https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Open Access Earth System Science Data Discussions

Reconstruction of Global 0.25° Land Lightning Density from 1979 to 2025 based on an ensemble machine learning

Hao Zheng[1, 2], Jun Wang[1, 2*], Hao Zhou[3], Jingfeng Ding[3], Haijin Dai[3], Zhi Huang[1, 2], Zishan Wang[1, 2], Meirong Wang[4, 5], Jianying Li[6], Hengmao Wang[1, 2], Fei Jiang[1, 2], Weimin Ju[1, 2]

[1]International Institute for Earth System Science, Nanjing University, Nanjing, Jiangsu 210023, China
[2]Jiangsu Provincial Key Laboratory for Advanced Remote Sensing and Geographic Information Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing, Jiangsu 210023, China
[3]College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China
[4]Joint Center for Data Assimilation Research and Applications/Key Laboratory of Meteorological Disaster, Ministry of Education/Joint International Research Laboratory of Climate and Environment Change (ILCEC)/Collaborative Innovation Center ON Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5]Mêdog Field Station for Scientific Observation and Research on Atmospheric Water Cycle/Xigazê and Mêdog National Climate Observatory, Tibet Meteorological Service, Lhasa, China
[6]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

Correspondence: **Jun Wang (wangjun@nju.edu.cn)**

Abstract. Lightning is a primary driver of severe convective hazards and wildfire ignitions, yet long-term, high-resolution gridded records have remained scarce due to the limited temporal coverage of ground-based networks and the sampling constraints of satellite observations. Here, we presented a new global 0.25° × 0.25° monthly land lightning stroke-density dataset spanning 1979–2025. To ensure robustness, we developed a ridge regression stacking ensemble that integrated four complementary machine learning architectures: eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Random Forest (RF), and Deep Neural Network (DNN). The ensemble achieved superior performance over each single model (test R^2 = 0.6895, RMSE = 0.0108, MAE = 0.0030), indicating that model blending effectively enhanced predictive stability. Individual validations confirmed high spatial fidelity, as the ensemble successfully reproduced the observed large-scale spatial distribution

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Open Access Earth System Science Data Discussions

and major tropical–subtropical continental lightning hotspots. Independent comparisons with the LIS/OTD gridded lightning climatology (±38°) further demonstrated strong spatiotemporal consistency, particularly in reproducing interannual variability. Our analysis revealed pronounced regional heterogeneity in multi-decadal trends: significant decreases were concentrated across several tropical convective centers, while localized increases emerged in specific mid-latitude regions. Attribution based on SHapley Additive exPlanations (SHAP) elucidated that these patterns were primarily governed by the coupling of thermodynamic instability (CAPE × TP), moisture availability, and ice-phase hydrometeor conditions. This dataset provided a physically constrained and spatially detailed basis for studying long-term lightning dynamics, offering practical inputs for natural-ignition modeling, lightning-produced NOx estimation, and the evaluation of lightning parameterizations in climate and Earth system models. The datasets of the 1979–2025 Global Land Lightning Density Reconstruction Version 1 (GLLDR v1) are publicly available at the Zenodo via the following DOI: https://doi.org/10.5281/zenodo.19722380 (Zheng et al., 2026a).

1 Introduction

Lightning is a fundamental atmospheric process with profound societal and ecological ramifications, primarily serving as a defining manifestation of severe convective weather and a critical driver of terrestrial disturbances (Schultz et al., 2011; Kalashnikov et al., 2023). Particularly in remote boreal and extratropical forests, lightning is the predominant natural ignition source, accounting for a disproportionate majority of the total burned area and shaping global fire regimes (Veraverbeke et al., 2017; Janssen et al., 2023). Under a warming climate, the projected escalation in lightning-caused ignitions threatens to trigger massive carbon releases from high-latitude peatlands, potentially fueling a positive feedback loop within the Earth system (Romps et al., 2014). Consequently, a long-term, spatially explicit record of global lightning activity is essential for establishing historical baselines and assessing future climate-driven risks.

Despite its significance, multi-decadal lightning monitoring remained constrained by observational gaps. Historically, spaceborne sensors like the Lightning Imaging Sensor and Optical Transient Detector (LIS/OTD) had provided the foundational satellite-based benchmark for lightning climatology (Cecil et al., 2014). However, these gridded products were typically provided at a relatively coarse spatial resolution (2.5°) and lacked the multi-decadal continuity required for long-term trend analysis.

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Furthermore, the TRMM-LIS component was largely confined to the tropics and subtropics (±38°), leaving the rapidly changing high-latitude regions—where lightning-fire interactions were most consequential—insufficiently sampled. Ground-based networks provided an important complement by enabling continuous global monitoring. For instance, global stroke detections from the World Wide Lightning Location Network (WWLLN) could be aggregated into gridded stroke-density fields, such as the open-access WWLLN Global Lightning Climatology (WGLC) products (Kaplan and Lau, 2021, 2022). However, these gridded products only began in 2010, and the initial years remained influenced by network build-out; accordingly, reprocessed subsets were required for applications demanding a temporally stable baseline (Kaplan and Lau, 2022). Consequently, existing products did not yet provide a long-term, globally complete, and fine-resolution gridded lightning dataset spanning the full satellite era (post-1979), hindering our capacity to detect long-term trends across diverse climate regimes.

To circumvent these limitations, researchers had primarily relied on two compromised approaches in climate and ecological modeling. First, many studies utilized simplified meteorological proxies, such as Convective Available Potential Energy multiplied by precipitation (CAPE × TP) or cloud-top height, to substitute for lightning activity in historical trend analysis and future risk projections (Williams et al., 1992; Romps et al., 2014; Romps et al., 2018; Wong et al., 2013). Second, global dynamic vegetation models (DGVMs) often prescribed natural ignitions using static gridded climatology derived from LIS/OTD (Thonicke et al., 2010; Li et al., 2012). While the climatology captured the mean spatial distribution of lightning, their static nature failed to represent inter-annual variability or multi-decadal trends, thereby masking the dynamic response of lightning-ignited fires to a changing climate.

Furthermore, these simplified proxies often failed to account for the multi-dimensional and non-linear nature of storm electrification. Beyond basic instability, other indices such as the Total Totals and K-index had been shown to capture specific convective environments conducive to lightning (Pérez-Invernón et al., 2023; Saleh et al., 2023). Lightning activity is also strongly modulated by moisture, dynamics, and cloud microphysics that extend beyond mere thermodynamic instability. Machine-learning studies had also been increasingly applied to lightning prediction and parameterization, showing that combinations of precipitation characteristics, low-level winds, and lower-tropospheric humidity could explain substantial fractions of large-scale lightning variability (Cavaiola et al., 2024; Verjans and Franzke, 2025). However, these efforts mainly focused on forecasting skill or parameterization

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

development, rather than on reconstructing a long-term, observation-constrained lightning dataset. The inclusion of cloud macrophysical and microphysical predictors—such as cloud fractions and column-integrated ice and liquid water—had been found to improve lightning classification, reflecting the central role of mixed-phase processes in storm electrification (Ukkonen and Mäkelä, 2019). Additionally, dynamical influences, particularly mid-tropospheric vertical motion near 500 hPa, further modulated updraft strength and storm development (Cheng et al., 2024). Finally, large-scale constraints on lightning variability were often represented through geographic and seasonal descriptors, including absolute latitude and the month of the year (Burrows et al., 2005; Cavaiola et al., 2024; Verjans and Franzke, 2025).

In this study, we bridged this gap by integrating these multi-dimensional environmental drivers into a comprehensive machine-learning framework, moving beyond both static climatology and single-proxy substitutions. We developed an observation-based reconstruction of global monthly lightning density at 0.25° resolution for the period 1979–2025. Leveraging ERA5 reanalysis data, we employed an ensemble approach—integrating XGBoost, RF, LightGBM, and DNN—trained against stable WWLLN observations (2013–2024). These models were blended through a ridge regression ensemble to minimize individual algorithmic biases and enhance spatial generalizability. Additionally, we applied the SHapley Additive exPlanations (SHAP) algorithm to provide a physically-grounded interpretation of the key drivers of lightning variability. This study delivered: (i) the longest high-resolution (0.25°) global gridded lightning dataset currently available, and (ii) a systematic analysis of environmental modulators to support wildfire modeling, atmospheric chemistry, and climate change assessment.

2 Data

2.1 Meteorological data

Meteorological predictors were obtained from the ECMWF Reanalysis v5 (ERA5), which incorporated significant advancements in model physics, dynamics, and 4D-Var data assimilation (Hersbach et al., 2020). In this study, we extracted monthly averaged data on both single levels and pressure levels from January 1979–December 2025 at a native spatial resolution of 0.25° × 0.25°. To comprehensively capture the multi-dimensional and non-linear drivers of storm electrification, we selected a suite of 30 environmental predictors and auxiliary descriptors, categorized into functional groups of thermodynamic instability, dynamical forcing, precipitation components, and moisture/cloud microphysics (Figure 1 and

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Tables S1). These included fundamental metrics such as CAPE and the CAPE × TP scaling proxy, mid-tropospheric vertical motion at 500 hPa, and column-integrated hydrometeor paths. The long-term stability and high fidelity of ERA5 across the post-1979 satellite era provided a robust physical foundation for our historical lightning reconstruction.

2.2 Lightning data for model training

Target lightning observations were derived from the WWLLN WGLC and time series dataset (Kaplan and Lau, 2022). WGLC reprocessed raw WWLLN stroke detections by applying detection-efficiency (DE) corrections and aggregating them into gridded density products. We utilized the monthly 5 arcmin stroke-density time series, which was conservatively aggregated to a 0.25° grid to align with ERA5 predictors. Although WGLC spanned 2010–2024, the 2010–2012 period remained influenced by the progressive build-out of the global sensor network; accordingly, we selected the 2013–2024 period for model training and evaluation to ensure a stable climatological baseline and minimize early-record inhomogeneity.

2.3 Existing lightning products for comparison

To provide an independent, large-scale benchmark for evaluating the spatial climatology of our reconstruction, we used the LIS/OTD 2.5 Degree Low Resolution Monthly Climatology Time Series (LRMTS) V2.3.2015, distributed by NASA's Global Hydrometeorology Resource Center DAAC (GHRC DAAC). This product merged observations from two spaceborne lightning sensors: the Optical Transient Detector (OTD) on Orbview-1 and the Lightning Imaging Sensor (LIS) onboard the TRMM satellite, and provided a monthly gridded total lightning flash-rate density series. While highly robust in the tropics and subtropics (±38°) where the LIS record was longest, high-latitude information in LRMTS was derived solely from OTD and should be interpreted with caution. This dataset served as a foundational reference for assessing the spatial fidelity of our high-resolution product.

3 Method

Figure 1 summarized the methodological framework used to reconstruct the global 0.25° monthly land lightning density dataset (1979–2025) and to analyze its primary drivers. The workflow began with the compilation of ERA5 meteorological predictors for 1979–2025 and WWLLN lightning density observations for 2013–2024, both at monthly resolution. All data underwent standardized preprocessing, including regridding, land masking, and the removal of missing values (NaNs). The processed samples

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Open Access Earth System Science Data Discussions

from the WWLLN-observation period (2013–2024) were randomly partitioned into training (80%) and testing (20%) sets to facilitate model development and independent evaluation. We employed four machine learning algorithms —XGBoost, RF, LightGBM, and DNN—as base learners. Their predictions were subsequently integrated via ridge regression stacking to generate the Global Land Lightning Density Reconstruction, Version 1 (GLLDR v1), at 0.25° monthly resolution. Finally, SHAP values and feature-importance metrics were applied to identify and quantify the key environmental factors influencing global lightning patterns.

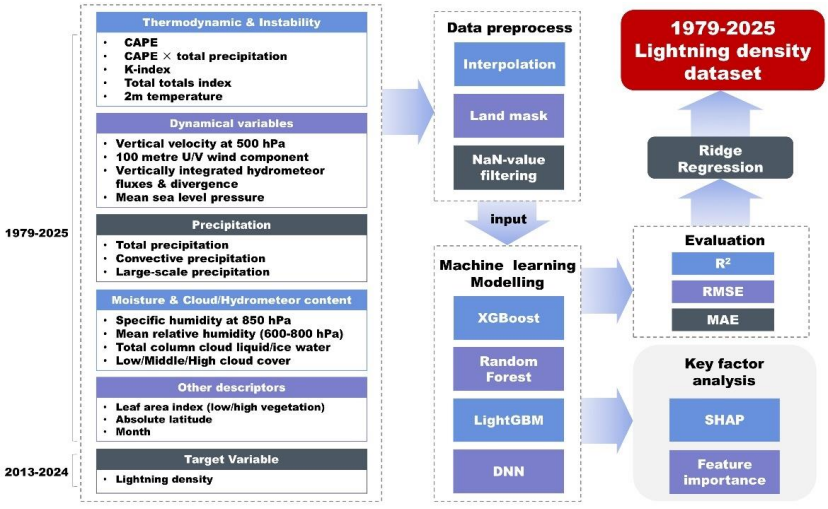


Figure 1. Workflow for the generation and analysis of the 0.25° global monthly land lightning density dataset (1979–2025).

3.1 Data processing

To construct a comprehensive predictor set, we selected a suite of ERA5 variables that represent the core environmental drivers of lightning, encompassing thermodynamics and atmospheric instability, dynamical fields, precipitation characteristics, moisture and cloud/hydrometeor properties, and several other descriptors. Beyond the native ERA5 variables, two physically-informed predictors were derived: the product of CAPE and total precipitation (CAPE × TP) and the mean relative humidity over the 600–800 hPa layer. The month-of-year was encoded using a cyclical transformation, and absolute latitude was included to account for latitudinal gradients. All input variables were summarized in Table S1 and Figure 1.

To ensure spatial consistency between observational and reanalysis data, the WWLLN lightning stroke

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

density (originally on a 5-arcmin grid) was regridded to the 0.25° × 0.25° ERA5 grid using area-weighted averaging. This conservative aggregation method preserved the areal mean within each target grid cell and minimized biases arising from sub-grid scale variability.

The analysis was restricted to continental regions; oceanic grid cells were masked, and Antarctica and Greenland were excluded due to the extreme rarity of lightning, providing insufficient constraints for model training and may introduce noise during evaluation. Following spatial interpolation and masking, samples lacking target values were removed to ensure the integrity of the supervised learning process. Strict data management was implemented during the random split (80% training, 20% testing) to prevent data leakage, ensuring that no information from the evaluation set influenced the model training phase.

3.2 Machine learning models

Machine learning (ML) has become a robust tool for Earth system estimation, excelling in tasks such as land-cover classification (Candido et al., 2021), land surface temperature prediction (Li et al., 2024), and soil moisture mapping (Zhang et al., 2023). To capture complex nonlinearities and leverage the complementary strengths of different architectures, we employed four base models: XGBoost, RF, LightGBM, and DNN.

XGBoost and LightGBM are both Gradient Boosting Decision Tree (GBDT) variants that construct additive ensembles in a stage-wise manner. XGBoost utilizes shrinkage and explicit regularization to mitigate overfitting (Chen and Guestrin, 2016), while LightGBM is optimized for large-scale datasets through histogram-based splitting and streamlined tree growth (Ke et al., 2017). In contrast, RF is a bagging-based ensemble that reduces predictive variance by averaging results from multiple independent trees grown via bootstrap resampling (Breiman, 2001). While tree-based ensembles are highly robust for interpolation, they can be limited in extrapolation (Hateffard et al., 2024). To address this, we incorporated a DNN, which excels at feature learning and generalization (Zhang et al., 2016). Integrating DNN with tree-based models within a stacking framework enhanced the overall robustness of the final estimates. Hyperparameters for all models were optimized using the Optuna automated framework (Akiba et al., 2019), selecting configurations that minimized validation error. In addition, to assess the robustness of model performance to data partitioning, we conducted a five-fold cross-validation experiment for the four base models over the 2013–2024 sample pool.

3.3 Model evaluation

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Model performance in reproducing monthly lightning density was quantified using several statistical metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). These metrics are defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (3)$$

where n is the sample size; y_i and $\hat{y}_i$ represent the observed and predicted values, respectively; and $\bar{y}$ is the mean of the observations.

3.4 Ridge regression stacking

To synthesize the strengths of the four base models, ridge regression was employed as a stacking meta-learner (Wolpert, 1992; Breiman, 1996). This meta-model learns the optimal linear combination of base-learner outputs while applying L_2 regularization to shrink coefficients, thereby reducing overfitting and enhancing generalization (Hoerl and Kennard, 1970).

Given that lightning density is physically constrained to be non-negative, all negative base-model predictions were truncated to zero prior to stacking. The ridge regularization parameter (α) was determined via 10-fold cross-validation on the training set. To ensure physical interpretability and prevent artificial compensation (sign-reversal) between models, we imposed a non-negativity constraint on the stacking weights (Breiman, 1996). The final ensemble output was similarly truncated at zero to ensure a physically consistent global dataset.

3.5 Driver analysis (Key factors)

To interpret the environmental factors of lightning, we utilized the SHAP algorithm, which provides a game-theoretic approach to quantifying predictor importance (Lundberg and Lee, 2017). SHAP values were calculated at the sample level; the sign indicates the direction of influence (increasing or decreasing the prediction relative to the baseline), the magnitude reflected the contribution strength. SHAP dependence plots were generated for the top-8 predictors to visualize how their contributions varied across their respective ranges. Additionally, traditional feature-importance rankings were computed as a

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

complementary measure of predictor relevance.

4 Results

4.1 Evaluation of model performance

We computed evaluation metrics for the four base learners and the ridge regression-based stacking ensemble using both training and independent test datasets to quantitatively assess model performance. Notably, all negative predictions were truncated to zero prior to metric calculation to maintain physical consistency with lightning occurrence. We first evaluated the predicted–observed scatterplots and associated R^2 values (Fig. 2), followed by a comprehensive quantitative assessment using RMSE and MAE (Table 1). Generally, all models captured the monotonic relationship between predicted and observed lightning density. However, increased dispersion was observed at higher densities, implying higher uncertainty in intense-lightning regimes—a phenomenon likely attributable to the inherent stochasticity and sub-grid variability of deep convection.

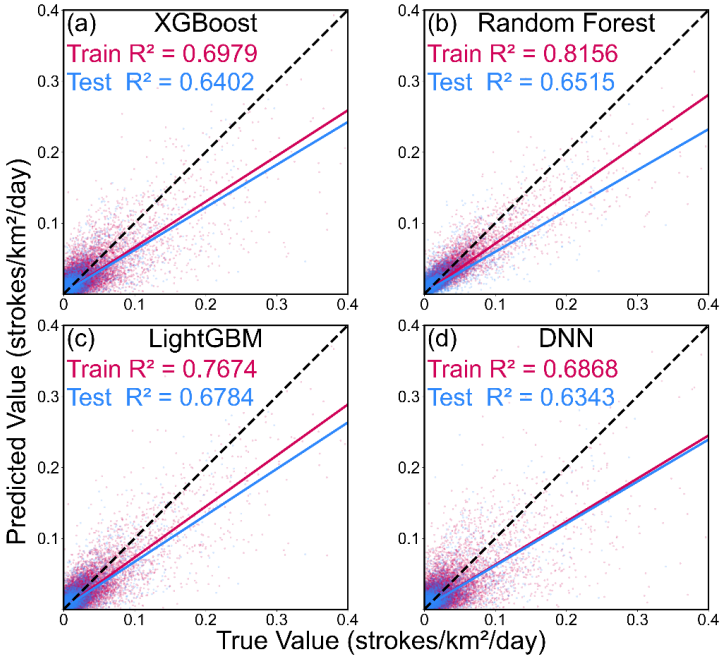


Figure 2. Scatterplots of predicted versus observed lightning density for the four machine-learning models. For visual clarify, a stratified random sampling strategy based on lightning density quantiles was employed, while the regression lines were fitted using the complete dataset.

Among the four base models, LightGBM demonstrated the optimal balance between predictive skill and

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Open Access Earth System Science Data Discussions

generalization, yielding the highest test R^2 (0.6784) and the minimum RMSE (0.0110), alongside a relatively low MAE (0.0032). While RF performed competitively, its performance gap between training and testing (R^2 = 0.8156 vs 0.6515) suggested a higher susceptibility to overfitting. In contrast, XGBoost maintained robust performance (R^2 = 0.6402, RMSE = 0.0116, and MAE = 0.0033), while DNN exhibited the most limited generalization capability, with the lowest test R^2 of 0.6343. Ultimately, the ridge regression ensemble outperformed all individual models on the independent test set (R^2 = 0.6895, RMSE = 0.0108), underscoring that the stacking strategy effectively mitigated individual model biases and enhanced the overall reliability of lightning density estimations. Supplementary five-fold cross-validation results for the four base models showed consistent model ranking and limited fold-to-fold variability (Table S2), further supporting the robustness of the hold-out test results.

Table 1. Evaluation metrics for the four base models (XGBoost, RF, LightGBM, and DNN) and their ridge regression stacking ensemble across training and test datasets.

Model	R^2		RMSE		MAE	
	train	test	train	test	train	test
XGBoost	0.6979	0.6402	0.0106	0.0116	0.0032	0.0033
RF	0.8156	0.6515	0.0083	0.0114	0.0020	0.0030
LightGBM	0.7674	0.6784	0.0093	0.0110	0.0029	0.0032
DNN	0.6868	0.6343	0.0108	0.0116	0.0034	0.0035
Ridge regression	0.8101	0.6895	0.0084	0.0108	0.0024	0.0030

We subsequently evaluated the spatial consistency of model performance across the test set to verify if the ensemble's superiority persisted across diverse geographical regions. Broadly, all five models exhibited coherent spatial patterns of grid-cell R^2 (Figs. 3a–e), with higher predictive skill concentrated over North America, Eurasia, and extensive tropical regions. Despite these qualitative similarities, the ensemble model demonstrated a distinct quantitative advantage in R^2 distributions (Fig. 3f). Specifically, 33.9% of global grid cells for the ensemble achieved R^2 values between 0.6 and 1.0, a preponderance that substantially exceeded those of XGBoost (20.6%), LightGBM (24.4%), and DNN (18.8%). RMSE diagnostics further characterized the spatial structure of estimation errors (Fig. S1). Errors remained consistently low for all models, with over 65% of grid cells yielding RMSE values within 0–0.005; for

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

the ensemble, this proportion increased to 69.0% (Fig. S1f). Notably, the ensemble exhibited a sharply truncated high-error tail, with only 5.1% of grid cells exceeding an RMSE of 0.02. Integrated with the global validation metrics (Table 1), these spatial diagnostics indicated that the stacking ensemble provided superior spatial robustness by effectively integrating the complementary strengths of individual learners, thereby justifying its selection for the final lightning reconstruction.

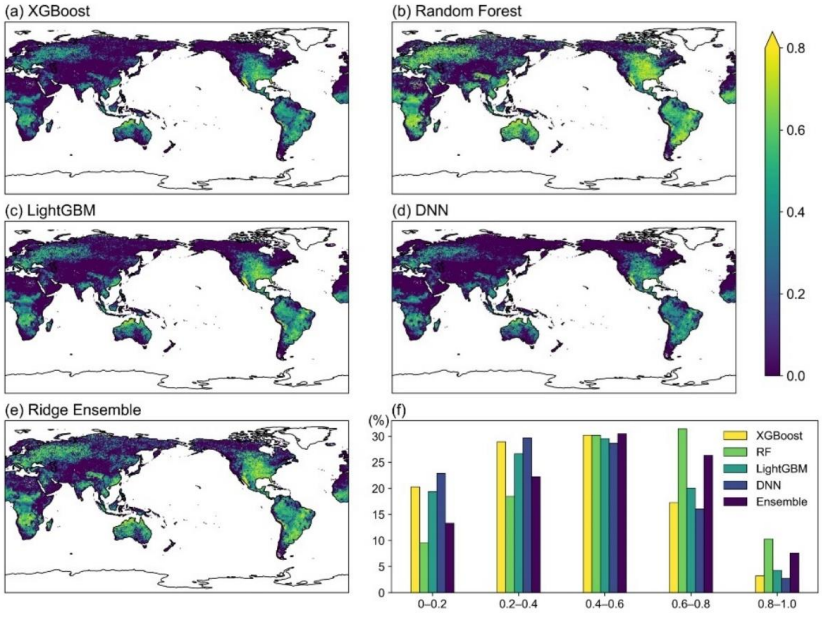


Figure 3. Spatial distribution of the coefficient of determination (R^2) for the four base models—**(a)** XGBoost, **(b)** RF, **(c)** LightGBM, **(d)** DNN—and **(e)** ridge regression stacking ensemble across the test datasets. The accompanying bar chart **(f)** shows the percentage of global grid cells falling within specific R^2 intervals for each model.

4.2 Reconstructed global lightning density

Based on the superior 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). The observed lightning-density climatology was characterized by major hotspots over the tropical Americas, equatorial Africa, South and Southeast Asia, and northern Australia (Fig. 4b). These major features were well-reproduced by the ridge regression ensemble (Fig. 4a), indicating that the reconstruction captured the principal global patterns of lightning activity. The bias map revealed that the systematic errors were not spatially uniform: overestimation was primarily concentrated in the low latitudes, particularly over

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

the tropical Americas, where deep convection and complex orographic–land–sea contrasts might enhance lightning-producing conditions (Fig. 4c). In contrast, negative biases were generally weaker and localized over parts of mid-latitude North America and the high-latitude Eurasia. Zonal-mean comparisons confirmed the close agreement between the reconstruction and observations ($r = 0.9969$, $p < 0.001$; Fig. 4d), with the dominant error being a tropical high bias within approximately ±20° latitude (Fig. 4e). These results demonstrated that the ensemble reconstruction reproduced the observed large-scale spatial distribution with high fidelity during the overlap period, providing a robust basis for extending the lightning density into earlier decades.

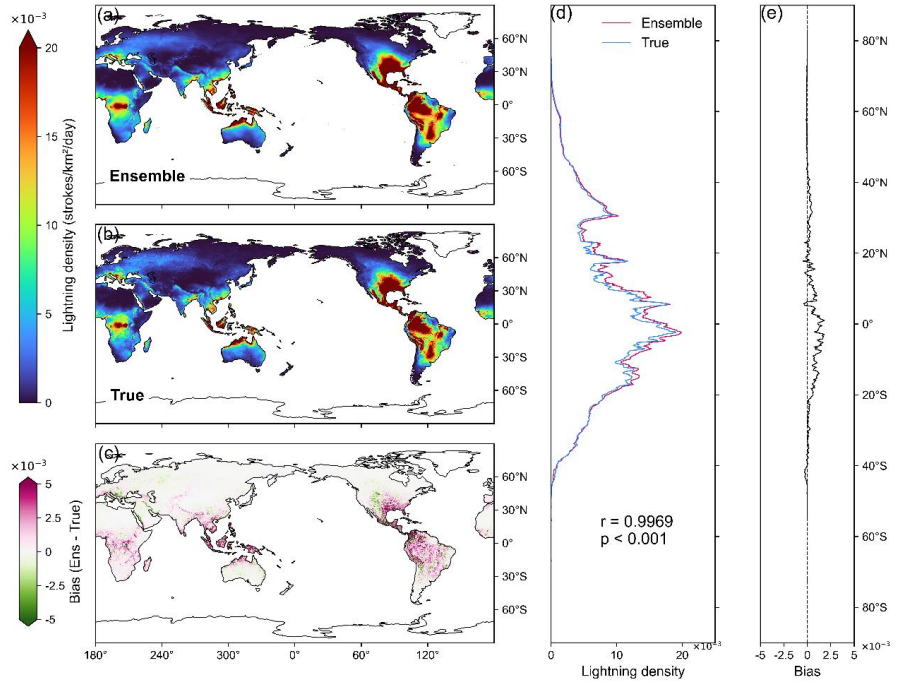


Figure 4. Spatial distributions of lightning density for **(a)** the ridge regression ensemble and **(b)** observations, alongside **(c)** the associated bias during 2013–2024. The corresponding zonal mean lightning densities for the ensemble and observations are shown in **(d)**, with the Pearson correlation coefficient and p-value indicated within the panel. The zonal mean bias is shown in **(e)**. Lightning density is expressed in strokes/km^2/day.

Having established its fidelity, we examined the reconstructed climatology for the 1979–2012 period

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

(Fig. 5). The multi-year mean lightning density for this earlier period retained the major hotspots seen in recent observations (Fig. 5a). The zonal-mean lightning density likewise exhibited a clear tropical maximum and a systematic decline toward higher latitudes (Fig. 5c), consistent with the principal meridional structure identified during 2013–2024. However, the difference field between 1979–2012 and 2013–2024 revealed that the earlier period was characterized by predominantly positive anomalies over large portions of the global land area, especially across South America, central and southern Africa, Southeast Asia, and northern Australia. Conversely, negative anomalies were more limited, occurring primarily over the southern United States (Fig. 5b). The zonal-mean differences further indicated that the positive anomalies were concentrated between the equator and about 45°S, while negative anomalies were most evident near 30°N (Fig. 5d). These findings suggested that while the large-scale spatial structure of lightning density remained broadly stable, noticeable regional and meridional shifts occurred between the two periods.

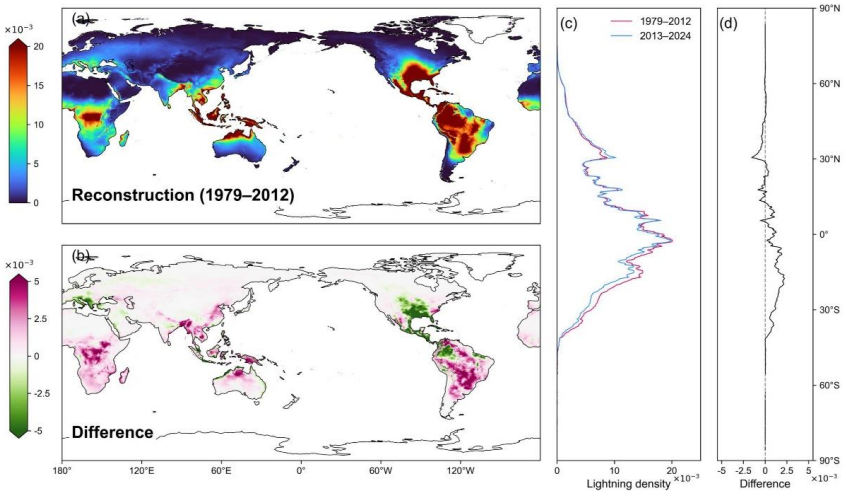


Figure 5. Global distribution and zonal-mean structure of the reconstructed lightning density. **(a)** Multi-year mean global lightning density from the ensemble reconstruction during 1979–2012. **(b)** Difference in multi-year mean lightning density between 1979–2012 and 2013–2024. **(c)** Zonal-mean lightning density for the two periods, and **(d)** the corresponding zonal-mean difference. The unit of the lightning density is strokes/km^2/day.

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

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

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 ($|lat| \leq 30°$) and the extratropical Northern Hemisphere ($30° < lat \leq 90°$), respectively. In the extratropical Northern Hemisphere, the reconstructed trend (slope = 0.63×10^{-4}, $p < 0.05$) was in excellent agreement with the observed trend (slope = 0.69×10^{-4}, $p < 0.1$), reflecting the high reliability of the ensemble in extratropical regions (Fig. 6c). Within the Tropics, while the reconstruction successfully captured the interannual fluctuations, the reconstructed trend (slope = 1.03×10^{-4}, $p < 0.05$) exhibited a slight deviation from the observations (slope = 1.98×10^{-4}, $p < 0.1$), likely due to the inherent stochasticity of intense tropical convection and the associated overestimation bias (Fig. 6b). 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). At the global scale, both the reconstruction (slope = 0.71×10^{-4}, $p < 0.05$) and the observations (slope = 1.24×10^{-4}, $p < 0.1$) showed weak upward trends during 2013–2024, which indicated that the ensemble reproduced the overall direction of recent global change. 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.

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

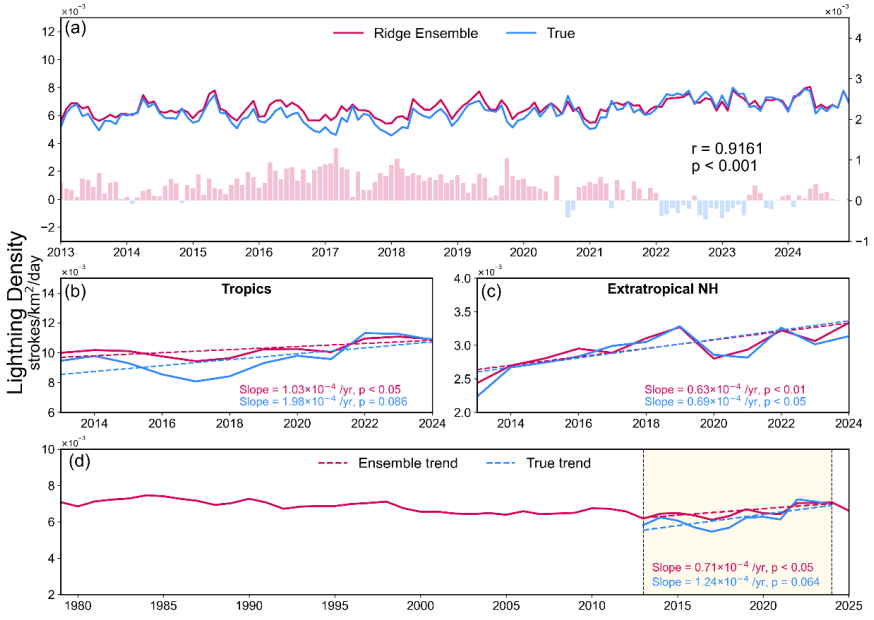


Figure 6. (a) Monthly global mean lightning density during 2013–2024 predicted by the ridge regression ensemble and observations, with residuals shown as bars. **(b)** and **(c)** present annual mean lightning density during 2013–2024 over the tropics (|lat| ≤ 30°) and the extratropical Northern Hemisphere (30° < lat ≤ 90°), 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 for 2013–2024 are shown for both the ensemble and the observations, together with the corresponding Mann–Kendall p values.

Using the reconstructed monthly lightning density for the entire 1979–2025 period, we investigated long-term changes by estimating grid-cell trends in annual lightning density using the Theil–Sen estimator, with statistical significance assessed by the Mann–Kendall test (Fig. 7). The resulting trend map exhibited pronounced regional heterogeneity across the globe. Broad and statistically significant negative trends dominated several major convective hotspots, including extensive areas of South America, central-to-southern Africa, Southeast Asia, and the Maritime Continent, with additional decreases observed over northern Australia. In North America, a marked decreasing band was evident across the western United States and Mexico, whereas positive trends were more localized and emerged primarily over the central United States. Over Europe, a coherent and significant increasing trend was observed over southern and

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

central Europe, particularly around the Mediterranean region. Collectively, these patterns suggested a general tendency toward decreasing lightning activity across several primary tropical land convection centers, contrasted by localized increases in specific mid-latitude regions. The clustering of significant trends within these key regions underscored strong regional contrasts in multi-decadal lightning dynamics, providing a comprehensive view of how global lightning activity evolved over the past nearly five decades.

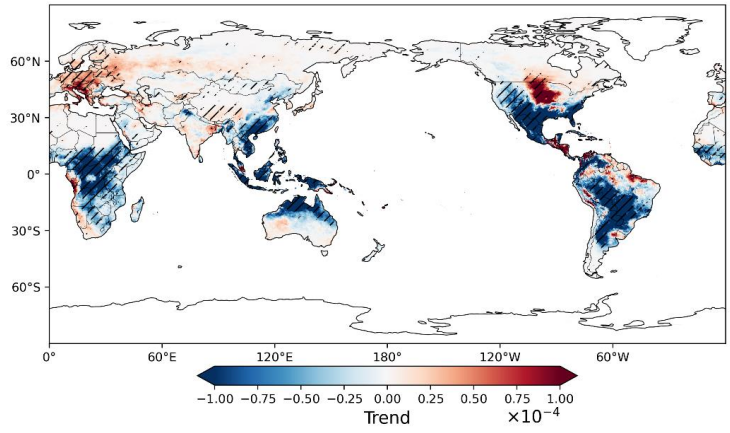


Figure 7. Spatial distribution of Theil–Sen trends in annual lightning density during 1979–2025. Hatched areas indicate regions where trends were significant at $p < 0.05$ based on the Mann–Kendall test.

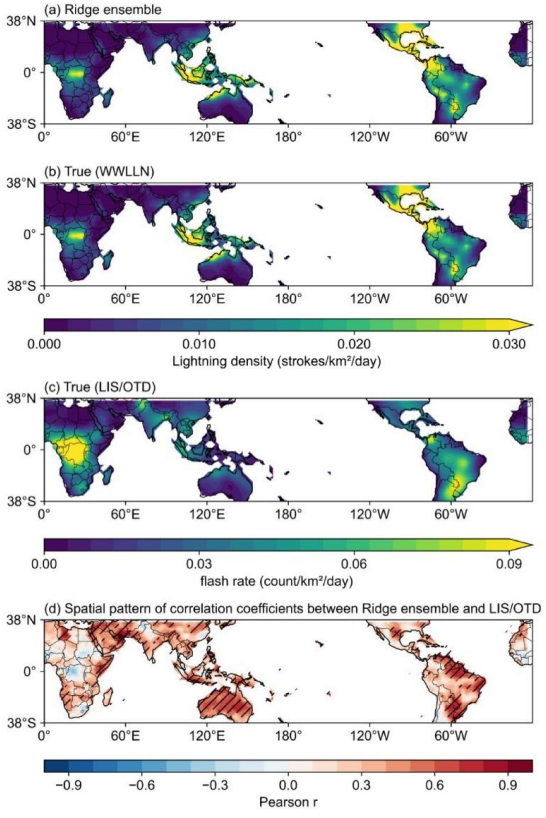
4.3 Comparison with other coarser products

To provide an external benchmark for our reconstruction, we selected the satellite-based LIS/OTD gridded lightning product as an independent reference. As described in Sect. 2.3, the merged LIS/OTD climatology was mainly available within the tropics and subtropics; thus, we restricted the comparison to the 38°S–38°N latitudinal band, where LIS sampling was robust and extensive.

Because LIS/OTD reported flash-rate density, whereas WWLLN and the ridge regression ensemble trained on it represented stroke density, direct agreement in absolute magnitude was not expected. We therefore emphasized (i) the relative spatial patterns of mean lightning activity and (ii) the consistency in temporal variability. Figures 8a–c show the monthly-mean lightning density during the 2013–2014 overlap period between WWLLN and LIS/OTD. To evaluate the agreement in variability, Fig. 8d presented grid-cell Pearson correlations between the annual mean time series of the ridge regression ensemble and LIS/OTD over the 1996–2014 period.

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

373 The ensemble reproduced the primary spatial structures present in WWLLN during 2013–2014 (Figs.
374 8a–b), which was expected given that WWLLN provided the observational constraint for model training.
375 When compared with LIS/OTD (Fig. 8c), the ensemble also captured broad tropical and subtropical
376 lightning hotspots, such as major continental convective centers, although differences in amplitude and
377 local contrasts remained evident. These discrepancies were consistent with the fundamental definition
378 difference between flashes (the complete discharge) and strokes (individual components within a flash),
379 as well as the distinct sampling/detection characteristics of spaceborne optical sensors versus ground-
380 based very-low-frequency (VLF) networks (Kaplan and Lau, 2021, 2022; Cecil et al., 2014; Rudlosky
381 and Shea, 2013). The correlation map (Fig. 8d) indicated that the ensemble exhibited predominantly
382 positive correlations with LIS/OTD across extensive tropical and subtropical regions, with many regions
383 reaching statistical significance (hatched). This suggested that, despite differing reported lightning metric,
384 the ensemble captured a substantial fraction of the interannual variability observed by independent
385 satellite sensors over key convective regions.



386

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Figure 8. Comparison of monthly mean lightning density within the 38°S–38°N latitudinal band during 2013–2014: **(a)** ridge regression ensemble, **(b)** WWLLN, and **(c)** LIS/OTD. Panel **(d)** shows the spatial pattern of Pearson correlation coefficients between the ridge regression ensemble and LIS/OTD based on the 1996–2014 annual mean time series. Hatched areas indicate regions where correlations were significant at $p < 0.05$.

4.4 Key factors analysis

We further examined the dominant environmental controls that the machine-learning models relied on to predict monthly lightning density. Given that SHAP analysis is particularly computationally efficient and well-suited for GBDT architectures, we conducted this attribution analysis specifically for XGBoost and LightGBM as representative boosting-based models. In addition, the feature importance of RF was evaluated through its built-in impurity-based metrics to provide a supplementary comparison. By interpreting these model attribution patterns, we aimed to clarify which large-scale conditions most consistently promoted or suppressed lightning activity in our reconstruction.

Figure 9 summarized the top eight predictors ranked by mean absolute SHAP values for XGBoost and LightGBM, identifying a highly consistent set of leading controls. The combined thermodynamic–precipitation indicator, CAPE × TP, was the dominant predictor in both models (34.4% for XGBoost; 28.2% for LightGBM), followed by CAPE (14.1% and 17.9%) and total column ice water (12.3% and 12.5%). Together, these three variables accounted for roughly 60% of the total attribution, indicating that the models primarily represented lightning variability through a coupled signal of atmospheric instability, convective triggering, and the availability of ice-phase hydrometeors relevant to storm electrification. The RF feature-importance ranking (Fig. S2) also placed CAPE × TP, CAPE, and convective precipitation among the leading predictors, alongside absolute latitude, confirming that the primary controls inferred from SHAP were not model-specific. RF further highlighted low-level moisture indicators (e.g. specific humidity at 850 hPa), complementing the GBDT-based emphasis on ice-related and precipitation-regime descriptors.

The SHAP summary plots further revealed the directionality of these effects. Higher values of CAPE × TP, CAPE, total column ice water, and convective precipitation were predominantly associated with positive SHAP contributions, implying enhanced lightning density under more unstable and convectively active conditions with stronger mixed-phase processes. In contrast, large-scale precipitation and medium

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

cloud cover tended to exhibit negative contributions at higher values, suggesting that environments characterized by stratiform precipitation or widespread mid-level cloudiness were generally less favorable for intense lightning activity. Absolute latitude showed a clear background modulation: lower latitudes were associated with positive contributions, while higher latitudes contributed weakly or negatively, consistent with the climatological confinement of frequent lightning to the tropics and subtropics.

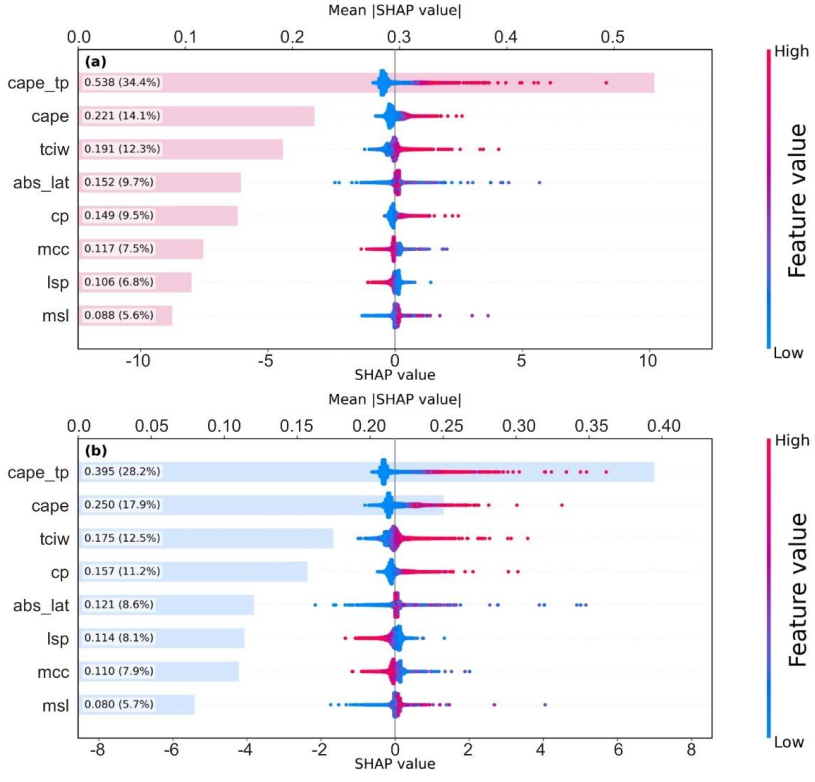


Figure 9. Key predictors of lightning density. SHAP summary plots illustrating the top eight variables for **(a)** XGBoost and **(b)** LightGBM. The colors represent the feature values (relative magnitude), while the horizontal bars denote the mean absolute SHAP values, indicating global feature importance.

Figure 10 provides more detailed response shapes using SHAP dependence plots from the XGBoost model. A similar analysis conducted for LightGBM yielded highly consistent results, and the corresponding dependence plots were provided in the Supplementary Information (Fig. S3). The relationships were strongly nonlinear for several key predictors. CAPE × TP and CAPE showed rapid

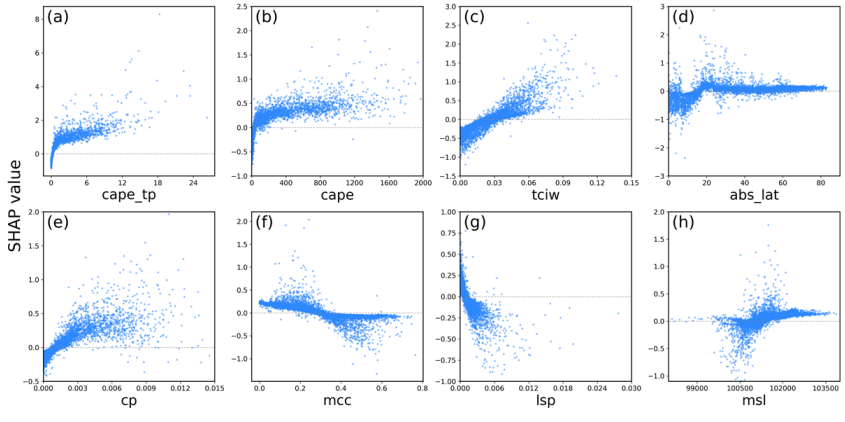
https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Open Access Earth System Science Data Discussions

increases in positive contributions from low to moderate values, followed by a tendency toward saturation at higher ranges. This pattern indicated that instability and precipitation activity were necessary but not sufficient: once a threshold was exceeded, additional increases yielded diminishing marginal contributions. This finding was consistent with the notion that other constraints such as moisture availability, storm organization, and microphysical state, increasingly regulated lightning intensity. The dependence for total column ice water was largely monotonic and positive, highlighting the importance of ice-phase hydrometeors in the electrification process.

Meanwhile, absolute latitude exhibited a transition from strong positive contributions at low latitudes to near-zero (or slightly negative) contributions at higher latitudes, reflecting a robust climatological background control. For variables linked to precipitation regime and cloud structure, the dependence plots showed consistent suppression signals. Larger large-scale precipitation was associated with increasingly negative contributions, indicating that a greater stratiform precipitation fraction was generally unfavorable for lightning-rich convection. Higher medium cloud cover also tended to reduce predicted lightning density, which was consistent with mid-level cloudiness acting as an indicator of broader, less electrification-efficient cloud systems.

The appreciable vertical spread of SHAP values at any given predictor value suggested that the effect of an individual predictor was modulated by other environmental variables, reinforcing the necessity for using multivariate predictors in the reconstruction. Taken together, the three model families consistently supported a physically coherent control chain in which lightning density was most strongly regulated by the coupling of atmospheric instability, convective triggering, moisture supply, and ice-phase microphysical conditions, all modulated by the latitude-dependent climatological background.



https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

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.

5 Discussion

Despite the availability of several lightning products, long-term, high-resolution, and globally consistent gridded lightning datasets remained limited. To bridge this gap, we developed an observation-constrained, machine-learning reconstruction of global monthly land lightning density at 0.25° resolution for 1979–2025, using a ridge regression stacking ensemble to enhance robustness. Beyond the dataset itself, we provided a key factors analysis that elucidated the dominant environmental controls identified by the model, thereby enhancing its physical interpretability.

5.1 Advantages of this reconstruction

A key advantage of our dataset was that it provided a long-term (1979–2025), fine-resolution (0.25°), and globally complete monthly lightning density record. This addressed the dual limitations of short temporal coverage in existing ground-based gridded products and the coarser spatial resolution and sampling constraints in widely used satellite climatologies. The ridge stacking framework improved robustness by integrating the complementary strengths of multiple models, yielding the best overall generalization among the tested models (test R^2 = 0.6895; RMSE = 0.0108; MAE = 0.0030). This performance, together with the spatially explicit R^2 diagnostics, supported the use of the ensemble reconstruction as the final product rather than any single base model.

During the observational overlap period (2013–2024), the ensemble reproduced the observed global-mean seasonal cycle and much of the short-term variability, while capturing the principal spatial climatology and major lightning hotspots. Importantly, the evaluation highlighted a regime-dependent error structure: dispersion increased at high lightning densities, consistent with stronger stochasticity and sub-grid variability in intense convection, while regional errors were smaller in many mid-latitude areas than in the tropics. These diagnostics provided practical guidance for interpretation, suggesting that the reconstruction was most reliable for representing large-scale spatiotemporal variability, while tropical intense-convection regimes remained comparatively more challenging to capture.

Comparison with LIS/OTD further supported the plausibility of the reconstruction beyond the WWLLN constraint. Although direct agreement in magnitude was not expected because LIS/OTD reported flash-

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

rate density whereas WWLLN (and our model) represented stroke density, the reconstructed fields captured the broad tropical and subtropical hotspots seen in LIS/OTD. Furthermore, the reconstruction exhibited predominantly positive correlations with LIS/OTD variability across extensive tropical and subtropical regions during the overlap period. In this sense, the reconstruction could be viewed as a physically grounded extension that preserved large-scale lightning variability signals at substantially finer spatial resolution and over a longer period than existing gridded benchmarks.

Finally, the accompanying attribution analyses reinforced the physical interpretability of the reconstruction. Across model families, the leading controls consistently emphasized coupled instability–precipitation forcing, ice-phase hydrometeor availability, and precipitation-regime descriptors, all modulated by latitude-dependent background conditions. The nonlinear and saturating response shapes in SHAP dependence plots were consistent with a regime view in which instability and convective activity were necessary but increasingly conditioned by moisture, storm organization, and microphysics once certain thresholds were exceeded. Together, these results indicated that the reconstructed dataset was not only statistically skillful but also aligned with a coherent set of large-scale environmental constraints on lightning.

5.2 Implications and potential applications

Beyond its methodological advantages, the GLLDR v1 dataset also had broad scientific value for studies of climate, atmospheric chemistry, and wildfire processes. By providing a spatially continuous, multi-decadal, and fine-resolution record of monthly lightning density, it expanded the opportunities for investigating long-term changes in lightning activity under climate variability and global warming. In particular, the dataset could support analyses of how lightning responds to changes in thermodynamic instability, moisture availability, and large-scale circulation across different climate regimes, thereby providing a useful observationally constrained benchmark for studies of lightning–climate relationships.

The dataset was also potentially valuable for atmospheric chemistry applications. Lightning is an important natural source of nitrogen oxides (NOx) in the free troposphere and therefore influences ozone formation, oxidizing capacity, and broader atmospheric chemical cycling (Schumann and Huntrieser, 2007; Mao et al., 2021). A long-term gridded lightning dataset could help constrain the spatial and temporal distribution of lightning-produced NOx emissions and improve their representation in chemical transport models and Earth system models. In this sense, GLLDR v1 may provide a useful data basis for

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

linking long-term changes in lightning activity with variability in atmospheric composition.

In addition, the dataset had clear relevance for wildfire research, as lightning is a major natural ignition source and an important driver of fire occurrence (Veraverbeke et al., 2017; Janssen et al., 2023). Temporally continuous and spatially explicit lightning information could improve the representation of natural ignition forcing in fire models, support analyses of lightning–fire coupling, and help assess how climate-driven changes in lightning might alter wildfire risk. More broadly, the dataset could also serve as a benchmark for evaluating lightning parameterizations in climate models and for supporting hazard-related analyses that require long-term historical lightning information.

5.3 Limitations and uncertainty

Despite the encouraging performance, two primary sources of uncertainty were particularly important for this reconstruction framework: (i) WWLLN detection efficiency and observational representativeness, and (ii) ERA5 predictor uncertainties and the inherent limits of large-scale environmental predictors.

WWLLN provided the observational constraint used to train the models, but its detection efficiency was known to vary across space and time, depending on network configuration and signal propagation conditions (Hutchins et al., 2012; Rudlosky and Shea, 2013). Such variability might have imprinted regional biases on the learned mapping, especially in regions where lightning characteristics and detection conditions differed substantially. In addition, the fundamental metric mismatch between stroke density (WWLLN) and flash-rate density (LIS/OTD) implied that discrepancies in magnitude and local contrasts were expected even when variability aligned. These differences were treated as a structural source of uncertainty rather than a simple model error. The tropical high-bias pattern diagnosed during 2013–2024 also indicated that systematic errors tended to concentrate in high-convection regimes, where both detection characteristics and convective heterogeneity were most pronounced.

Moreover, the reconstruction was ultimately driven by ERA5-based predictors, and variables tied to convection and microphysics were subject to reanalysis uncertainties (Hersbach et al., 2020). For convective parameters specifically, the reliance on convective parameterization at reanalysis resolution could introduce biases in thermodynamic indices such as CAPE (Taszarek et al., 2021). These uncertainties could propagate into the learned relationships and might have contributed to regional error structures, particularly in the tropics where lightning was tightly linked to localized convective organization, a process only indirectly represented by monthly mean large-scale predictors.

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

Furthermore, applying relationships learned over the 2013–2024 period to earlier decades might have been affected by shifts in the joint distribution of predictors over time (Sugiyama et al., 2007). Finally, while ridge regularization yielded lower-variance estimates, it was associated with a known shrinkage effect that might have slightly dampened the amplitude of variability and extremes. This shrinkage was taken into account when interpreting the reconstructed long-term changes (Hoerl and Kennard, 1970). Future improvements could involve incorporating additional independent constraints, such as alternative satellite-era sensors or regional networks, to better diagnose and mitigate region-dependent biases. In addition, future efforts might include: (i) time-split or regime-split validation to quantify sensitivity to climate background shifts; (ii) perturbation-based sensitivity tests for key predictors; and (iii) expanded uncertainty estimates based on ensemble spread or alternative reanalysis inputs. More broadly, incorporating predictors that better represented storm organization (where feasible) and evaluating performance across distinct convective regimes would help refine the interpretation of trends and regional changes.

6 Data availability

Data described in this manuscript can be accessed at Zenodo under https://doi.org/10.5281/zenodo.19722380 (Zheng et al., 2026a). The ERA5 monthly meteorological predictors used in this study were obtained from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Monthly averaged data on pressure levels from 1940 to present are available at https://doi.org/10.24381/cds.6860a573, and monthly averaged data on single levels from 1940 to present are available at https://doi.org/10.24381/cds.f17050d7 (last access: 07 April 2026). The lightning observations used for model training and evaluation were obtained from the World Wide Lightning Location Network (WWLLN) Global Lightning Climatology (WGLC) monthly time-series dataset. We used the monthly 5 arcmin stroke-density product and conservatively aggregated it to a 0.25° grid to match the ERA5 predictors (Kaplan and Lau, 2021, 2022). The LIS/OTD 2.5 Degree Low Resolution Monthly Climatology Time Series (LRMTS) V2.3.2015 used for independent comparison is distributed by NASA's Global Hydrometeorology Resource Center DAAC (GHRC DAAC) and is available at https://doi.org/10.5067/LIS/LIS-OTD/DATA309 (last access: 07 April 2026).

7 Code availability

The python code used to create the figures included in this paper is provided at

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

https://doi.org/10.5281/zenodo.19723880 (Zheng et al., 2026b).

8 Conclusions

This study delivered a new global, high-resolution (0.25°) monthly lightning stroke-density reconstruction for the 1979–2025 period, filling a long-standing gap in spatially continuous, multi-decadal lightning information. By providing a record that extended beyond the satellite-era gridded products and the limited duration of ground-based constraints, this work established a robust, long-term baseline for lightning-related hazard and impact studies. In addition, we provided a model-consistent attribution of the dominant environmental controls, enabling a physically interpretable use of the reconstruction and advancing the understanding of lightning spatiotemporal variability and its underlying drivers.

During the evaluation period (2013–2024), the multi-model stacking ensemble demonstrated robust predictive skill and reproduced the leading spatiotemporal structures constrained by WWLLN, supporting its application to the earlier decades. 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 diagnosing where lightning-related hazards and convective activity shifted over the past nearly five decades and for benchmarking climate-model simulations against a spatially explicit historical reference.

The key factors attribution further strengthened the utility of the dataset. By identifying a coherent set of leading controls—primarily the coupling of thermodynamic instability (CAPE × TP), moisture, and ice-phase hydrometeor availability—and by revealing strong nonlinearity and interactions, the analysis not only improved model interpretability but also offered a physical lens for interpreting the reconstructed regional trend contrasts. In this sense, the attribution results acted as a “bridge” between the reconstructed fields and process-oriented questions, connecting changes in lightning density to the broader evolution of the large-scale environment.

Several limitations were noted. WWLLN detection efficiency varied in space and time, and the reconstruction inherited uncertainties from the ERA5 reanalysis associated with the evolving observing system and data assimilation. Future work that incorporates multi-sensor lightning constraints, explicitly accounts for time-varying observational biases, and propagates reanalysis uncertainties into an ensemble

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

framework would further enhance the confidence and broaden the applicability of this global lightning record.

Supplement.

The link to the supplement will be included by Copernicus, if applicable.

Author contributions.

HZ and JW conceived and designed this study. HZ processed the datasets, performed the statistical analysis, and prepared the figures. HZ and JW drafted the manuscript, with all authors contributing to the review and editing process.

Competing interests.

The contact author has declared that none of the authors has any competing interests.

Acknowledgements.

The calculations in this paper have been done on the computing facilities in the High Performance Computing Center of Nanjing University. This research has been supported by the National Natural Science Foundation of China (grant nos. 42475129); and the Tibet science and technology innovation base construction project (XZ202401YD0008).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, Kdd'19: Proceedings of the 25th Acm Sigkdd International Conferencce on Knowledge Discovery and Data Mining, 2623–2631, 10.1145/3292500.3330701, 2019.

Breiman, L.: Stacked regressions, Mach Learn, 24, 49–64, 1996.

Breiman, L.: Random forests, Mach Learn, 45, 5–32, Doi 10.1023/A:1010933404324, 2001.

Burrows, W. R., Price, C., and Wilson, L. J.: Warm season lightning probability prediction for Canada and the northern United States, Weather and Forecasting, 20, 971–988, Doi 10.1175/Waf895.1, 2005.

Candido, C., Blanco, A. C., Medina, J., Gubatanga, E., Santos, A., Ana, R. S., and Reyes, R. B.: Improving the consistency of multi-temporal land cover mapping of Laguna lake watershed using light gradient boosting machine (LightGBM) approach, change detection analysis, and Markov chain, Remote Sens Appl, 23, ARTN 100565

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

10.1016/j.rsase.2021.100565, 2021.

Cavaiola, M., Cassola, F., Sacchetti, D., Ferrari, F., and Mazzino, A.: Hybrid AI-enhanced lightning flash prediction in the medium-range forecast horizon, Nat Commun, 15, ARTN 1188 10.1038/s41467-024-44697-2, 2024.

Cecil, D. J., Buechler, D. E., and Blakeslee, R. J.: Gridded lightning climatology from TRMM-LIS and OTD: Dataset description, Atmos Res, 135, 404–414, 10.1016/j.atmosres.2012.06.028, 2014.

Chen, T. Q. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–794, 10.1145/2939672.2939785, 2016.

Cheng, W. Y., Kim, D., Henderson, S., Ham, Y. G., Kim, J. H., and Holzworth, R. H.: Machine Learning-Based Lightning Parameterizations for the CONUS, Artif Intell Earth S, 3, ARTN e230024 10.1175/AIES-D-23-0024.1, 2024.

Hateffard, F., Steinbuch, L., and Heuvelink, G. B. M.: Evaluating the extrapolation potential of random forest digital soil mapping, Geoderma, 441, ARTN 116740 10.1016/j.geoderma.2023.116740, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q J Roy Meteor Soc, 146, 1999–2049, 10.1002/qj.3803, 2020.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression - Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55–&, Doi 10.1080/00401706.1970.10488634, 1970.

Hutchins, M. L., Holzworth, R. H., Brundell, J. B., and Rodger, C. J.: Relative detection efficiency of the World Wide Lightning Location Network, Radio Sci, 47, Artn Rs6005 10.1029/2012rs005049, 2012.

Janssen, T. A. J., Jones, M. W., Finney, D., Van der Werf, G. R., van Wees, D., Xu, W. X., and Veraverbeke, S.: Extratropical forests increasingly at risk due to lightning fires, Nat Geosci, 16, 1136–+,

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

10.1038/s41561-023-01322-z, 2023.

Kalashnikov, D. A., Abatzoglou, J. T., Loikith, P. C., Nauslar, N. J., Bekris, Y., and Singh, D.: Lightning-Ignited Wildfires in the Western United States: Ignition Precipitation and Associated Environmental Conditions, Geophys Res Lett, 50, ARTN e2023GL103785 10.1029/2023GL103785, 2023.

Kaplan, J. O. and Lau, K. H. K.: The WGLC global gridded lightning climatology and time series, Earth Syst Sci Data, 13, 3219–3237, 10.5194/essd-13-3219-2021, 2021.

Kaplan, J. O. and Lau, K. H. K.: World Wide Lightning Location Network (WWLLN) Global Lightning Climatology (WGLC) and time series, 2022 update, Earth Syst Sci Data, 14, 5665–5670, 10.5194/essd-14-5665-2022, 2022.

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., Ye, Q. W., and Liu, T. Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Adv Neur In, 30, 2017.

Li, B., Liang, S. L., Ma, H., Dong, G. P., Liu, X. B., He, T., and Zhang, Y. F.: Generation of global 1 km all-weather instantaneous and daily mean land surface temperatures from MODIS data, Earth Syst Sci Data, 16, 3795–3819, 10.5194/essd-16-3795-2024, 2024.

Li, F., Zeng, X. D., and Levis, S.: A process-based fire parameterization of intermediate complexity in a Dynamic Global Vegetation Model (vol 9, pg 2761, 2012), Biogeosciences, 9, 4771–4772, 10.5194/bg-9-4771-2012, 2012.

Lundberg, S. M. and Lee, S. I.: A Unified Approach to Interpreting Model Predictions, Adv Neur In, 30, 2017.

Mao, J. Q., Zhao, T. L., Keller, C. A., Wang, X., McFarland, P. J., Jenkins, J. M., and Brune, W. H.: Global Impact of Lightning-Produced Oxidants, Geophys Res Lett, 48, ARTN e2021GL095740 10.1029/2021GL095740, 2021.

Pérez-Invernón, F. J., Gordillo-Vázquez, F. J., Huntrieser, H., and Jöckel, P.: Variation of lightning-ignited wildfire patterns under climate change, Nat Commun, 14, ARTN 739 10.1038/s41467-023-36500-5, 2023.

Romps, D. M., Seeley, J. T., Vollaro, D., and Molinari, J.: Projected increase in lightning strikes in the United States due to global warming, Science, 346, 851–854, 10.1126/science.1259100, 2014.

Romps, D. M., Charn, A. B., Holzworth, R. H., Lawrence, W. E., Molinari, J., and Vollaro, D.: CAPE

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

682 Times P Explains Lightning Over Land But Not the Land-Ocean Contrast, Geophys Res Lett, 45, 12623–
683 12630, 10.1029/2018gl080267, 2018.
684 Rudlosky, S. D. and Shea, D. T.: Evaluating WWLLN performance relative to TRMM/LIS, Geophys Res
685 Lett, 40, 2344–2348, 10.1002/grl.50428, 2013.
686 Saleh, N., Gharaylou, M., Farahani, M. M., and Alizadeh, O.: Performance of Lightning Potential Index,
687 Lightning Threat Index, and the Product of CAPE and Precipitation in the WRF Model, Earth Space Sci,
688 10, ARTN e2023EA003104
689 10.1029/2023EA003104, 2023.
690 Schultz, C. J., Petersen, W. A., and Carey, L. D.: Lightning and Severe Weather: A Comparison between
691 Total and Cloud-to-Ground Lightning Trends, Weather and Forecasting, 26, 744–755, 10.1175/waf-d-
692 10-05026.1, 2011.
693 Schumann, U. and Huntrieser, H.: The global lightning-induced nitrogen oxides source, Atmos Chem
694 Phys, 7, 3823–3907, DOI 10.5194/acp-7-3823-2007, 2007.
695 Sugiyama, M., Krauledat, M., and Müller, K. R.: Covariate shift adaptation by importance weighted cross
696 validation, J Mach Learn Res, 8, 985–1005, 2007.
697 Taszarek, M., Pilguj, N., Allen, J. T., Gensini, V., Brooks, H. E., and Szuster, P.: Comparison of
698 Convective Parameters Derived from ERA5 and MERRA-2 with Rawinsonde Data over Europe and
699 North America, J Climate, 34, 3211–3237, 10.1175/Jcli-D-20-0484.1, 2021.
700 Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., and Carmona-Moreno, C.: The
701 influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions:
702 results from a process-based model, Biogeosciences, 7, 1991–2011, 10.5194/bg-7-1991-2010, 2010.
703 Ukkonen, P. and Mäkelä, A.: Evaluation of Machine Learning Classifiers for Predicting Deep Convection,
704 J Adv Model Earth Sy, 11, 1784–1802, 10.1029/2018ms001561, 2019.
705 Veraverbeke, S., Rogers, B. M., Goulden, M. L., Jandt, R. R., Miller, C. E., Wiggins, E. B., and Randerson,
706 J. T.: Lightning as a major driver of recent large fire years in North American boreal forests, Nat Clim
707 Change, 7, 529–+, 10.1038/Nclimate3329, 2017.
708 Verjans, V. and Franzke, C. L. E.: Development of a Data-Driven Lightning Scheme for Implementation
709 in Global Climate Models, J Adv Model Earth Sy, 17, ARTN e2024MS004464
710 10.1029/2024MS004464, 2025.

https://doi.org/10.5194/essd-2026-318
Preprint. Discussion started: 29 May 2026

711 Williams, E. R., Rutledge, S. A., Geotis, S. G., Renno, N., Rasmussen, E., and Rickenbach, T.: A Radar
712 and Electrical Study of Tropical Hot Towers, J Atmos Sci, 49, 1386–1395, Doi 10.1175/1520-
713 0469(1992)049<1386:Araeso>2.0.Co;2, 1992.
714 Wolpert, D. H.: Stacked Generalization, Neural Networks, 5, 241–259, Doi 10.1016/S0893-
715 6080(05)80023-1, 1992.
716 Wong, J., Barth, M. C., and Noone, D.: Evaluating a lightning parameterization based on cloud-top height
717 for mesoscale numerical model simulations, Geosci Model Dev, 6, 429–443, 10.5194/gmd-6-429-2013,
718 2013.
719 Zhang, J. B., Zheng, Y., Qi, D. K., Li, R. Y., and Yi, X. W.: DNN-Based Prediction Model for Spatio-
720 Temporal Data, 24th Acm Sigspatial International Conference on Advances in Geographic Information
721 Systems (Acm Sigspatial Gis 2016), Artn 92
722 10.1145/2996913.2997016, 2016.
723 Zhang, Y. F., Liang, S. L., Ma, H., He, T., Wang, Q., Li, B., Xu, J. L., Zhang, G. D., Liu, X. B., and Xiong,
724 C. H.: Generation of global 1 km daily soil moisture product from 2000 to 2020using ensemble learning,
725 Earth Syst Sci Data, 15, 2055–2079, 10.5194/essd-15-2055-2023, 2023.
726 Zheng, H., Wang, J., Zhou, H., Ding, J., Dai, H., Huang, Z., Wang, Z., Wang, M., Wang, H., Jiang, F.,
727 and Ju, W.: Global Land Lightning Density Reconstruction version 1 (GLLDR v1) (v1.0), Zenodo
728 [dataset], 10.5281/zenodo.19722380, 2026a.
729 Zheng, H., Wang, J., Zhou, H., Ding, J., Dai, H., Huang, Z., Wang, Z., Wang, M., Wang, H., Jiang, F.,
730 and Ju, W.: GLLDR v1 Reconstruction Code (v1.0), Zenodo [code], 10.5281/zenodo.19723880, 2026b.
731