## Response to Community Comments

The study addresses an interesting question. However, I wonder why the authors decided to randomly split the dataset into training and testing subsets. As reviewer #2 noted, machine learning model evaluation based on random splits may be overly optimistic in the presence of spatial and temporal dependence (e.g. Ploton et al., 2020). In my view, using complete years exclusively for either training or testing would provide a more robust assessment of the model's ability to generalize to unseen time periods, which is consistent with its intended use for hindcasting. To assess inter-annual variability while making use of the entire dataset, a leave-one-year-out cross validation approach could also be considered.

Additionally, given the fundamental differences between WWLLN and LIS/OTD, I am surprised by the strong correlation between the Ridge ensemble and LIS/OTD in most regions (Fig. 8d). Is the correlation indeed computed on annual mean values?

Finally, I believe the manuscript would benefit from a comparison or discussion of the reported trends in stroke density with previously published trends in thunder days over the last decades (e.g. Lavigne et al., 2019).

Reply: 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.

First, we added Ploton et al. (2020) to explain that random partitions of spatially structured data may not ensure independence between training and test observations and may therefore produce overly optimistic performance estimates: "*Random sample partitions may not ensure independence between training and test observations when dependence structures exist in geospatial data and may therefore yield overly optimistic estimates of predictive performance (Ploton et al., 2020).*". We also added a year-blocked validation in which 2015, 2019, and 2021 were withheld as complete test years, with all withheld samples excluded from preprocessing, hyperparameter optimization, early stopping, and model training. The full procedure and results are provided in Text S2 and Tables S12–S13.

Second, we clarified that the correlations shown in Fig. 8d were calculated between the annual mean lightning-density time series of GLLDR v1 and LIS/OTD over 1996–2014.

Third, we added a discussion in Sect. 5.2 comparing the reconstructed stroke-density trends with the ground-based thunder-day trends reported by Lavigne et al. (2019): "*Previous analyses of ground-based thunder-day records also revealed pronounced regional heterogeneity in long-term thunderstorm activity (Lavigne et al., 2019). Some regional patterns, such as the decrease over Australia, were qualitatively consistent with the reconstructed stroke-density trends in this study, whereas the trend directions differed across several tropical convective regions. These differences are not necessarily contradictory because thunder days primarily characterize thunderstorm-occurrence frequency, while flash or stroke density also depends on lightning production per storm; differences in the analysis period, spatial sampling, and lightning metrics further limit direct quantitative comparison between the two studies.*".

**References:**

Lavigne, T., Liu, C. T., and Liu, N. N.: How Does the Trend in Thunder Days Relate to the Variation of Lightning Flash Density?, J Geophys Res-Atmos, 124, 4955–4974, 10.1029/2018jd029920, 2019.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat Commun, 11, ARTN 4540 10.1038/s41467-020-18321-y, 2020.