Development of a data-driven lightning model for implementation in Global Climate Models

Vincent Verjans¹ and Christian L.E Franzke²

¹IBS Center for Climate Physics
²Center for Climate Physics, Institute for Basic Science

Abstract

This study proposes a new global-scale lightning model, predicting lightning rates from large-scale climatic variables. Using satellite lightning records spanning a period of 29 years, we apply machine learning methods to derive a functional relationship between lightning and climate reanalysis data. In particular, we design a model tree, representing different lightning regimes with separate single hidden layer neural networks of low dimensionality. We apply multiple complexity constraints in the model development stages, which makes the lightning model straightforward to implement as a lightning scheme for global climate models (GCMs). We demonstrate that, for years not used for model training, our lightning model captures 70.6% of the daily global spatio-temporal lightning variability, which corresponds to a >42% relative improvement compared to well-established lightning schemes. Similarly, the model correlates well with lightning observations for the monthly climatology (r>0.92), inter-annual variability (r>0.90), and latitudinal and longitudinal distributions (r>0.86). Most notably, the model brings a critical improvement in representing lightning magnitude and variability in the three tropical lightning chimney regions: central Africa, the Amazon, and the Maritime Continent. We implement the lightning model in the Community Earth System Model to verify its stability and performance as a GCM component, and we provide detailed implementation guidelines. As an intermediate approach between high-dimensional machine learning models and first-order lightning parameterizations, our model offers GCMs a straightforward and efficient tool to improve lightning simulation, which is critical for representing atmospheric chemistry and naturally-ignited wildfires.
Development of a data-driven lightning model for implementation in Global Climate Models

Verjans Vincent$^{1,2}$ and Christian L. E. Franzke$^{1,2}$

$^1$Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea
$^2$Pusan National University, Busan, Republic of Korea

Key Points:

- Using lightning records spanning 29 years, we develop a global model capturing the relationship between lightning and large-scale climate.
- Model evaluation shows improved representation of spatio-temporal variability at all scales compared to established lightning schemes.
- Our model uses machine learning with complexity constraints, facilitating implementation in climate models, with verified stability.

Corresponding author: Verjans Vincent, vverjans@pusan.ac.kr
Abstract
This study proposes a new global-scale lightning model, predicting lightning rates from large-scale climatic variables. Using satellite lightning records spanning a period of 29 years, we apply machine learning methods to derive a functional relationship between lightning and climate reanalysis data. In particular, we design a model tree, representing different lightning regimes with separate single hidden layer neural networks of low dimensionality. We apply multiple complexity constraints in the model development stages, which makes the lightning model straightforward to implement as a lightning scheme for global climate models (GCMs). We demonstrate that, for years not used for model training, our lightning model captures 70.6% of the daily global spatio-temporal lightning variability, which corresponds to a > 42% relative improvement compared to well-established lightning schemes. Similarly, the model correlates well with lightning observations for the monthly climatology ($r > 0.92$), inter-annual variability ($r > 0.90$), and latitudinal and longitudinal distributions ($r > 0.86$). Most notably, the model brings a critical improvement in representing lightning magnitude and variability in the three tropical lightning chimney regions: central Africa, the Amazon, and the Maritime Continent. We implement the lightning model in the Community Earth System Model to verify its stability and performance as a GCM component, and we provide detailed implementation guidelines. As an intermediate approach between high-dimensional machine learning models and first-order lightning parameterizations, our model offers GCMs a straightforward and efficient tool to improve lightning simulation, which is critical for representing atmospheric chemistry and naturally-ignited wildfires.

Plain Language Summary
Lightning is a worldwide phenomenon, which affects atmospheric chemistry and can cause wildfire ignitions. However, representing lightning in climate models at the global scale is challenging, because it depends on small-scale physical processes not explicitly represented in global models. In this study, we develop a lightning model, which estimates lightning rates from large-scale climate variables. We use machine learning methods to extract differences in the relationship between lightning and climate in different lightning regimes. We impose several constraints to keep the lightning model simple, which facilitates implementation and use of the model as a component of global climate models. We show that predictions of the lightning model reproduce temporal and spatial variability of lightning observations. In addition, the match to observed lightning rates is improved compared to currently widely used lightning schemes.

1 Introduction
Lightning is a prevalent phenomenon on Earth, with an estimated mean global lightning flash rate of 44±5 fl. s$^{-1}$ (Christian et al., 2003; Cecil et al., 2014; Blakeslee et al., 2020). Lightning is unevenly distributed, occurring predominantly over land and in the tropics (Christian et al., 2003). It also exhibits temporal variability across timescales: sub-daily, seasonal, annual, and inter-annual (Williams et al., 2005). This spatio-temporal variability of lightning is driven by its sensitivity to climate conditions. Yet, the climate-lightning relationship is not understood in detail. For example, the scientific community has yet to reach a consensus about the mechanisms behind the land-ocean contrast, about the causes for differences between tropical regions, and even about future changes in lightning regimes under anthropogenic climate change (Williams & Stanfill, 2002; Finney et al., 2018; Romps, 2019).

There are important reasons to understand the sensitivity of lightning to climate. First, lightning is estimated to produce ~ 10% of the global atmospheric emissions of nitrogen oxides ($\text{NO}_x$), and up to ~ 23% in the tropics (Schumann & Huntrieser, 2007). Lightning emissions of $\text{NO}_x$ predominate in the middle- and upper-troposphere, where
other NO\textsubscript{x} emissions are mostly absent, and where the NO\textsubscript{x} lifetime is longer than at
the surface. As a consequence, lightning plays a disproportionate role on tropospheric
chemistry (Wild, 2007). Since ozone (O\textsubscript{3}) net production depends non-linearly on the
NO\textsubscript{x} mixing ratio, lightning affects the concentration and distribution of tropospheric
O\textsubscript{3}, which is both an oxidant and a greenhouse gas (Schumann & Huntrieser, 2007). Fur-
thermore, lightning-produced NO\textsubscript{x} also impacts the oxidising capacity of the atmosphere
through their effects on the concentration of hydroxyl radicals (OH), which are the pri-
mary regulators of methane (CH\textsubscript{4}) losses (Schumann & Huntrieser, 2007). Second, light-
ning is almost exclusively responsible for all natural wildfire ignitions (Veraverbeke et
al., 2017; Janssen et al., 2023). Observations suggest that human- and lightning-ignited
fires have different characteristics; the latter are more sensitive to climatic conditions through
fuel moisture, and generally occur in more remote locations (Balch et al., 2017; Veraver-
beke et al., 2017). Lightning-ignited fires therefore explain most of the temporal vari-
ability in burned area in some specific ecosystems, such as boreal and intact forests (Ve-
raverbeke et al., 2017; Janssen et al., 2023). While changing climatic conditions lead to
increased frequency and severity of lightning-ignited wildfires in some regions, there re-
main major uncertainties in the future distribution of climate-lightning-wildfire inter-
actions (Janssen et al., 2023; Pérez-Invernón et al., 2023).

GCMs do not explicitly simulate cloud electrification and lightning, because of the
sub-kilometer resolution required to simulate these fine-scale processes (Fierro et al., 2015).
Instead, GCMs rely on lightning parameterizations, which empirically relate large-scale
atmospheric conditions to lightning flash rates (e.g., Price & Rind, 1992; Lopez, 2016;
Gordillo-Vázquez et al., 2019). Alternatively, GCMs use a pre-processed observation-based
lightning climatology as input forcing, such as in the most recent Fire Modeling Inter-
comparison Project (Rabin et al., 2017). With the latter approach, GCMs ignore the cause-
effect relation between climate and lightning. Advances in representing convection- and
cloud-related variables in GCMs (e.g., Peters et al., 2017) offer opportunities to, in turn,
 improve the accuracy of GCM lightning schemes. Such progress would enable consistency
between modeled climate and both atmospheric chemistry and wildfires.

The advent of lightning flash rate observations from satellites (Christian et al., 2003;
Cecil et al., 2014) has enabled major improvements in recent large-scale lightning pa-
parameterizations (e.g., Finney et al., 2014; Lopez, 2016; Stolz et al., 2017; Etten-Bohm
et al., 2021). Also, the pioneering parameterization of Price & Rind (1992) is still ex-
tensively used in GCMs (Thornhill et al., 2021). Despite, important scientific efforts, large
uncertainties remain in lightning parameterizations, which in turn translate in uncertain-
ties in atmospheric chemistry and wildfire impacts. For example, most parameterizations
use a multiplicative scaling factor to approximately match the observed global total light-
ing flash rate. For commonly used parameterizations, this factor typically spans the range
0.05 to 4.00 (Gordillo-Vázquez et al., 2019). The use of such a factor is necessary to yield
modeled lightning-produced NO\textsubscript{x} predictions in agreement with observation-based es-
timates (5±3 Tg nitrogen yr\textsuperscript{-1}, Schumann & Huntrieser, 2007). As another example, the
different sensitivities of state-of-the-art lightning parameterizations to anthropogenic cli-
mate change imply that future changes in lightning are essentially unknown: predictions
of global lightning change by 2100 in high-emission scenarios span -15% to +43% (Finney
et al., 2018; Romps, 2019). Finally, most parameterizations are calibrated to data from
restricted areas, generally the tropics or North-America, where most lightning data are
available (e.g., Finney et al., 2014; Romps et al., 2014; Stolz et al., 2017). While the trop-
ics account for \( \frac{3}{4} \) of global lightning (Christian et al., 2003), globally-accurate light-
ing predictions are important, for example to estimate wildfire risks in sensitive boreal
forests (Janssen et al., 2023).

Improvements in observational and data assimilation techniques are driving increas-
ing quality and quantity of climate reanalysis products (e.g., Hersbach et al., 2020). In
parallel, advances in statistical modeling and machine learning methods, as well as com-

putational power, are offering new opportunities in data-driven climate sciences (Bracco et al., 2018; Reichstein et al., 2019). Such techniques are increasingly used to predict lightning from weather conditions (Ukkonen & Mäkelä, 2019; Cheng et al., 2024). However, the implementation of machine learning methods within state-of-the-art numerical climate models faces many challenges, mostly related to codebase compatibility issues (Partee et al., 2022). In this study, we develop a data-driven lightning model based on climate reanalyses and satellite lightning measurements. We exploit satellite records of lightning spanning a period of 29 years, which, to the best of our knowledge, is longer than any previous study for development of a global-scale lightning scheme. We use a parsimonious machine learning approach, with the objective to make the lightning model straightforward to implement as a GCM component. We detail the model calibration, demonstrate its fidelity with respect to observations, including comparisons with other lightning parameterizations, and implement it in the Community Earth System Model (CESM) (Danabasoglu et al., 2020).

2 Methods

2.1 Lightning Data

We calibrate our lightning parameterization to satellite observations of lightning flashes from three separate missions. The first spaceborne optical sensor used to measure lightning flash rates from space was the Optical Transient Detector (OTD) (Christian et al., 2003). The OTD mission covered the period 1995-2000, with an extensive latitudinal range of ±75°. OTD was a prototype for the subsequent Lightning Imaging Sensor (LIS), launched in 1997 onboard the Tropical Rainfall Measuring Mission (TRMM). The LIS-TRMM exclusively covered low-latitude regions (±38° latitude), and was fully functional until early 2014. Combining measurements from OTD and LIS-TRMM has allowed the production of lightning data sets, which have provided unprecedented details about lightning spatio-temporal variability (Cecil et al., 2014). Both OTD and LIS being onboard low-orbit satellites, they performed ≥14 orbits per day, but with a viewing duration of any location at a given pass of 1 to 3 minutes. OTD and LIS-TRMM had a 10 and 5 km resolution, respectively. Their detection efficiencies varied depending on local time, ranging between 0.38 and 0.52 for OTD and 0.69 and 0.88 for LIS-TRMM. We refer to Cecil et al. (2014) for a thorough description of OTD and LIS-TRMM measurement capabilities, and for the production of lightning flash rates data sets. These lightning data sets have been extensively used to develop statistical relationships between lightning flash rates and climatological variables (e.g., Finney et al., 2014; Lopez, 2016; Stolz et al., 2017).

Since 2017, the LIS mission has been extended by being set to work on the International Space Station (ISS) platform (LIS-ISS, Blakeslee et al., 2020). Compared to LIS-TRMM, LIS-ISS has improved latitudinal coverage (±55°) and horizontal resolution (4 km). Importantly, LIS-ISS allows continuing spaceborne lightning measurements across all longitudes beyond the LIS-TRMM mission. Yet, to the best of our knowledge, this new record, spanning >7 years at the time of writing, has not been used in the development of any lightning model.

In this study, we use the gridded daily time series of flash rate and viewtime from the combined OTD and LIS-TRMM data, referred to as LRTS in Cecil et al. (2014). This data set was established by applying spatio-temporal smoothing to raw counts of flash measurements in order to alleviate sampling biases induced by viewing time limitations (Cecil et al., 2014). LRTS provides daily flash rate density values at a 2.5° × 2.5° resolution from May 1995 until February 2014. In contrast, no post-processed, ready-to-use gridded data set of the LIS-ISS measurements is available at the time of this study. Instead, only files of flash counts and satellite viewtimes are available. Here, we process the LIS-ISS data from 2017 to 2023 included in order to extend the daily 2.5° × 2.5°
LRTS time series, albeit with a gap from February 2014 to February 2017. For converting the LIS-ISS data from raw flash counts and viewtimes into gridded time series, we follow the description of Cecil et al. (2014) for LRTS. First, we divide any flash count value by the corresponding satellite viewtime. Second, following the information from the LIS-ISS user guide, we scale this flash rate by the LIS detection efficiency as a function of local time (Table 2 of Cecil et al., 2014). Third, the scaled effective flash rates are aggregated on a 2.5°×2.5° grid, and converted to flash densities \([\text{fl. km}^{-2} \text{yr}^{-1}]\). We exclude outliers using a 5-sigma rule, and values are adjusted with a small latitude-dependent factor to match total lightning flash rates provided in Table 1 of Blakeslee et al. (2020). Fourth, we apply both a 7.5°×7.5° window moving average in space, and a 100-day window moving average in time. Note that for simplicity, we do not perform further digital filtering, as the smoothing procedure resulted in a well-filtered signal.

### 2.2 Climate data

For relating measured lightning density to climatic variables at the global scale, we use reanalysis data from ERA5 (Hersbach et al., 2020). ERA5 offers a broad range of atmospheric variables at 31 km horizontal resolution, on 137 vertical levels, and at a hourly temporal resolution. We refer to Hersbach et al. (2020) for a detailed description of ERA5. In this study, we use ERA5 daily mean data, by averaging any climatic fields at the 0:00, 6:00, 12:00, and 18:00 UTC time steps. We consider a large set of potential climatic variables that can serve as input features for our lightning model. In line with our objective to develop a model scheme that can be implemented in GCMs, we restrict the potential climatic features to those which are typical GCM variables. The potential ERA5 input features used are listed in Table 1. We note that cloud top height and cloud cover are taken from the Advanced Very High Resolution Radiometer data set (Karlsson et al., 2023). Also in Table 1 are two time-constant geographical features: surface elevation and absolute latitude. Every input feature is gridded on the 2.5°×2.5° grid of LRTS.

The selection of these 25 potential features (Table 1) is motivated by previous research. For example, Convective Available Potential Energy (CAPE), cloud-top-height, and precipitation are some of the most commonly-used climatic features to predict lightning density values (e.g., Price & Rind, 1992; Romps et al., 2014; Lopez, 2016). We include features quantifying ice and liquid water contents in clouds, as they govern cloud electrification and charge transfer mechanisms (Saunders, 1993). Wind can modulate the regime of convective storms (Weisman & Klemp, 1982). Here, in addition to including 10m wind magnitude and vertical wind velocity at 500 hPa, we also follow Etten-Bohm et al. (2021) by computing deep wind shear and low-level wind shear between the 900 to 300 hPa and 900 to 700 hPa levels, respectively. Also, while surface latent and sensible heat fluxes are usually ignored in lightning parameterizations, Williams & Stanford (2002) have theorized that they exert a thermodynamical control on the convection regime, and therefore on lightning.

#### 2.2.1 Training and test data separation

Our data set consists of all daily lightning density values from the aggregation of the LRTS and LIS-ISS data, spanning 1995-2023, and the corresponding daily climatic variables. We separate the data in two subsets: training, and test data. We split the data based on entire years. We choose the years 2000, 2005, 2010, 2020, and 2023 as the test years. Data from these years are completely ignored when calibrating our lightning model. Accounting for the lightning data gap in 2015-2017, the remaining 22 years are used for training, but lightning data are available only for part of the years 1995, 2014, and 2017. In total, the combined training and test data sets include \(>50\times10^6\) daily lightning flash density values on the 2.5°×2.5° grid.
Table 1. Potential input features considered for the global lightning model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Units</th>
<th>Formulation details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convective Available Potential Energy</td>
<td>CAPE</td>
<td>J kg s(^{-2})</td>
<td></td>
</tr>
<tr>
<td>Cloud-top-height</td>
<td>cth</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>Cloud-base-height</td>
<td>cbh</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>Cloud cover</td>
<td>c(_c)</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>2m temperature</td>
<td>(T(_{2m}))</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>2m dew point temperature</td>
<td>(T(_{d,2m}))</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>Specific humidity (850 hPa)</td>
<td>(q(_{850hPa}))</td>
<td>kg kg(^{-1})</td>
<td></td>
</tr>
<tr>
<td>Cloud ice content</td>
<td>c(_{ice})</td>
<td>kg km(^{-2})</td>
<td></td>
</tr>
<tr>
<td>Cloud liquid water content</td>
<td>c(_{liq})</td>
<td>kg km(^{-2})</td>
<td></td>
</tr>
<tr>
<td>Column total water vapor</td>
<td>(m(_{vap}))</td>
<td>kg km(^{-2})</td>
<td></td>
</tr>
<tr>
<td>Column total water</td>
<td>(m(_{w}))</td>
<td>kg km(^{-2})</td>
<td>(c(<em>{ice})+c(</em>{liq})+m(_{vap}))</td>
</tr>
<tr>
<td>K-index</td>
<td>(K(_I))</td>
<td>°C</td>
<td>((T(<em>{850hPa}) - T(</em>{500hPa})) + T(<em>{d,850hPa}) - (T(</em>{700hPa}) - T(_{d,700hPa})))</td>
</tr>
<tr>
<td>Total totals index</td>
<td>(t(_T))</td>
<td>K</td>
<td>(T(<em>{850hPa}) + T(</em>{d,850hPa}) - 2T(_{500hPa}))</td>
</tr>
<tr>
<td>Wind speed magnitude at 10m</td>
<td>(w(_{10m}))</td>
<td>m s(^{-1})</td>
<td>((u(<em>{10m})^2 + v(</em>{10m})^2)^\frac{1}{2})</td>
</tr>
<tr>
<td>Lagrangian pressure tendency (500 hPa)</td>
<td>(\omega(_{500hPa}))</td>
<td>Pa s(^{-1})</td>
<td>negative of the vertical wind speed (500 hPa)</td>
</tr>
<tr>
<td>Low-level wind shear</td>
<td>(s(_{w,low}))</td>
<td>m s(^{-1})</td>
<td>([((u(<em>{700hPa}) - u(</em>{700hPa}))^2 + (v(<em>{700hPa}) - v(</em>{700hPa}))^2])^\frac{1}{2})</td>
</tr>
<tr>
<td>Deep-level wind shear</td>
<td>(s(_{w,deep}))</td>
<td>m s(^{-1})</td>
<td>([((u(<em>{300hPa}) - u(</em>{300hPa}))^2 + (v(<em>{300hPa}) - v(</em>{300hPa}))^2])^\frac{1}{2})</td>
</tr>
<tr>
<td>Total precipitation</td>
<td>(t(_p))</td>
<td>m s(^{-1})</td>
<td></td>
</tr>
<tr>
<td>Convective precipitation</td>
<td>(c(_p))</td>
<td>m s(^{-1})</td>
<td></td>
</tr>
<tr>
<td>Zero degree level</td>
<td>(b(_{0C}))</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>Surface pressure</td>
<td>(p(_s))</td>
<td>Pa</td>
<td></td>
</tr>
<tr>
<td>Surface latent heat flux</td>
<td>(Q(_l))</td>
<td>W m(^{-2})</td>
<td>positive downwards</td>
</tr>
<tr>
<td>Surface sensible heat flux</td>
<td>(Q(_e))</td>
<td>W m(^{-2})</td>
<td>positive downwards</td>
</tr>
<tr>
<td>Surface elevation above sea-level</td>
<td>(h(_s))</td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>Absolute latitude</td>
<td>(</td>
<td>\phi</td>
<td>)</td>
</tr>
</tbody>
</table>

\(u\) and \(v\) denote the longitudinal and latitudinal component of the wind speed [m s\(^{-1}\)], respectively.

hPa and m subscripts denote a feature value at a given pressure level [hPa] and height level [m], respectively.
2.3 Tree model

For our global lightning scheme, we calibrate a model with lightning flash density \( L \) as response variable [fl. km\(^{-2}\) yr\(^{-1}\)]. Our model uses a decision tree architecture. The general principle of decision trees is to partition the feature space into a set of separate entities, referred to as the leaves of the tree (Hastie et al., 2009; Costa & Pedreira, 2023). The decision tree splits input data following feature-dependent decision rules, starting from the root node, and until each data sample is assigned to a leaf node. A decision tree can be further refined into a model tree, which uses separate regression models in each leaf to capture leaf-specific dependencies of the response variable to the input features. The tree architecture offers numerous advantages: separation between different lightning regimes, interpretable structures, and ease of implementation as a scheme in GCMs.

The most common approach for model trees is to use linear regression models in the leaves (Costa & Pedreira, 2023). However, there is flexibility possible in this choice (Zeileis et al., 2008). In this study, we design our leaf models as Extreme Learning Machines (ELMs, Huang et al., 2006). An ELM is a generalized version of a single hidden layer feedforward neural network (SLFN), and it can be used in both classification and regression settings (Huang et al., 2012). It has been proposed that ELMs can be incorporated as part of model trees, and such ELM-trees have shown good performance on a variety of data sets, with notable efficiency for large data sets (Wang et al., 2015; Zhou & Yan, 2019). ELMs also offer the advantage of being simple to formulate, and thus facilitate prospects of implementation in GCMs.

In this section, we explain the process of tree induction, i.e., the building of the ELM-tree. Deep trees with many branches and leaves can provide more flexible models of \( L \). On the other hand, excessive tree size can lead to overfitting, or to large increases in model complexity for little gain in model skill. The tree induction is therefore an optimization process with the goal to partition the feature space such that our model correctly predicts lightning density values, while using few partitions and ELMs of low complexity. The entire tree induction process consists of three separate steps, detailed in the next three sub-sections.

2.3.1 Tree growing

In the first step of the tree induction, we grow an oversized tree with the commonly-used Classification And Regression Tree algorithm (CART, Breiman et al., 1984). CART builds the tree starting from the root node. At each node it finds the optimal splitting feature, referred to as a split-feature, and its splitting value. In this sense, CART is a greedy algorithm: at each split, it aims to minimize the weighted standard deviation in the target variable in the two child nodes. In principle, this process could continue until the number of leaves equals the number of data samples. In our tree induction procedure, we grow the oversized tree with a maximum depth of 8, corresponding to 129 leaf nodes and 256 nodes in total.

The distribution of the global daily lightning data is characterized by a large proportion of values close to 0. Due to the spatio-temporal smoothing applied to the lightning measurements, it is likely that most instances of lightning density close to zero correspond to lightning-free days. In order to deal with this particularity of satellite lightning density data, we force the first split of the tree to separate a zero-branch from a non-zero-branch. In the zero-branch, the tree immediately ends into a leaf that always returns the zero value, and thus incorporates no leaf model. In contrast, tree growing continues in the non-zero-branch. Using the zero-branch allows the ELM calibration to focus on non-negligible lightning values. Here, we set a threshold of 0.1 fl. km\(^{-2}\) yr\(^{-1}\) as a limit between daily lightning values to be classified as zero or not. In contrast to the other splitting decisions of the ELM-tree, the zero-branch split is a classification task: we need to find the optimal decision rule for separating \( L \leq 0.1 \) versus \( L > 0.1 \) fl. km\(^{-2}\) yr\(^{-1}\).
We use the receiver operating characteristic curve (ROC curve) to optimize both the split-feature, and its splitting threshold for the zero-branch.

### 2.3.2 ELM calibration

In the second step of tree induction, we fit an ELM in each node of the tree, except the zero-branch. This includes both the leaves and the interior nodes of the oversized tree. The ELMs are fit only to those data samples falling into the corresponding node. As a SLFN, for any given node, an ELM model for lightning is formulated as:

\[ \hat{L}_i = \sum_{j=1}^{\tilde{N}} \beta_j g \left( w_j^T x_i \right), \]  

(1)

where \( i \) denotes the \( i \)th data sample, \( \tilde{N} \) is the number of hidden neurons in the SLFN, \( x_i \) is the vector of input features with an intercept term, \( w_j \) is the weight vector of the \( j \)th hidden node, \( g() \) is the activation function, and \( \beta \) is the weight vector connecting the hidden layer to the output, \( \hat{L}_i \). In ELMs, the hidden weights \( w \) do not need to be calibrated. Instead, they are generated randomly, and only the output weights \( \beta \) are calibrated. This has the major advantage that an analytical solution exists for \( \beta \) (Huang et al., 2006), making the model calibration fast and scalable to large data sets. Assuming that there are \( N \) data samples, Equation (1) can be rewritten in matrix notation:

\[ \hat{L} = H\beta, \]  

(2)

where \( \beta \) is of size \((\tilde{N} \times 1)\), and the hidden layer matrix \( H \) of size \((N \times \tilde{N})\) is given by:

\[ H = \begin{bmatrix}
g(w_1^T x_1) & \ldots & g(w_{\tilde{N}}^T x_1) \\
\vdots & \ddots & \vdots \\
g(w_1^T x_N) & \ldots & g(w_{\tilde{N}}^T x_N)
\end{bmatrix}. \]  

(3)

We fix the number of hidden neurons to 4, i.e., \( \tilde{N} = 4 \). This small number of hidden neurons ensures that the lightning ELM-tree remains easy to implement in GCMs, using Equation (1). We verify (Fig. S1) that the rate of model performance increase reduces once the number of hidden neurons reaches 4. For \( g() \), we use the rectified linear unit (ReLU) activation function:

\[ g \left( w_j^T x_i \right) = \max \left( 0, w_j^T x_i \right). \]  

(4)

ELMs can accommodate a large variety of activation functions (Huang et al., 2006). In prior investigations, we experimented with the use of the logistic and softplus activation functions, which showed comparable performance but with larger numbers of fit-features. We decided to use the ReLU to maintain ELMs of lower dimensionality, as well as for the straightforward implementation of Equation (4) in GCMs. At each ELM calibration, we sample the hidden weights randomly from a uniform distribution bounded between \([-1; 1]\). This random sampling is performed 10 times, and the ELM is calibrated with each random sample. Only the random sample of \( \{w_j\}_{j=1: \tilde{N}} \) leading to the lowest root mean square error (RMSE) is retained. This avoids forcing the ELM to use weights that were, by chance, a poor choice for fitting Equation (1). It should also be noted that we scale each feature to the range \([0, 1]\) for the calibration process, but we rescale the ELM coefficients after calibration such that computation on the true scale of feature values is strictly equivalent.

Fitting coefficients \( \beta_j \) for given hidden weights \( w_j \) is an analytical step. The minimal norm least squares solution can be found using the Moore-Penrose generalized inverse, \( H^+ \), of the hidden matrix \( H \). Here, we add a small L2 regularizer \( \lambda \) to make the solution for \( \beta \) more stable (Huang et al., 2012):

\[ H^+ = (H^T H + \lambda I)^{-1} H^T, \]  

(5)
where \( I \) is the identity matrix, and we set \( \lambda = 0.1 \). The solution for \( \beta \) is given by:

\[
\beta = H^+ L. \tag{6}
\]

At each node, \( H^+ \) and \( L \) in Equation (6) use only a subset of the daily lightning and climatic data samples, corresponding to those samples falling into the given tree node.

While we use a zero-branch calibrated to lightning events \( \leq 0.1 \text{ fl. km}^{-2} \text{ yr}^{-1} \), nothing prevents an ELM to output values \( \leq 0.1 \text{ fl. km}^{-2} \text{ yr}^{-1} \) as well. This motivates the choice of the small threshold \( 0.1 \text{ fl. km}^{-2} \text{ yr}^{-1} \) for the zero-branch classification: the classifier does not need to capture all the low lightning events, because \( \hat{L} \) can also be small in the non-zero-branch. Due to the formulation of SLFNs (Eq. (1)), ELM output could range on the entire real line. As such, we set any predicted negative \( \hat{L} \) value to 0.

Initially, there are 25 features available to each ELM (Table 1). Including all the features would result in very high-dimensional ELMs, as the size of each hidden weight vector \( w_j \) scales linearly with the number of features. We use the Bayesian Information Criterion (BIC, Schwarz, 1978) to find the optimal number of features to include in each ELM. Assuming normally-distributed errors, the BIC can be formulated as:

\[
\text{BIC} = B \log(N) + N \log \left( \frac{1}{N} \sum_{i=1}^{N} (L_i - \hat{L}_i)^2 \right), \tag{7}
\]

where \( B \) is the number of model parameters, and \( N \) the number of data samples. Minimizing the BIC thus corresponds to minimizing the residual sum of squared errors (RSS), with a penalty coefficient proportional to the number of parameters \( B \). In the case of an ELM with \( \tilde{N} \) hidden neurons and using \( M \) features, \( B = \tilde{N} \times (M + 2) \).

Here, we use forward stepwise selection of features (e.g., Hastie et al., 2009). Each ELM model starts with only an intercept term, i.e., \( w_j \) is a scalar. Then, we in turn include each single potential feature individually (Table 1), and recalibrate the ELM. We consider the feature that has lead to the largest reduction in RSS as the potential extra-feature. If the BIC has decreased, we keep the extra-feature in the model, and continue to iterate through the remaining features. Once the BIC stops decreasing, we stop the forward stepwise process, and keep only those features included up to that step. Features included in the ELMs are referred to as fit-features, in contrast to the split-features used to separate the different branches in the tree growing process (Sect. 2.3.1). The sets of ELM fit-features from the different tree nodes, as well as the set of split-features, can overlap but need not be the same. As a last step of the ELM calibration, in each tree node, we also fit a simple linear regression model with the fit-features. If the linear regression model achieves a lower BIC compared to the ELM due to the reduction in parameter numbers from dropping the hidden layer, we swap the ELM model for the linear model. The entire ELM calibration procedure is summarized in Algorithm 1. The size of the oversized tree, the large number of data samples and potential features (Table 1), and the embedded loops in Algorithm 1 make clear the benefits of the analytical solution for \( \beta \) (Eq. (6)).

### 2.3.3 Tree pruning

In the third step of tree induction, after having grown the oversized tree and fitted the ELM models at each node, we prune the ELM-tree to decrease its complexity. Several pruning strategies exist, and we adopt weakest-link pruning, which is a form of cost-complexity pruning (Hastie et al., 2009). The idea behind cost-complexity pruning is to find a balance between tree size and residual errors, where tree size \(|\Gamma|\) is evaluated as the number of leaf nodes. We formulate a complexity-penalized cost function (Hastie
Algorithm 1 ELM calibration procedure at any given tree node.
This Algorithm uses Equations (2, 3, 4, 5, 6, 7).

1: at the given tree node, find all data samples $i$ of the node
2: consider all potential features $1 : M_{\text{max}}$
3: Features $\leftarrow \{\}$
4: $X \leftarrow 1_N$
5: $M \leftarrow 1$
6: $\text{BIC}^{(1)}, \text{BIC}^{(2)} \leftarrow +\infty$
7: while $\text{BIC}^{(2)} \leq \text{BIC}^{(1)}$ and $M < M_{\text{max}}$ do
8: $M \leftarrow M + 1$
9: $\text{BIC}^{(2)} \leftarrow +\infty$
10: for $m = 1 : M_{\text{max}}$ do
11: if $m$ not in Features then
12: $\text{RSS}_{\text{min}} \leftarrow +\infty$
13: $X^{(m)} \leftarrow [X, \text{Feature}(m)]$
14: for $k = 1 : 10$ do
15: $w^{(k)} \leftarrow U(-1, 1)$, size: $(M \times \tilde{N})$
16: $\beta^{(k)} \leftarrow H^+ L$
17: $\text{RSS}^{(k)} \leftarrow \sum_i \left( L_i - \hat{L}_i \right)^2$
18: if $\text{RSS}^{(k)} < \text{RSS}_{\text{min}}$ then
19: $w, \beta, \text{RSS}_{\text{min}} \leftarrow w^{(k)}, \beta^{(k)}, \text{RSS}^{(k)}$
20: $\text{BIC}^{(m)} \leftarrow \text{BIC}(M, \text{RSS}_{\text{min}})$
21: if $\text{BIC}^{(m)} < \text{BIC}^{(2)}$ then
22: $m^*, \text{BIC}^{(2)} \leftarrow m, \text{BIC}^{(m)}$
23: if $\text{BIC}^{(2)} < \text{BIC}^{(1)}$ then
24: Features $\leftarrow$ Features$+m^*$
25: $X \leftarrow [X, \text{Feature}(m^*)]$
26: $\text{BIC}^{(1)} \leftarrow \text{BIC}^{(2)}$
27: $M^* \leftarrow \text{size(Features)}$
28: calibrate Linear Model (LM) to $X, L \rightarrow \beta^{(\text{LM})}$
29: $\text{RSS}^{(\text{LM})} \leftarrow \sum_i \left( L_i - \hat{L}_i^{(\text{LM})} \right)^2$
30: $\text{BIC}^{(\text{LM})} \leftarrow \text{BIC}(M^*, \text{RSS}^{(\text{LM})})$
31: if $\text{BIC}^{(\text{LM})} < \text{BIC}^{(1)}$ then
32: swap ELM for LM
C(Γ, α_c) = \frac{1}{N} \sum_{i=1}^{N} (L_i - \hat{L}_i)^2 + α_c|Γ|,
(8)

where α_c is a regularization parameter. Weakest-link pruning consists of successively finding and pruning the internal node of which the removal causes the smallest per leaf increase in the RSS. This process continues until the tree is reduced to the single root tree. For a given α_c penalty, there is a single tree that minimizes C(Γ, α_c). It can be shown that this optimal tree is always in the sequence of trees generated by weakest-link pruning (Ripley & Hjort, 1996). Weakest-link pruning is summarized in Algorithm 2.

Algorithm 2 Weakest-link pruning.

This Algorithm uses Equation (8).

1: choose range(α_c)
2: Γ_init ← initial oversized tree
3: for all nodes i in Γ_init do
4: compute RSS_i
5: Γ ← Γ_init
6: Sequence(Γ) ← {Γ}
7: while Γ ≠ Γ_root do
8: for all non-leaf nodes i in Γ do
9: j(i) ← set of leaf nodes of i
10: n(j(i)) ← size (j(i))
11: ∆(RSS)_i = RSS_i - \sum_{j(i)} RSS_{j(j(i))}
12: find non-leaf node i’ minimizing ∆(RSS)_i / n(j(i))
13: Γ ← Γ pruned below i’
14: Sequence(Γ) ← Sequence(Γ) + Γ
15: for α_c in range(α_c) do
16: find Γ’(α_c) in Sequence(Γ) minimizing C(Γ, α_c)

The value of α_c governs the trade-off between minimizing the training data RSS, and minimizing the tree size. It can be tuned to optimize the generalization performance of the ELM-tree, or to find an ELM-tree of desired complexity, which is important for the objective of implementation in GCMs. We use cross-validation to select the value of α_c (Hastie et al., 2009). To limit the computational expense of the cross-validation, we sub-sample the data at 5-day resolution, but we do not sub-sample spatially. Then, we split the training data annually and keep only the 19 training years with full lightning data coverage. We opt for annual splits because of the temporal smoothing applied to the lightning measurements; if the data were split monthly instead, there would be a high degree of artificial correlation between data from different cross-validation samples. We perform leave-one-out cross-validation. That is, we perform 19 individual ELM-tree model training procedures, each leaving out one specific year of training data, and then being evaluated on the left-out year. We measure the sensitivity of the cross-validation error to α_c to guide our choice of tree-complexity penalization strength. Once we select our optimal α_c, we calibrate the ELM-tree with the entire training data set, and without temporal sub-sampling.

2.4 Implementation in a Global Climate Model

We implement the ELM-tree lightning scheme in the Community Earth System Model version 2.2 (CESM2.2, Danabasoglu et al., 2020). The lightning scheme affects atmospheric chemistry through the generation of NOx compounds. The atmospheric light-
ning output is also coupled with the land model to provide prognostic lightning flash densities at each model time step. In the land model, lightning serves as a source of natural wildfire ignitions. This atmosphere-to-land coupling of lightning flash rate is not available in the latest CESM2.2 public release, and has been developed as part of this study.

We perform a simple and idealized model simulation to demonstrate the capabilities offered by the ELM-tree implementation with an interactive atmosphere-to-land coupling. Our CESM2.2 set-up uses the Community Atmosphere Model version 6 for the atmosphere, and the the Community Land Model Version 5 with the fire model of Li et al. (2012) for the land (Lawrence et al., 2019; Danabasoglu et al., 2020). The ocean and sea-ice components are set as inactive to save computational expense. We perform a single 1995-2015 run, which is initialized from the CESM2.2 historical run performed for CMIP6. We use a coarse 1.9° × 2.5° resolution, and with interactive biogenic emissions from fires. We use default CESM2.2 parameterizations for the vertical distribution of lightning-produced NO$_x$ (Pickering et al., 1998), the energy per flash (Price et al., 1997), and the energy difference between intra-cloud and cloud-to-ground flashes (Ridley et al., 2005). The objective of this CESM2.2 simulation is not to perform a detailed comparison with observations, but rather to demonstrate that the ELM-tree can be implemented in state-of-the-art GCMs, and that it produces realistic and stable estimates of lightning, even when driven exclusively with model fields.

### 2.5 Comparison with existing parameterizations

For evaluation, we compare the performance of the ELM-tree to that of two other state-of-the-art lighting flash density parameterizations. In particular, we use the lightning schemes of Price & Rind (1992) (PR92) and of Finney et al. (2014) (F14).

PR92 is the lightning parameterization used to simulate lightning-produced NO$_x$ in all the GCMs taking part to the CMIP6 experiment and simulating tropospheric chemistry (Thornhill et al., 2021). As such, it is also the parameterization of the latest CESM2.2 release (Danabasoglu et al., 2020), although wildfire ignitions use a pre-processed input lightning climatology. The PR92 lightning density is predicted using two separate formulations for the land and ocean, both following a power-dependence on cloud-top-height. In PR92, $L$ scales with cloud-top-height to the power 4.9 and 1.73 over land and ocean, respectively (Price & Rind, 1992). In addition, we apply the correction of Price & Rind (1994) to account for different grid resolutions, as all lightning predictions are computed on the same 2.5° × 2.5° grid as the lightning LRTS product.

F14 is based on cloud ice flux, which relates it in a physical manner to the non-inductive charging mechanism in thunderstorms (Reynolds et al., 1957). It depends only on mid-tropospheric values, and computations are performed at the specific 440 hPa vertical level. At this level, cloud ice flux is the product of specific cloud ice content and updraught mass flux, rescaled by the fractional cloud area. We download all these quantities from ERA5, and interpolate them linearly at the 440 hPa level. We compute daily mean values by averaging all variables at the 0:00, 6:00, 12:00, and 18:00 UTC time steps, similar to our treatment of all the climate data used in this study. Similarly to PR92, maritime and continental lightning have separate formulations in F14. Following Finney et al. (2014), we set lightning to 0 if the 440 hPa fractional cloud cover is < 0.01.

### 3 Results

#### 3.1 ELM-tree induction and cross-validation

We perform the cross-validation experiment to guide our choice of the ELM-tree size. Figure 1a shows the change in cross-validated RMSE with the number of leaf nodes (|Γ|). As expected, the RMSE increases as |Γ| decreases, because smaller trees have both
less partitions of the feature space and less leaf ELM models. However, the ±1 standard deviation (±1σ) intervals overlap from ELM-trees having just 3 up to 46 leaf nodes (Fig. 1a). A very parsimonious approach could therefore use only the zero-branch split, and two ELMs to compute non-zero lightning. Here instead, we identify a clear slope break in the rate of RMSE decrease with |Γ|, occurring at |Γ| ≈ 7 (Fig. 1a, right y-axis). The red curve shows that a given RMSE decrease becomes abruptly more expensive in |Γ| once |Γ| > 7. As such, we select the complexity penalty αc corresponding to |Γ| = 7, which provides an appropriate balance between ELM-tree predictive performance and size (Fig. 1a, green dotted line).

![Figure 1.](image)

**Figure 1.** Results from the cross-validation experiment. (a) The change of the cross-validation root mean square error (RMSE) with the number of leaf nodes (|Γ|) used in the ELM-tree. The red line shows the derivative of the RMSE with respect to |Γ|, evaluated using centered differencing. The dotted green line shows the selected ELM-tree size, corresponding to |Γ| = 7. (b) The selection probability of each feature when |Γ| = 7, averaged across the cross-validation samples. The fit-feature selection probability is the proportion of leaf models using that feature. The split-feature selection probability is the proportion of leaf nodes for which the set of decision splitting rules includes that feature at least once. Selection probabilities are averaged across the 19 cross-validation samples. Figure 1 shows that split decision rules are based on a relatively small num-

We show the selection probability of the 25 different features, both as fit-features (Fig. 1c) and split-features (Fig. 1d). The fit-feature selection probability is computed as the proportion of leaf models using a given feature. Similarly, the split-feature selection probability is the proportion of leaf nodes having been split in at least one upper branch on a given feature. Selection probabilities are averaged across the 19 cross-validation samples. Figure 1 shows that split decision rules are based on a relatively small num-
ber of features. In contrast, the fit-feature selection probabilities are more spread across
the different potential features. Although, as the number of leaf nodes decreases, selection probabilities tend to concentrate on a smaller set of fit-features, which are important to the ELM models close to the ELM-tree root.

In Figure 1b, we show selection probabilities at the selected number of leaf nodes $|\Gamma| = 7$. We only show those features reaching at least 10% fit- or split-feature probability. Five split-features are used with probability $\geq 10\%$, and three dominate with probability values $> 70\%$: $K_I$, $h_s$, and CAPE. Notably, $K_I$ is always used at the root node as split-feature for the zero-branch. In addition to its role as a split-feature, CAPE is the fit-feature with the highest selection probability (51%). Figure 1b demonstrates that more fit-features have selection probabilities $\geq 10\%$ compared to the split-features. These fit-features are associated with geography ($|\phi|$), surface heat fluxes ($Q_l$, $Q_e$), cloud properties (e.g., $cbh$, $c_{liq}$), or other climatic features (e.g., CAPE, $tt_I$, $w_{10m}$).

Figure 2. Receiver Operating Characteristic (ROC) curve for the selection of the zero-branch split criterion. The selected optimal classifier corresponds to the value $K_I=12.24^\circ C$. Specificity and recall are defined in Equation (9). $K_I$ is chosen as split feature because the distance from its optimal classifier to the perfect classifier is the smallest among all potential features considered (Table 1).

Concerning the zero-branch split, Figure 2 shows the ROC curve as the $K_I$ splitting value changes. The ROC curve shown is computed with the entire training data set. Data samples with $K_I$ less than the splitting value are classified as zero-lightning events. The ROC curve shows the trade-off between recall and specificity:

$$ \begin{align*}
\text{recall} &= \frac{TP_0}{P_0}, \\
\text{specificity} &= \frac{TN_0}{N_0},
\end{align*} $$

where $P_0$ and $N_0$ are the total number of zero-lightning and non-zero-lightning data samples, as defined by the 0.1 fl. km$^{-2}$ yr$^{-1}$ threshold. $TP_0$ is the number of zero-lightning data samples correctly classified by the split criterion, and similarly $TN_0$ is the number of non-zero-lightning data samples correctly classified. As illustrated in Figure 2, the perfect classifier would satisfy recall=specificity=1. For all the potential features, their respective optimal split value is taken along their ROC curve where the distance to the ideal classifier is minimized. We select $K_I$ as optimal split-feature because its optimal split-value achieves the smallest distance to the perfect classifier. This is valid not only on the full training data set, but also for each cross-validation fold (Fig. 1b). The $K_I$
zero-branch split value is $12.24^\circ C$, which yields values for recall and specificity of 0.72 and 0.63, respectively.

Imposing the cross-validated tree size $|\Gamma| = 7$, we fit the final ELM-tree to the full training data set. This results in an ELM-tree with a maximum depth of 4. A schematic of the model is shown in Figure 3. Notably, all the non-zero branch leaf models use an ELM, as the simpler linear regression model is never favored by the BIC (see Algorithm 1). The leaf ELMs use on average 3.5 features, but with some overlap between the leaf ELMs. As such, our ELM-tree uses only 12 features in total: $CAPE$, $cbh$, $T_{d,2m}$, $c_{sec}$, $m_w$, $K_I$, $tt$, $h_0$, $Q_l$, $h_s$, and $|\phi|$ (see Table 1). The detailed description of the ELM-tree and an implementation pseudo-code are provided in the Supporting Information.

![Figure 3. Schematic of the architecture of the calibrated ELM-tree. Split-features are given in their tree node, along with their respective splitting value. Inequality symbols show the direction of the splitting rules. The left-split at the tree root corresponds to the zero-branch leaf node, which always returns 0 fl. km$^{-2}$ yr$^{-1}$. In the non-zero-branch, each leaf node consists of an Extreme Learning Machine (ELM) model, as illustrated with the sketches of single hidden layer feedforward neural networks. Note that the second split decision $h_s \leq 0.5$ m effectively separates the maritime and continental domains, which can be used as an equivalent split decision.]

### 3.2 Evaluation and comparison with other lightning schemes

We use the data from the test years (2000, 2005, 2010, 2020, 2023) to evaluate the out-of-sample ELM-tree performance. All performance metrics in this Section are computed with respect to the test years only, and no re-scaling to match the observed global total lightning is applied. In order to provide a meaningful evaluation, we perform spatio-temporal smoothing of the outputs from the ELM-tree, PR92, and F14 in the same manner as was done for the LIS lightning product (Cecil et al., 2014).

Figure 4 shows maps of the mean lightning rate from observations, from the ELM-tree, and from the PR92 and F14 parameterizations. The ELM-tree reproduces the observed spatial lightning patterns well. For example, it captures the lightning peak in central-Africa relative to the other tropical regions, the longitudinal gradient across Eurasia, the...
Figure 4. Mean lightning flash density over the test years 2000, 2005, 2010, 2020, and 2023. From (a), the LIS observations, (b) the ELM-tree, (c) PR92, and (d) F14. Note the logarithmic color scale.

East-West contrast in North-America, and lightning over Australia. PR92 and F14 reproduce part of the spatial distribution, but display large deviations from observed lightning density values in several regions. For example, PR92 underestimates lightning rates in Europe, central Africa, as well as Central- and North-America, but overestimates lightning over the Amazon. F14 overestimates lightning in central Asia and western North-America, and exhibits a strong under-estimation in Africa.

The improved fidelity of the ELM-tree to observations is clearer and better quantified in Figure 5, showing key performance metrics for the three lightning schemes compared to the LIS observations: mean bias, RMSE, and Pearson correlation. Note here that both the RMSE and the Pearson correlation are computed after aggregating the daily time series into monthly lightning density values. In the equatorial regions, PR92 has a positive bias over the Amazon and the Maritime Continent, but a strongly negative bias over central Africa (Fig. 5b). In the northern and southern extra-tropics, it shows an almost ubiquitous negative bias. F14 displays an even stronger negative bias in central Africa, as well as under-estimations in West-Africa, India, and East-Asia (Fig. 5c). The over-estimations in central Asia and western North-America are also evident. In all these regions, the ELM-tree bias is much smaller, even though it still under-estimates central African and North-American lightning rates (Fig. 5a). The monthly RMSE maps further demonstrate the improved performance of the ELM-tree (Fig. 5 d,e,f). In particular, errors of the ELM-tree are consistently smaller in East-Asia, Europe, Australia, over the Amazon, and most prominently, in central Africa. Finally, analyzing the monthly correlation (Fig. 5 g,h,i), all three lightning schemes capture monthly variability better over land than over the ocean, with correlation ($r$) generally $>0.6$ and statistically significant at the 0.01 level. On average, PR92 shows the weakest temporal correlation with observed values, being very low in Europe, India, and South-America for example. This highlights that PR92 is prone to compensating errors, artificially improving its bias metric in Figure 5b. The ELM-tree improvements in monthly correlation compared to both PR92 and F14 are ubiquitous over all land areas, except the Middle-East where F14 performs better. We identify particularly higher correlation values of the ELM-tree in Central- and North-America, India, Australia, Europe, and over the Amazon (Fig. 5g). Over the latter, the ELM-tree yields statistically significant monthly correlation, in contrast to PR92 and F14. Such a global-scale improvement in the correlation metric demonstrates...
Figure 5. Evaluation of the ELM-tree (top row), PR92 (middle row), and F14 (bottom row) lightning schemes with respect to the LIS observations: (a, b, c) mean bias, (d, e, f) monthly root mean square error (RMSE), (g, h, i) monthly Pearson correlation coefficient. In (g, h, i), note that the color scale lower bound is 0.6, and that hatched areas denote absence of statistical significance at the 0.01 level using a two-tailed t-test. All performance statistics are computed only over the test years (2000, 2005, 2010, 2020, 2023).

that the ELM-tree captures seasonal lightning patterns across a range of different climates with greater fidelity. In summary, Figure 5 shows that the ELM-tree performs better than PR92 and F14 in terms of bias, RMSE, and/or correlation in many regions. However, we note a slightly larger bias over the tropical oceans than the two other schemes (Fig. 5a).

Figures 5 a,b,c also show that all three schemes exhibit a negative bias in the same four regions: central Africa, North-West India, the south-eastern USA, and South-East South-America. The latter region is not discussed here because it coincides with the South-Atlantic anomaly and, therefore, high observational uncertainties (Christian et al., 2003). Central Africa, North-West India, and the south-eastern USA are major lightning hotspots (Fig. 4a), and are underestimated by other lightning schemes as well (e.g., Lopez, 2016; Stolz et al., 2017; Etten-Bohm et al., 2021). Still, we underline the large performance improvement of the ELM-tree in central Africa, where both bias and RMSE are more than halved compared to PR92 and F14 (Fig. 5 a,d).

Following this spatial evaluation, we also evaluate the skill of the lightning schemes in reproducing all single daily lightning density values from LIS. Figure 6 shows a density plot for each scheme, where each observed value is compared with the corresponding model estimate. Figure 6 includes every single observation during test years from both land and ocean. The ELM-tree demonstrates a coefficient of determination ($R^2$) of 0.706. This means that our model is able to reproduce $>70\%$ of the daily global-scale spatio-temporal variability. This is a relative improvement of $>42\%$ compared to PR92 and F14, which capture only 49.5% and 46.1% of the observed variability, respectively. In absolute terms, this corresponds to an RMSE decrease of $>1.3$ fl. km$^{-2}$ yr$^{-1}$ and a re-
Figure 6. Density plots showing all daily LIS observations during the test years (2000, 2005, 2010, 2020, 2023), and the corresponding model estimates from (a) the ELM-tree, (b) PR92, and (c) F14. Performance statistics for each lightning scheme are provided, where $N$ is the number of data samples, $R^2$ is the coefficient of determination, and RMSE is the root mean square error. The dashed line denotes the 1:1 perfect fit line. Note the logarithmic color scale.

duction of the negative bias of $>0.6$ fl. km$^{-2}$ yr$^{-1}$ ($>64\%$). Since we show counts with a logarithmic color scale in Figure 6, it is clear that the vast majority of pairs of observations and ELM-tree estimates are aligned along the 1:1 line. However, the ELM-tree $\hat{L}$ estimates deviate more strongly from the 1:1 line at very high lightning values $L$, which causes part of its negative bias. In Figure 6b, PR92 shows some very large $\hat{L}$ estimates (up to 86 fl. km$^{-2}$ yr$^{-1}$) for $L$ values very close to 0. This is likely attributable to its dependence on cloud top height to the power 4.9 over land, i.e., it is very sensitive to unusually large input values. In contrast, we note that the ELM-tree sensitivity to any climatic input feature is at most linear (Eqs. (1,4)), strongly reducing the risk of such unrealistic estimates during GCM runs.

Figure 7. Average lightning flash density (a) latitudinal distribution and (b, c) longitudinal distribution from the LIS observations and from the ELM-tree, PR92, and F14 lightning schemes. In (b) and (c), only areas in the extra-tropics and tropics are considered, respectively. The limit for the tropics is $\pm 23.44$ latitude. RMSE is the root mean square error and $r$ the Pearson correlation coefficient. Note that the y-axis in (c) spans twice the range of the y-axes in (a, b). Averaging is performed over the test years (2000, 2005, 2010, 2020, 2023).
Figure 7 focuses again on the spatial evaluation, showing the averaged latitudinal and longitudinal distributions of lightning density values. The latitudinal distribution (Fig. 7a) is best reproduced by the ELM-tree, with a higher correlation with respect to observations ($r = 0.94$) than PR92 ($r = 0.87$) and F14 ($r = 0.80$). The RMSE of F14 (1.33 fl. km$^{-2}$) is lower than that of PR92 (1.45 fl. km$^{-2}$), but the RMSE of the ELM-tree is further reduced by 50% (0.67 fl. km$^{-2}$). We separate the longitudinal distribution in two separate components: the extra-tropics (Fig. 7b) and the tropics (Fig. 7c). In the extra-tropics, the ELM-tree yields a correlation coefficient of 0.87, compared to 0.81 and 0.72 for PR92 and F14, respectively. In terms of RMSE, ELM-tree and F14 perform comparably (1.53 and 1.71 fl. yr$^{-1}$ km$^{-2}$, respectively), while PR92 confirms its strong negative extra-tropical bias (RMSE of 2.33 fl. yr$^{-1}$ km$^{-2}$, see also Fig. 5b). The better performance of the ELM-tree in the tropics is illustrated in Figure 7c. Its correlation coefficient is 0.95, substantially higher than PR92 (0.87) and F14 (0.81). Its RMSE of 1.90 fl. yr$^{-1}$ km$^{-2}$ is 36% lower than that of PR92 (2.97 fl. yr$^{-1}$ km$^{-2}$) and 52% lower than that of F14 (3.97 fl. yr$^{-1}$ km$^{-2}$). Figure 7c highlights once more the clear improvement in central African lightning representation of the ELM-tree, even though this longitudinal lightning peak is still underestimated. The ELM-tree RMSE in the tropics is close to that in the extra-tropics, despite lightning flash rate values being typically twice as large (Fig. 7 b,c). Paired with its high correlation ($r = 0.95$), this underlines the ELM-tree ability to simulate differences between the three tropical lightning chimneys: central Africa, the Amazon, and the Maritime Continent.

Figure 8. Temporal variability performance of the ELM-tree, PR92, and F14 lightning schemes with respect to LIS observations. Monthly climatology in the (a) northern and (b) southern hemispheres. Inter-annual variability in the (c) northern and (d) southern hemispheres. In (a, b), averaging is performed over the test years (2000, 2005, 2010, 2020, 2023). RMSE is the root mean square error and $r$ the Pearson correlation coefficient. Note that y-axes are different in (a, b) and (c, d).

Finally, Figure 8 focuses on performance in terms of temporal variability aggregated at the hemispheric scale. In particular, we show the monthly climatology and inter-annual variability, both computed only over the test years, in the northern and southern hemispheres separately. Figure 8 a,b shows that all three schemes reproduce the climatology well, as they demonstrate high correlation with observations ($r \geq 0.92$). However, Figure 8a demonstrates that both PR92 and F14 strongly underestimate the summer light-
ning peak in the northern hemisphere. The ELM-tree largely reduces this underestimation, resulting in a > 60% decrease in the northern hemisphere monthly climatology RMSE.

In the southern hemisphere, correlation coefficients of the three schemes for the monthly climatology are close, but the ELM-tree reaches an RMSE > 40% lower compared to PR92 and F14. Concerning the inter-annual variability (Fig. 8 c,d), we first acknowledge that 5 test years is a small sample to evaluate model skill. Over these 5 years, PR92 is not able to reproduce year-to-year variations, both in the northern ($r = 0.28$) and southern ($r = 0.19$) hemispheres. F14 performs better concerning this metric, yielding $r = 0.86$ and $r = 0.80$, respectively. Nevertheless, we find that the ELM-tree outperforms PR92 and F14 in both hemispheres, with $r > 0.90$. The RMSE decrease using the ELM-tree is even more pronounced, being > 70% and > 56% lower in the northern and southern hemispheres, respectively.

### 3.3 Simulations with the Community Earth System Model

We have implemented the ELM-tree as a lightning scheme in CESM2.2. The goal of the CESM2.2 simulation is not to compare in detail the model run to lightning, wildfire, and NO$_x$ observations. Such variables also depend on the simulated climate as well as parameterizations such as the ignition efficiency, and the amount of nitrogen generation per lightning flash. Furthermore, we use a coarse model set-up, with only the atmosphere and land components active, while the ocean and sea-ice components are inactive. The purpose here is to demonstrate that the ELM-tree can be implemented within a GCM source code, that modeled lightning values are stable, and that the ELM-tree can be fully-coupled to the wildfire ignition scheme.

Our 1995-2015 simulation yields an annual mean global lightning rate of 46.24±0.79 fl. s$^{-1}$, where ± denotes the standard deviation of inter-annual variability. This agrees well with the observed values from LRTS over 1996-2014 of 44.01±1.86 fl. s$^{-1}$, although the model underestimates inter-annual variability. Note that the years without entire annual coverage in the LRTS record are discarded in the latter calculation. According to OTD and LIS uncertainty estimates of ~ 5 fl. s$^{-1}$ (Christian et al., 2003; Blakeslee et al., 2020), the CESM2.2 global lightning output is within the observational uncertainty range.

The CESM2.2 mean lightning density spatial distribution, without spatio-temporal smoothing, is shown in Figure 9a. The spatial patterns are realistic compared to observations (see Fig. 4a), reproducing the contrasts between tropics and extra-tropics and between land and ocean, as well as high lightning regions such as eastern North-America and eastern Asia. On the other hand, some model biases appear, such as lightning overestimation in northwestern North-America and Australia, and underestimation in central Africa. Still, in general, we find very good agreement in both the main spatial patterns as well as the global total lightning.

Figure 9b shows the map of standard deviation in monthly variability, as modeled by the ELM-tree implemented in CESM2.2. Unsurprisingly, variability is highest in regions of high lightning rates. Furthermore, high-latitude regions, both in the northern and southern hemispheres, show a stronger relative variability, driven by more pronounced seasonal patterns. We also investigate the NO$_x$ production in our CESM2.2 run. The total annual mean production is 3.27 ± 0.05 Tg N yr$^{-1}$, which is within the estimated range of 2-8 Tg N yr$^{-1}$ (Schumann & Huntrieser, 2007). In the current CESM2.2 parameterization, each lightning flash is assumed to produce an equal and fixed amount of nitrogen (Price et al., 1997; Ridley et al., 2005). As such, there is a one-to-one relationship between modeled flash density and lightning-produced NO$_x$ (Fig. 9c). Finally, we also show the monthly correlation between lightning flash rates and burned area (Fig. 9d). While lightning serves as a natural ignition source, its relation to burned area is also strongly modulated by weather conditions. For example, we find that in the Sahel,
Figure 9. Output from the 1995-2015 CESM2.2 simulation with the ELM-tree implemented as the lightning model. (a) The mean lightning flash density, (b) the monthly standard deviation in lightning flash density, (c) the mean lightning-produced NO\textsubscript{x} rate, and (d) the monthly Pearson correlation between lightning flash density and burned area. In (a), the CESM2.2 lightning output has not been smoothed, in contrast to Figure 4. In (c), lightning-produced NO\textsubscript{x} scales linearly with lightning density from (a). In (d), note that the color scale saturates at 0.6, and that hatched areas denote absence of statistical significance at the 0.05 level using a two-tailed t-test.

nning is anti-correlated with burned area, as it occurs mostly during the wet season. In contrast, the correlation is positive in high-latitude regions, where most lightning events occur in summer.

The good agreement between our CESM2.2 test run using the ELM-tree implementation and observations of global total lightning rates, lightning spatial patterns, and global lightning-produced NO\textsubscript{x} generation is encouraging. This demonstrates that the ELM-tree provides a stable and realistic lightning scheme for GCMs, and further suggests that good fidelity with respect to observations can be achieved. However, we underline that the performance depends on GCM-specific aspects, such as the convection scheme, as well as the parameterization of cloud-to-ground versus intra-cloud flashes, and their respective efficiencies in producing NO\textsubscript{x} compounds.

4 Discussion

Lightning schemes developed for long spatial- and temporal-scales have traditionally used simple mathematical relationships and few climatic input features (e.g., Finney et al., 2014; Romps et al., 2014; Lopez, 2016). In contrast, in regional studies typically focusing on very short-term forecasts, machine learning methods have demonstrated the benefits of using more flexible models and a larger set of features (e.g., Mostajabi et al., 2019; Ukkonen & Mäkelä, 2019; Cavaiola et al., 2024). The latter approach poses the problem of potential implementation in GCMs. For example, Ukkonen & Mäkelä (2019) used a random forest model of 800 different trees and a neural network with multiple 30-neuron hidden layers; such architectures pose substantial implementation challenges in GCM code bases, or are expensive in terms of computing memory. In this study, we have used an intermediate approach by developing a data-driven and flexible model, while us-
ing several constraints to preserve a parsimonious model structure as well as improve generalization skills. We have demonstrated the improved fidelity of the ELM-tree with respect to observations compared to traditional lightning parameterizations. Implementation of the ELM-tree in GCM source code only requires a few if-conditions to reproduce the shallow tree-based architecture (Fig. 3), dot products between prescribed parameter values and common climatic variables (Eq. (1)), and a simple max() operation that we use as activation function (Eq. (4)). As such, our ELM-tree can be implemented in any GCM with minimal coding effort, as we have demonstrated for the case of CESM2.2. In the Supporting Information, we provide detailed pseudo-code and all parameter values to guide implementation of the ELM-tree in other GCMs.

The quality of our data-driven approach depends directly on the input data sets used. The ELM-tree has been calibrated with reanalysis data from ERA5 (Hersbach et al., 2020). While ERA5 is generally recognized as a state-of-the-art estimate of the atmosphere state, it inevitably includes errors, for example in the representation of convective systems (Lavers et al., 2022). Another challenge is the lightning data. Here, we have post-processed the available LIS-ISS data (Blakeslee et al., 2020) to create a data set which, in conjunction with LRTS (Cecil et al., 2014), spans a 29-year period. The smoothing procedure applied to all the LIS and OTD lightning data is required to compensate for satellite sampling biases. However, smoothing also affects the relationships between climatic variables and lightning. As both the diversity and quality of lightning data sets increase (Blakeslee et al., 2020; Kaplan & Lau, 2021), combining different data products could minimize their individual shortcomings (e.g., Bitzer et al., 2016), and thus further enhance global lightning modeling prospects.

The list of potential climatic features used here is non-exhaustive (Table 1). We have focused on features that are commonly available from both GCMs and reanalysis data. Previous work included other features, such as convective mass flux or convective ice mass flux at a given pressure level (Allen & Pickering, 2002; Finney et al., 2014), and within-cloud vertical profiles of concentration of snow, graupel, and cloud condensates (Lopez, 2016). Another non-climatic predictor of considerable importance is aerosol concentration, because it is supposed to exert an influence complementary to thermodynamics on cloud dynamics and lightning (Williams & Sátori, 2004). Aerosol concentration has been included in the parameterization of Stolz et al. (2017) for example, but was derived from a global chemical-transport model. Their use of model output was motivated by large uncertainties in observation-based aerosol concentration products. Retrieval from satellite measurements of aerosol properties is particularly challenging near clouds (Marshall et al., 2021), which are the conditions most relevant to lightning. Furthermore, the performance of GCMs in simulating atmospheric aerosol concentrations remains questionable (Vogel et al., 2022). Our objective here is to develop an observation-based lightning model to be implemented in GCMs. As such, given these observation- and GCM-related shortcomings with respect to aerosols, we have preferred not to include aerosols in the lightning model. Overall, the input features used here are easy to obtain from standard reanalyses and GCM outputs, without additional processing or code modification requirements.

We have demonstrated that the ELM-tree lightning model implemented in a state-of-the-art GCM produces stable and realistic lightning rates. In particular, the global total lightning rate and lightning-produced NO$_x$ generation are within observational uncertainty ranges, and spatial lightning patterns reproduce the main characteristics from observations. Nevertheless, for improving the simulation of lightning-produced NO$_x$ and lightning-ignited wildfires in GCMs, the improved lightning flash model presented here is only a first, yet necessary step. For example, both of these processes are sensitive to the ratio of intra-cloud to cloud-to-ground lightning. This ratio is often assumed to depend on the sub-freezing cloud thickness (Price & Rind, 1993), but this relationship is still uncertain at the global scale (Gordillo-Vázquez et al., 2019). Generally, the sensi-
ivity of NO$_x$ production to multiple factors is poorly constrained, including to latitude, to intra-cloud versus cloud-to-ground flashes, and to continental versus maritime environments (Pickering et al., 2016). The vertical distribution of lightning-produced NO$_x$ is another factor of great importance in atmospheric chemistry modeling, yet it is also estimated with uncertain parameterizations (Pickering et al., 1998; Gordillo-Vázquez et al., 2019). Similarly, concerning wildfires, the ignition probability from a lightning discharge depends on complex factors such as the duration of the continuing current phase (Pérez-Invernón et al., 2023), rendering ignition parameterizations in GCMs poorly constrained (Li et al., 2012).

We have quantified the ELM-tree improved performance compared to the well-established benchmarks PR92 (Price & Rind, 1992) and F14 (Finney et al., 2014). In particular, the ELM-tree achieves a $>42\%$ improvement in terms of spatio-temporal variability at the global scale (Fig. 6). On the one hand, first-order formulations such as PR92 and F14 have the advantage that lightning can, as a first approximation, be intuitively related to a small number of climatic proxies. On the other hand, given the complexity of the lightning process, more flexible approaches can capture relationships that are hard to constrain from physical principles (Ukkonen & Mäkelä, 2019; Cheng et al., 2024). The ELM-tree proposed in this study is more flexible than traditional GCM lightning parameterizations, still its interpretable architecture illustrated in Figure 3 offers insights into the model behavior. As an example, we analyze the model dependence on CAPE, because this feature has a predominant importance on lightning in our ELM-tree (Fig. 1b), which is also recognized in the literature (e.g., Romps et al., 2014; Lopez, 2016; Cheng et al., 2024). The ELM-tree uses CAPE as split-feature twice, effectively separating the tree in different CAPE regimes (Fig. 3). We find that CAPE is used as fit-feature in three of the six leaf node ELMs. However, it is not used in any of the two leaf node ELMs in the high CAPE regime, i.e., the two right-most leaf node ELMs in Figure 3. The ELM-tree thus identifies that, in high CAPE conditions, the lightning rate becomes insensitive to CAPE. This is supported by prior studies, showing that past a certain CAPE threshold, convective activation shows a weak or absent dependence on CAPE (e.g., Sherwood, 1999). This effect is also found by analyzing the marginal sensitivity of the ELM-tree to all CAPE values from our training data set (Fig. S2). In particular we find good agreement between the model and observed marginal sensitivities, but the model insensitivity to CAPE initiates at lower CAPE values. More generally, we find that the model reproduces the observed marginal sensitivities to any of the ELM-tree input features with good skill (Fig. S2).

5 Conclusions

We have developed a machine learning method to design a global lightning model. Our ELM-tree lightning model uses Extreme Learning Machine models embedded in a decision tree. This approach is more flexible than traditional lightning parameterizations for Global Climate Models (GCMs), but we have enforced several constraints to keep the model parsimonious, and therefore facilitate its implementation in any climate model. Our implementation of the ELM-tree in CESM2.2 demonstrates that it yields stable and realistic lightning predictions, agreeing within uncertainty ranges with observations without requirements of further scaling factors. In the Supporting Information, we provide a detailed pseudo-code as well as all parameter values to give a clear guide to the implementation of the ELM-tree in GCMs. As an intermediate approach between high-dimensional machine learning models and first-order parameterizations, our model offers a pragmatic and efficient solution for an improved lightning representation in GCMs.

Evaluation on test years not used in the ELM-tree calibration has demonstrated the improved fidelity of the ELM-tree with respect to observations when compared to well-established lightning schemes. In particular, the ELM-tree explains 70.2% of the daily global spatio-temporal lightning variability, which is a $>42\%$ relative improvement com-
pared to these benchmarks. When scaled for the entire globe, such reduced errors could be of large impact in, for example, closure of the global ozone budget (Wild, 2007). A notable improvement is the ability to capture the differences between the three main lightning chimneys: central Africa, the Amazon, and the Maritime Continent. Furthermore, the ELM-tree reaches high correlation with observed lightning across all land masses, demonstrating its ability to reproduce seasonal patterns in a range of different climates. In addition to improved regional lightning representation, the ELM-tree performance is validated at the global scale: observed variations along latitude and along longitude are both represented with a correlation coefficient > 0.86. Similarly, at the scale of the northern and southern hemispheres, the ELM-tree represents the monthly lightning climatology and inter-annual variability with correlation coefficients > 0.92 and > 0.90, respectively.

Our proposed novel lightning model for GCMs opens the door to numerous research avenues. Building upon previous work, it can help the lightning modeling community to further study unsolved problems, such as lightning changes under different anthropogenic climate change scenarios, the climate-lightning-wildfire three-way interactions, and future changes in tropospheric concentrations of ozone and methane.

6 Open Research

The LRTS gridded lightning data product Cecil et al. (2014) and the LIS-ISS flash count data are available via the NASA Earthdata portal (https://www.earthdata.nasa.gov). The reanalysis data are available via the Copernicus Climate Data Store portal (https://cds.climate.copernicus.eu/). The CESM model can be downloaded from https://github.com/ESCOMP/CESM. Detailed guidelines for the implementation of the ELM-tree in climate models and all parameter values are provided in the Supporting Information. In addition, the corresponding author can be contacted for any question related to the ELM-tree model. Source code of the ELM-tree model, as well as scripts to reproduce evaluation figures of this study are available in a Zenodo repository (Verjans, 2024).

Acknowledgments

We thank colleagues at the IBS Center for Climate Physics for their comments and insights about this study, in particular Ja Yeon Moon and Jiaoyang Ruan. The simulations were conducted on the IBS/ICCP supercomputer “Aleph,” a 1.43 petaflops high-performance Cray XC50-LC Skylake computing system with 18 720 processor cores, 9.59 PB storage, and 43 PB tape archive space. We also acknowledge the support of KREONET. VV and CF were supported by the Institute for Basic Science (IBS), Republic of Korea, under IBS-R028-D1, and CF by the National Research Fund of Korea (NRF-2022M3K3A1097082).

References


-27-


Development of a data-driven lightning model for implementation in Global Climate Models

Verjans Vincent¹,² and Christian L. E. Franzke¹,²

¹Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea
²Pusan National University, Busan, Republic of Korea

Key Points:

• Using lightning records spanning 29 years, we develop a global model capturing the relationship between lightning and large-scale climate.
• Model evaluation shows improved representation of spatio-temporal variability at all scales compared to established lightning schemes.
• Our model uses machine learning with complexity constraints, facilitating implementation in climate models, with verified stability.

Corresponding author: Verjans Vincent, vverjans@pusan.ac.kr
Abstract
This study proposes a new global-scale lightning model, predicting lightning rates from large-scale climatic variables. Using satellite lightning records spanning a period of 29 years, we apply machine learning methods to derive a functional relationship between lightning and climate reanalysis data. In particular, we design a model tree, representing different lightning regimes with separate single hidden layer neural networks of low dimensionality. We apply multiple complexity constraints in the model development stages, which makes the lightning model straightforward to implement as a lightning scheme for global climate models (GCMs). We demonstrate that, for years not used for model training, our lightning model captures 70.6% of the daily global spatio-temporal lightning variability, which corresponds to a > 42% relative improvement compared to well-established lightning schemes. Similarly, the model correlates well with lightning observations for the monthly climatology (r > 0.92), inter-annual variability (r > 0.90), and latitudinal and longitudinal distributions (r > 0.86). Most notably, the model brings a critical improvement in representing lightning magnitude and variability in the three tropical lightning chimney regions: central Africa, the Amazon, and the Maritime Continent. We implement the lightning model in the Community Earth System Model to verify its stability and performance as a GCM component, and we provide detailed implementation guidelines. As an intermediate approach between high-dimensional machine learning models and first-order lightning parameterizations, our model offers GCMs a straightforward and efficient tool to improve lightning simulation, which is critical for representing atmospheric chemistry and naturally-ignited wildfires.

Plain Language Summary
Lightning is a worldwide phenomenon, which affects atmospheric chemistry and can cause wildfire ignitions. However, representing lightning in climate models at the global scale is challenging, because it depends on small-scale physical processes not explicitly represented in global models. In this study, we develop a lightning model, which estimates lightning rates from large-scale climate variables. We use machine learning methods to extract differences in the relationship between lightning and climate in different lightning regimes. We impose several constraints to keep the lightning model simple, which facilitates implementation and use of the model as a component of global climate models. We show that predictions of the lightning model reproduce temporal and spatial variability of lightning observations. In addition, the match to observed lightning rates is improved compared to currently widely used lightning schemes.

1 Introduction
Lightning is a prevalent phenomenon on Earth, with an estimated mean global lightning flash rate of 44±5 fl. s⁻¹ (Christian et al., 2003; Cecil et al., 2014; Blakeslee et al., 2020). Lightning is unevenly distributed, occurring predominantly over land and in the tropics (Christian et al., 2003). It also exhibits temporal variability across timescales: sub-daily, seasonal, annual, and inter-annual (Williams et al., 2005). This spatio-temporal variability of lightning is driven by its sensitivity to climate conditions. Yet, the climate-lightning relationship is not understood in detail. For example, the scientific community has yet to reach a consensus about the mechanisms behind the land-ocean contrast, about the causes for differences between tropical regions, and even about future changes in lightning regimes under anthropogenic climate change (Williams & Stanfill, 2002; Finney et al., 2018; Romps, 2019).

There are important reasons to understand the sensitivity of lightning to climate. First, lightning is estimated to produce ~ 10% of the global atmospheric emissions of nitrogen oxides (NOₓ), and up to ~ 23% in the tropics (Schumann & Huntrieser, 2007). Lightning emissions of NOₓ predominate in the middle- and upper-troposphere, where
other NO\textsubscript{x} emissions are mostly absent, and where the NO\textsubscript{x} lifetime is longer than at
the surface. As a consequence, lightning plays a disproportionate role on tropospheric
chemistry (Wild, 2007). Since ozone (O\textsubscript{3}) net production depends non-linearly on the
NO\textsubscript{x} mixing ratio, lightning affects the concentration and distribution of tropospheric
O\textsubscript{3}, which is both an oxidant and a greenhouse gas (Schumann & Huntrieser, 2007). Fur-
thermore, lightning-produced NO\textsubscript{x} also impacts the oxidising capacity of the atmosphere
through their effects on the concentration of hydroxyl radicals (OH), which are the pri-
mary regulators of methane (CH\textsubscript{4}) losses (Schumann & Huntrieser, 2007). Second, light-
ning is almost exclusively responsible for all natural wildfire ignitions (Veraverbeke et
al., 2017; Janssen et al., 2023). Observations suggest that human- and lightning-ignited
fires have different characteristics; the latter are more sensitive to climatic conditions through
fuel moisture, and generally occur in more remote locations (Balch et al., 2017; Veraver-
beke et al., 2017). Lightning-ignited fires therefore explain most of the temporal vari-
ability in burned area in some specific ecosystems, such as boreal and intact forests (Ve-
raverbeke et al., 2017; Janssen et al., 2023). While changing climatic conditions lead to
increased frequency and severity of lightning-ignited wildfires in some regions, there re-
main major uncertainties in the future distribution of climate-lightning-wildfire inter-
actions (Janssen et al., 2023; Pérez-Invernón et al., 2023).

GCMs do not explicitly simulate cloud electrification and lightning, because of the
sub-kilometer resolution required to simulate these fine-scale processes (Fierro et al., 2015).
Instead, GCMs rely on lightning parameterizations, which empirically relate large-scale
atmospheric conditions to lightning flash rates (e.g., Price & Rind, 1992; Lopez, 2016;
Gordillo-Vázquez et al., 2019). Alternatively, GCMs use a pre-processed observation-based
lightning climatology as input forcing, such as in the most recent Fire Modeling Inter-
comparison Project (Rabin et al., 2017). With the latter approach, GCMs ignore the cause-
effect relation between climate and lightning. Advances in representing convection- and
cloud-related variables in GCMs (e.g., Peters et al., 2017) offer opportunities to, in turn,
 improve the accuracy of GCM lightning schemes. Such progress would enable consistency
between modeled climate and both atmospheric chemistry and wildfires.

The advent of lightning flash rate observations from satellites (Christian et al., 2003;
Cecil et al., 2014) has enabled major improvements in recent large-scale lightning pa-
parameterizations (e.g., Finney et al., 2014; Lopez, 2016; Stolz et al., 2017; Etten-Bohm
et al., 2021). Also, the pioneering parameterization of Price & Rind (1992) is still ex-
tensively used in GCMs (Thornhill et al., 2021). Despite, important scientific efforts, large
uncertainties remain in lightning parameterizations, which in turn translate in uncertain-
ties in atmospheric chemistry and wildfire impacts. For example, most parameterizations
use a multiplicative scaling factor to approximately match the observed global total light-
ing flash rate. For commonly used parameterizations, this factor typically spans the range
0.05 to 4.00 (Gordillo-Vázquez et al., 2019). The use of such a factor is necessary to yield
modeled lightning-produced NO\textsubscript{x} predictions in agreement with observation-based es-
timates (5±3 Tg nitrogen yr\textsuperscript{-1}, Schumann & Huntrieser, 2007). As another example, the
different sensitivities of state-of-the-art lightning parameterizations to anthropogenic cli-
mate change imply that future changes in lightning are essentially unknown: predictions
of global lightning change by 2100 in high-emission scenarios span -15% to +43% (Finney
et al., 2018; Romps, 2019). Finally, most parameterizations are calibrated to data from
restricted areas, generally the tropics or North-America, where most lightning data are
available (e.g., Finney et al., 2014; Romps et al., 2014; Stolz et al., 2017). While the trop-
ics account for \(\frac{2}{3}\) of global lightning (Christian et al., 2003), globally-accurate light-
ing predictions are important, for example to estimate wildfire risks in sensitive boreal
forests (Janssen et al., 2023).

Improvements in observational and data assimilation techniques are driving increas-
ing quality and quantity of climate reanalysis products (e.g., Hersbach et al., 2020). In
parallel, advances in statistical modeling and machine learning methods, as well as com-
putational power, are offering new opportunities in data-driven climate sciences (Bracco et al., 2018; Reichstein et al., 2019). Such techniques are increasingly used to predict lightning from weather conditions (Ukkonen & Mäkelä, 2019; Cheng et al., 2024). However, the implementation of machine learning methods within state-of-the-art numerical climate models faces many challenges, mostly related to codebase compatibility issues (Partee et al., 2022). In this study, we develop a data-driven lightning model based on climate reanalyses and satellite lightning measurements. We exploit satellite records of lightning spanning a period of 29 years, which, to the best of our knowledge, is longer than any previous study for development of a global-scale lightning scheme. We use a parsimonious machine learning approach, with the objective to make the lightning model straightforward to implement as a GCM component. We detail the model calibration, demonstrate its fidelity with respect to observations, including comparisons with other lightning parameterizations, and implement it in the Community Earth System Model (CESM) (Danabasoglu et al., 2020).

2 Methods

2.1 Lightning Data

We calibrate our lightning parameterization to satellite observations of lightning flashes from three separate missions. The first spaceborne optical sensor used to measure lightning flash rates from space was the Optical Transient Detector (OTD) (Christian et al., 2003). The OTD mission covered the period 1995-2000, with an extensive latitudinal range of ±75°. OTD was a prototype for the subsequent Lightning Imaging Sensor (LIS), launched in 1997 onboard the Tropical Rainfall Measuring Mission (TRMM). The LIS-TRMM exclusively covered low-latitude regions (±38° latitude), and was fully functional until early 2014. Combining measurements from OTD and LIS-TRMM has allowed the production of lightning data sets, which have provided unprecedented details about lightning spatio-temporal variability (Cecil et al., 2014). Both OTD and LIS being onboard low-orbit satellites, they performed ≥14 orbits per day, but with a viewing duration of any location at a given pass of 1 to 3 minutes. OTD and LIS-TRMM had a 10 and 5 km resolution, respectively. Their detection efficiencies varied depending on local time, ranging between 0.38 and 0.52 for OTD and 0.69 and 0.88 for LIS-TRMM. We refer to Cecil et al. (2014) for a thorough description of OTD and LIS-TRMM measurement capabilities, and for the production of lightning flash rates data sets. These lightning data sets have been extensively used to develop statistical relationships between lightning flash rates and climatological variables (e.g., Finney et al., 2014; Lopez, 2016; Stolz et al., 2017).

Since 2017, the LIS mission has been extended by being set to work on the International Space Station (ISS) platform (LIS-ISS, Blakeslee et al., 2020). Compared to LIS-TRMM, LIS-ISS has improved latitudinal coverage (±55°) and horizontal resolution (4 km). Importantly, LIS-ISS allows continuing spaceborne lightning measurements across all longitudes beyond the LIS-TRMM mission. Yet, to the best of our knowledge, this new record, spanning >7 years at the time of writing, has not been used in the development of any lightning model.

In this study, we use the gridded daily time series of flash rate and viewtime from the combined OTD and LIS-TRMM data, referred to as LRTS in Cecil et al. (2014). This data set was established by applying spatio-temporal smoothing to raw counts of flash measurements in order to alleviate sampling biases induced by viewing time limitations (Cecil et al., 2014). LRTS provides daily flash rate density values at a 2.5° × 2.5° resolution from May 1995 until February 2014. In contrast, no post-processed, ready-to-use gridded data set of the LIS-ISS measurements is available at the time of this study. Instead, only files of flash counts and satellite viewtimes are available. Here, we process the LIS-ISS data from 2017 to 2023 included in order to extend the daily 2.5° × 2.5°
LRTS time series, albeit with a gap from February 2014 to February 2017. For converting the LIS-ISS data from raw flash counts and viewtimes into gridded time series, we follow the description of Cecil et al. (2014) for LRTS. First, we divide any flash count value by the corresponding satellite viewtime. Second, following the information from the LIS-ISS user guide, we scale this flash rate by the LIS detection efficiency as a function of local time (Table 2 of Cecil et al., 2014). Third, the scaled effective flash rates are aggregated on a 2.5°×2.5° grid, and converted to flash densities [fl. km⁻² yr⁻¹]. We exclude outliers using a 5-sigma rule, and values are adjusted with a small latitude-dependent factor to match total lightning flash rates provided in Table 1 of Blakeslee et al. (2020). Fourth, we apply both a 7.5°×7.5° window moving average in space, and a 100-day window moving average in time. Note that for simplicity, we do not perform further digital filtering, as the smoothing procedure resulted in a well-filtered signal.

2.2 Climate data

For relating measured lightning density to climatic variables at the global scale, we use reanalysis data from ERA5 (Hersbach et al., 2020). ERA5 offers a broad range of atmospheric variables at 31 km horizontal resolution, on 137 vertical levels, and at a hourly temporal resolution. We refer to Hersbach et al. (2020) for a detailed description of ERA5. In this study, we use ERA5 daily mean data, by averaging any climatic fields at the 0:00, 6:00, 12:00, and 18:00 UTC time steps. We consider a large set of potential climatic variables that can serve as input features for our lightning model. In line with our objective to develop a model scheme that can be implemented in GCMs, we restrict the potential climatic features to those which are typical GCM variables. The potential ERA5 input features used are listed in Table 1. We note that cloud top height and cloud cover are taken from the Advanced Very High Resolution Radiometer data set (Karlsson et al., 2023). Also in Table 1 are two time-constant geographical features: surface elevation and absolute latitude. Every input feature is gridded on the 2.5°×2.5° grid of LRTS.

The selection of these 25 potential features (Table 1) is motivated by previous research. For example, Convective Available Potential Energy (CAPE), cloud-top-height, and precipitation are some of the most commonly-used climatic features to predict lightning density values (e.g., Price & Rind, 1992; Romps et al., 2014; Lopez, 2016). We include features quantifying ice and liquid water contents in clouds, as they govern cloud electrification and charge transfer mechanisms (Saunders, 1993). Wind can modulate the regime of convective storms (Weisman & Klemp, 1982). Here, in addition to including 10m wind magnitude and vertical wind velocity at 500 hPa, we also follow Etten-Bohm et al. (2021) by computing deep wind shear and low-level wind shear between the 900 to 300 hPa and 900 to 700 hPa levels, respectively. Also, while surface latent and sensible heat fluxes are usually ignored in lightning parameterizations, Williams & Stanford (2002) have theorized that they exert a thermodynamical control on the convection regime, and therefore on lightning.

2.2.1 Training and test data separation

Our data set consists of all daily lightning density values from the aggregation of the LRTS and LIS-ISS data, spanning 1995-2023, and the corresponding daily climatic variables. We separate the data in two subsets: training, and test data. We split the data based on entire years. We choose the years 2000, 2005, 2010, 2020, and 2023 as the test years. Data from these years are completely ignored when calibrating our lightning model. Accounting for the lightning data gap in 2015-2017, the remaining 22 years are used for training, but lightning data are available only for part of the years 1995, 2014, and 2017. In total, the combined training and test data sets include >50×10⁶ daily lightning flash density values on the 2.5° × 2.5° grid.
Table 1. Potential input features considered for the global lightning model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Units</th>
<th>Formulation details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convective Available Potential Energy</td>
<td>CAPE</td>
<td>J kg s^{-2}</td>
<td>/</td>
</tr>
<tr>
<td>Cloud-top-height</td>
<td>cth</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Cloud-base-height</td>
<td>cbh</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Cloud cover</td>
<td>c_f</td>
<td>%</td>
<td>/</td>
</tr>
<tr>
<td>2m temperature</td>
<td>T_{2m}</td>
<td>K</td>
<td>/</td>
</tr>
<tr>
<td>2m dew point temperature</td>
<td>T_{d,2m}</td>
<td>K</td>
<td>/</td>
</tr>
<tr>
<td>Specific humidity (850 hPa)</td>
<td>q_{850hPa}</td>
<td>kg kg^{-1}</td>
<td>/</td>
</tr>
<tr>
<td>Cloud ice content</td>
<td>c_{ice}</td>
<td>kg km^{-2}</td>
<td>/</td>
</tr>
<tr>
<td>Cloud liquid water content</td>
<td>c_{liq}</td>
<td>kg km^{-2}</td>
<td>/</td>
</tr>
<tr>
<td>Column total water vapor</td>
<td>m_{vap}</td>
<td>kg km^{-2}</td>
<td>/ c_{ice} + c_{liq} + m_{vap}</td>
</tr>
<tr>
<td>Column total water</td>
<td>m_{w}</td>
<td>kg km^{-2}</td>
<td>(T_{850hPa} - T_{500hPa}) + T_{d,850hPa} - (T_{700hPa} - T_{d,700hPa})</td>
</tr>
<tr>
<td>K-index</td>
<td>K_{I}</td>
<td>°C</td>
<td>(T_{850hPa} - T_{500hPa}) + T_{d,850hPa} - (T_{700hPa} - T_{d,700hPa})</td>
</tr>
<tr>
<td>Total totals index</td>
<td>t_{t}</td>
<td>K</td>
<td>(T_{850hPa} + T_{d,850hPa} - 2T_{500hPa})</td>
</tr>
<tr>
<td>Wind speed magnitude at 10m</td>
<td>U_{10m}</td>
<td>m s^{-1}</td>
<td>(a_{10m}^2 + v_{10m}^2)^{1/2}</td>
</tr>
<tr>
<td>Lagrangian pressure tendency (500 hPa)</td>
<td>ω_{500hPa}</td>
<td>Pa s^{-1}</td>
<td>negative of the vertical wind speed (500 hPa)</td>
</tr>
<tr>
<td>Low-level wind shear</td>
<td>s_{w,low}</td>
<td>m s^{-1}</td>
<td>[(U_{900hPa} - U_{700hPa})^2 + (V_{900hPa} - V_{700hPa})^2]^{1/2}</td>
</tr>
<tr>
<td>Deep-level wind shear</td>
<td>s_{w,deep}</td>
<td>m s^{-1}</td>
<td>[(U_{900hPa} - U_{300hPa})^2 + (V_{900hPa} - V_{300hPa})^2]^{1/2}</td>
</tr>
<tr>
<td>Total precipitation</td>
<td>t_{p}</td>
<td>m s^{-1}</td>
<td>/</td>
</tr>
<tr>
<td>Convective precipitation</td>
<td>c_{p}</td>
<td>m s^{-1}</td>
<td>/</td>
</tr>
<tr>
<td>Zero degree level</td>
<td>h_{OC}</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Surface pressure</td>
<td>p_{s}</td>
<td>Pa</td>
<td>/</td>
</tr>
<tr>
<td>Surface latent heat flux</td>
<td>Q_{l}</td>
<td>W m^{-2}</td>
<td>positive downwards</td>
</tr>
<tr>
<td>Surface sensible heat flux</td>
<td>Q_{e}</td>
<td>W m^{-2}</td>
<td>positive downwards</td>
</tr>
<tr>
<td>Surface elevation above sea-level</td>
<td>h_{s}</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Absolute latitude</td>
<td>[\varphi]</td>
<td>°</td>
<td>/</td>
</tr>
</tbody>
</table>

u and v denote the longitudinal and latitudinal component of the wind speed [m s^{-1}], respectively.

hPa and m subscripts denote a feature value at a given pressure level [hPa] and height level [m], respectively.
2.3 Tree model

For our global lightning scheme, we calibrate a model with lightning flash density $L$ as response variable [fl. km$^{-2}$ yr$^{-1}$]. Our model uses a decision tree architecture. The general principle of decision trees is to partition the feature space into a set of separate entities, referred to as the leaves of the tree (Hastie et al., 2009; Costa & Pedreira, 2023). The decision tree splits input data following feature-dependent decision rules, starting from the root node, and until each data sample is assigned to a leaf node. A decision tree can be further refined into a model tree, which uses separate regression models in each leaf to capture leaf-specific dependencies of the response variable to the input features. The tree architecture offers numerous advantages: separation between different lightning regimes, interpretable structures, and ease of implementation as a scheme in GCMs.

The most common approach for model trees is to use linear regression models in the leaves (Costa & Pedreira, 2023). However, there is flexibility possible in this choice (Zeileis et al., 2008). In this study, we design our leaf models as Extreme Learning Machines (ELMs, Huang et al., 2006). An ELM is a generalized version of a single hidden layer feedforward neural network (SLFN), and it can be used in both classification and regression settings (Huang et al., 2012). It has been proposed that ELMs can be incorporated as part of model trees, and such ELM-trees have shown good performance on a variety of data sets, with notable efficiency for large data sets (Wang et al., 2015; Zhou & Yan, 2019). ELMs also offer the advantage of being simple to formulate, and thus facilitate prospects of implementation in GCMs.

In this section, we explain the process of tree induction, i.e., the building of the ELM-tree. Deep trees with many branches and leaves can provide more flexible models of $L$. On the other hand, excessive tree size can lead to overfitting, or to large increases in model complexity for little gain in model skill. The tree induction is therefore an optimization process with the goal to partition the feature space such that our model correctly predicts lightning density values, while using few partitions and ELMs of low complexity. The entire tree induction process consists of three separate steps, detailed in the next three sub-sections.

2.3.1 Tree growing

In the first step of the tree induction, we grow an oversized tree with the commonly-used Classification And Regression Tree algorithm (CART, Breiman et al., 1984). CART builds the tree starting from the root node. At each node it finds the optimal splitting feature, referred to as a split-feature, and its splitting value. In this sense, CART is a greedy algorithm: at each split, it aims to minimize the weighted standard deviation in the target variable in the two child nodes. In principle, this process could continue until the number of leaves equals the number of data samples. In our tree induction procedure, we grow the oversized tree with a maximum depth of 8, corresponding to 129 leaf nodes and 256 nodes in total.

The distribution of the global daily lightning data is characterized by a large proportion of values close to 0. Due to the spatio-temporal smoothing applied to the lightning measurements, it is likely that most instances of lightning density close to zero correspond to lightning-free days. In order to deal with this particularity of satellite lightning density data, we force the first split of the tree to separate a zero-branch from a non-zero-branch. In the zero-branch, the tree immediately ends into a leaf that always returns the zero value, and thus incorporates no leaf model. In contrast, tree growing continues in the non-zero-branch. Using the zero-branch allows the ELM calibration to focus on non-negligible lightning values. Here, we set a threshold of 0.1 fl. km$^{-2}$ yr$^{-1}$ as a limit between daily lightning values to be classified as zero or not. In contrast to the other splitting decisions of the ELM-tree, the zero-branch split is a classification task: we need to find the optimal decision rule for separating $L \leq 0.1$ versus $L > 0.1$ fl. km$^{-2}$ yr$^{-1}$.
We use the receiver operating characteristic curve (ROC curve) to optimize both the split-feature, and its splitting threshold for the zero-branch.

### 2.3.2 ELM calibration

In the second step of tree induction, we fit an ELM in each node of the tree, except the zero-branch. This includes both the leaves and the interior nodes of the oversized tree. The ELMs are fit only to those data samples falling into the corresponding node. As a SLFN, for any given node, an ELM model for lightning is formulated as:

\[
\hat{L}_i = \sum_{j=1}^{\tilde{N}} \beta_j g(\mathbf{w}_j^T \mathbf{x}_i),
\]

where \(i\) denotes the \(i\)th data sample, \(\tilde{N}\) is the number of hidden neurons in the SLFN, \(\mathbf{x}_i\) is the vector of input features with an intercept term, \(\mathbf{w}_j\) is the weight vector of the \(j\)th hidden node, \(g()\) is the activation function, and \(\beta\) is the weight vector connecting the hidden layer to the output, \(\hat{L}_i\). In ELMs, the hidden weights \(\mathbf{w}\) do not need to be calibrated. Instead, they are generated randomly, and only the output weights \(\beta\) are calibrated. This has the major advantage that an analytical solution exists for \(\beta\) (Huang et al., 2006), making the model calibration fast and scalable to large data sets. Assuming that there are \(N\) data samples, Equation (1) can be rewritten in matrix notation:

\[
\hat{L} = H\beta,
\]

where \(\beta\) is of size \((\tilde{N} \times 1)\), and the hidden layer matrix \(H\) of size \((N \times \tilde{N})\) is given by:

\[
H = \begin{bmatrix}
g(\mathbf{w}_1^T \mathbf{x}_1) & \ldots & g(\mathbf{w}_N^T \mathbf{x}_1) \\
\vdots & \ddots & \vdots \\
g(\mathbf{w}_1^T \mathbf{x}_N) & \ldots & g(\mathbf{w}_N^T \mathbf{x}_N)
\end{bmatrix}.
\]

We fix the number of hidden neurons to 4, i.e., \(\tilde{N} = 4\). This small number of hidden neurons ensures that the lightning ELM-tree remains easy to implement in GCMs, using Equation (1). We verify (Fig. S1) that the rate of model performance increase reduces once the number of hidden neurons reaches 4. For \(g()\), we use the rectified linear unit (ReLU) activation function:

\[
g \left( \mathbf{w}_j^T \mathbf{x}_i \right) = \max \left( 0, \mathbf{w}_j^T \mathbf{x}_i \right).
\]

ELMs can accommodate a large variety of activation functions (Huang et al., 2006). In prior investigations, we experimented with the use of the logistic and softplus activation functions, which showed comparable performance but with larger numbers of fit-features. We decided to use the ReLU to maintain ELMs of lower dimensionality, as well as for the straightforward implementation of Equation (4) in GCMs. At each ELM calibration, we sample the hidden weights randomly from a uniform distribution bounded between \([-1; 1]\). This random sampling is performed 10 times, and the ELM is calibrated with each random sample. Only the random sample of \(\{\mathbf{w}_j\}_{j=1,\tilde{N}}\) leading to the lowest root mean square error (RMSE) is retained. This avoids forcing the ELM to use weights that were, by chance, a poor choice for fitting Equation (1). It should also be noted that we scale each feature to the range \([0, 1]\) for the calibration process, but we rescale the ELM coefficients after calibration such that computation on the true scale of feature values is strictly equivalent.

Fitting coefficients \(\beta_j\) for given hidden weights \(\mathbf{w}_j\) is an analytical step. The minimal norm least squares solution can be found using the Moore-Penrose generalized inverse, \(H^+\), of the hidden matrix \(H\). Here, we add a small L2 regularizer \(\lambda\) to make the solution for \(\beta\) more stable (Huang et al., 2012):

\[
H^+ = (H^T H + \lambda I)^{-1} H^T,
\]
where $I$ is the identity matrix, and we set $\lambda = 0.1$. The solution for $\beta$ is given by:

$$\beta = H^+ L.$$  \hspace{1cm} (6)

At each node, $H^+$ and $L$ in Equation (6) use only a subset of the daily lightning and climatic data samples, corresponding to those samples falling into the given tree node.

While we use a zero-branch calibrated to lightning events $\leq 0.1$ fl. km$^{-2}$ yr$^{-1}$, nothing prevents an ELM to output values $\leq 0.1$ fl. km$^{-2}$ yr$^{-1}$ as well. This motivates the choice of the small threshold 0.1 fl. km$^{-2}$ yr$^{-1}$ for the zero-branch classification: the classifier does not need to capture all the low lightning events, because $L$ can also be small in the non-zero-branch. Due to the formulation of SLFNs (Eq. (1)), ELM output could range on the entire real line. As such, we set any predicted negative $L$ value to 0.

Initially, there are 25 features available to each ELM (Table 1). Including all the features would result in very high-dimensional ELMs, as the size of each hidden weight vector $\mathbf{w}$ scales linearly with the number of features. We use the Bayesian Information Criterion (BIC, Schwarz, 1978) to find the optimal number of features to include in each ELM. Assuming normally-distributed errors, the BIC can be formulated as:

$$\text{BIC} = B \log(N) + N \log \left( \frac{1}{N} \sum_{i=1}^{N} \left(L_i - \hat{L}_i\right)^2 \right),$$  \hspace{1cm} (7)

where $B$ is the number of model parameters, and $N$ the number of data samples. Minimizing the BIC thus corresponds to minimizing the residual sum of squared errors (RSS), with a penalty coefficient proportional to the number of parameters $B$. In the case of an ELM with $\tilde{N}$ hidden neurons and using $M$ features, $B = \tilde{N} \times (M + 2)$.

Here, we use forward stepwise selection of features (e.g., Hastie et al., 2009). Each ELM model starts with only an intercept term, i.e., $\mathbf{w}_j$ is a scalar. Then, we in turn include each single potential feature individually (Table 1), and recalibrate the ELM. We consider the feature that has lead to the largest reduction in RSS as the potential extra-feature. We compare the BIC of the ELM without the extra-feature and with the extra-feature. If the BIC has decreased, we keep the extra-feature in the model, and continue to iterate through the remaining features. Once the BIC stops decreasing, we stop the forward stepwise process, and keep only those features included up to that step. Features included in the ELMs are referred to as fit-features, in contrast to the split-features used to separate the different branches in the tree growing process (Sect. 2.3.1). The sets of ELM fit-features from the different tree nodes, as well as the set of split-features, can overlap but need not be the same. As a last step of the ELM calibration, in each tree node, we also fit a simple linear regression model with the fit-features. If the linear regression model achieves a lower BIC compared to the ELM due to the reduction in parameter numbers from dropping the hidden layer, we swap the ELM model for the linear model. The entire ELM calibration procedure is summarized in Algorithm 1. The size of the oversized tree, the large number of data samples and potential features (Table 1), and the embedded loops in Algorithm 1 make clear the benefits of the analytical solution for $\beta$ (Eq. (6)).

### 2.3.3 Tree pruning

In the third step of tree induction, after having grown the oversized tree and fitted the ELM models at each node, we prune the ELM-tree to decrease its complexity. Several pruning strategies exist, and we adopt weakest-link pruning, which is a form of cost-complexity pruning (Hastie et al., 2009). The idea behind cost-complexity pruning is to find a balance between tree size and residual errors, where tree size $|\Gamma|$ is evaluated as the number of leaf nodes. We formulate a complexity-penalized cost function (Hastie...
Algorithm 1 ELM calibration procedure at any given tree node. This Algorithm uses Equations (2, 3, 4, 5, 6, 7).

1: at the given tree node, find all data samples $i$ of the node
2: consider all potential features $1: M_{\text{max}}$
3: Features ← {}  
4: $X ← 1_N$
5: $M ← 1$
6: $\text{BIC}^{(1)}, \text{BIC}^{(2)} ← +\infty$
7: while $\text{BIC}^{(2)} ≤ \text{BIC}^{(1)}$ and $M < M_{\text{max}}$ do
8:    $M ← M + 1$
9:    $\text{BIC}^{(2)} ← +\infty$
10:     for $m = 1 : M_{\text{max}}$ do
11:         if $m$ not in Features then
12:             $\text{RSS}_{\text{min}} ← +\infty$
13:             $X^{(m)} ← [X, \text{Feature}(m)]$
14:             for $k = 1 : 10$ do
15:                 $w^{(k)} ← U(-1, 1)$, size: $(M × \tilde{N})$
16:                 $\beta^{(k)} ← H^T L$
17:                 $\text{RSS}^{(k)} ← \sum_i \left( L_i - \hat{L}_i \right)^2$
18:                 if $\text{RSS}^{(k)} < \text{RSS}_{\text{min}}$ then
19:                     $w, \beta, \text{RSS}_{\text{min}} ← w^{(k)}, \beta^{(k)}, \text{RSS}^{(k)}$
20:             $\text{BIC}^{(m)} ← \text{BIC}(M, \text{RSS}_{\text{min}})$
21:             if $\text{BIC}^{(m)} < \text{BIC}^{(2)}$ then
22:                 $m^*, \text{BIC}^{(2)} ← m, \text{BIC}^{(m)}$
23:         if $\text{BIC}^{(2)} < \text{BIC}^{(1)}$ then
24:             Features ← Features+m*
25:             $X ← [X, \text{Feature}(m^*)]$
26:             $\text{BIC}^{(1)} ← \text{BIC}^{(2)}$
27:     if $\text{BIC}^{(2)} < \text{BIC}^{(1)}$ then
28:         calibrate Linear Model (LM) to $X, L → \beta^{(\text{LM})}$
29:         $\text{RSS}^{(\text{LM})} ← \sum_i \left( L_i - \hat{L}_i^{(\text{LM})} \right)^2$
30:         $\text{BIC}^{(\text{LM})} ← \text{BIC}(M^*, \text{RSS}^{(\text{LM})})$
31:     if $\text{BIC}^{(\text{LM})} < \text{BIC}^{(1)}$ then
32:         swap ELM for LM
manuscript submitted to Journal of Advances in Modeling Earth Systems (JAMES)

\[ C(\Gamma, \alpha_c) = \frac{1}{N} \sum_{i=1}^{N} (L_i - \hat{L}_i)^2 + \alpha_c|\Gamma|, \]  

(8)

where \( \alpha_c \) is a regularization parameter. Weakest-link pruning consists of successively finding and pruning the internal node of which the removal causes the smallest per leaf increase in the RSS. This process continues until the tree is reduced to the single root tree. For a given \( \alpha_c \) penalty, there is a single tree that minimizes \( C(\Gamma, \alpha_c) \). It can be shown that this optimal tree is always in the sequence of trees generated by weakest-link pruning (Ripley & Hjort, 1996). Weakest-link pruning is summarized in Algorithm 2.

**Algorithm 2** Weakest-link pruning.
This Algorithm uses Equation (8).

1: choose range(\( \alpha_c \))
2: \( \Gamma_{init} \leftarrow \) initial oversized tree
3: for all nodes \( i \) in \( \Gamma_{init} \) do
4: compute \( \text{RSS}_i \)
5: \( \Gamma \leftarrow \Gamma_{init} \)
6: \( \text{Sequence}(\Gamma) \leftarrow \{ \Gamma \} \)
7: while \( \Gamma \neq \Gamma_{root} \) do
8: for all non-leaf nodes \( i \) in \( \Gamma \) do
9: \( j(i) \leftarrow \) set of leaf nodes of \( i \)
10: \( n(j(i)) \leftarrow \) size \((j(i))\)
11: \( \Delta(\text{RSS})_i = \text{RSS}_i - \sum_{j(i)} \text{RSS}_{j(i)} \)
12: find non-leaf node \( i' \) minimizing \( \frac{\Delta(\text{RSS})_i}{n(j(i))} \)
13: \( \Gamma \leftarrow \Gamma \) pruned below \( i' \)
14: \( \text{Sequence}(\Gamma) \leftarrow \text{Sequence}(\Gamma) + \Gamma \)
15: for \( \alpha_c \) in range(\( \alpha_c \)) do
16: find \( \Gamma'(\alpha_c) \) in \( \text{Sequence}(\Gamma) \) minimizing \( C(\Gamma, \alpha_c) \)

2.4 Implementation in a Global Climate Model

We implement the ELM-tree lightning scheme in the Community Earth System Model version 2.2 (CESM2.2, Danabasoglu et al., 2020). The lightning scheme affects atmospheric chemistry through the generation of NO\(_x\) compounds. The atmospheric light-
ning output is also coupled with the land model to provide prognostic lightning flash densities at each model time step. In the land model, lightning serves as a source of natural wildfire ignitions. This atmosphere-to-land coupling of lightning flash rate is not available in the latest CESM2.2 public release, and has been developed as part of this study.

We perform a simple and idealized model simulation to demonstrate the capabilities offered by the ELM-tree implementation with an interactive atmosphere-to-land coupling. Our CESM2.2 set-up uses the Community Atmosphere Model version 6 for the atmosphere, and the the Community Land Model Version 5 with the fire model of Li et al. (2012) for the land (Lawrence et al., 2019; Danabasoglu et al., 2020). The ocean and sea-ice components are set as inactive to save computational expense. We perform a single 1995-2015 run, which is initialized from the CESM2.2 historical run performed for CMIP6. We use a coarse 1.9° × 2.5° resolution, and with interactive biogenic emissions from fires. We use default CESM2.2 parameterizations for the vertical distribution of lightning-produced NO\textsubscript{x} (Pickering et al., 1998), the energy per flash (Price et al., 1997), and the energy difference between intra-cloud and cloud-to-ground flashes (Ridley et al., 2005). The objective of this CESM2.2 simulation is not to perform a detailed comparison with observations, but rather to demonstrate that the ELM-tree can be implemented in state-of-the-art GCMs, and that it produces realistic and stable estimates of lightning, even when driven exclusively with model fields.

2.5 Comparison with existing parameterizations

For evaluation, we compare the performance of the ELM-tree to that of two other state-of-the-art lighting flash density parameterizations. In particular, we use the lightning schemes of Price & Rind (1992) (PR92) and of Finney et al. (2014) (F14).

PR92 is the lightning parameterization used to simulate lightning-produced NO\textsubscript{x} in all the GCMs taking part to the CMIP6 experiment and simulating tropospheric chemistry (Thornhill et al., 2021). As such, it is also the parameterization of the latest CESM2.2 release (Danabasoglu et al., 2020), although wildfire ignitions use a pre-processed input lightning climatology. The PR92 lightning density is predicted using two separate formulations for the land and ocean, both following a power-dependence on cloud-top-height. In PR92, \( L \) scales with cloud-top-height to the power 4.9 and 1.73 over land and ocean, respectively (Price & Rind, 1992). In addition, we apply the correction of Price & Rind (1994) to account for different grid resolutions, as all lightning predictions are computed on the same 2.5° × 2.5° grid as the lightning LRTS product.

F14 is based on cloud ice flux, which relates it in a physical manner to the non-inductive charging mechanism in thunderstorms (Reynolds et al., 1957). It depends only on mid-tropospheric values, and computations are performed at the specific 440 hPa vertical level. At this level, cloud ice flux is the product of specific cloud ice content and updraught mass flux, rescaled by the fractional cloud area. We download all these quantities from ERA5, and interpolate them linearly at the 440 hPa level. We compute daily mean values by averaging all variables at the 0:00, 6:00, 12:00, and 18:00 UTC time steps, similar to our treatment of all the climate data used in this study. Similarly to PR92, maritime and continental lightning have separate formulations in F14. Following Finney et al. (2014), we set lightning to 0 if the 440 hPa fractional cloud cover is < 0.01.

3 Results

3.1 ELM-tree induction and cross-validation

We perform the cross-validation experiment to guide our choice of the ELM-tree size. Figure 1a shows the change in cross-validated RMSE with the number of leaf nodes (|\( \Gamma \)|). As expected, the RMSE increases as |\( \Gamma \)| decreases, because smaller trees have both
less partitions of the feature space and less leaf ELM models. However, the ±1 standard deviation (±1σ) intervals overlap from ELM-trees having just 3 up to 46 leaf nodes (Fig. 1a). A very parsimonious approach could therefore use only the zero-branch split, and two ELMs to compute non-zero lightning. Here instead, we identify a clear slope break in the rate of RMSE decrease with |Γ|, occurring at |Γ| ≈ 7 (Fig. 1a, right y-axis). The red curve shows that a given RMSE decrease becomes abruptly more expensive in |Γ| once |Γ| > 7. As such, we select the complexity penalty αc corresponding to |Γ| = 7, which provides an appropriate balance between ELM-tree predictive performance and size (Fig. 1a, green dotted). 

Figure 1. Results from the cross-validation experiment. (a) The change of the cross-validation root mean square error (RMSE) with the number of leaf nodes (|Γ|) used in the ELM-tree. The red line shows the derivative of the RMSE with respect to |Γ|, evaluated using centered differencing. The dotted green line shows the selected ELM-tree size, corresponding to |Γ| = 7. (b) The selection probability of each feature when |Γ| = 7, averaged across the cross-validation samples. The fit-feature selection probability is the proportion of leaf models using that feature. The split-feature selection probability is the proportion of leaf nodes for which the set of decision splitting rules includes that feature at least once. Selection probabilities are averaged across the 19 cross-validation samples. Figure 1 shows that split decision rules are based on a relatively small num-

We show the selection probability of the 25 different features, both as fit-features (Fig. 1c) and split-features (Fig. 1d). The fit-feature selection probability is computed as the proportion of leaf models using a given feature. Similarly, the split-feature selection probability is the proportion of leaf nodes having been split in at least one upper branch on a given feature. Selection probabilities are averaged across the 19 cross-validation samples. Figure 1 shows that split decision rules are based on a relatively small num-
number of features. In contrast, the fit-feature selection probabilities are more spread across
the different potential features. Although, as the number of leaf nodes decreases, selection probabilities tend to concentrate on a smaller set of fit-features, which are important to the ELM models close to the ELM-tree root.

In Figure 1b, we show selection probabilities at the selected number of leaf nodes $| \Gamma | = 7$. We only show those features reaching at least 10% fit- or split-feature probability. Five split-features are used with probability $\geq 10\%$, and three dominate with probability values $> 70\%$: $K_I$, $h_m$, and $CAPE$. Notably, $K_I$ is always used at the root node as split-feature for the zero-branch. In addition to its role as a split-feature, $CAPE$ is the fit-feature with the highest selection probability (51%). Figure 1b demonstrates that more fit-features have selection probabilities $\geq 10\%$ compared to the split-features. These fit-features are associated with geography ($| \varphi |$), surface heat fluxes ($Q_I, Q_e$), cloud properties (e.g., $\text{cbh}, \text{c}_\text{liq}$), or other climatic features (e.g., $CAPE, tt_I, w_{10m}$).

![Figure 2](image.png)

**Figure 2.** Receiver Operating Characteristic (ROC) curve for the selection of the zero-branch split criterion. The selected optimal classifier corresponds to the value $K_I=12.24^\circ \text{C}$. Specificity and recall are defined in Equation (9). $K_I$ is chosen as split feature because the distance from its optimal classifier to the perfect classifier is the smallest among all potential features considered (Table 1).

Concerning the zero-branch split, Figure 2 shows the ROC curve as the $K_I$ splitting value changes. The ROC curve shown is computed with the entire training data set. Data samples with $K_I$ less than the splitting value are classified as zero-lightning events. The ROC curve shows the trade-off between recall and specificity:

$$
\begin{align*}
\text{recall} &= \frac{TP_0}{P_0} \\
\text{specificity} &= \frac{TN_0}{N_0},
\end{align*}
$$

where $P_0$ and $N_0$ are the total number of zero-lightning and non-zero-lightning data samples, as defined by the 0.1 fl. km$^{-2}$ yr$^{-1}$ threshold. $TP_0$ is the number of zero-lightning data samples correctly classified by the split criterion, and similarly $TN_0$ is the number of non-zero-lightning data samples correctly classified. As illustrated in Figure 2, the perfect classifier would satisfy $\text{recall}=\text{specificity}=1$. For all the potential features, their respective optimal split value is taken along their ROC curve where the distance to the ideal classifier is minimized. We select $K_I$ as optimal split-feature because its optimal split-value achieves the smallest distance to the perfect classifier. This is valid not only on the full training data set, but also for each cross-validation fold (Fig. 1b). The $K_I$
zero-branch split value is 12.24°C, which yields values for recall and specificity of 0.72 and 0.63, respectively.

Imposing the cross-validated tree size $|\Gamma| = 7$, we fit the final ELM-tree to the full training data set. This results in an ELM-tree with a maximum depth of 4. A schematic of the model is shown in Figure 3. Notably, all the non-zero branch leaf models use an ELM, as the simpler linear regression model is never favored by the BIC (see Algorithm 1). The leaf ELMs use on average 3.5 features, but with some overlap between the leaf ELMs. As such, our ELM-tree uses only 12 features in total: CAPE, cbb, $T_{d,2m}$, $c_{ice}$, $m_w$, $K_I$, $tt_l$, $h_{GC}$, $p_s$, $Q_l$, $h_s$, and $|\varphi|$ (see Table 1). The detailed description of the ELM-tree and an implementation pseudo-code are provided in the Supporting Information.

**Figure 3.** Schematic of the architecture of the calibrated ELM-tree. Split-features are given in their tree node, along with their respective splitting value. Inequality symbols show the direction of the splitting rules. The left-split at the tree root corresponds to the zero-branch leaf node, which always returns 0 fl. km$^{-2}$ yr$^{-1}$. In the non-zero-branch, each leaf node consists of an Extreme Learning Machine (ELM) model, as illustrated with the sketches of single hidden layer feedforward neural networks. Note that the second split decision $h_s \leq 0.5$ m effectively separates the maritime and continental domains, which can be used as an equivalent split decision.

### 3.2 Evaluation and comparison with other lightning schemes

We use the data from the test years (2000, 2005, 2010, 2020, 2023) to evaluate the out-of-sample ELM-tree performance. All performance metrics in this Section are computed with respect to the test years only, and no re-scaling to match the observed global total lightning is applied. In order to provide a meaningful evaluation, we perform spatio-temporal smoothing of the outputs from the ELM-tree, PR92, and F14 in the same manner as was done for the LIS lightning product (Cecil et al., 2014).

Figure 4 shows maps of the mean lightning rate from observations, from the ELM-tree, and from the PR92 and F14 parameterizations. The ELM-tree reproduces the observed spatial lightning patterns well. For example, it captures the lightning peak in central-Africa relative to the other tropical regions, the longitudinal gradient across Eurasia, the
Figure 4. Mean lightning flash density over the test years 2000, 2005, 2010, 2020, and 2023. From (a), the LIS observations, (b) the ELM-tree, (c) PR92, and (d) F14. Note the logarithmic color scale.

East-West contrast in North-America, and lightning over Australia. PR92 and F14 reproduce part of the spatial distribution, but display large deviations from observed lightning density values in several regions. For example, PR92 underestimates lightning rates in Europe, central Africa, as well as Central- and North-America, but overestimates lightning over the Amazon. F14 overestimates lightning in central Asia and western North-America, and exhibits a strong under-estimation in Africa.

The improved fidelity of the ELM-tree to observations is clearer and better quantified in Figure 5, showing key performance metrics for the three lightning schemes compared to the LIS observations: mean bias, RMSE, and Pearson correlation. Note here that both the RMSE and the Pearson correlation are computed after aggregating the daily time series into monthly lightning density values. In the equatorial regions, PR92 has a positive bias over the Amazon and the Maritime Continent, but a strongly negative bias over central Africa (Fig. 5b). In the northern and southern extra-tropics, it shows an almost ubiquitous negative bias. F14 displays an even stronger negative bias in central Africa, as well as under-estimations in West-Africa, India, and East-Asia (Fig. 5c).

The over-estimations in central Asia and western North-America are also evident. In all these regions, the ELM-tree bias is much smaller, even though it still under-estimates central African and North-American lightning rates (Fig. 5a). The monthly RMSE maps further demonstrate the improved performance of the ELM-tree (Fig. 5 d,e,f). In particular, errors of the ELM-tree are consistently smaller in East-Asia, Europe, Australia, over the Amazon, and most prominently, in central Africa. Finally, analyzing the monthly correlation (Fig. 5 g,h,i), all three lightning schemes capture monthly variability better over land than over the ocean, with correlation ($r$) generally $> 0.6$ and statistically significant at the 0.01 level. On average, PR92 shows the weakest temporal correlation with observed values, being very low in Europe, India, and South-America for example. This highlights that PR92 is prone to compensating errors, artificially improving its bias metric in Figure 5b. The ELM-tree improvements in monthly correlation compared to both PR92 and F14 are ubiquitous over all land areas, except the Middle-East where F14 performs better. We identify particularly higher correlation values of the ELM-tree in Central- and North-America, India, Australia, Europe, and over the Amazon (Fig. 5g). Over the latter, the ELM-tree yields statistically significant monthly correlation, in contrast to PR92 and F14. Such a global-scale improvement in the correlation metric demonstrates —16—
that the ELM-tree captures seasonal lightning patterns across a range of different climates with greater fidelity. In summary, Figure 5 shows that the ELM-tree performs better than PR92 and F14 in terms of bias, RMSE, and/or correlation in many regions. However, we note a slightly larger bias over the tropical oceans than the two other schemes (Fig. 5a).

Figures 5 a,b,c also show that all three schemes exhibit a negative bias in the same four regions: central Africa, North-West India, the south-eastern USA, and South-East South-America. The latter region is not discussed here because it coincides with the South-Atlantic anomaly and, therefore, high observational uncertainties (Christian et al., 2003). Central Africa, North-West India, and the south-eastern USA are major lightning hotspots (Fig. 4a), and are underestimated by other lightning schemes as well (e.g., Lopez, 2016; Stolz et al., 2017; Etten-Bohm et al., 2021). Still, we underline the large performance improvement of the ELM-tree in central Africa, where both bias and RMSE are more than halved compared to PR92 and F14 (Fig. 5 a,d).

Following this spatial evaluation, we also evaluate the skill of the lightning schemes in reproducing all single daily lightning density values from LIS. Figure 6 shows a density plot for each scheme, where each observed value is compared with the corresponding model estimate. Figure 6 includes every single observation during test years from both land and ocean. The ELM-tree demonstrates a coefficient of determination ($R^2$) of 0.706. This means that our model is able to reproduce > 70% of the daily global-scale spatio-temporal variability. This is a relative improvement of > 42% compared to PR92 and F14, which capture only 49.5% and 46.1% of the observed variability, respectively. In absolute terms, this corresponds to an RMSE decrease of > 1.3 fl. km$^{-2}$ yr$^{-1}$ and a re-
Figure 6. Density plots showing all daily LIS observations during the test years (2000, 2005, 2010, 2020, 2023), and the corresponding model estimates from (a) the ELM-tree, (b) PR92, and (c) F14. Performance statistics for each lightning scheme are provided, where $N$ is the number of data samples, $R^2$ is the coefficient of determination, and RMSE is the root mean square error. The dashed line denotes the 1:1 perfect fit line. Note the logarithmic color scale.

Figure 6. Density plots showing all daily LIS observations during the test years (2000, 2005, 2010, 2020, 2023), and the corresponding model estimates from (a) the ELM-tree, (b) PR92, and (c) F14. Performance statistics for each lightning scheme are provided, where $N$ is the number of data samples, $R^2$ is the coefficient of determination, and RMSE is the root mean square error. The dashed line denotes the 1:1 perfect fit line. Note the logarithmic color scale.

Figure 7. Average lightning flash density (a) latitudinal distribution and (b, c) longitudinal distribution from the LIS observations and from the ELM-tree, PR92, and F14 lightning schemes. In (b) and (c), only areas in the extra-tropics and tropics are considered, respectively. The limit for the tropics is ±23.44 latitude. RMSE is the root mean square error and $r$ the Pearson correlation coefficient. Note that the y-axis in (c) spans twice the range of the y-axes in (a, b). Averaging is performed over the test years (2000, 2005, 2010, 2020, 2023).
Figure 7 focuses again on the spatial evaluation, showing the averaged latitudinal and longitudinal distributions of lightning density values. The latitudinal distribution (Fig. 7a) is best reproduced by the ELM-tree, with a higher correlation with respect to observations ($r = 0.94$) than PR92 ($r = 0.87$) and F14 ($r = 0.80$). The RMSE of F14 (1.33 fl. km$^{-2}$) is lower than that of PR92 (1.45 fl. km$^{-2}$), but the RMSE of the ELM-tree is further reduced by 50% (0.67 fl. km$^{-2}$). We separate the longitudinal distribution in two separate components: the extra-tropics (Fig. 7b) and the tropics (Fig. 7c). In the extra-tropics, the ELM-tree yields a correlation coefficient of 0.87, compared to 0.81 and 0.72 for PR92 and F14, respectively. In terms of RMSE, ELM-tree and F14 perform comparably (1.53 and 1.71 fl. yr$^{-1}$ km$^{-2}$, respectively), while PR92 confirms its strong negative extra-tropical bias (RMSE of 2.33 fl. yr$^{-1}$ km$^{-2}$, see also Fig. 5b). The better performance of the ELM-tree in the tropics is illustrated in Figure 7c. Its correlation coefficient is 0.95, substantially higher than PR92 (0.87) and F14 (0.81). Its RMSE of 1.90 fl. yr$^{-1}$ km$^{-2}$ is 36% lower than that of PR92 (2.97 fl. yr$^{-1}$ km$^{-2}$) and 52% lower than that of F14 (3.97 fl. yr$^{-1}$ km$^{-2}$). Figure 7c highlights once more the clear improvement in central African lightning representation of the ELM-tree, even though this longitudinal lightning peak is still underestimated. The ELM-tree RMSE in the tropics is close to that in the extra-tropics, despite lightning flash rate values being typically twice as large (Fig. 7b,c). Paired with its high correlation ($r = 0.95$), this underlines the ELM-tree ability to simulate differences between the three tropical lightning chimneys: central Africa, the Amazon, and the Maritime Continent.

Figure 8. Temporal variability performance of the ELM-tree, PR92, and F14 lightning schemes with respect to LIS observations. Monthly climatology in the (a) northern and (b) southern hemispheres. Inter-annual variability in the (c) northern and (d) southern hemispheres. In (a, b), averaging is performed over the test years (2000, 2005, 2010, 2020, 2023). RMSE is the root mean square error and $r$ the Pearson correlation coefficient. Note that y-axes are different in (a, b) and (c, d).

Finally, Figure 8 focuses on performance in terms of temporal variability aggregated at the hemispheric scale. In particular, we show the monthly climatology and inter-annual variability, both computed only over the test years, in the northern and southern hemispheres separately. Figure 8 a,b shows that all three schemes reproduce the climatology well, as they demonstrate high correlation with observations ($r \geq 0.92$). However, Figure 8a demonstrates that both PR92 and F14 strongly underestimate the summer light-
ning peak in the northern hemisphere. The ELM-tree largely reduces this underestimation, resulting in a > 60% decrease in the northern hemisphere monthly climatology RMSE.

In the southern hemisphere, correlation coefficients of the three schemes for the monthly climatology are close, but the ELM-tree reaches an RMSE > 40% lower compared to PR92 and F14. Concerning the inter-annual variability (Fig. 8 c,d), we first acknowledge that 5 test years is a small sample to evaluate model skill. Over these 5 years, PR92 is not able to reproduce year-to-year variations, both in the northern ($r = 0.28$) and southern ($r = 0.19$) hemispheres. F14 performs better concerning this metric, yielding $r = 0.86$ and $r = 0.80$, respectively. Nevertheless, we find that the ELM-tree outperforms PR92 and F14 in both hemispheres, with $r < 0.90$. The RMSE decrease using the ELM-tree is even more pronounced, being > 70% and > 56% lower in the northern and southern hemispheres, respectively.

3.3 Simulations with the Community Earth System Model

We have implemented the ELM-tree as a lightning scheme in CESM2.2. The goal of the CESM2.2 simulation is not to compare in detail the model run to lightning, wildfire, and NO\textsubscript{x} observations. Such variables also depend on the simulated climate as well as parameterizations such as the ignition efficiency, and the amount of nitrogen generation per lightning flash. Furthermore, we use a coarse model set-up, with only the atmosphere and land components active, while the ocean and sea-ice components are inactive. The purpose here is to demonstrate that the ELM-tree can be implemented within a GCM source code, that modeled lightning values are stable, and that the ELM-tree can be fully-coupled to the wildfire ignition scheme.

Our 1995-2015 simulation yields an annual mean global lightning rate of $46.24 \pm 0.79$ fl. s\textsuperscript{-1}, where ± denotes the standard deviation of inter-annual variability. This agrees well with the observed values from LRTS over 1996-2014 of $44.01 \pm 1.86$ fl. s\textsuperscript{-1}, although the model underestimates inter-annual variability. Note that the years without entire annual coverage in the LRTS record are discarded in the latter calculation. According to OTD and LIS uncertainty estimates of ~ 5 fl. s\textsuperscript{-1} (Christian et al., 2003; Blakeslee et al., 2020), the CESM2.2 global lightning output is within the observational uncertainty range.

The CESM2.2 mean lightning density spatial distribution, without spatio-temporal smoothing, is shown in Figure 9a. The spatial patterns are realistic compared to observations (see Fig. 4a), reproducing the contrasts between tropics and extra-tropics and between land and ocean, as well as high lightning regions such as eastern North-America and eastern Asia. On the other hand, some model biases appear, such as lightning overestimation in northwestern North-America and Australia, and underestimation in central Africa. Still, in general, we find very good agreement in both the main spatial patterns as well as the global total lightning.

Figure 9b shows the map of standard deviation in monthly variability, as modeled by the ELM-tree implemented in CESM2.2. Unsurprisingly, variability is highest in regions of high lightning rates. Furthermore, high-latitude regions, both in the northern and southern hemispheres, show a stronger relative variability, driven by more pronounced seasonal patterns. We also investigate the NO\textsubscript{x} production in our CESM2.2 run. The total annual mean production is $3.27 \pm 0.05$ Tg N yr\textsuperscript{-1}, which is within the estimated range of 2-8 Tg N yr\textsuperscript{-1} (Schumann & Huntrieser, 2007). In the current CESM2.2 parameterization, each lightning flash is assumed to produce an equal and fixed amount of nitrogen (Price et al., 1997; Ridley et al., 2005). As such, there is a one-to-one relationship between modeled flash density and lightning-produced NO\textsubscript{x} (Fig. 9c). Finally, we also show the monthly correlation between lightning flash rates and burned area (Fig. 9d). While lightning serves as a natural ignition source, its relation to burned area is also strongly modulated by weather conditions. For example, we find that in the Sahel, light-
Figure 9. Output from the 1995-2015 CESM2.2 simulation with the ELM-tree implemented as the lightning model. (a) The mean lightning flash density, (b) the monthly standard deviation in lightning flash density, (c) the mean lightning-produced NO$_x$ rate, and (d) the monthly Pearson correlation between lightning flash density and burned area. In (a), the CESM2.2 lightning output has not been smoothed, in contrast to Figure 4. In (c), lightning-produced NO$_x$ scales linearly with lightning density from (a). In (d), note that the color scale saturates at 0.6, and that hatched areas denote absence of statistical significance at the 0.05 level using a two-tailed t-test.

ning is anti-correlated with burned area, as it occurs mostly during the wet season. In contrast, the correlation is positive in high-latitude regions, where most lightning events occur in summer.

The good agreement between our CESM2.2 test run using the ELM-tree implementation and observations of global total lightning rates, lightning spatial patterns, and global lightning-produced NO$_x$ generation is encouraging. This demonstrates that the ELM-tree provides a stable and realistic lightning scheme for GCMs, and further suggests that good fidelity with respect to observations can be achieved. However, we underline that the performance depends on GCM-specific aspects, such as the convection scheme, as well as the parameterization of cloud-to-ground versus intra-cloud flashes, and their respective efficiencies in producing NO$_x$ compounds.

4 Discussion

Lightning schemes developed for long spatial- and temporal-scales have traditionally used simple mathematical relationships and few climatic input features (e.g., Finney et al., 2014; Romps et al., 2014; Lopez, 2016). In contrast, in regional studies typically focusing on very short-term forecasts, machine learning methods have demonstrated the benefits of using more flexible models and a larger set of features (e.g., Mostajabi et al., 2019; Ukkonen & Mäkelä, 2019; Cavaiola et al., 2024). The latter approach poses the problem of potential implementation in GCMs. For example, Ukkonen & Mäkelä (2019) used a random forest model of 800 different trees and a neural network with multiple 30-neuron hidden layers; such architectures pose substantial implementation challenges in GCM code bases, or are expensive in terms of computing memory. In this study, we have used an intermediate approach by developing a data-driven and flexible model, while us-
ing several constraints to preserve a parsimonious model structure as well as improve generalization skills. We have demonstrated the improved fidelity of the ELM-tree with respect to observations compared to traditional lightning parameterizations. Implementation of the ELM-tree in GCM source code only requires a few if-conditions to reproduce the shallow tree-based architecture (Fig. 3), dot products between prescribed parameter values and common climatic variables (Eq. (1)), and a simple max() operation that we use as activation function (Eq. (4)). As such, our ELM-tree can be implemented in any GCM with minimal coding effort, as we have demonstrated for the case of CESM2.2.

In the Supporting Information, we provide detailed pseudo-code and all parameter values to guide implementation of the ELM-tree in other GCMs.

The quality of our data-driven approach depends directly on the input data sets used. The ELM-tree has been calibrated with reanalysis data from ERA5 (Hersbach et al., 2020). While ERA5 is generally recognized as a state-of-the-art estimate of the atmosphere state, it inevitably includes errors, for example in the representation of convective systems (Lavers et al., 2022). Another challenge is the lightning data. Here, we have post-processed the available LIS-ISS data (Blakeslee et al., 2020) to create a data set which, in conjunction with LRTS (Cecil et al., 2014), spans a 29-year period. The smoothing procedure applied to all the LIS and ODT lightning data is required to compensate for satellite sampling biases. However, smoothing also affects the relationships between climatic variables and lightning. As both the diversity and quality of lightning data sets increase (Blakeslee et al., 2020; Kaplan & Lau, 2021), combining different data products could minimize their individual shortcomings (e.g., Bitzer et al., 2016), and thus further enhance global lightning modeling prospects.

The list of potential climatic features used here is non-exhaustive (Table 1). We have focused on features that are commonly available from both GCMs and reanalysis data. Previous work included other features, such as convective mass flux or convective ice mass flux at a given pressure level (Allen & Pickering, 2002; Finney et al., 2014), and within-cloud vertical profiles of concentration of snow, graupel, and cloud condensates (Lopez, 2016). Another non-climatic predictor of considerable importance is aerosol concentration, because it is supposed to exert an influence complementary to thermodynamics on cloud dynamics and lightning (Williams & Sátori, 2004). Aerosol concentration has been included in the parameterization of Stolz et al. (2017) for example, but was derived from a global chemical-transport model. Their use of model output was motivated by large uncertainties in observation-based aerosol concentration products. Retrieval from satellite measurements of aerosol properties is particularly challenging near clouds (Marshak et al., 2021), which are the conditions most relevant to lightning. Furthermore, the performance of GCMs in simulating atmospheric aerosol concentrations remains questionable (Vogel et al., 2022). Our objective here is to develop an observation-based lightning model to be implemented in GCMs. As such, given these observation- and GCM-related shortcomings with respect to aerosols, we have preferred not to include aerosols in the lightning model. Overall, the input features used here are easy to obtain from standard reanalyses and GCM outputs, without additional processing or code modification requirements.

We have demonstrated that the ELM-tree lightning model implemented in a state-of-the-art GCM produces stable and realistic lightning rates. In particular, the global total lightning rate and lightning-produced NO\textsubscript{x} generation are within observational uncertainty ranges, and spatial lightning patterns reproduce the main characteristics from observations. Nevertheless, for improving the simulation of lightning-produced NO\textsubscript{x} and lightning-ignited wildfires in GCMs, the improved lightning flash model presented here is only a first, yet necessary step. For example, both of these processes are sensitive to the ratio of intra-cloud to cloud-to-ground lightning. This ratio is often assumed to depend on the sub-freezing cloud thickness (Price & Rind, 1993), but this relationship is still uncertain at the global scale (Gordillo-Vázquez et al., 2019). Generally, the sensi-
tivity of NO\textsubscript{x} production to multiple factors is poorly constrained, including to latitude, to intra-cloud versus cloud-to-ground flashes, and to continental versus maritime environments (Pickering et al., 2016). The vertical distribution of lightning-produced NO\textsubscript{x} is another factor of great importance in atmospheric chemistry modeling, yet it is also estimated with uncertain parameterizations (Pickering et al., 1998; Gordillo-Vázquez et al., 2019). Similarly, concerning wildfires, the ignition probability from a lightning discharge depends on complex factors such as the duration of the continuing current phase (Pérez-Invernón et al., 2023), rendering ignition parameterizations in GCMs poorly constrained (Li et al., 2012).

We have quantified the ELM-tree improved performance compared to the well-established benchmarks PR92 (Price & Rind, 1992) and F14 (Finney et al., 2014). In particular, the ELM-tree achieves a > 42% improvement in terms of spatio-temporal variability at the global scale (Fig. 6). On the one hand, first-order formulations such as PR92 and F14 have the advantage that lightning can, as a first approximation, be intuitively related to a small number of climatic proxies. On the other hand, given the complexity of the lightning process, more flexible approaches can capture relationships that are hard to constrain from physical principles (Ukkonen & Mäkelä, 2019; Cheng et al., 2024). The ELM-tree proposed in this study is more flexible than traditional GCM lightning parameterizations, still its interpretable architecture illustrated in Figure 3 offers insights into the model behavior. As an example, we analyze the model dependence on CAPE, because this feature has a predominant importance on lightning in our ELM-tree (Fig. 1b), which is also recognized in the literature (e.g., Romps et al., 2014; Lopez, 2016; Cheng et al., 2024). The ELM-tree uses CAPE as split-feature twice, effectively separating the tree in different CAPE regimes (Fig. 3). We find that CAPE is used as fit-feature in three of the six leaf node ELMs. However, it is not used in any of the two leaf node ELMs in the high CAPE regime, i.e., the two right-most leaf node ELMs in Figure 3. The ELM-tree thus identifies that, in high CAPE conditions, the lightning rate becomes insensitive to CAPE. This is supported by prior studies, showing that past a certain CAPE threshold, convective activation shows a weak or absent dependence on CAPE (e.g., Sherwood, 1999). This effect is also found by analyzing the marginal sensitivity of the ELM-tree to all CAPE values from our training data set (Fig. S2). In particular we find good agreement between the model and observed marginal sensitivities, but the model insensitivity to CAPE initiates at lower CAPE values. More generally, we find that the model reproduces the observed marginal sensitivities to any of the ELM-tree input features with good skill (Fig. S2).

5 Conclusions

We have developed a machine learning method to design a global lightning model. Our ELM-tree lightning model uses Extreme Learning Machine models embedded in a decision tree. This approach is more flexible than traditional lightning parameterizations for Global Climate Models (GCMs), but we have enforced several constraints to keep the model parsimonious, and therefore facilitate its implementation in any climate model. Our implementation of the ELM-tree in CESM2.2 demonstrates that it yields stable and realistic lightning predictions, agreeing within uncertainty ranges with observations without requirements of further scaling factors. In the Supporting Information, we provide a detailed pseudo-code as well as all parameter values to give a clear guide to the implementation of the ELM-tree in GCMs. As an intermediate approach between high-dimensional machine learning models and first-order parameterizations, our model offers a pragmatic and efficient solution for an improved lightning representation in GCMs.

Evaluation on test years not used in the ELM-tree calibration has demonstrated the improved fidelity of the ELM-tree with respect to observations when compared to well-established lightning schemes. In particular, the ELM-tree explains 70.2% of the daily global spatio-temporal lightning variability, which is a > 42% relative improvement com-
pared to these benchmarks. When scaled for the entire globe, such reduced errors could be of large impact in, for example, closure of the global ozone budget (Wild, 2007). A notable improvement is the ability to capture the differences between the three main lightning chimneys: central Africa, the Amazon, and the Maritime Continent. Furthermore, the ELM-tree reaches high correlation with observed lightning across all land masses, demonstrating its ability to reproduce seasonal patterns in a range of different climates. In addition to improved regional lightning representation, the ELM-tree performance is validated at the global scale: observed variations along latitude and along longitude are both represented with a correlation coefficient $> 0.86$. Similarly, at the scale of the northern and southern hemispheres, the ELM-tree represents the monthly lightning climatology and inter-annual variability with correlation coefficients $> 0.92$ and $> 0.90$, respectively.

Our proposed novel lightning model for GCMs opens the door to numerous research avenues. Building upon previous work, it can help the lightning modeling community to further study unsolved problems, such as lightning changes under different anthropogenic climate change scenarios, the climate-lightning-wildfire three-way interactions, and future changes in tropospheric concentrations of ozone and methane.

6 Open Research

The LRTS gridded lightning data product Cecil et al. (2014) and the LIS-ISS flash count data are available via the NASA Earthdata portal (https://www.earthdata.nasa.gov). The reanalysis data are available via the Copernicus Climate Data Store portal (https://cds.climate.copernicus.eu/). The CESM model can be downloaded from https://github.com/ESCOMP/CESM. Detailed guidelines for the implementation of the ELM-tree in climate models and all parameter values are provided in the Supporting Information. In addition, the corresponding author can be contacted for any question related to the ELM-tree model. Source code of the ELM-tree model, as well as scripts to reproduce evaluation figures of this study are available in a Zenodo repository (Verjans, 2024).

Acknowledgments

We thank colleagues at the IBS Center for Climate Physics for their comments and insights about this study, in particular Ja Yeon Moon and Jiaoyang Ruan. The simulations were conducted on the IBS/ICCP supercomputer “Aleph,” a 1.43 petaflops high-performance Cray XC50-LC Skylake computing system with 18 720 processor cores, 9.59 PB storage, and 43 PB tape archive space. We also acknowledge the support of KRE-ONET. VV and CF were supported by the Institute for Basic Science (IBS), Republic of Korea, under IBS-R028-D1, and CF by the National Research Fund of Korea (NRF-2022M3K3A1097082).

References


Huang, G., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for...


-27-


Supporting Information for "Development of a data-driven lightning model for implementation in Global Climate Models."

Verjans Vincent¹, Christian Franzke¹,²

¹Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea
²Pusan National University, Busan, Republic of Korea

Contents of this file

1. Text S1
2. Tables S1 to S4
3. Figures S1 to S2

Text S1.

In Algorithm 1, we provide the detailed pseudo-code to implement the ELM-tree in Global Climate Models (GCMs). All the input features required are defined in Table S1, as well as their respective units, sign conventions, and other calculation details. All the ELM-tree parameters are defined in Tables S2, S3, S4. Table S2 includes the split-feature thresholds, Table S3 includes the output weight vectors of the single hidden layer feedforward neural networks (SLFNs), and Table S4 includes the hidden weight vectors of the SLFNs. As a reminder, we provide here the equations defining the Extreme Learning

v.verjans@pusan.ac.kr

May 16, 2024, 6:42am
Machine (ELM) models. For any given ELM model, the lightning prediction \( \hat{L}_i \) [fl. km\(^{-2}\) yr\(^{-1}\)] for a data sample \( i \) is given by:

\[
\hat{L}_i = \sum_{j=1}^{\tilde{N}} \beta_j g \left( w^T_j x_i \right),
\]

where \( \tilde{N} \) is the number of hidden neurons in the SLFN (here, \( \tilde{N}=4 \)), \( x_i \) is the vector of input features with an intercept term, \( w_j \) is the vector of hidden weights for the \( j \)th hidden node, \( \beta_j \) is the \( j \)th element of the vector of output weights \( \beta \). The activation function \( g() \) is the rectified linear unit (ReLU):

\[
g \left( w^T_j x_i \right) = \max \left( 0, w^T_j x_i \right).
\]

Finally, we set any potential negative \( \hat{L}_i \) prediction from Eq. (1) to 0. These steps are also detailed in Algorithm 1.

Figure S1 shows the sensitivity of the ELM-tree to the number of hidden neurons used in the ELM models. To evaluate this sensitivity, the ELM-tree is trained with 1, 2, 3, 4, 5, and 6 hidden neurons in the SLFNs. The training is performed with a subset of the training data years: all training years except 1999, 2004, 2009, 2014, 2019, 2022. The performance is then measured as the root mean squared error (RMSE) on all observations from the gridded daily lightning data in these six excluded years.

Figure S2 shows the marginal sensitivity of the ELM-tree lightning estimates (\( \hat{L} \)) and observed lightning values (\( L \)) to each input feature of the ELM-tree. The marginal sensitivities are computed using the entire training data.
Algorithm 1 ELM-tree pseudo-code.

Computation for the estimated lightning density \( \hat{L} \) [fl. km\(^{-2}\) yr\(^{-1}\)]. Input features are defined in Table S1, and parameter values are given in Tables S2, S3, S4.

1: Define \( S_0(K_I), S_1(h_s), S_2(CAPE), S_3(CAPE), S_4(b_{d,2m}), S_5(h_s) \)
2: Define \( \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \)
3: Define \( w_{(1),1}, w_{(2),1}, w_{(3),1}, w_{(4),1}, w_{(1),2}, w_{(2),2}, w_{(3),2}, w_{(4),2}, w_{(1),3}, w_{(2),3}, w_{(3),3}, w_{(4),3} \)
4: Define \( w_{(1),4}, w_{(2),4}, w_{(3),4}, w_{(4),4}, w_{(1),5}, w_{(2),5}, w_{(3),5}, w_{(4),5}, w_{(1),6}, w_{(2),6}, w_{(3),6}, w_{(4),6} \)
5: if \( K_I \leq S_0(K_I) \) then
6: \( \hat{L} = 0 \)
7: else
8: if \( h_s \leq S_1(h_s) \) then
9: \( \hat{L} = \beta_{T}^1 \left[ \max(0, w_{(1),1}^T, x), \max(0, w_{(2),1}^T, x), \max(0, w_{(3),1}^T, x), \max(0, w_{(4),1}^T, x) \right] \)
10: else
11: if \( CAPE \leq S_2(CAPE) \) then
12: \( x = [1, CAPE, cbh] \)
13: \( \hat{L} = \beta_{T}^2 \left[ \max(0, w_{(1),2}^T, x), \max(0, w_{(2),2}^T, x), \max(0, w_{(3),2}^T, x), \max(0, w_{(4),2}^T, x) \right] \)
14: else
15: if \( CAPE \leq S_3(CAPE) \) then
16: \( x = [1, Q_i, [c], m_w] \)
17: \( \hat{L} = \beta_{T}^3 \left[ \max(0, w_{(1),3}^T, x), \max(0, w_{(2),3}^T, x), \max(0, w_{(3),3}^T, x), \max(0, w_{(4),3}^T, x) \right] \)
18: else
19: \( x = [1, [c], tt, p_s, Q_i, CAPE, hoc] \)
20: \( \hat{L} = \beta_{T}^4 \left[ \max(0, w_{(1),4}^T, x), \max(0, w_{(2),4}^T, x), \max(0, w_{(3),4}^T, x), \max(0, w_{(4),4}^T, x) \right] \)
21: else
22: if \( h_s \leq S_4(b_{d,2m}) \) then
23: \( x = [1, [c], tt, Q_i] \)
24: \( \hat{L} = \beta_{T}^5 \left[ \max(0, w_{(1),5}^T, x), \max(0, w_{(2),5}^T, x), \max(0, w_{(3),5}^T, x), \max(0, w_{(4),5}^T, x) \right] \)
25: else
26: \( x = [1, [c], m_w] \)
27: \( \hat{L} = \beta_{T}^6 \left[ \max(0, w_{(1),6}^T, x), \max(0, w_{(2),6}^T, x), \max(0, w_{(3),6}^T, x), \max(0, w_{(4),6}^T, x) \right] \)
28: \( \hat{L} = \max(0, \hat{L}) \)
29: \( \hat{L} = \max(0, \hat{L}) \)
30: \( \hat{L} = \max(0, \hat{L}) \)
31: end
Table S1. Input features for the ELM-tree

<table>
<thead>
<tr>
<th>Feature</th>
<th>Symbol</th>
<th>Units</th>
<th>Formulation details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convective Available Potential Energy</td>
<td>CAPE</td>
<td>J kg s(^{-1})</td>
<td>/</td>
</tr>
<tr>
<td>Cloud-base-height</td>
<td>cbh</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>2m dew point temperature</td>
<td>(T_{d,2m})</td>
<td>K</td>
<td>/</td>
</tr>
<tr>
<td>Cloud ice content</td>
<td>(c_{ice})</td>
<td>kg km(^{-2})</td>
<td>/</td>
</tr>
<tr>
<td>Column total water</td>
<td>(m_w)</td>
<td>kg km(^{-2})</td>
<td>(c_{ice} + c_{liq} + m_{vap})</td>
</tr>
<tr>
<td>K-index</td>
<td>(K_I)</td>
<td>°C</td>
<td>((T_{850\text{hPa}} - T_{500\text{hPa}}) + T_{d,850\text{hPa}} - (T_{700\text{hPa}} - T_{d,700\text{hPa}}))</td>
</tr>
<tr>
<td>Total totals index</td>
<td>(tt)</td>
<td>K</td>
<td>(T_{850\text{hPa}} + T_{d,850\text{hPa}} - 2T_{500\text{hPa}})</td>
</tr>
<tr>
<td>Zero degree level</td>
<td>(h_{0C})</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Surface pressure</td>
<td>(p_s)</td>
<td>Pa</td>
<td>/</td>
</tr>
<tr>
<td>Surface latent heat flux</td>
<td>(Q_l)</td>
<td>W m(^{-2})</td>
<td>positive downwards</td>
</tr>
<tr>
<td>Surface elevation above sea-level</td>
<td>(h_s)</td>
<td>m</td>
<td>/</td>
</tr>
<tr>
<td>Absolute latitude</td>
<td>(</td>
<td>\varphi</td>
<td>)</td>
</tr>
</tbody>
</table>

\(T\), \(c_{liq}\), \(m_{vap}\) are temperature [K], cloud liquid water content [kg km\(^{-2}\)], and column total water vapor [kg km\(^{-2}\)], respectively.

hPa and m subscripts denote a feature value at a given pressure level [hPa] and height level [m], respectively.

Table S2. Split parameters for the ELM-tree

<table>
<thead>
<tr>
<th>Split parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_0(K_I))</td>
<td>12.24°C</td>
</tr>
<tr>
<td>(S_1(h_s))</td>
<td>0.5 m</td>
</tr>
<tr>
<td>(S_2(\text{CAPE}))</td>
<td>146.9 J kg s(^{-1})</td>
</tr>
<tr>
<td>(S_3(\text{CAPE}))</td>
<td>924.6 J kg s(^{-1})</td>
</tr>
<tr>
<td>(S_4a(T_{d,2m}))</td>
<td>295.5 K</td>
</tr>
<tr>
<td>(S_4b(h_s))</td>
<td>304.5 m</td>
</tr>
</tbody>
</table>

Note that the split criterion \(S_1(h_s) \leq 0.5\) m effectively separates the maritime and continental domains.

The land mask can be used as an alternative, and equivalent, split criterion in climate models.

May 16, 2024, 6:42am
Table S3. Output weights for the ELM-tree

<table>
<thead>
<tr>
<th>Output weight vector</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>$[-8.48, -59.85, 8.57, 27.61]^T$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$[34.33, 115.49, -58.15, -202.14]^T$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$[-57.46, 96.79, -57.07, 2.24]^T$</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$[-88.76, 36.46, 29.22, 64.53]^T$</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>$[-94.61, 67.99, 20.95, -110.88]^T$</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>$[61.48, 30.38, -661.70, 157.11]^T$</td>
</tr>
</tbody>
</table>
Table S4. Hidden weights for the ELM-tree

<table>
<thead>
<tr>
<th>Hidden weight vector</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{(1),1}$</td>
<td>$[6.75 \times 10^{-1}, -8.48 \times 10^{-5}, -2.68 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(2),1}$</td>
<td>$[1.14 \times 10^{-1}, -1.73 \times 10^{-5}, -3.56 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(3),1}$</td>
<td>$[-1.93 \times 10^{-1}, 7.22 \times 10^{-5}, 5.58 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(4),1}$</td>
<td>$[4.41 \times 10^{-1}, 7.17 \times 10^{-5}, -3.88 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(1),2}$</td>
<td>$[6.34 \times 10^{-1}, 1.16 \times 10^{-4}, -1.06 \times 10^{-2}, 8.94 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(2),2}$</td>
<td>$[1.36, 3.02 \times 10^{-4}, 1.22 \times 10^{-2}, -7.93 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(3),2}$</td>
<td>$[-5.35 \times 10^{-1}, 3.71 \times 10^{-4}, 1.30 \times 10^{-2}, 3.72 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(4),2}$</td>
<td>$[9.08 \times 10^{-1}, 6.63 \times 10^{-4}, 3.74 \times 10^{-3}, -4.01 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(1),3}$</td>
<td>$[-1.41, -6.65 \times 10^{-5}, 6.38 \times 10^{-5}, 7.45 \times 10^{-3}, 1.69 \times 10^{-5}, -2.62 \times 10^{-7}]^T$</td>
</tr>
<tr>
<td>$w_{(2),3}$</td>
<td>$[1.97 \times 10^{-2}, 5.67 \times 10^{-5}, -6.27 \times 10^{-4}, 5.82 \times 10^{-3}, 4.79 \times 10^{-6}, -1.04 \times 10^{-7}]^T$</td>
</tr>
<tr>
<td>$w_{(3),3}$</td>
<td>$[1.41, -3.91 \times 10^{-5}, 5.44 \times 10^{-4}, -1.02 \times 10^{-2}, -4.06 \times 10^{-6}, -3.78 \times 10^{-7}]^T$</td>
</tr>
<tr>
<td>$w_{(4),3}$</td>
<td>$[1.85, -4.31 \times 10^{-5}, 6.33 \times 10^{-4}, -1.00 \times 10^{-4}, -7.52 \times 10^{-6}, -1.66 \times 10^{-8}]^T$</td>
</tr>
<tr>
<td>$w_{(1),4}$</td>
<td>$[-1.06, -1.11 \times 10^{-2}, -3.15 \times 10^{-3}, 1.21 \times 10^{-5}, -2.19 \times 10^{-4}, -1.37 \times 10^{-5}, 1.38 \times 10^{-4}]^T$</td>
</tr>
<tr>
<td>$w_{(2),4}$</td>
<td>$[1.87 \times 10^{-1}, -1.31 \times 10^{-2}, 4.67 \times 10^{-3}, 2.97 \times 10^{-6}, 4.93 \times 10^{-4}, -5.83 \times 10^{-5}, 4.42 \times 10^{-6}]^T$</td>
</tr>
<tr>
<td>$w_{(3),4}$</td>
<td>$[9.64 \times 10^{-1}, 1.08 \times 10^{-2}, 9.97 \times 10^{-3}, -1.45 \times 10^{-5}, -3.04 \times 10^{-4}, 5.43 \times 10^{-5}, 3.10 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(4),4}$</td>
<td>$[-1.62, -4.66 \times 10^{-3}, 7.75 \times 10^{-3}, 1.34 \times 10^{-5}, -3.43 \times 10^{-4}, 8.99 \times 10^{-5}, 1.17 \times 10^{-4}]^T$</td>
</tr>
<tr>
<td>$w_{(1),5}$</td>
<td>$[6.99 \times 10^{-1}, -1.26 \times 10^{-2}, -2.96 \times 10^{-3}, 7.50 \times 10^{-5}]^T$</td>
</tr>
<tr>
<td>$w_{(2),5}$</td>
<td>$[1.80 \times 10^{-1}, 1.93 \times 10^{-3}, 7.40 \times 10^{-3}, -6.25 \times 10^{-4}]^T$</td>
</tr>
<tr>
<td>$w_{(3),5}$</td>
<td>$[1.57, 5.77 \times 10^{-3}, 4.92 \times 10^{-3}, 4.86 \times 10^{-4}]^T$</td>
</tr>
<tr>
<td>$w_{(4),5}$</td>
<td>$[4.32 \times 10^{-1}, 8.80 \times 10^{-3}, -7.50 \times 10^{-3}, 5.11 \times 10^{-4}]^T$</td>
</tr>
<tr>
<td>$w_{(1),6}$</td>
<td>$[2.44 \times 10^{-1}, -4.32 \times 10^{-3}, 8.10 \times 10^{-10}]^T$</td>
</tr>
<tr>
<td>$w_{(2),6}$</td>
<td>$[6.91 \times 10^{-1}, 7.23 \times 10^{-3}, -7.40 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(3),6}$</td>
<td>$[1.84 \times 10^{-1}, -5.00 \times 10^{-3}, -4.08 \times 10^{-9}]^T$</td>
</tr>
<tr>
<td>$w_{(4),6}$</td>
<td>$[2.66 \times 10^{-1}, -1.36 \times 10^{-2}, -2.37 \times 10^{-9}]^T$</td>
</tr>
</tbody>
</table>
Figure S1. Sensitivity of the ELM-tree root mean square error (RMSE) to the number of hidden neurons ($\tilde{N}$) used in the Extreme Learning Machine leaf node models. For this sensitivity analysis, the years 1999, 2004, 2009, 2014, 2019, 2022 have been removed from the training data, and the RMSE is evaluated on data from these 6 years.
Figure S2. Marginal sensitivities of lightning flash density ($L$) to each input feature of the ELM-tree, from observed (black) and model-estimated (green) lightning. Marginal sensitivities have been computed with respect to the entire training data set. The x-axes span the 0.1% to 99.9% quantiles of each input feature.