Review of "Reconstruction of Global 0.25° Land Lightning Density from 1979 to 2025 based on an ensemble machine learning" (essd-2026-318)

Summary

This manuscript presents a global, 0.25°, monthly lightning stroke-density reconstruction for 1979–2025. It uses a ridge-regression stacking ensemble of four ML models trained on WWLLN observations (2013–2024) and ERA5 predictors. The dataset fills a real gap: no existing product combines long spatio-temporal coverage and global completeness. The workflow is clearly described and the SHAP-based attribution adds interpretive value. However, several evaluation choices clearly overstate model skill and some interpretive claims go beyond what the figures support resulting in misleading conclusions. I recommend major revision. My comments are grouped by topic, roughly following the order of the manuscript.

Major Comments

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).

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.

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.

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.

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.

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.

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.

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.

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 are unvalidated 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.

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 the electrification 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).

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.

Minor Comments

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

- Redundant text: ERA5 variable descriptions are repeated in Sect. 2.1 and 3.1. Remove the duplication.

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

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

- 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.

Recommendation

The analysis addresses an important need and the overall approach is reasonable and very interesting. But the current validation does not fully support several central claims, particularly on spatial/temporal generalization, accuracy in typical (non-hotspot) conditions and the reliability of the pre-2013 reconstruction. I recommend **major revision, before it can be considered for publication.**