Response to Reviewer #2 Comments

Thank you very much for your valuable comments on our manuscript. These comments are highly appreciated and have been very helpful in improving the quality of our manuscript and strengthening the overall research.

Major Comments

Comment 1: Validation might overestimates model skill

The 80/20 split is random, not spatial or temporal. Lightning density is strongly correlated in space and time. Neighboring cells and sequential days/months can therefore appear in both training and test sets. This inflates reported skill.

Fix: Add a spatial-block and/or temporal-block validation (e.g., withhold regions, or train/test on separate years).

Reply: We thank the reviewer for this important comment. We agree that the sample-wise random split may provide an optimistic estimate because spatially and temporally correlated samples can occur in both subsets. We therefore added a year-blocked validation in which 2015, 2019, and 2021 were withheld as complete test years and excluded from all stages of model development. Under this more conservative setting, the four base learners retained test R^2 values of 0.507–0.541 and Pearson correlations of 0.712–0.736, with broadly consistent performance across the three withheld years. The full methods and results have been added to Text. S2 and Tables S12–S13.

Text S2. *Year-blocked validation*

To evaluate temporal transferability while reducing the influence of temporal autocorrelation between the training and test samples, we conducted an additional year-blocked validation for the four base learners. The years 2015, 2019, and 2021 were withheld as complete temporal blocks, whereas 2013, 2014, 2016–2018, 2020, and 2022–2024 were used for model development. The same year assignment was applied to XGBoost, RF, LightGBM, and DNN. Samples from the withheld years were excluded from hyperparameter optimization, early stopping, feature and target standardization, and final model fitting. Model performance was evaluated using R^2, Pearson correlation, RMSE, and MAE.

All four models retained meaningful predictive skill when evaluated against the fully withheld years (Table S12). XGBoost achieved the highest test performance (R^2 = 0.5407, r = 0.7357), followed by RF (R^2 = 0.5203, r = 0.7228), LightGBM (R^2 = 0.5189, r = 0.7213), and DNN (R^2 = 0.5073, r = 0.7123). Test RMSE values ranged from 0.0132 to 0.0137, and MAE values ranged from 0.0038 to 0.0040. Performance was broadly consistent across 2015, 2019, and 2021 (Table S13), although DNN showed slightly lower skill in 2015. These results confirmed that the sample-wise random split provided more optimistic estimates, while the models still retained cross-year predictive capability under the more conservative validation setting.

Table S12. *Overall performance under year-blocked validation*

Model	*Test R^2*	*Test Pearson r*	*Test RMSE*	*Test MAE*
XGBoost	*0.5407*	*0.7357*	*0.01321*	*0.00378*
RF	*0.5203*	*0.7228*	*0.0135*	*0.00394*
LightGBM	*0.5189*	*0.7213*	*0.01352*	*0.00386*
DNN	*0.5073*	*0.7123*	*0.01369*	*0.00398*

Table S13. *Performance of the four base learners for each fully withheld test year.*

Model	*Year*	*R^2*	*Pearson r*	*RMSE*	*MAE*
XGBoost	*2015*	*0.5342*	*0.731*	*0.013*	*0.00377*
	2019	*0.5503*	*0.742*	*0.01335*	*0.00388*
	2021	*0.5369*	*0.7347*	*0.01329*	*0.0037*
RF	*2015*	*0.5166*	*0.7195*	*0.01324*	*0.00395*
	2019	*0.5309*	*0.7299*	*0.01364*	*0.00404*
	2021	*0.5129*	*0.7195*	*0.01363*	*0.00383*
LightGBM	*2015*	*0.5095*	*0.7141*	*0.01333*	*0.00385*
	2019	*0.5323*	*0.7302*	*0.01362*	*0.00395*
	2021	*0.5138*	*0.7197*	*0.01362*	*0.00377*
DNN	*2015*	*0.4797*	*0.6927*	*0.01373*	*0.00399*
	2019	*0.5175*	*0.7194*	*0.01383*	*0.00407*
	2021	*0.5228*	*0.7242*	*0.01349*	*0.00388*

A brief summary was added to Sects. 3.2, 4.1, and 5.3 at Lines 206-209, Lines 292-294, Lines 659-662 and Lines 686-689: "*We further performed a year-blocked validation by withholding 2015, 2019, and 2021 as complete test years, while the remaining nine years were used for model development. The withheld years were excluded from hyperparameter optimization, early stopping, preprocessing, and model training. Full details are provided in Text S2.*" "*Although performance decreased relative to the sample-wise random holdout, all four base learners retained positive and broadly consistent predictive skill across the fully withheld years (Tables S12–S13).*" "*Compared with the sample-wise random split, the year-blocked evaluation provided a more conservative estimate of predictive skill. Nevertheless, the broadly consistent performance across the fully withheld years indicated that the learned meteorology–lightning relationships remained transferable across unseen years within the observational period.*" and "*In addition, future efforts might include: (i) spatial-block or regime-split validation to further quantify sensitivity to spatial dependence and climate-background shifts; (ii) perturbation-based sensitivity tests for key predictors; and (iii) expanded uncertainty estimates based on ensemble spread or alternative reanalysis inputs.*".

Comment 2: Stacking procedure is underspecified

It is not ultimately clear whether the ridge regression stacking ensemble used out-of-fold base-model predictions, or in-sample predictions from models already fit on the same data. The latter would leak information and could favor the base learner that overfits most. Random Forest shows the largest train vs. test gap (R^2 = 0.82 vs. 0.65), which is consistent with this risk.

Fix: State clearly how stacking predictions were generated. Use out-of-fold predictions if not already done.

Reply: We thank the reviewer for highlighting this important methodological detail. We apologize that the out-of-fold procedure used in the stacking ensemble was not clearly described in the original manuscript. We have now clarified this procedure in Sect. 3.4 of the revised manuscript by adding the following text at Lines 238-243: "*To avoid information leakage in the stacking procedure, the ridge regression meta-learner was trained using five-fold out-of-fold predictions from the four base models rather than in-sample predictions. The same fold assignment was used for all base models, and each training sample was predicted only by a model that had not been*

fitted using that sample. After the stacking weights were determined, the base models were refitted using the complete training set and applied to the independent test set.".

Comment 3: Pre-2013 reconstruction is not validated

The core contribution, i.e., the 1979–2012 reconstruction, is pure extrapolation and it should be highlighted as such. All quantitative validation uses 2013–2024 data. Furthermore, the LIS/OTD comparison adds only qualitative support: it uses a different lightning metric (flash rate vs. stroke density) and only the 38°S–38°N band.

Fix: State these limitation prominently in the Abstract, Introduction, Results, and Conclusions, not only in the limitations section 5.3.

Reply: We thank the reviewer for highlighting the need to distinguish clearly between model evaluation during the WWLLN observational period and the reconstruction for the earlier period. We agree that the 1979–2012 component represents a model-based temporal extrapolation beyond the 2013–2024 WWLLN training and evaluation period. We also agree that the comparison with LIS/OTD should be interpreted as an independent consistency assessment rather than a direct quantitative validation, because LIS/OTD reports flash-rate density rather than stroke density and the comparison is restricted to 38°S–38°N. Accordingly, we have revised the Abstract, Introduction, Results, and Conclusions to clarify the extrapolative nature of the pre-2013 reconstruction and to define more precisely the scope of the LIS/OTD comparison.

Specifically, we revised the Abstract at Lines 30-37: "*The 1979–2012 portion represents a model-based temporal extrapolation beyond the 2013–2024 WWLLN observational period and is therefore evaluated through large-scale consistency checks rather than direct contemporaneous validation. Individual validations indicated spatial agreement, as the ensemble successfully reproduced the observed large-scale spatial distribution and major tropical–subtropical continental lightning hotspots. Comparative assessment with the LIS/OTD flash-rate climatology over 38° S–38° N provides an additional consistency check, showing agreement in broad spatial patterns and interannual variability.*".

In the Introduction, we added: "*Because contemporaneous global stroke-density observations are unavailable before 2013, the 1979–2012 portion represents a model-based temporal extrapolation of the relationships learned during the 2013–2024 WWLLN observational period. Accordingly, quantitative evaluation using the 2013–2024 data assesses model performance within the observational period but does not directly validate the pre-2013 reconstruction.*" at Lines 108-112.

We also revised Sect. 4.2 of the Results at Lines 350-352: "*Building on this overlap-period evaluation, we examined the reconstructed climatology for 1979–2012, which represents a model-based temporal extrapolation beyond the 2013–2024 WWLLN observational period (Fig. 5).*".

The caption of Fig. 7 was correspondingly revised to specify: "*Figure 7. Spatial distribution of Theil–Sen trends in reconstructed annual lightning density during 1979–2025. Hatched areas indicate regions where trends were significant at $p < 0.05$ based on the Mann–Kendall test. Theil–Sen slope values are expressed in strokes km^{-2} day^{-1} yr^{-1}.*".

In Sect. 4.3, we revised the role of the LIS/OTD comparison at Lines 452-453 and Lines 478-480 as follows: "*To provide an independent consistency assessment for our reconstruction, we selected the satellite-based LIS/OTD gridded lightning product as an independent reference.*" and "*This agreement provided independent support for the broad spatial patterns and interannual variability represented by the reconstruction over key tropical and subtropical convective regions, despite the difference between*

flash-rate and stroke-density metrics.".

And we revised the Conclusions at Lines 719-729: "*During the evaluation period (2013–2024), the multi-model stacking ensemble demonstrated robust predictive skill and reproduced the leading spatiotemporal structures constrained by WWLLN, providing the observational basis for constructing the long-term reconstruction. The 1979–2012 component represents a model-based temporal extrapolation of the relationships learned during the 2013–2024 observational period and is not directly validated by contemporaneous global stroke-density observations. The reconstructed long-term series revealed pronounced regional heterogeneity in lightning trends: significant decreases were concentrated over several tropical land convection centers, whereas increases were more localized and emerged over parts of midlatitude regions, such as the Mediterranean and the central United States. These results provided a practical basis for investigating reconstructed changes in lightning-related hazards and convective activity over the past nearly five decades and a historical reference for broad-scale comparisons with lightning simulations from climate models.*".

Comment 4: Reported error metrics do not have clear context

$R^2 \approx 0.69$, RMSE = 0.0108, MAE = 0.0030 look reasonable if they are considered without context. But lightning density is extremely skewed: most land cells have monthly density far below 0.005 strokes/km²/day. Global R^2 is therefore largely influenced by the contrast between high- and low-density regions, partly encoded directly via the latitude predictor (abs_lat). MAE of 0.0030 may be quite large relative to typical (non-hotspot as in many parts of Europe for example) values.

Fix: Report the target distribution, as the other reviewer also suggested (mean, median, percentiles, maximum). You could also stratify RMSE/MAE by stroke density regime, separating "hotspots" from "typical" cells.

Reply: We thank the reviewer for this important comment. We agree that the global R^2, RMSE, and MAE values should be interpreted in the context of the highly zero-inflated and right-skewed distribution of monthly lightning stroke density, rather than as uniformly representative of all density conditions. In response, we added two complementary analyses. First, we summarized the observed target distribution for the full 2013–2024 sample pool and separately for the training and test subsets, including the mean, median, selected percentiles, maximum, and the proportions of zero-density and low-density samples (Table S4). Second, we evaluated test-set RMSE and MAE separately across four observed lightning-density regimes for all four base models and the ridge regression ensemble (Table S5).

The added results show that the test-set distribution is strongly imbalanced: 64.98% of the samples have zero observed lightning density, and 82.93% have densities below 0.005 strokes km^{-2} day^{-1}. Regime-specific errors increase in absolute magnitude with observed lightning density, and the relative performance of the models varies among density regimes. The ridge regression ensemble yields the lowest RMSE and MAE in the zero-density and extreme-density regimes, whereas RF yields the lowest errors in the low-to-moderate- and high-density regimes. These additions provide the requested context for the global metrics and clarify that the values reported in Table 1 represent aggregate performance across density regimes with markedly different sample frequencies and error magnitudes.

We added the following sentence at Lines 221-223 to describe the additional distributional and regime-specific evaluations: "*To account for the strongly zero-inflated and right-skewed target*

distribution, we additionally summarized the observed lightning-density distribution and evaluated test-set RMSE and MAE across four observed lightning-density regimes (Supplementary Text S2 and Tables S4–S5).".

And we added the following concise description at Lines 301-305: "*The observed target distribution was strongly zero-inflated and right-skewed, with 64.98% of the test samples having zero lightning density and 82.93% having densities below 0.005 strokes km^{-2} day^{-1} (Table S4). Regime-specific evaluation showed that absolute errors increased with observed lightning density and that relative model performance varied among density regimes (Table S5 and Supplementary Text S2).*".

We added a new supplementary subsection S2 and Table S4-S5:

Text S2. *Target distribution and regime-specific model errors*

The observed monthly lightning stroke density exhibited a strongly zero-inflated and right-skewed distribution (Table S4). In the test set, the mean density was 0.005016 strokes km^{-2} day^{-1}, whereas the median was zero. Zero-density samples accounted for 64.98% of the test set, and 82.93% of the samples had densities below 0.005 strokes km^{-2} day^{-1}. The 90th, 95th, and 99th percentiles were 0.01250, 0.02629, and 0.08099 strokes km^{-2} day^{-1}, respectively, while the maximum reached 1.82587 strokes km^{-2} day^{-1}. The training and test subsets showed nearly identical distributional characteristics, indicating that the random partition preserved the overall target distribution.

Table S4. *Distribution statistics of the observed monthly lightning stroke density used for model training and evaluation during 2013–2024, with units of strokes km^{-2} day^{-1}. P75, P90, P95, and P99 denote the 75th, 90th, 95th, and 99th percentiles, respectively.*

Split	*N*	*Mean*	*Median*	*P75*	*P90*	*P95*	*P99*	*Maximum*
All	*34184448*	*0.0050246*	*0*	*0.0016025*	*0.0125046*	*0.0263342*	*0.0812396*	*1.8258656*
Train	*27347558*	*0.0050268*	*0*	*0.0016029*	*0.0125052*	*0.026346*	*0.0813*	*1.7960455*
Test	*6836890*	*0.005016*	*0*	*0.0016008*	*0.0125017*	*0.0262898*	*0.0809879*	*1.8258656*

To examine how this imbalanced distribution affected the reported error metrics, test-set RMSE and MAE were evaluated separately for zero-density, low-to-moderate-density, high-density, and extreme-density samples (Table S5). Absolute errors increased progressively with observed lightning density, and the ridge regression ensemble produced the lowest RMSE and MAE in the zero-density and extreme-density regimes. These results show that the global metrics reported in Table 1 represent aggregate performance across density regimes with markedly different sample frequencies and error magnitudes.

Table S5. *Test-set RMSE and MAE for the four base models and the ridge regression ensemble across observed lightning-density regimes. The regimes were defined using the observed test-set lightning density as follows: zero density, $y_{obs} = 0$ (N = 4,442,737, 64.98%); low-to-moderate density, $0 < y_{obs} \leqslant P_{90}$ (N = 1,710,464, 25.02%); high density, $P_{90} < y_{obs} \leqslant P_{99}$ (N = 615,320, 9.00%); and extreme density, $y_{obs} > P_{99}$ (N = 68,369, 1.00%). Here, P_{90} = 0.01250167 and P_{99} = 0.08098790 strokes km^{-2} day^{-1}. RMSE and MAE are reported in strokes km^{-2} day^{-1}.*

Observed lightning-density regime	***Model***	***RMSE***	***MAE***
Zero-density regime	*XGBoost*	*0.001358*	*0.000427*

	RF	*0.001152*	*0.000338*
	LightGBM	*0.001308*	*0.000420*
	DNN	*0.001428*	*0.000657*
	Ridge regression	*0.001122*	*0.000263*
Low-to-moderate-density regime	*XGBoost*	*0.007244*	*0.004413*
	RF	*0.006527*	*0.003943*
	LightGBM	*0.006861*	*0.004154*
	DNN	*0.007317*	*0.004331*
	Ridge regression	*0.006932*	*0.004184*
High-density regime	*XGBoost*	*0.019116*	*0.013975*
	RF	*0.016572*	*0.011961*
	LightGBM	*0.019050*	*0.013845*
	DNN	*0.020659*	*0.014723*
	Ridge regression	*0.018061*	*0.012960*
Extreme-density regime	*XGBoost*	*0.093235*	*0.068053*
	RF	*0.096808*	*0.067820*
	LightGBM	*0.086228*	*0.062358*
	DNN	*0.091242*	*0.067778*
	Ridge regression	*0.085786*	*0.060722*

Comment 5: High-density lightning is systematically underestimated, not just "noisy"

In Fig. 2, regression lines fall below the 1:1 line at high true values for all four models. The text attributes this to "stochasticity," but it is actually a consistent bias. This matters because wildfire ignition and lightning-NOx applications depend on exactly these extreme events (see your section 5.2).

Fix: Quantify the existing bias and at least discuss this explicitly as a real limitation for the stated applications and do not "smooth" this problem out.

Reply: We thank the reviewer for this important observation. We agree that the behavior at high observed lightning densities reflects a consistent negative bias rather than increased random dispersion alone. We therefore quantified the signed bias in the upper tail of the observed lightning-density distribution for all four base models and the ridge regression ensemble.

In response, We added the bias-evaluation method to Sect. 3.3 (Lines 223-226): "*Building on this stratification, we further quantified the signed bias in the upper tail by calculating the mean and relative biases of the four base models and the ridge regression ensemble for the high-density ($P_{90} < y_{obs} \leqslant P_{99}$) and extreme-density ($y_{obs} > P_{99}$) regimes (Table S6). Mean bias was defined as the mean of predicted minus observed lightning density, and relative bias was calculated by dividing the mean bias by the mean observed density and multiplying by 100%; negative values therefore indicate underestimation.*".

And we revised the interpretation of Fig. 2 at Lines 268-273 and Lines 288-290: "*At higher observed densities, the scatter became broader and the fitted regression lines fell below the 1:1 line, indicating that the increased uncertainty was accompanied by a systematic underestimation of the upper tail. The regime-specific bias analysis showed relative biases ranging from −15.49% to −6.90% in the high-density regime and from −42.51% to −33.73% in the extreme-density regime. Notably, the ridge regression ensemble exhibited the smallest negative bias in both*

regimes, with relative biases of −6.90% and −33.73%, respectively (Table S6).” and *“The ridge ensemble also exhibited the smallest negative bias among the five models in both the high- and extreme-density regimes, although substantial underestimation remained under extreme conditions (Table S6).*”.

And we added the quantitative bias results to Table S6:

Table S6. *Mean and relative biases of the four base models and the ridge regression ensemble for the upper tail of the observed test-set lightning-density distribution. The high-density regime was defined as $P_{90} < y_{obs} \leqslant P_{99}$ (N = 615,320, 9.00%); and extreme density, $y_{obs} > P_{99}$ (N = 68,369, 1.00%), where P_{90} = 0.01250167 and P_{99} = 0.08098790 strokes km^{-2} day^{-1}. Mean bias was calculated as the mean of predicted minus observed lightning density, and relative bias was calculated as the mean bias divided by the mean observed density. Negative values indicate systematic underestimation.*

Observed lightning-density regime	***Model***	***Mean bias***	***Relative bias (%)***
High-density regime $P_{90} < y_{obs} \leqslant P_{99}$	*XGBoost*	*−0.004580*	*−15.49*
	RF	*−0.003222*	*−10.90*
	LightGBM	*−0.004190*	*−14.17*
	DNN	*−0.004515*	*−15.27*
	Ensemble	*−0.002040*	*−6.90*
Extreme-density regime $y_{obs} > P_{99}$	*XGBoost*	*−0.058766*	*−40.20*
	RF	*−0.062138*	*−42.51*
	LightGBM	*−0.050022*	*−34.22*
	DNN	*−0.052705*	*−36.06*
	Ensemble	*−0.049297*	*−33.73*

We added an explicit application-related caution to Sect. 5.2 (Lines 630-635): “*However, because the reconstruction systematically underestimates high and extreme lightning densities, applications that depend on the absolute magnitude of intense lightning, such as lightning-ignited wildfire and lightning-produced NOx estimation, should be interpreted cautiously. For these applications, GLLDR v1 is more suitable for examining large-scale climatological patterns and long-term variability than for precisely quantifying the intensity of the most extreme lightning conditions.*”.

We replaced the previous general discussion of ridge shrinkage in Sect. 5.3 with a more explicit discussion of the detected limitation at Lines 670-677: “*A further model-related limitation was the systematic underestimation of high lightning densities. Although the ridge regression ensemble showed the smallest negative bias among the five models, it still underestimated the mean observed density in the extreme-density regime by 33.73%. This upper-tail compression may be related to the rarity of extreme samples and the limited representation of storm-scale convective organization by monthly large-scale predictors; regularization-induced shrinkage may additionally contribute to the ridge ensemble. The resulting limitation primarily concerns the accuracy of reconstructed absolute magnitudes under the most intense lightning conditions.*”.

We also added a corresponding future improvement direction at Lines 689-692: “*More broadly, incorporating predictors that better represented storm organization (where feasible), evaluating performance across distinct convective regimes, and exploring density-aware training strategies to reduce upper-tail bias would help refine the interpretation of trends and regional changes.*”.

Comment 6: Zonal-mean correlation (Fig. 4d) is not strong independent evidence

r = 0.9969 mainly reflects the known tropical-to-pole lightning gradient, which is already encoded via the abs_lat predictor. Zonal averaging over many grid cells almost entirely smooths out local disagreement.

Fix: Soften the interpretive language. Consider comparing zonal anomalies instead of raw absolute values.

Reply: We thank the reviewer for this important clarification. We agree that the very high correlation between the reconstructed and observed zonal means primarily reflects the dominant tropical-to-polar lightning gradient and is further enhanced by the smoothing inherent in zonal averaging. We have therefore softened the interpretation of Fig. 4d and no longer present this correlation as independent evidence of local-scale spatial agreement. Instead, we interpret it only as descriptive evidence that the reconstruction reproduces the broad meridional gradient.

We also clarify that the absolute- and percentage-difference profiles in Figs. 4e–f provide complementary information on the magnitude and direction of latitude-dependent discrepancies, although they do not remove the background meridional gradient. Local and regional differences are therefore evaluated primarily from the spatial comparison and bias maps rather than from the zonal-mean correlation alone.

We revised the interpretation of the zonal-mean correlation in Sect. 4.2 (at Lines 340-352) as follows: "*The reconstructed and observed zonal-mean profiles showed a high correlation ($r=0.9969$, $p<0.001$; Fig. 4d), indicating that the reconstruction reproduced the broad tropical-to-polar gradient in lightning density. However, this high correlation was partly driven by the dominant meridional gradient and the smoothing inherent in zonal averaging and should therefore not be interpreted as independent evidence of local-scale spatial agreement. The absolute and percentage differences further characterized latitude-dependent discrepancies in magnitude, with the dominant error being a tropical positive bias within approximately ±20° latitude (Figs. 4e–f). Taken together, these comparisons indicate that the ensemble reconstruction captured the principal large-scale spatial distribution during the 2013–2024 observational period, while the bias maps and zonal-difference profiles further characterized the remaining regional and latitude-dependent discrepancies. Building on this overlap-period evaluation, we examined the reconstructed climatology for 1979–2012, which represents a model-based temporal extrapolation beyond the 2013–2024 WWLLN observational period (Fig. 5).*".

Comment 7: Improvement from stacking is not statistically convincing

The ensemble beats the best single model (LightGBM) by only ~0.01 R^2 (0.6895 vs. 0.6784), with no uncertainty estimate. The added complexity of a four-model ensemble is not clearly justified by this, yet the text calls it repeatedly "superior performance."

Fix: Report for example a bootstrap confidence interval for this difference. Adjust the language accordingly to be more cautious.

Reply: We thank the reviewer for this important comment. We agree that the uncertainty in the relatively small performance difference should be quantified. We therefore conducted a paired month-block bootstrap comparison between the ridge regression ensemble and LightGBM, the best-performing individual model based on test R^2 and RMSE. Based on 10,000 bootstrap replicates, the ensemble increased R^2 by 0.0110, with a 95% confidence interval of 0.0095 –

0.0127, while the confidence intervals for the RMSE and MAE differences also consistently favored the ensemble. These results indicate a modest but consistent improvement. We have added the bootstrap method and results, provided the full comparison in Table S7, and revised the manuscript to avoid repeatedly using the term "superior performance."

We revised the corresponding statement in the Abstract (at Lines 28-30) as follows: "*Among the evaluated models, the ridge regression ensemble achieved the best overall test performance (R^2 = 0.6895, RMSE = 0.0108, MAE = 0.0030), indicating that model blending effectively enhanced predictive stability.* ".

We added the paired bootstrap method in Sect. 3.3 (at Lines 227-233) as follows: "*To quantify the uncertainty in the performance gain of the ridge regression ensemble relative to the strongest individual-model benchmark, we conducted a paired month-block bootstrap comparison with LightGBM, which achieved the highest test R^2 and the lowest test RMSE among the four individual models. The 144 months during 2013–2024 were sampled with replacement while retaining all test samples and paired model predictions within each selected month. Differences in R^2, RMSE, and MAE were recalculated for 10,000 bootstrap replicates, and the 95% confidence intervals were defined by the 2.5th and 97.5th percentiles of the resulting bootstrap distributions (Table S7).*. ".

We revised the model-comparison results in Sect. 4.1 (Lines 283-287) as follows: "*Ultimately, the ridge regression ensemble achieved the highest test R^2 (0.6895) and the lowest RMSE (0.0108) among the five models (Tabel 1). Compared with LightGBM, the best-performing individual model based on test R^2 and RMSE, the ensemble increased R^2 by 0.0110 (95% paired month-block bootstrap CI: 0.0095–0.0127) and also consistently reduced RMSE and MAE (Table S7).*".

We revised the interpretation of the spatial model comparison in Sect. 4.1 (Lines 305-306 and Lines 321-324) as follows: "*We subsequently evaluated the spatial consistency of model performance across the test set to compare how the five models performed across diverse geographical regions.*" and "*Together with the global evaluation metrics and the paired bootstrap comparison, these spatial diagnostics supported the selection of the ridge regression ensemble for the final lightning reconstruction, while characterizing its improvement over the best-performing individual model as modest but consistent.*".

We replaced the stronger wording at the beginning of Sect. 4.2 (Lines 331-332) as follows: "*Based on the overall test-set performance of the ridge regression ensemble, we reconstructed the global lightning density and first evaluated its spatial fidelity during the 2013–2024 period (Fig. 4).*".

We added a new Supplementary Table S7, which presents the test-set performance of LightGBM and the ridge regression ensemble, their paired differences in R^2, RMSE, and MAE, and the corresponding 95% month-block bootstrap confidence intervals:

Table S7. *Paired month-block bootstrap comparison between LightGBM and the ridge regression ensemble on the test set.*

Metric	***LightGBM***	***Ensemble***	***Difference*** *(Ensemble − LightGBM)*	***95% bootstrap CI***
R^2	0.678443	0.689484	0.011041	0.009459 to 0.012660
RMSE	0.01095	0.01076	−0.000190	−0.000215 to −0.000165
MAE	0.003182	0.002991	−0.000191	−0.000201 to −0.000180

Comment 8: Figure 6 trend comparisons rely on a short record and show mixed up reconstructed-trend significance

Trends in panels (b)–(d) use only 12 annual values (2013–2024), limiting statistical power. In all three sub-analyses, the reconstructed trend is more significant than the observed one (e.g., Tropics: $p<0.05$ vs. $p<0.1$). This is consistent with the reconstruction having lower interannual variance than the noisier observations, likely due to ensemble smoothing. This inflates significance without necessarily improving accuracy. The tropical trend is also underestimated by about half (reconstructed 1.03×10^{-4}/yr vs. observed 1.98×10^{-4}/yr), described in the text only as "a slight deviation". Finally, panel (d) shows the full 1979–2025 series, but the trend line and significance are computed only for 2013–2024; the pre-2013 portion carries no statistical support in this figure.
Fix: (I) Discuss how ensemble smoothing affects trend-test power. (ii) Report confidence intervals on trend slopes, not just p-values and use consistent test statistics and thresholds. (iii) Revise "slight deviation" to reflect the actual factor-of-two (!) gap. (iv) You might also clarify in the caption that no trend test covers the pre-2013 period.
Reply: Thank you for this helpful comment. We revised Fig.6 and its caption to report numerical Mann–Kendall p-values and to clarify that all trend analyses in panels (b)–(d) were restricted to 2013–2024:

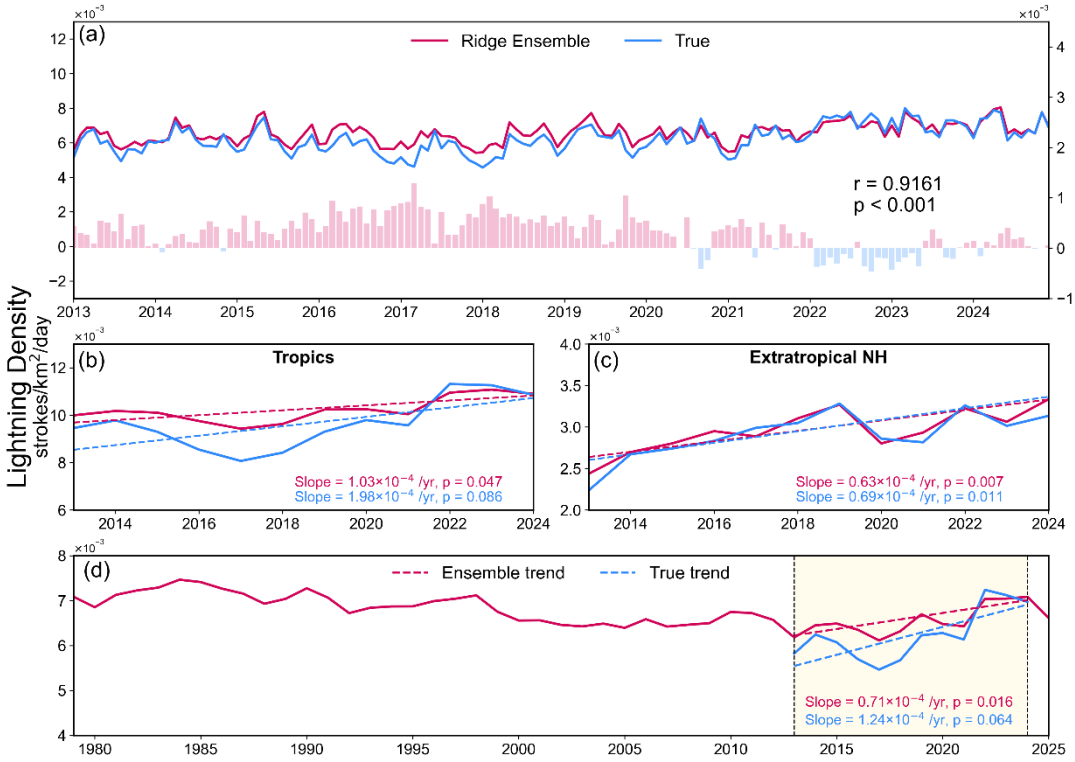

Figure 6. ***(a)*** *Monthly global mean lightning density during 2013–2024 predicted by the ridge regression ensemble and observations, with residuals shown as bars.* ***(b)*** *and* ***(c)*** *present annual mean lightning density during 2013–2024 over the tropics ($|lat| \leqslant 30^\circ$) and the extratropical Northern Hemisphere ($30^\circ < lat \leqslant 90^\circ$), respectively.* ***(d)*** *Annual global mean lightning density from the ridge regression ensemble (1979–2025), with the observational period (2013–2024) highlighted. For panels* ***(b)–(d)****, Theil–Sen trend lines and the corresponding Mann–Kendall p-values were calculated exclusively for 2013–2024 and are shown for both the ensemble and the observations.*

We also added Table S8 reporting the Theil–Sen slopes and their 95% confidence intervals:

***Table S8**. Theil–Sen trend estimates and 95% confidence intervals for the reconstructed and observed annual lightning density series. All trend estimates were calculated for the 2013–2024 period. Trend slopes and their 95% confidence intervals were estimated using the Theil–Sen method, and two-sided Mann–Kendall p-values are reported. Slopes and confidence intervals are expressed in units of 10^{-4} strokes km^{-2} day^{-1} yr^{-1}.*

Region	Series	Theil–Sen slope	95% CI	Mann–Kendall (p)
Tropics	Ensemble	1.03	0.06 to 2.00	0.047
	Observed	1.98	−0.30 to 3.98	0.086
Extratropical NH	Ensemble	0.63	0.29 to 1.13	0.007
	Observed	0.69	0.21 to 1.06	0.011
Global	Ensemble	0.71	0.24 to 1.15	0.016
	Observed	1.24	−0.04 to 2.16	0.064

The Results section was revised to quantify the tropical difference in trend magnitude, interpret the slope estimates together with their confidence intervals, and acknowledge that the generally lower p-values of the reconstructed trends may partly reflect the smoother interannual variability of the ensemble series. We further clarified that no trend test was applied to the pre-2013 portion shown in Fig. 6d at Lines 385-411: "*The temporal behavior of the reconstructed lightning density was further analyzed at monthly and annual scales (Fig. 6). During the observational period, the monthly global mean lightning density produced by the ridge regression ensemble closely tracked the observations and successfully reproduced the seasonal cycle and short-term variability, albeit with a slight tendency toward overestimation (Fig. 6a). To assess latitudinal differences in performance, we analyzed the annual mean lightning density for the Tropics ($|lat| \leqslant 30^\circ$) and the extratropical Northern Hemisphere ($30° < lat \leqslant 90^\circ$), respectively. To characterize the uncertainty associated with these trend estimates based on 12 annual values, the corresponding 95% confidence intervals of the Theil–Sen slopes are reported in Table S8.*

In the extratropical Northern Hemisphere, the reconstructed trend (slope = 0.63×10^{-4}, $p = 0.007$) was similar in magnitude to the observed trend (slope = 0.69×10^{-4}, $p = 0.011$), and the 95% confidence intervals of both slopes remained above zero and showed substantial overlap, supporting good agreement in trend direction and magnitude during 2013–2024 (Fig. 6c and Table S8). Within the Tropics, the reconstruction captured the interannual fluctuations, but the reconstructed trend (slope = 1.03×10^{-4}, $p = 0.047$) was approximately 48% lower than the observed trend estimate (slope = 1.98×10^{-4}, $p = 0.086$), indicating a lower reconstructed trend magnitude. The wider 95% confidence interval of the observed slope included zero, reflecting the limited precision of trend estimation over the 12-year period (Fig. 6b and Table S8).

Finally, placing these recent years into a long-term context, the reconstructed global mean lightning density for 1979–2025 exhibited pronounced interannual and decadal variability (Fig. 6d). During the common 2013–2024 period, both the reconstruction (slope = 0.71×10^{-4}, $p = 0.016$) and the observations (slope = 1.24×10^{-4}, $p = 0.064$) showed weak upward trends, indicating consistency in the direction of recent global change, despite differences in the estimated trend magnitude. The generally lower p-values of the reconstructed trends may partly reflect the smoother interannual evolution of the ensemble series. The trend comparisons were therefore interpreted using both the slope estimates and their confidence intervals. Overall, these results indicated that the reconstruction captured the observed temporal variability well during the overlap period, while providing a historical perspective that revealed how recent variations fit into the broader patterns of variations since 1979.".

Comment 9: Figure 7's long-term trend map is presented with more confidence than the extrapolation supports

The 1979–2012+2025 trend map rests on 34 unvalidated years. The dominant predictor, CAPE×TP, is derived from ERA5 fields known to have long-term trends linked to changes in the assimilated observing system. Any such artifact would propagate directly into this map. The manuscript mentions ERA5 uncertainty in general terms, which is good, (Sect. 5.3, lines 532-538) but does not test its effect on Fig. 7, nor report whether all four base models agree on trend sign.

Fix: (i) I have already mentioned it before, but it is important to state explicitly that pre-2013 trends areunvalidated and which implications this has. (ii) Test trend sensitivity to known ERA5 inhomogeneities. You might analyse how ERA5 predictor distributions (e.g., CAPE, TP) differ between 1979–2012 and 2013–2024 and briefly discuss your results (you do not need to show a figure I guess). (iii) Concerning the agreement of the trend sign, you might report the fraction of grid cells where all four base models agree on trend sign, as an uncertainty diagnostic (similar to Figure 3f). (iv) Discuss how the demonstrated tropical trend underestimation in Fig. 6 affects confidence in Fig. 7's magnitudes.

Reply: We thank the reviewer for this important comment. We have revised the manuscript at Lines 422-425 to clarify that the trends shown in Fig. 7 for the pre-2013 period and 2025 are model-based extrapolations that have not been directly validated against contemporaneous lightning observations: "*Because the pre-2013 period and 2025 fall outside the WWLLN-constrained training and evaluation period, the trend magnitudes and statistical significance shown in Fig. 7 represent characteristics of a model-based extrapolation rather than independently validated observational trends.*".

Following the reviewer's suggestions, we compared the distributions of CAPE, TP, and CAPE×TP between 1979–2012 and 2013–2024 and added the results to Table S9.

Table S9. *Comparison of the distributions of key ERA5 predictors between the extrapolation and training periods. The extrapolation period refers to 1979–2012, and the training period refers to 2013–2024. Mean difference was calculated as* $(Mean_{1979–2012}–Mean_{2013–2024})/Mean_{2013–2024}\times100\%$*. The distribution-overlap coefficient ranges from 0 to 1, with higher values indicating greater similarity between the two distributions. The final column reports the percentage of pre-2013 values falling outside the 1st–99th percentile range of the corresponding predictor during 2013–2024.*

Predictor	*1979–2012 mean*	*2013–2024 mean*	*Mean difference (%)*	*1979–2012 median (P5–P95)*	*2013–2024 median (P5–P95)*	*Distribution overlap*	*Pre-2013 values outside training P1–P99 (%)*
CAPE	*151.63*	*134.46*	*12.77*	*7.50 (0.125–927.75)*	*7.50 (0.125–803.38)*	*0.971*	*1.52*
TP	*0.002246*	*0.002217*	*1.32*	*0.001221 (0–0.008348)*	*0.001196 (0–0.008369)*	*0.984*	*0.89*
CAPE×TP	*0.8332*	*0.7175*	*16.12*	*0.008401 (0–6.2069)*	*0.008183 (0–5.2006)*	*0.984*	*1.5*

The distributions showed strong overall overlap, although the upper tails of CAPE and CAPE×TP were higher during the earlier period (Lines 425-429): "*The distributions of CAPE, TP, and CAPE×*

TP showed strong overlap between 1979–2012 and the 2013–2024 training period, with 98.48%–99.11% of the pre-2013 values remaining within the training-period 1st–99th percentile ranges, although the higher upper tails of CAPE and CAPE×TP indicate that some influence on local trend magnitudes cannot be excluded (Table S9).".

We therefore added a corresponding discussion of possible ERA5 temporal inhomogeneities and the limits of this diagnostic in Sect. 5.3 (Lines 664-670): "*Our comparison showed strong overall distributional overlap for CAPE, TP, and CAPE×TP between 1979–2012 and the training period, indicating that most pre-2013 predictor values remained within the input ranges represented during model training (Table S9). Nevertheless, the upper tails of CAPE and CAPE×TP were higher during 1979–2012, indicating some temporal differences in high-value convective conditions. This diagnostic cannot exclude localized or time-specific ERA5 inhomogeneities or distinguish possible observing-system effects from genuine climate variability.* ".

We also quantified the agreement in trend sign among the four base models and added the results to Table S10.

Table S10. *Agreement in the sign of 1979–2025 grid-cell trends among the four base models. Trend slopes were estimated independently for XGBoost, LightGBM, RF, and DNN using the Theil–Sen estimator. Agreement denotes grid cells where all four models produced either positive slopes or negative slopes. Ridge-significant grid cells refer to locations where the Mann–Kendall test for the ridge-ensemble trend yielded $p<0.05$. Percentages were calculated relative to the number of common-valid grid cells in each subset. This diagnostic evaluates trend direction rather than trend magnitude.*

Grid-cell subset	***Valid grid cells***	***All four models agree on trend sign (%)***
All common-valid land grid cells	*237,392*	*61.59*
Ridge-significant grid cells ($p<0.05$)	*84,016*	*93.92*

Accordingly, we added the following statement to Sect. 4.3 at Lines 436-439: "*The four base models agreed on the trend sign in 61.59% of all common-valid land grid cells and 93.92% of grid cells with significant ridge-ensemble trends, providing an additional uncertainty diagnostic for the mapped trend directions (Table S10).*".

Finally, we clarified at Lines 439-441 that the tropical trend magnitudes in Fig. 7 should be interpreted more cautiously than their signs and broad spatial patterns, given the attenuation identified in Fig. 6: "*Given the attenuation of the reconstructed tropical trend magnitude during 2013–2024 (Fig. 6), the tropical trend magnitudes in Fig. 7 should be interpreted more cautiously than their signs and broad spatial patterns.*".

Comment 10: Figure 10 interpretation goes beyond what the plots show

The "saturation" pattern at high CAPE×TP/CAPE values could reflect data sparsity at the tail rather than a real physical regime shift, sample density is not shown. Further, CAPE and CAPE×TP are clearly collinear, so panels (a) and (b) may partly reflect a shared signal.

Additionally, SHAP is computed for XGBoost and LightGBM individually, not for the final ensemble model.

Fix: (I) Add a sample-density indicator to each panel. (ii) Note the CAPE/CAPE×TP collinearity explicitly and soften causal language (e.g., "consistent with" instead of "highlighting the role of ... in theelectrification process"). (iii) Use full names in the plot to enhance readability (not just

abbreviations like abs_lat). (v) Furthermore, state clearly that attribution reflects base learners, not the ensemble. If feasible, compute SHAP or permutation importance on the ensemble output directly (you could also do a rough estimate manually by leaving variables out during prediction and assess the performance).

Reply: We thank the reviewer for this helpful comment. We added log-scaled sample-density histograms to each panel of the XGBoost and LightGBM SHAP dependence plots (Figs. S4 and S5), allowing the sampling support and sparsely populated upper tails to be visualized. Predictor abbreviations in the plots were also replaced with full names to improve readability (Lines 526-529): "*To indicate the sampling support across the predictor ranges, gray marginal histograms showing log-scaled sample frequencies were added to the XGBoost and LightGBM dependence plots provided in Figs. S4 and S5, respectively.* ".

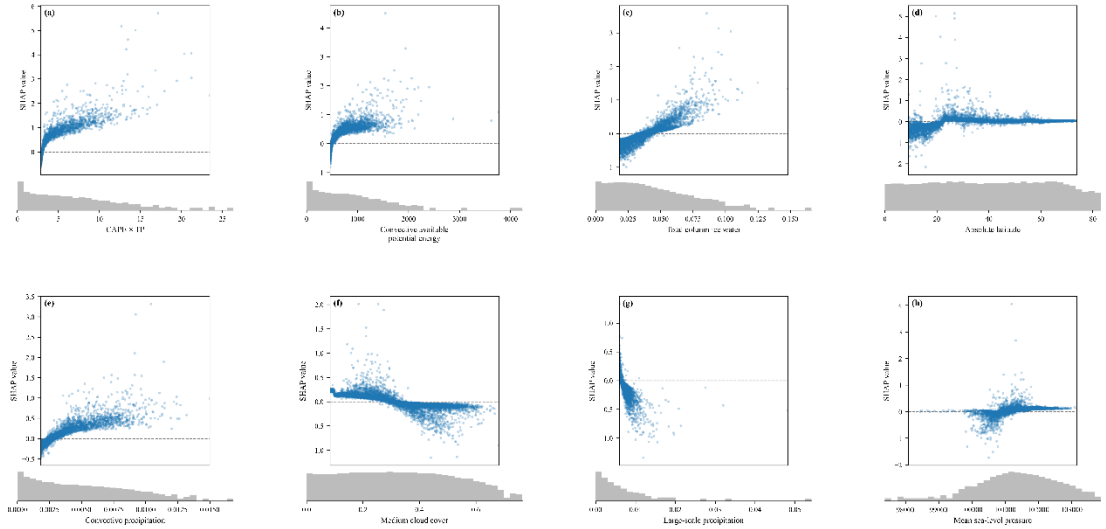

Figure S3. *SHAP dependence plots for the eight leading predictors identified by the LightGBM base model. The x-axis shows the predictor value, and the y-axis shows its SHAP contribution to the LightGBM prediction relative to the model baseline. The gray marginal histograms below each panel show the corresponding sample frequencies on a logarithmic scale, allowing sparsely sampled tails to remain visible. These attribution patterns characterize the LightGBM base model rather than the final ridge regression ensemble.*

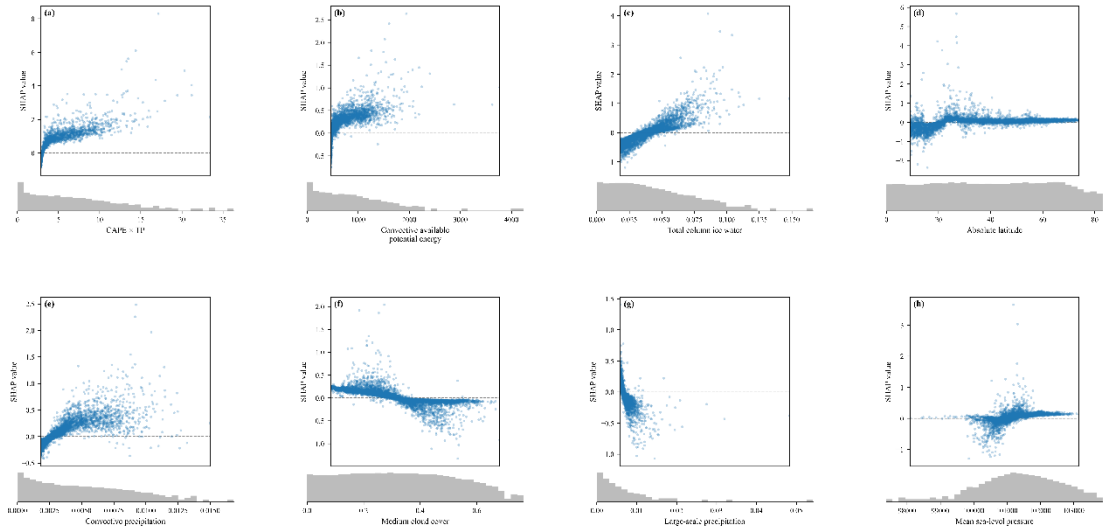

Figure S4. *SHAP dependence plots for the eight leading predictors identified by the XGBoost base model. The predictors and plotting procedure are the same as those described for Fig. S3. These attribution patterns characterize the XGBoost base model rather than the final ridge regression ensemble.*

Sect. 4.4 was revised (Lines 529-539) to interpret the apparent flattening at high CAPE × TP and CAPE values more cautiously, explicitly account for sparse upper-tail sampling and the collinearity arising because CAPE × TP contains CAPE, and avoid causal or threshold-based interpretations unsupported by the plots: "*The relationships were strongly nonlinear for several key predictors. CAPE × TP and CAPE showed rapid increases in positive contributions from low to moderate values, followed by an apparent flattening at higher ranges. However, the marginal distributions in Figs. S4 and S5 show that the upper ranges of both predictors were relatively sparsely sampled. Therefore, this apparent flattening should be interpreted cautiously and should not, by itself, be regarded as evidence of a distinct physical saturation threshold. Moreover, because CAPE × TP contains CAPE by construction, the two predictors are inherently dependent and share substantial information. Their SHAP dependence patterns may therefore partly reflect predictor collinearity rather than independent effects. Subject to these limitations, the dependence patterns may be consistent with high lightning-density predictions being jointly conditioned by moisture availability, storm organization, and microphysical state, rather than by instability or precipitation alone.*"

We clarified in Sect. 3.5 (Lines 252-255) and the relevant figure captions that the SHAP attribution reflects the XGBoost and LightGBM base models rather than the final ridge regression ensemble: "*SHAP values were calculated separately for XGBoost and LightGBM and therefore characterize how these two base learners used the predictors, rather than providing direct predictor-level attribution for the final ridge regression ensemble.*" and "***Figure 10.*** *SHAP dependence plots for the top eight influencing factors identified* ***by the XGBoost model****. The y-axis represented the SHAP value for a specific feature, reflecting its marginal contribution to the predicted lightning density, while the x-axis denoted the feature value. Predictor abbreviations are same as in Fig. 9.*".

Comment 11: Inconsistent value ranges across figures

Figure 2 axes span 0–0.4 strokes/km²/day; Figs. 4 and 5 use colorbars to ~0.02; Fig. 8 uses ~0.03. Fig. 4b and Fig. 8b both show "true" WWLLN density, yet differ in maximum value, despite largely overlapping domains. The Fig. 2 range likely results from the stated stratified sampling of scatter points (rare high values intentionally over-represented for visualization). None of this is explained and the target-variable distribution needed to judge it is never reported.

Fix: (i) Report the target's percentile distribution (e.g., 50th, 90th, 99th, maximum) in text or supplement. (ii) State what fraction of Fig. 2's plotted points exceed the range shown in Figs. 4/5/8. (iii) Clarify why Fig. 4b and Fig. 8b maxima differ, and whether this reflects interannual variability, a different averaging method or independent colorbar choices.

Reply: We thank the reviewer for identifying the unclear presentation of value ranges across the figures. We added Table S11 to report the distributional statistics of the complete monthly WWLLN target dataset.

Table S11. *Distributional statistics of the monthly WWLLN target variable during 2013–2024. Lightning density is expressed in strokes km^{-2} day^{-1}. Statistics were calculated from all valid monthly 0.25° WWLLN target samples used for model development during 2013–2024.*

Statistic	***Value***

Number of valid samples	*34,184,448*
Zero-valued samples (%)	*64.9727*
Mean	*0.005025*
Standard deviation	*0.019313*
50th percentile	*0*
75th percentile	*0.001602*
90th percentile	*0.012505*
95th percentile	*0.026334*
99th percentile	*0.08124*
99.5th percentile	*0.117686*
99.9th percentile	*0.236613*
Maximum	*1.825866*

We also added Text S3 to document the requested proportions for the stratified sample used in Fig. 2 and to explain the different value ranges in Figs. 4, 5, and 8. In particular, Text S3 clarifies the different temporal and spatial aggregations used for Figs. 4b and 8b and distinguishes the independently selected colorbar limits from the actual maxima of the underlying data fields. These additions clarify that the wider range in Fig. 2 reflects individual monthly samples and stratified visualization sampling, whereas the map panels show temporally and/or spatially averaged fields.

Text S3. *Distribution of the WWLLN target variable and explanation of figure-specific value ranges*

The complete monthly 0.25° WWLLN target dataset for 2013–2024 was strongly zero-inflated and right-skewed. Among the valid samples, 64.97% were zero. The 50th, 90th, 99th, and 99.9th percentiles were 0, 0.01250, 0.08124, and 0.23661 strokes km^{-2} day^{-1}, respectively, while the maximum was 1.82587 strokes km^{-2} day^{-1}. Among the 40,000 observations selected through stratified sampling for visualization in Fig. 2, 16.65% and 10.68% exceeded 0.02 and 0.03 strokes km^{-2} day^{-1}, respectively. The corresponding proportions in the complete target dataset were 6.67% and 4.31%. The thresholds of 0.02 and 0.03 strokes km^{-2} day^{-1} correspond to the upper colorbar limits used for the lightning-density fields in Figs. 4–5 and Fig. 8, respectively. These colorbar endpoints are visualization limits rather than the actual maximum values of the underlying data fields.

Although Figs. 4b and 8b both display WWLLN lightning density, they represent different temporal and spatial aggregations. Figure 4b shows the 2013–2024 multiyear mean on the original 0.25° grid, whereas Fig. 8b shows the 2013–2014 mean after aggregation to 2.5° and alignment with the LIS/OTD grid. The shorter temporal averaging period allows interannual differences to remain, while the coarser spatial aggregation smooths localized high-density values. Consequently, the underlying maxima of the two fields are not expected to be identical. The apparent difference between the plotted maxima is additionally influenced by the independently selected colorbar limits and does not indicate an inconsistency in the WWLLN data.

Minor Comments

Comment 12: WWLLN uncertainty

Discuss uncertainty in WWLLN detection-efficiency corrections more directly, since these data form the training target.

Reply: We thank the reviewer for this helpful suggestion. We revised Sect. 5.3 (Lines 642-644) to clarify that the WWLLN detection-efficiency correction reduces but does not eliminate observational uncertainty: "*Although detection-efficiency correction reduces known observational*

biases, residual uncertainty in the corrected WGLC target may still propagate into the reconstructed dataset."

.

Comment 13: Redundant text

ERA5 variable descriptions are repeated in Sect. 2.1 and 3.1. Remove the duplication.

Reply: We thank the reviewer for pointing out this duplication. We removed the repeated overview of the ERA5 variable categories from Sect. 3.1: "*The predictor set described in Sect. 2.1 and summarized in Table S1 and Fig. 1 was used as model input.*".

Comment 14: Applications language

Soften claims about wildfire modeling, lightning-NOx estimation and climate model benchmarking with this type of analysis.

Reply: We thank the reviewer for this helpful comment. We softened the application-related language in the Abstract (Lines 41-44), Sect. 5.2 (Lines 620-621), and the Conclusions (Lines 727-729): "*This dataset provided a physically constrained and spatially detailed basis for studying long-term lightning dynamics, and may provide a historical lightning reference for studies of natural ignition, lightning-produced NOx, and lightning representations in climate and Earth system models.*", "*A long-term gridded lightning dataset could provide a reference for investigating the spatial and temporal variability of lightning-produced NOx.* " and "*These results provided a practical basis for investigating reconstructed changes in lightning-related hazards and convective activity over the past nearly five decades and a historical reference for broad-scale comparisons with lightning simulations from climate models.*".

Comment 15: Overstated language generally

Terms like "superior performance," "high spatial fidelity," and "excellent agreement" appear stronger than the evidence supports in several places.

Reply: We thank the reviewer for this helpful comment. Most instances of potentially overstated language had already been revised in response to the other comments. We additionally replaced the remaining expression "high spatial fidelity" with the more objective wording "spatial agreement" at Lines 33-35: "*Individual validations indicated spatial agreement, as the ensemble successfully reproduced the observed large-scale spatial distribution and major tropical–subtropical continental lightning hotspots.*".

Comment 16:

ERA5 predictors run through December 2025 and the dataset is labeled 1979–2025, but WWLLN training data stop at 2024 (line 129). So 2025, like 1979–2012, is model-extrapolated. Please clarify: (i) whether 2025 WGLC data existed at the time of analysis and could have been used for training, and (ii) whether the 2025 ERA5 data are final or preliminary (ERA5 expver), since preliminary fields can differ from the final product. At minimum, 2025 should carry the same "unvalidated" status as the pre-2013 period.

Reply: We thank the reviewer for highlighting this point. The WGLC data available during model development extended only through 2024; therefore, 2025 observations could not be included in model training or contemporaneous validation. We have explicitly identified the 2025 reconstruction, together with the pre-2013 period, as model-based temporal extrapolation without

direct contemporaneous WGLC validation. We also verified that all ERA5 predictor fields used for 2025 were from the final ERA5 product (expver = 0001), rather than the preliminary ERA5T stream.