

Response to Reviewer 2

Dear Reviewer,

We thank you for the detailed review. We have revised the manuscript and Supplementary Information to clarify measurement metadata, quality control, model development and evaluation, figure captions, interpretation, and the data product. Below we respond to each comment and identify the corresponding locations in the current documents.

Detailed Point-by-Point Response

Comment 1: "The introduction lacks systematic comparison with existing ground-based NH₃ networks over the North China Plain. Please clarify how your 10-site dense county network complements previous regional sparse monitoring networks. Readers cannot easily tell the novelty boundary between your dataset and published in-situ datasets."

[Response]: We agree that the contribution of the county network needs to be clear. The revised Introduction compares it with the six-site NCP network of [Shen et al. \(2011\)](#), the 43-site Nationwide Nitrogen Deposition Monitoring Network, the 53-site AMoN-China campaign, and high-frequency observations at Gucheng and in Beijing.

The Introduction distinguishes the information available from these networks. [Shen et al. \(2011\)](#) included one site in Quzhou, and monthly integrated observations cannot resolve diurnal variability. Our record includes hourly data from 10 sites over four years, allowing comparisons among nearby agricultural environments and analyses from diurnal to interannual scales.

The comparison is included in the second paragraph of the Introduction.

Revised location: Manuscript, Introduction, second paragraph (p. 2, lines 41–61). The revised text states: "*However, [Shen et al. \(2011\)](#) included only one site in Quzhou, limiting the characterization of spatial differences among agricultural environments within the county.*"

Comment 2: "When discussing satellite NH₃ products (IASI, CrIS), the manuscript only mentions their footprint and overpass limitations. It is better to briefly summarise typical bias characteristics of these satellite retrievals over intensive agricultural regions, to further emphasise why high-quality ground reference data like this study is urgently required."

[Response]: We agree and have added representative IASI and CrIS validation findings to the Introduction alongside the limitations associated with footprints and overpass times.

An evaluation of an earlier IASI NH₃ product against ground-based FTIR columns reported a mean relative difference of -32.4% across 547 comparisons ([Dammers et al., 2016](#)). CrIS validation found that near-surface retrievals tended to overestimate low concentrations and underestimate high concentrations ([Dammers et al., 2017](#)).

These examples motivate the use of independent ground observations to assess local spatial representativeness and temporal sampling. They are presented as findings from the cited validation studies, rather than as bias estimates specific to Quzhou.

Revised location: Manuscript, Introduction, third paragraph (p. 3, lines 62–76). The revised text states: “*CrIS validation found that near-surface retrievals tended to overestimate low concentrations and underestimate high concentrations (Dammers et al., 2017).*”

Comment 3: "Section 2.1 mentions monitoring sites but provides very limited details of NH₃ analysers: instrument model, measurement principle, detection limit, calibration frequency, filter maintenance schedule and uncertainty range of raw hourly NH₃ observations are missing. These critical metadata are essential for a measurement-report-style paper."

[Response]: We agree that the measurement metadata need to be reported. Section 2.1 and **Table S1** now describe the analyzer, measurement principle, averaging and logging intervals, manufacturer specifications, and scheduled calibration and maintenance. Section 2.2 describes the NH₃ quality-control and hourly-aggregation rules.

NH₃ was measured with Thermo Scientific Model 17i analyzers at 2 m. The analyzer uses chemiluminescence and calculates NH₃ from the difference between the total nitrogen channel (N_t) and the NO_x channel. Each logged value was a 120 s moving average updated every second, with values recorded at 1 min intervals.

The manufacturer specifies a detection limit of 1 ppb at a 120 s averaging time. **Table S1** also reports precision of ±0.4 ppb on the 500 ppb range, zero noise of 0.5 ppb RMS at 120 s averaging, zero drift <1 ppb per 24 h, span drift of ±1% of full scale per 24 h, and linearity of ±1% of full scale. The 0 ppb screening lower limit used in QA/QC is reported separately from the manufacturer detection limit. Hourly means required at least 36 valid 1-min records (60% coverage).

Calibration was scheduled monthly. **Table S1** gives the engineer-supplied schedule for monthly calibration, filter replacement, inlet cleaning, and desiccant replacement, quarterly internal cleaning, linearity checks, and flow calibration, and annual reaction-chamber cleaning, O-ring and consumable replacement, and leak checks. The table states that this schedule is not a complete record of actual maintenance at each station.

A station-specific empirical uncertainty range for the complete four-year hourly record is not available. We therefore report the manufacturer specifications separately. **Table S1** notes that a field uncertainty assessment should include calibration-gas traceability, conversion efficiency, inlet losses, and drift between maintenance events.

Revised location: Manuscript, Section 2.1, second paragraph (pp. 4–5, lines 106–117), and Section 2.2, first paragraph (pp. 5–6, lines 126–150). SI, **Table S1 (p. 7)**. The revised text states: “*The manufacturer specifies a detection limit of 1 ppb at an averaging time of 120 s (Thermo Fisher Scientific, 2012, 2014).*”

Comment 4: "The study uses ERA5-Land precipitation to fill rain-gauge gaps. Although it gives statistical metrics ($R^2 = 0.7$, bias and RMSE), it does not discuss potential impacts of ERA5-Land precipitation uncertainty on subsequent machine-learning imputation for NH_3 concentrations."

[Response]: We agree that precipitation uncertainty should be discussed. All precipitation values used in this study were obtained from ERA5-Land, as stated in Section 2.2. **Fig. S1** provides an independent comparison with 327040 reference station-hours, yielding $R^2 = 0.69$, RMSE = 0.12 mm, and MAE = 0.02 mm.

Section 3.4 states that ERA5-Land may not fully represent local rainfall and that this can introduce uncertainty into NH_3 predictions. It also states that the low SHAP importance of precipitation does not quantify that uncertainty. The comparison in **Fig. S1** evaluates precipitation agreement. It does not directly quantify uncertainty in the reconstructed NH_3 series.

Revised location: Manuscript, Section 2.2, first paragraph (pp. 5–6, lines 126–150), and Section 3.4, third paragraph (p. 17, lines 396–398). SI, **Fig. S1 (p. 2)**. The revised text states: "*ERA5-Land precipitation may not fully represent local rainfall, introducing uncertainty into NH_3 predictions.*"

Comment 5: "Machine-learning input features include station ID as a categorical spatial variable. Please explicitly explain how you encoded the station ID for Random Forest/XGBoost/LightGBM models. Is one-hot encoding applied? Different encoding strategies may affect model performance."

[Response]: We agree and have clarified the encoding in Section 2.2. All three models use one-hot encoding for station categories during training and gap filling within the monitoring network. Section 2.4 explains that the signed SHAP values of the 10 station indicators are summed for each sample to obtain one Site contribution.

Revised location: Manuscript, Section 2.2, last paragraph (p. 6, lines 156–164), and Section 2.4, last paragraph (p. 9, lines 219–231). The revised text states: "*All three models used one-hot encoding for station categories.*"

Comment 6: "Model comparison focuses on test-set statistical metrics and extreme-value residuals. However, there is no cross-site evaluation: the authors did not perform site-held-out cross-validation (leave-one-site-out test). This cannot fully verify whether XGBoost can generalise to unseen sites within this county, which weakens the robustness of model selection conclusion."

[Response]: We agree that the random train/test split does not assess transfer to an unobserved station. We added leave-one-site-out validation for all three models and report station-specific and pooled results in **Table S5**. Station indicator variables were omitted in this validation, as stated in the table note.

The pooled RMSEs were 28.21 ppb for Random Forest, 29.46 ppb for XGBoost, and 29.59 ppb for LightGBM. Random Forest had the lowest pooled RMSE, and errors varied among stations.

XGBoost was selected for gap filling at the existing stations because it gave the lowest overall errors and the lowest MAE for observations above 60 ppb under the common random split. Section 3.1 distinguishes this application from transfer to unobserved sites. Section 2.3 also states that the random split does not test temporal extrapolation.

Revised location: Manuscript, Section 2.3, last paragraph (p. 8, lines 205–210), and Section 3.1, third paragraph (p. 10, lines 250–259). SI, **Table S5 (p. 12)**. The revised text states: “*To assess transfer to unobserved stations, we also conducted leave-one-site-out validation, fitting each model to measured NH₃ at the other nine stations and reporting sample counts, RMSE, MAE, and bias (Table S5).*”

Comment 7: “The manuscript does not report hyper-parameter tuning details: search range, grid-search space, final optimal hyper-parameters for three tree-based models. For a measurement report providing public dataset, these model configuration details should be supplied either in main text or supplementary materials.”

[Response]: We agree and have documented the preliminary parameter searches and final configurations in Section 2.3 and **Table S4**. The current models use one-hot station indicators while retaining the parameters and iteration counts selected in the earlier analysis with numeric station labels. No new hyperparameter search was performed after the encoding change.

Random Forest compared nine combinations of predictor fractions (0.5, 0.75, and 1) and minimum leaf sizes (4, 8, and 16), using 300 trees per candidate on an 80:20 inner split. The final model used 1000 trees, a predictor fraction of 0.75, a minimum leaf size of 4, and unrestricted depth. XGBoost compared four prespecified configurations on an 80:20 inner split, with early stopping followed by refitting. The selected configuration used a maximum depth of 12, a learning rate of 0.03, and 3277 boosting rounds. LightGBM compared 20 randomly sampled configurations using fivefold cross-validation and used 5000 boosting rounds in the final model. **Table S4** gives the search values and final settings. The outer test set was excluded from the searches.

Revised location: Manuscript, Section 2.3, parameter-selection paragraph (pp. 7–8, lines 185–204). SI, **Table S4 (pp. 10–11)**. The revised text states: “*Preliminary search settings and final parameter values are listed in Table S4.*”

Comment 8: “**Figure 2:** Panel (a-f) compares model performance of training and test datasets. The colour bar uses log-scale point density, but no description about point count distribution. I suggest adding sample numbers for training/test subsets in figure caption. Besides, high-NH₃ samples (>60 ppb) are critical for agricultural hotspot research; please mark the number count of these high-concentration samples in caption or supplementary table.”

[Response]: We agree and have added the sample counts to the **Fig. 2** caption and **Table S6**. The model-development sample contains 274218 station-hours with non-missing NH₃ after QA/QC. The random split uses 205662 training samples and 68556 test samples. The full record contains 350640 station-hours, including 76422 hours with missing NH₃. The caption also explains that point color

represents the number of samples in each cell of a 100×100 observed–predicted grid on a logarithmic scale.

Observed NH_3 exceeded 60 ppb in 20804 training samples and 6938 test samples (10.12% of each subset). **Table S6** reports the corresponding test-set MAEs of 26.43 ppb for Random Forest, 23.38 ppb for LightGBM, and 21.9 ppb for XGBoost.

Revised location: Manuscript, **Fig. 2** caption (p. 11, lines 262–270) and Section 3.1, second paragraph (p. 9, lines 242–249). SI, **Tables S3** and **S6** (pp. 9 and 13). The revised text states: “*The top row (a–c) shows training results ($n = 205662$), and the bottom row (d–f) shows test results ($n = 68556$) under the common random split.*”

Comment 9: “**Figure 3:** Radial plot visualises inter-annual station mean values. This graphic style is not conventional in atmospheric-science publications; many readers may struggle to interpret the orange segment representing excess concentration above county mean. Consider supplementing a conventional grouped bar chart in supplementary materials for easier reading.”

[Response]: We agree that a conventional chart helps readers compare the station means. **Fig. S3** provides a grouped bar chart with the same annual means as **Fig. 3** for the 10 stations and the county monitoring network.

Revised location: Manuscript, Section 3.2, first paragraph (pp. 11–12, lines 272–287). SI, **Fig. S3** caption (p. 4). The revised text states: “*Annual mean NH_3 concentrations at the 10 stations and across the county monitoring network.*”

Comment 10: “**Figure 4:** Monthly time series for ten sites. The four monitoring periods overlap in one subplot without distinct marker styles; some lines are visually tangled. Please optimise line colour/marker differentiation or simplify visualisation. In addition, the caption should define the exact time windows of four monitoring periods concisely.”

[Response]: We have revised **Fig. 4** to distinguish the four monitoring years with different colors and markers. The caption defines each year as 13 May to 12 May of the following year, explains how the two partial May periods are combined, and identifies the seasons separated by the vertical dashed lines.

Revised location: Manuscript, **Fig. 4** caption (p. 14, lines 321–330). The revised text states: “*Monthly means are calculated from the complete hourly series. Because each monitoring year runs from 13 May to 12 May, the May value combines the partial May periods at the beginning and end of that monitoring year. Vertical dashed lines separate the seasonal groups (spring: March–May, summer: June–August, autumn: September–November, winter: December–February).*”

Comment 11: “Section 3.2 observes a county-scale trend: NH_3 rose first then decreased, and authors link the decline to fertiliser reduction and livestock manure management policies. However, the dataset cannot provide direct causal proof. Please moderate your statement and explicitly

acknowledge other confounding factors (inter-annual meteorological variability) that may also contribute to concentration trends."

[Response]: We agree and have moderated the causal interpretation in Section 3.2. The observed decline is no longer attributed directly to fertilizer-reduction or manure-management policies.

The revised text acknowledges temperature, precipitation, humidity, wind, regional transport, and gas–particle partitioning as possible contributors. It states that the four-year record alone cannot separate management effects from interannual meteorological variability.

Revised location: Manuscript, Section 3.2, second paragraph (p. 12, lines 288–299). The revised text states: *"The four-year record alone cannot separate management effects from interannual meteorological variability."*

Comment 12: "North-south spatial gradient is attributed to cropland versus livestock sources. Yet the manuscript lacks quantitative source apportionment (e.g. correlation between site-level NH₃ and local livestock density / fertiliser application statistics). Supporting statistical analysis would strengthen the argument of source-driven spatial heterogeneity."

[Response]: We agree that the source interpretation needs supporting evidence and a clear limit. We have not conducted a new site-level correlation analysis or quantitative source apportionment for this network. The revised interpretation instead uses published evidence and acknowledges other controls on ambient concentrations.

Section 3.2 compares our concentration gradient with the independent Quzhou emission inventory of [Kang et al. \(2023a\)](#). The inventory estimated 6559.7 t of NH₃ emissions in 2019, with livestock production contributing 61% and fertilizer application 28%, and showed higher emissions in the south with Nanliyue and Baizhai as hotspots. Section 3.3 also cites published isotope observations in Quzhou ([Feng et al., 2022](#)).

The spatial agreement supports the interpretation of the concentration gradient, but does not establish a site-specific source contribution for the current monitoring period. The text explicitly acknowledges meteorology, atmospheric chemistry, and transport.

Revised location: Manuscript, Section 3.2, first paragraph (pp. 11–12, lines 272–287). The revised text states: *"These differences are consistent with the spatial heterogeneity of agricultural emissions, although concentrations are also affected by meteorology, atmospheric chemistry, and transport."*

Comment 13: "Section 3.3 reports bimodal seasonal peaks (March and June) linked to fertilisation events. Southern livestock-dominated sites maintain high NH₃ in autumn-winter. The discussion could go deeper: compare your seasonal pattern with previous published NH₃ observations from other North-China agricultural counties, to discuss similarities and discrepancies."

[Response]: We agree and have expanded the comparison with published agricultural observations in Section 3.3.

The summer enhancement in Quzhou is compared with the six-site NCP observations of [Shen et al. \(2011\)](#), including Wuqiao, Huimin, and Shouguang, and the Gucheng observations of [Meng et al. \(2018\)](#), where elevated NH_3 followed urea application in July. The discussion explains that local farming schedules and weather may shift the timing of concentration peaks.

Revised location: Manuscript, Section 3.3, seasonal-comparison paragraph (p. 15, lines 339–348). The revised text states: “*The March peak in Quzhou and relatively high winter concentrations at NLY complement existing regional observations and show differences in NH_3 seasonality among agricultural counties.*”

Comment 14: "Diurnal cycle analysis mentions morning peak around 08:00 local time. Please note: local standard time (Beijing time) vs solar time difference in this region should be mentioned; boundary-layer height observational evidence is absent when explaining diurnal variation mechanism."

[Response]: We agree with both points. The **Fig. S5** caption identifies China Standard Time (CST, UTC+8) and uses the longitudes in **Table S3** to explain that CST is approximately 19–20 min ahead of local mean solar time. Thus, 08:00 CST corresponds to about 07:40–07:41 local mean solar time.

Boundary-layer height and emission-source contributions were not measured in this network. Section 3.3 therefore presents nighttime accumulation, dew evaporation, and boundary-layer development as plausible contributors supported by previous studies, rather than directly demonstrated causes of the Quzhou morning peak.

Revised location: Manuscript, Section 3.3, last paragraph (p. 15, lines 349–360). SI, **Fig. S5** caption (p. 6) and **Table S3 (p. 9)**. The revised text states: “*The horizontal axis uses China Standard Time (CST, UTC+8). Based on the station longitudes in Table S3, CST is approximately 19–20 min ahead of local mean solar time, so 08:00 CST corresponds to about 07:40–07:41 local mean solar time.*”

Comment 15: "SHAP analysis shows spatial-temporal features contribute up to 59.3% of total importance, much higher than meteorological variables. The discussion should elaborate what this implies: NH_3 variability in this county is dominated by emission source timing and location rather than meteorology alone. This key conclusion is not sufficiently highlighted."

[Response]: We agree that the relative importance of the spatiotemporal predictors needs interpretation. They account for 59.44% of mean absolute SHAP importance in the current model. Section 3.4 explains how day of year, month, hour, station identity, and year can represent temporal activities and differences among local environments.

This percentage describes the predictors' contribution to model predictions. It does not quantify the fraction of ambient concentration variability explained or the contribution of particular emission sources.

Revised location: Manuscript, Section 3.4, first paragraph (pp. 15–16, lines 362–373). The revised text states: *“It is neither the proportion of ambient concentration variability explained nor a quantitative apportionment of emission sources.”*

Comment 16: "The nonlinear meteorological thresholds (vapour pressure, temperature, wind speed) are derived from gap-filled dataset. Please add a short discussion on the uncertainty of these threshold values induced by machine-learning gap-filling."

[Response]: Thank you for raising this point. The XGBoost model was trained on measured NH_3 , and the SHAP analysis used test samples with measured NH_3 . No model-filled NH_3 target values were used in either step.

Some meteorological predictors required gap filling before model development, as described in Section 2.2. Missing station observations were replaced by the mean of valid observations of the same variable at other stations in the same hour. This required at least one other station with a valid observation. All precipitation inputs came from ERA5-Land. These inputs may differ from local conditions at an individual station and may therefore affect the fitted SHAP responses.

In Section 3.4, we interpret the transitions near 1.02 kPa vapor pressure, 13.78 °C air temperature, and 1.11 m s⁻¹ wind speed as approximate model response points rather than fixed physical thresholds. Their locations may vary with model fitting, sampling, correlations among predictors, and uncertainty in the meteorological inputs. We have not separately quantified how meteorological imputation uncertainty propagates to these response points, so we do not report numerical uncertainty intervals.

Revised location: Manuscript, Section 2.2, first paragraph (pp. 5–6, lines 126–150), and Section 3.4, first and third paragraphs (pp. 15–16, lines 362–373, and p. 17, lines 396–398). The revised text states: *“The XGBoost model was trained on measured NH_3 , and SHAP analysis used test samples with measured NH_3 .”*

Comment 17: "In Conclusions section, limitations are briefly mentioned (only valid for Quzhou County, uncertainty for long data gaps). Please expand this part: discuss representativeness limitation when extrapolating these thresholds and spatial-temporal features to other counties of the North China Plain."

[Response]: We agree and have expanded the limitations in Section 4. The dataset represents the environments sampled within Quzhou County. Other NCP counties may differ in cropping systems, fertilization schedules, livestock systems, soils, irrigation, background pollution, and transport conditions.

The Conclusions state that transferring the observed patterns or fitted responses requires independent local observations and validation. They also explain that the monitoring framework can be extended to other agricultural counties.

Revised location: Manuscript, Section 4, last paragraph (p. 19, lines 445–454). The revised text states: “Therefore, transferring the observed patterns or fitted responses requires independent local observations and validation.”

Comment 18: "The dataset is shared via Figshare repository. Please briefly note what quality-control levels the published dataset contains: raw observational data only, or the final gap-filled hourly product? Users need clear guidance on dataset usage."

[Response]: We have clarified the released data product in the Data availability section. The four annual hourly files contain station coordinates, meteorological predictors, and separate observed and complete NH₃ series. Across 350640 station-hours, NH₃_Observed contains 274218 retained hourly measurements and NA for 76422 missing hours. These missing records include original gaps and observations treated as missing during QA/QC, as described in **Table S3**. NH₃_Imputed preserves the retained measurements and fills missing hours with XGBoost estimates. NH₃_Source distinguishes observed and model-filled values. The files provide quality-controlled hourly observations and the complete series after gap filling.

Revised location: Manuscript, Data availability (p. 19, lines 458–467). SI, **Table S3** note (p. 9). The revised text states: “Across 350640 station-hours, NH₃_Observed contains 274218 retained hourly measurements and NA for 76422 missing hours. NH₃_Imputed preserves those measurements and fills the missing hours with XGBoost estimates. NH₃_Source distinguishes observed values from model-filled values.”

Summary of Revisions

Section	Revision
Introduction (Section 1)	Compared the county network with existing ground observations and reported representative IASI and CrIS validation findings.
Methods (Sections 2.1-2.3)	Reported analyzer specifications, averaging and logging intervals, screening and hourly-coverage rules, maintenance information, station encoding, parameter searches, and leave-one-site-out validation.
Model evaluation (Section 3.1)	Reported the common training/test sample counts, counts above 60 ppb, and validation errors, and explained the scope of model selection.
Results and discussion (Sections 3.2-3.4)	Moderated causal and source-attribution claims, compared published observations, interpreted SHAP importance, and discussed ERA5-Land uncertainty.
Figures and Supplement	Added grouped-bar Fig. S3 , clarified the Fig. 4 periods and Fig. S5 time basis, and reported model settings and validation results in Tables S4–S6 .
Conclusions and Data availability	Expanded the limits of transfer beyond Quzhou and described the observed and complete hourly series in the released data product.

Concluding Remarks

We thank the reviewer for the suggestions. The responses above identify the changes to measurement metadata, model development and evaluation, figure captions, interpretation, and data availability.

The revised manuscript distinguishes measured observations from model estimates and states the limits of statistical, causal, and spatial interpretation.