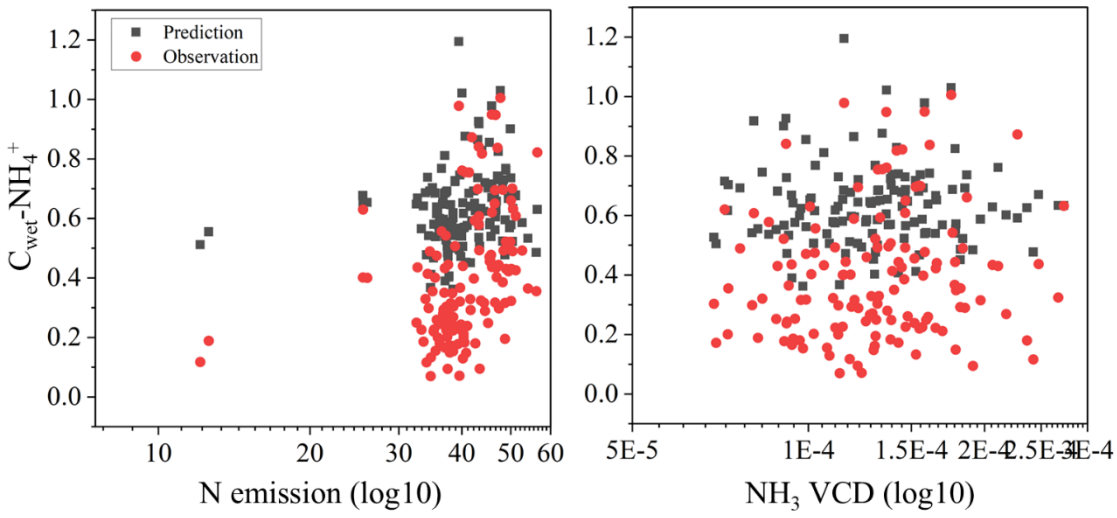

Figure S14. Observed and predicted relationship between C_{wet}-NH_4^+ and N emission (left) and NH_3 VCD (right) under Case S3-Remote.

Table S1

Comparison of model performance between using emission data and satellite data as features

Targets	Species	C_{wet} - NH_4^+		C_{wet} - NO_3^-	
	Features	Emission	Satellite[#]	Emission	Satellite
10-CV validation	R	0.76	0.75	0.72	0.77
	RMSE	0.7	0.7	0.45	0.44
	MAE	0.37	0.36	0.24	0.23
	NMB	0.12%	0.31%	-0.02%	-0.05%
Independent test	R	0.74	0.77	0.71	0.70
	RMSE	0.71	0.72	0.49	0.5
	MAE	0.38	0.37	0.27	0.26
	NMB	-0.61%	-1.01%	3.05%	-2.45%

Note: We used satellite derived NO_2 vertical column density (VCD) and NH_3 VCD to replace the NO_x emission and NH_3 emission used in developing the C_{wet} models in the Base case; while other features kept the same. The NO_2 VCD data came from OMI QA4ECV version 1.1 OFFLINE (2004-2021) (https://www.temis.nl/airpollution/no2.php, last access:2025/10/14). The NH_3 VCD data came from Metop-A NH_3 total column Level 3 data (2007-2021) (https://iasi.aeris-data.fr/, last access:2025/10/14).

Table S2

Hyperparameters used in the Base case

Hyperparameter	Grid search ranges	Optimal options
n_estimators	100,2000,num=10	1000
gamma	1,110,num=11	1
learning-rate	0,0.05,0.1,0.2,0.5,1	0.1
max_depth	1,100,num=10	80

Table S3

Comparison of ML performance between using concentration (mg N L^{-1}) and deposition (kg N ha^{-1} $month^{-1}$) as target

Processes	Species	NH_4^+		NO_3^-	
	Targets	C_{wet}	D_{wer}	C_{wet}	D_{wet}
Training	R	0.97	0.99	0.94	0.98
	sMAPE	70.39%	60.72%	71.36%	61.77%
10-CV validation	R	0.73	0.78	0.72	0.78
	sMAPE	82.16%	80.53%	79.69%	76.99%
Independent test	R	0.74	0.79	0.71	0.78
	sMAPE	82.03%	80.81%	79.78%	75.37%

Table S4

Spatial density of observational sites in six regions.

Regions		Site Count	Land Area ($\times 10^4$ km^2)	Site density (sites/ million km^2)
Inside China	NC	97	90.0	1.1
	NENWTP	92	594.4	0.2
	SE	156	125.8	1.2
	SW	96	140.4	0.7
Outside China	EACN	23	215.9	0.1
	SEA	32	443.8	0.1

Note: Site density is calculated by dividing number of sites with land area)