Response to reviewers' comments

Reviewer #1

This manuscript addresses an important issue: how input-data-induced dataset shift affects ML-based estimates of reactive nitrogen wet deposition. The topic is relevant and the case study is potentially useful. However, I recommend reconsideration after major revisions. The main concerns are:

Response: We appreciate the opportunity to revise our manuscript and are grateful for the insightful comments provided by reviewer#1. In the following, we have provided detailed responses to each of the reviewer's comments. **Our responses (in blue) and the corresponding edits in the manuscript (in red) are shown below.** All the page and line numbers mentioned below are referred to the revised manuscript.

(1) the contribution to GMD needs to be more clearly framed as an assessment framework/case study rather than a new modelling system;

Response:

We sincerely appreciate this critical comment that helps us reposition and sharpen the core novelty and contribution of this work for the GMD audience, whose scope centers on model development, evaluation, uncertainty and intercomparison frameworks. We fully acknowledge the reviewer's concern that the original manuscript over-emphasized the XGBoost predictive model itself, while downplaying our central innovation: a systematic quantitative uncertainty assessment framework for input-induced dataset shift in geospatial machine learning deposition modelling. We have implemented comprehensive revisions across the abstract, introduction, discussion & conclusion and all section headings to reframe the paper as a methodological assessment framework demonstrated via a reactive nitrogen wet deposition case study, not a novel standalone ML modelling system. The key revisions are detailed below:

(1) Revision in Abstract

We have fully revised the abstract to re-centre the narrative around our transferable input dataset shift assessment framework, rather than the XGBoost predictive model. We explicitly state that the unmodified XGBoost algorithm only serves as a fixed baseline tool to implement our sensitivity testing workflow, and clearly distinguish the secondary regional N_r wet deposition case results from the primary methodological contribution of this generalizable evaluation framework, fully aligning the manuscript framing with the journal's focus on model assessment methodology.

Lines 15-40:

Abstract: Machine learning (ML) has been extensively applied to studies on spatial distribution characteristics of atmospheric composition, but quantitative assessments of

uncertainties arising from input data properties are still lacking. We develop a generalizable quantitative assessment framework for input dataset shift in geospatial machine learning, and demonstrate its implementation via a case study of reactive nitrogen wet deposition (D_{wet}) across East and Southeast Asia. We adopt the standard Extreme Gradient Boosting (XGBoost) as a fixed baseline predictive tool to carry out three targeted sensitivity experiments under this framework with a systematical quantification of uncertainties from sample size, uneven spatial distribution and imbalanced site types. Key findings revealed that: (1) Sample size below 6,000–9,000 led to a maximum 12% accuracy loss, while exceeding this threshold provided marginal performance improvement. (2) Uneven spatial distribution of monitoring sites caused 9–51% deviations from baseline performance, with >50% variability in D_{wet} estimations in data-scarce regions (e.g., western China and SEA). (3) Imbalanced site types lead to insufficient representation of remote sites, resulting in 9–40% overall accuracy loss and a high risk of severe overestimation (100%) in remote areas. The bias was attributed to both data range shifts and altered feature-target relationships (e.g., NH_3 emission vs. D_{wet}-NH_4^+). Additionally, inconsistencies among multi-source datasets and limitations of ML structure further introduced uncertainties. This study quantified previously unaddressed input data-induced uncertainties in ML-based N_r deposition research, providing critical insights for reliable application of ML-derived data in N_r management. Beyond this regional D_{wet} case findings, the transferable uncertainty assessment workflow constitutes methodological advance for model assessment and uncertainty evaluation, which can be widely applied to other geospatial interpolation tasks limited by sparse monitoring data.

(2) Revision in Introduction

We have fully revised the introduction part. (1) We emphasized the major contribution of this study as a newly proposed quantifiable framework to isolate and rank uncertainties driven by sample size, spatial site distribution and imbalanced site types. (2) We deleted all sentences positioning the XGBoost model as the core innovation. We clarify that XGBoost is selected as a well-established ML algorithm solely to implement our framework, rather than presenting a newly modified modelling architecture. By these revisions, we clearly articulate the paper's two-tier GMD-relevant contribution: (1) A transferable quantitative assessment workflow to disentangle input-induced dataset shift in geospatial ML interpolation, fully aligned with GMD's focus on model evaluation and uncertainty methodology. (2) A regional case study of reactive nitrogen wet deposition that delivers actionable empirical thresholds and regional uncertainty hotspots for nitrogen modelling practitioners.

Lines 57-61: In addition, more attention has been paid to model optimization, for instance developing new model architectures (Zhao et al., 2022; Chen et al., 2023; Zhou et al., 2023), modifying existing models (Li et al., 2020a), and implementing ensemble techniques (Rui Li, 2021; Lu et al., 2020), while systematic quantitative evaluation frameworks targeting input-data-driven uncertainties remain absent.

Lines 85-87: To our knowledge, no systematic investigation has quantified these biases due to lack of unified quantitative assessment framework for these three types of input-induced bias simultaneously.

Lines 90-96: To fill this gap, this study proposed a quantifiable framework to isolate and rank uncertainties driven by sample size, spatial site distribution and imbalanced site types, and demonstrated its implementation via a case study of reactive nitrogen wet deposition (D_{wet}). Figure 1 presents the research framework. We employed the well-established extreme gradient boosting (XGBoost) as a baseline prediction tool within our newly proposed assessment framework to calculate regional D_{wet} across East Asia (EA) and Southeast Asia (Section 3.1).

Line 104-106: The regional D_{wet} findings serve as a case demonstrating the proposed assessment framework, which is broadly applicable to ML applications targeting analogous topics, particularly those involving data scarcity challenges and geospatial interpolation tasks.

(3) Revision in Methodology

We changed Sect. 2.2 heading from "Development of the ML model" to "Configuration of the baseline ML tool" to avoid wording that implies we propose a newly developed ML architecture. In addition, we made the following changes to clarify that the XGBoost model serve as the baseline across all experiments, and explicitly distinguishes it from our core contribution, the dataset-shift assessment framework.

Lines 150-151: The standard XGBoost algorithm is configured here as a prediction tool for predicting C_{wet} values.

Lines 227-230: Throughout all sensitivity cases, the XGBoost model remains fixed as a consistent baseline predictor. The primary outputs of these tests are quantitative uncertainty metrics and spatial bias estimates, which form the core results of our dataset-shift assessment framework.

(4) Revision in Results

Subsection headings within Section 3 were refined to clearly distinguish two layers of outputs: predictive performance metrics (Section 3.1–3.3) and spatially explicit deposition estimation biases (Section 3.4). The standard structure starting from the base case and progressing to sensitivity tests is fully retained. The revised wording avoids

overemphasis on machine learning model development and consistently frames the model only as a fixed benchmark for our uncertainty assessment framework.

Subheadings:

3.1 Prediction and performance under the base case

3.2 Influence of sample size on predictive performance

3.3 Sensitivity of predictive performance to input dataset shift

3.4 Impacts on spatial estimations of nitrogen deposition amounts

Revisions in manuscript:

Line 283-285: The prediction and performance under the base case serve as a fixed benchmark. All subsequent sensitivity analyses use this reference to quantify uncertainties triggered by variations in input datasets within our assessment framework.

Lines 451-453: Variations in predictive performance identified above further propagate and generate spatial biases in regional nitrogen deposition estimates. This section evaluates such spatially distributed estimation deviations under different input scenarios.

(5) Revision in Discussion and conclusion

We revised Section 4.1 by adding two introductory sentences at the very beginning to emphasize on our core methodological contribution.

Lines 500-503: This work establishes a quantitative assessment framework to systematically isolate, rank and visualize input-driven dataset-shift uncertainties for geospatial machine learning interpolation. This analytical workflow is generalizable for modelling tasks with sparse, uneven ground monitoring observations.

Lines 682-687: Beyond N_r deposition, the proposed assessment workflow can be readily adapted to other geospatial modelling applications including $PM_{2.5}$ and ozone spatial interpolation. For any environmental interpolation study relying on limited, unevenly distributed monitoring records, this framework offers a standardized route to test how sampling bias propagates into final spatial estimates. This transferability makes our approach a reusable evaluation tool for atmospheric modelling research.

(2) reproducibility is not yet sufficient, as the archived notebooks lack a complete executable workflow, environment information, and full figure-generation scripts;

Response: We have supplemented a README file that provides detailed descriptions of all contents within this reproducibility archive, including a sequential executable workflow, Python environment specifications, folder structure breakdown, and a complete index mapping each manuscript figure to its corresponding source files.

Recommended Execution Workflow

(1) Run Code/Baseline.ipynb to train the baseline model, generate baseline evaluation metrics, and export baseline Cwet gridded estimations for 2000, 2005, 2010, 2015 and 2020. All outputs are saved under Data/Outputs/Base.
(2) Execute Code/CaseS1-sample_size.ipynb for sensitivity case S1. This script loads baseline datasets from Data/Baseline, generates test sub-datasets saved in Data/CaseS1, runs predictions and exports statistical metrics into Data/Outputs/CaseS1.
(3) Execute Code/CaseS2-site_distribution.ipynb and Code/CaseS3-site_type.ipynb for sensitivity cases S2 and S3. Each notebook loads baseline input datasets, performs model predictions, and exports statistical results together with gridded Cwet CSV files into Data/Outputs/CaseS2 and Data/Outputs/CaseS3.
(4) After completing all model-related computations, run Python plotting scripts stored under Figures/code/. These scripts adopt relative file paths to load precomputed CSV results from Data/Outputs/ and reproduce Python-generated statistical figures. Spatial maps, fitted trend panels and composite charts are post-processed via ArcGIS 10.8.1, Origin 2021 and Microsoft Excel separately.
Note: Full source tables, geospatial files and software project files (.opj, .xlsx, .mxd) for ArcGIS/Origin/Excel figures are archived in corresponding subfolders under Figures. Custom visual adjustments (colour ramps, layout, annotations) were manually configured in third-party software and cannot be fully reproduced by Python code alone. Readers can reconstruct all manuscript figures using the archived raw data and project files.

We exported our unified Anaconda computing environment and provide both environment.yml (full conda environment specification) and requirements.txt (pip-locked package versions) at repository root for reproducibility of model training and sensitivity experiments.
1. Environment Setup
All notebooks and scripts are developed and tested under Python 3.13.7.
Key required packages: xgboost, scikit-learn, pandas, numpy, matplotlib, scipy.
Users can recreate the identical runtime environment using either the provided environment.yml or requirements.txt stored in the repository root directory.

We have supplemented a README file that provides detailed descriptions of all contents within this reproducibility archive.
2. Code Folder
This folder stores all modelling and Python plotting scripts for baseline simulation and three sensitivity experiments:

(1) Baseline.ipynb: Baseline XGBoost model construction, performance evaluation and full-domain gridded Cwet prediction.
(2) CaseS1-sample_size.ipynb: Sensitivity test for varying training sample volumes
(3) CaseS2-site_distribution.ipynb: Sensitivity test for biases caused by uneven spatial monitoring site distribution
(4) CaseS3-site_type.ipynb: Sensitivity test for biases caused by imbalanced urban/rural/remote site composition

3. Data Folder
All input training datasets and auto-generated output results are stored here:
(1) Baseline: Core training dataset for the baseline model
(2) CaseS1 / CaseS2 / CaseS3: Sub-datasets automatically generated by corresponding sensitivity test scripts
(3) Inputs: Static gridded feature inputs used to predict Cwet for years 2000, 2005, 2010, 2015 and 2020
(4) Outputs: Automatically generated CSV outputs including model performance metrics and gridded Cwet predictions for all simulation scenarios

4. Figures Folder
This directory archives all source materials required to reproduce all main-text and supplementary figures, categorized by visualization software:
(1) code: Python plotting scripts for statistical scatter and comparison figures
(2) origin: Origin 2021 project files (.opj) and supporting CSV data
(3) gis: ArcGIS geospatial resources, including region shapefiles, 0.25° grid reference, prediction CSV tables and ArcGIS project file (.mxd). Subfolders inside gis: /regions (study & sub-region shapefiles), /grids (0.25° modelling grid), /csv (Dwet prediction tables for spatial mapping)
(4) excel: Microsoft Excel tables and chart source files
(5) Fig_list.txt: Index table mapping every manuscript figure to its corresponding source path inside this directory

We further archive all source materials required to reconstruct every main-text and supplementary figure. A complete figure-to-source mapping table is provided as Fig_list.txt in the /Figures folder
Figures in Main Text
Fig.1 | Concept framework | created by PowerPoint
Fig.2 | created by ArcMap 10.8.1 | underlying data: /Figures/gis/

Fig.3 a-f | created by Python script: /Figures/code/ScatterPlot_Fig3_Figs5.ipynb
Fig.4a-b | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.4c-d | created by Microsoft Excel | source: /Figures/excel/Fig4.xlsx
Fig.5 | created by Microsoft Excel | source: /Figures/excel/Fig5.xlsx
Fig.6 | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.7a-c | created by Microsoft Excel | source: /Figures/excel/Fig7a-c.xlsx
Fig.7d-e | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.8 | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.9 | created by ArcMap 10.8.1 | underlying data: /Figures/gis/
Fig.10 | created by Microsoft Excel | source: /Figures/excel/Fig10.xlsx

Figures in SI
Fig.S1 | created by Microsoft Excel | source: /Figures/excel/Figs1.xlsx
Fig.S2a-b | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.S2c | created by Microsoft Excel | source: /Figures/excel/Figs2c.xlsx
Fig.S3 | created by Python script: /Figures/code/RandomSample_Figs3.py
Fig.S4 | created by Microsoft Excel | source: /Figures/excel/Figs4.xlsx
Fig.S4a-b and d-e | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.S5 | created by Python script: /Figures/code/ScatterPlot_Fig3_Figs5.ipynb
Fig.S6 | created by Microsoft Excel | source: /Figures/excel/Figs6.xlsx
Fig.S7 | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.S8 | created by Microsoft Excel | source: /Figures/excel/Figs8.xlsx
Fig.S9 | created by Microsoft Excel | source: /Figures/excel/Figs9.xlsx
Fig.S10 | created by ArcMap 10.8.1 | underlying data: /Figures/gis/
Fig.S11 | created by ArcMap 10.8.1 | underlying data: /Figures/gis/
Fig.S12 | created by Origin 2021 | source: /Figures/origin/Figs.opju
Fig.S13 | created by Microsoft Excel | source: /Figures/excel/Figs13.xlsx
Fig.S14 | created by Origin 2021 | source: /Figures/origin/Figs.opju

Note for ArcGIS-generated figures: The gis.mxd ArcGIS project file in /Figures/gis/ is provided for reference only. This project file is sensitive to ArcGIS software version and local absolute file-paths and may fail to open on other machines. All underlying geospatial datasets (reference grid, region shapefiles, model-prediction CSV tables) are fully archived, so spatial maps can be rebuilt from these raw datasets in ArcGIS or QGIS.

(3) the validation strategy based on random train/test splits and random 10-fold CV may overestimate generalization because the data have station-level, temporal, and seasonal dependence;

Response: We thank the reviewer for this critical methodological comment, and we fully agree with this concern. Conventional random train-test splitting and random 10-fold cross-validation generate information leakage across training and validation folds, as station monitoring data carry inherent spatial, interannual and seasonal autocorrelation, which inevitably produces optimistically biased predictive metrics.

To empirically demonstrate this overestimation risk, we rely on supporting results from our previously published work using the identical East & Southeast Asia dataset and consistent modelling framework (Tan et al., 2026). That study implemented temporal out-of-sample validation with independent 2000–2007 EANET observations (outside our 2008–2018 training window), plus cross-network validation against separate CERN and CFAND forest monitoring sites. As shown in Figs. 3 and 4 of Tan et al. (2026), model R values drop and RMSE rises when predicting unseen time periods and independent site groups, which directly proves random CV cannot reflect real-world extrapolation capability.

Reference: Tan, J., Zhang, Y., Liu, X., et al: Data driven analysis on changing patterns of reactive nitrogen deposition in East and Southeast Asia, Atmos. Environ., 374, 121962, https://doi.org/10.1016/j.atmosenv.2026.121962, 2026.

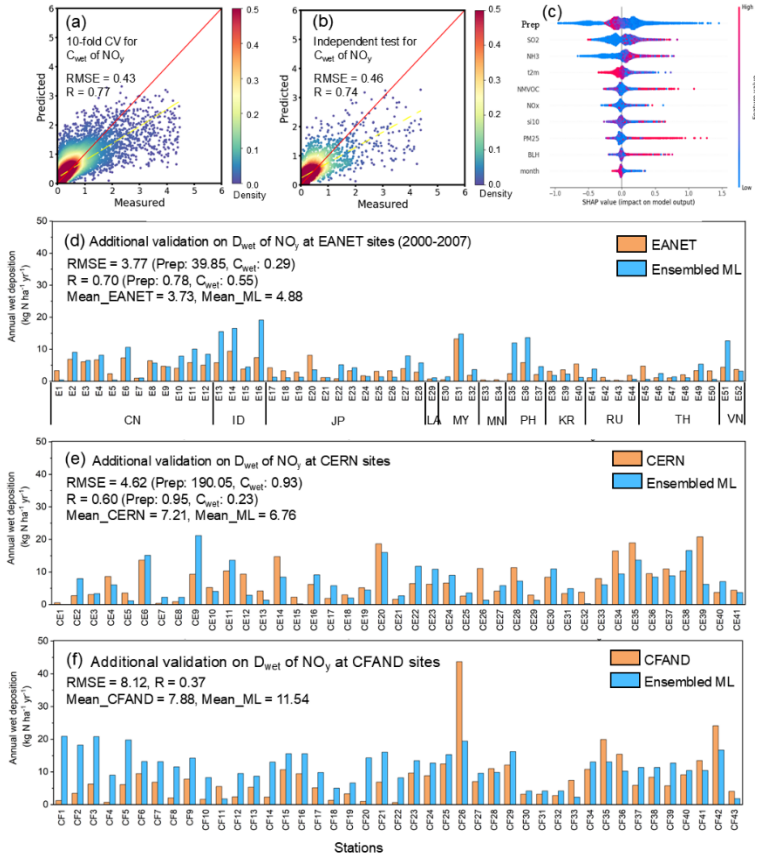

Figure 3 from Tan et al., 2026. Model performances on simulating wet deposition of NO_y. (a-b) Scatter plots for the observation and ensemble prediction of C_{wet}-NO_y (unit: mg(N) L^{-1}). (c) SHAP summary plot by XGB model. (d-f) Additional evaluation of model performance on D_{wet}-NOy (unit: kg N ha^{-1} yr^{-1}). Note for abbreviations: CN-China, KR-South Korea, JP-Japan, MN-Mongolia, ID-Indonesia, LA- Lao PDR, MY-Malaysia, PH-Philippines, TH-Thailand, VN-Vietnan, RU-Russia, CE-CERN sites, CF-CFAND sites. See locations of sites in Fig. S1 and information in Tables S1-S4.

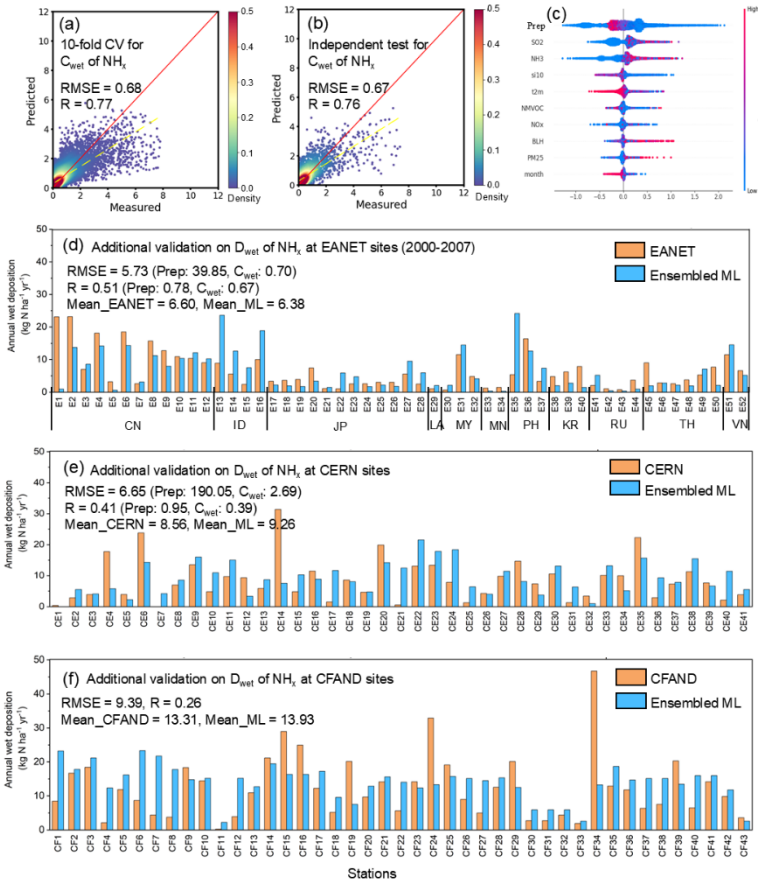

Figure 4 from Tan et al., 2026. Same as Fig. 3 but for D_{wet} of NH_x.

Motivated by the clear performance decay observed in temporal and cross-site extrapolation from our prior work, we designed three exclusive sensitivity tests within the current manuscript to disentangle the three core bias sources matching the reviewer's concern (station spatial distribution, uneven site types, limited sample volume): sample-size sensitivity analysis, spatial leave-one-region-out cross-validation, and stratified evaluation grouped by urban/rural/remote background sites. These experiments separately quantify how each data characteristic amplifies extrapolation error, especially in data-sparse Southeast Asian subdomains.

Taken together, the temporal extrapolation evidence from our published paper and the three novel quantitative sensitivity analyses in this revision jointly confirm that standard random cross-validation overestimates generalization due to station, temporal and seasonal data dependence. To our knowledge, this work is the first to simultaneously quantify all three categories of input-induced bias under a unified assessment workflow, filling a key methodological gap in geospatial nitrogen deposition modelling. In the revised manuscript, we supplemented the following information:

Lines 81-89 (Introduction): Conventional random train-test splitting and random k-fold cross-validation risk overoptimistic assessment of model performance (Tan et al., 2026), as monitoring observations exhibit strong spatial, seasonal and interannual dependence. Complementary sensitivity tests are hence conducted in this work to disentangle the drivers of extrapolation uncertainty for wet nitrogen deposition modelling. To our knowledge, no systematic investigation has quantified these biases due to lack of unified quantitative assessment framework for these three types of input-induced bias simultaneously. And lacking such uncertainty assessment may lead to overconfidence in data reliability (Zhu et al., 2025), potentially misinforming critical environmental policies and ecological risk assessments.

Lines 613-629: Standard random train-test splitting and random 10-fold cross-validation tend to overestimate model predictive skill, since monitoring observations carry inherent spatial, seasonal and interannual autocorrelation that creates information leakage between training and validation subsets. Our companion study using the identical East and Southeast Asian dataset (Tan et al., 2026) confirmed evident predictive decay during temporal and cross-network out-of-sample validation (Figs. 3–4 in Tan et al., 2026), solidifying this overoptimistic bias inherent to random splitting. Nevertheless, the primary drivers behind such extrapolation performance degradation remained unquantified in prior work.

Building upon those findings, this study designed three targeted sensitivity experiments to disentangle the core sources of extrapolation uncertainty (sample-size testing, spatial leave-one-region-out cross-validation, and stratified evaluation grouped by site category). We adopted a single-variable controlled experimental framework, where nuisance factors including data source and temporal coverage are held nearly constant to minimize extraneous confounding signals. This controlled setup enables us to separately attribute extrapolation biases to three independent dataset-shift drivers: inadequate sample volume, uneven spatial sampling distribution, and imbalanced urban/rural/remote site composition.

Reference: Tan, J., Zhang, Y., Liu, X., et al: Data driven analysis on changing patterns of reactive nitrogen deposition in East and Southeast Asia, Atmos. Environ., 374, 121962, https://doi.org/10.1016/j.atmosenv.2026.121962, 2026.

(4) the attribution of S2/S3 biases to dataset shift is partly confounded by data source, region, site type, and time coverage;

Response: We appreciate this thoughtful comment on confounding factors during bias attribution. To disentangle the independent impacts of sample volume, spatial

separation and site-type imbalance, all three sensitivity tests follow a single-variable controlled design that holds other interfering covariates constant:

-S2 spatial cross-validation fixes the composition of site types, observation sources and temporal coverage across training folds, only withholding an entire independent geographic region to isolate spatial extrapolation bias;

-S3 stratified validation is performed within the same regional domains and matching multi-year time windows, removing geographic and temporal confounding when quantifying bias induced by unbalanced urban/rural/remote monitoring records.

In the revised manuscript, we clarify this single-variable controlled strategy, which minimizes mutual interference and supports reasonable independent attribution of bias sources.

Lines 222-227 (method): All three sensitivity tests follow a single-variable controlled design. When isolating uncertainty arising from one target factor (sample volume, spatial distribution, or site-type composition), we hold all other variables nearly constant to minimize extraneous confounding signals, thereby supporting separate quantification of uncertainties induced by sample volume, spatial distribution and imbalanced site composition.

Lines 621-629: Building upon those findings, this study designed three targeted sensitivity experiments to disentangle the core sources of extrapolation uncertainty (sample-size testing, spatial leave-one-region-out cross-validation, and stratified evaluation grouped by site category). We adopted a single-variable controlled experimental framework, where nuisance factors including data source and temporal coverage are held nearly constant to minimize extraneous confounding signals. This controlled setup enables us to separately attribute extrapolation biases to three independent dataset-shift drivers: inadequate sample volume, uneven spatial sampling distribution, and imbalanced urban/rural/remote site composition.

(5) the 6000-9000 sample threshold and the Dwet uncertainty statements should be presented more cautiously.

Response: We have softened the wording for the 6000–9000 sample threshold and D_{wet} uncertainty ranges in Abstract, Section 4.1 and 4.3. All quantitative observations are now explicitly constrained to the EA and SEA monitoring dataset and sensitivity test configurations used in this study, avoiding overgeneralization of these numerical thresholds to other study domains or deposition modelling systems.

Lines 24-31 (Abstract): Key findings revealed that: (1) Within the dataset used for EA and SEA in this work, sample size below 6,000–9,000 led to a maximum 12% accuracy loss, while exceeding this threshold provided only marginal performance improvement.

(2) Uneven spatial distribution of monitoring sites caused 9–51% deviations from baseline performance, with >50% variability in D_{wet} estimations in extreme data-sparse test scenarios (e.g., western China and SEA). (3) Imbalanced site types lead to insufficient representation of remote sites, resulting in 9–40% overall accuracy loss and a high risk of severe overestimation (100%) under certain test configurations for remote areas.

Lines 512-518: In our Case S1 (section 3.2), within the observational dataset covering East and Southeast Asia adopted in this work, we observe a marked saturation trend in model predictive performance when the training sample volume reaches 6,000–9,000 records. Reducing sample size below this range yields up to a 12% drop in overall predictive accuracy. For our specific study domain and monitoring data composition, further increases in training volume beyond 9,000 samples produce only marginal performance gains for N_r wet deposition modelling.

Lines 526-528: Under these spatial extrapolation test scenarios, domain-averaged D_{wet} deviations span −13% to 22%, while data-sparse subregions (western China, EACN and SEA) exhibit local prediction deviations exceeding 50% under extreme out-of-sample conditions.

Lines 533-537: Relative deviations from the baseline model output fall between 9% and 40% across stratified site-type tests, translating to domain-wide D_{wet} estimation uncertainties of −14% to 10%. For remote areas with limited monitoring records in western China and SEA, extreme prediction biases exceeding 50% are observed when training data lack representative remote-site observations.

Line 630-631: Across our three core sensitivity test designs, modelled D_{wet} estimation biases range from 9% up to over 50% under extreme sampling constraints.

I suggest that the authors improve the reproducibility package, clarify the feature set and code-method consistency, add grouped or spatial/temporal validation tests, and narrow the conclusions accordingly.

Response: We have upgraded the reproducibility material with documented executable scripts and environment specifications, clearly documented the feature set to ensure code–manuscript consistency. Our spatial leave-one-region-out and site-stratified sensitivity experiments act as the grouped/spatial extrapolation validation tests recommended by the reviewer. Meanwhile, we have bounded all quantitative conclusions to our study domain and experimental setup to avoid overgeneralization.