JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000501

Interpretable Machine Learning-Based Radiation Emulation for ICON

Katharina Hafner[1,2], Fernando Iglesias-Suarez[2], Sara Shamekh[3], Pierre Gentine[4,5], Marco A. Giorgetta[6], Robert Pincus[7], and Veronika Eyring[1,2]

[1]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany, [2]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Weßling, Germany, [3]Courant Institute of Mathematical Sciences, New York University (NYU), New York, NY, USA, [4]Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA, [5]Earth Institute and Data Science Institute, Columbia University, New York, NY, USA, [6]Max-Planck-Institut für Meteorologie, Hamburg, Germany, [7]Lamont-Doherty Earth Observatory, Palisades, NY, USA

Abstract The radiation parameterization is one of the computationally most expensive components of Earth system models (ESMs). To reduce computational cost, radiation is often calculated on coarser spatial or temporal scales, or both, than other physical processes in ESMs, leading to uncertainties in cloud-radiation interactions and thereby in radiative temperature tendencies. One way to address this issue is to emulate the radiation parameterization using machine learning (ML), which is typically faster and has good accuracy in high-dimensional parameter spaces. This study investigates the development and interpretation of an ML-based radiation emulator using the ICOsahedral Non-hydrostatic model with the RTE+RRTMGP radiation code, which calculates radiative fluxes based on the atmospheric state and its optical properties. With a Bidirectional Long Short-Term Memory architecture, which can account for vertical bidirectional auto-correlation, we can accurately emulate shortwave and longwave heating rates with a mean absolute error of 0.045 K/d (2.77%) and 0.060 K/d (4.50%), respectively. Further, we analyze the trained neural networks using Shapley Additive exPlanations and confirm that the networks have learned physically meaningful relationships among the inputs and outputs. It is worth noting that we observe that the local temperature is used as a predictive source for the longwave heating, consistent with physical models of radiation. For shortwave heating, we find that clouds reflect radiation, leading to reduced heating below the cloud. In contrast, an architecture that is not inspired by the underlying physics, such as a multilayer perceptron, tends to rely on spurious or less physically meaningful correlations to make its predictions.

Plain Language Summary To estimate future impacts of climate change, we rely on climate projections generated by Earth System Models (ESMs). Radiation plays a crucial role in driving the climate system and is among the most computationally intensive components of ESMs. To save computing resources, radiation calculations are often performed less frequently or with lower detail, which introduces uncertainties in how clouds and radiation interact. Here, we develop an ML model to accelerate radiation calculations while maintaining accuracy. Specifically, we utilize this model to mimic how radiation is calculated in a well-known climate model, the ICON model. Our ML-based model reliably predicts heating rates for both sunlight (shortwave radiation) and heat from Earth and atmosphere (longwave radiation). We analyze the predictions of the ML-based emulator, which is motivated by the underlying physics, and demonstrate that it successfully captures physical relationships. Further, the interpretability analysis shows that a simpler ML model that is not inspired by the underlying physics uses non-causal relationships to make the predictions.

Key Points:

- The energy consistency of the machine learning-based emulator is improved by explicitly enforcing energy conservation during training
- Bidirectional Long Short-Term Memory learn physically meaningful relationships related to locality such as thermal emission and non-locality such as reflection by clouds
- The interpretability analysis shows that BiLSTMs are consistent with physical principles unlike MLPs

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. Hafner,
hafner@iup.physik.uni-bremen.de

Citation:

Hafner, K., Iglesias-Suarez, F., Shamekh, S., Gentine, P., Giorgetta, M. A., Pincus, R., & Eyring, V. (2025). Interpretable machine learning-based radiation emulation for ICON. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000501. https://doi.org/10.1029/2024JH000501

Received 12 NOV 2024
Accepted 13 SEP 2025

1. Introduction

Climate change is already negatively impacting the current conditions, making it essential to accurately model the complex Earth system for effective adaptation. Climate models, particularly Earth System Models (ESMs), are crucial for predicting global and regional changes, but key uncertainties limit their accuracy (Eyring et al., 2016). ESMs integrate many components representing the atmosphere, ocean, and land, which interact with each other. However, due to the computational cost of projecting a changing climate over multiple decades, ESMs operate at a large horizontal resolution of 40–160 km per grid cell (Chen et al., 2021). Despite these coarse grid scales and the long model time steps that require coarse scales, the representation of radiation in the shortwave and longwave spectrum would be overwhelmingly expensive without considerable simplifications (Hogan & Matricardi, 2020),

as outlined below. Still, such schemes would be too expensive if applied at every model time step, and hence it is common to compute the radiative transfer only at multiples of the model time step, so that the diurnal cycle of the shortwave radiation can be reasonably resolved and the interaction with the clouds can be represented to some degree. These compromises driven by computational constraints entail uncertainties with possible effects on the projected climates.

Radiation is the driver of many atmospheric processes, but it is very expensive to compute. To reduce the computational time, many approximations and simplifications are made, such as reduced spectral resolution, and computation of radiation at lower frequency and sometimes also coarser horizontal resolution than other physical processes (Mlawer et al., 1997; Morcrette et al., 2007). This requires scaling across time and space. Usually, the radiative flux is computed for a given state and cloud distribution. If the radiation time step is coarser than the model time step, then the shortwave flux for model time steps between radiation time steps can be scaled by the change in incoming radiation at the top of the atmosphere, and the longwave flux can be scaled by the change in surface temperature. However, effects of changes in the simulated atmospheric composition or cloud distribution on radiative fluxes and heating remain unaccounted for between radiation time steps. Cloud-radiation interactions, however, are important because clouds reflect shortwave radiation, leading to less heating. Additionally, clouds absorb and emit longwave radiation, leading to less or more heating depending on the surrounding temperature (Wallace & Hobbs, 2006). Therefore, an accurate treatment of cloud-radiation interactions is important. One option to increase accuracy is to compute radiation on the same horizontal and temporal scale as other physical processes, which would take up 10 times more computing time compared to the standard setup with infrequent radiation calls.

A promising possibility is the use of machine learning (ML), which has been successfully used in various applications including ML-based parameterizations for physical processes in ESMs. The ML-based emulation of radiation was historically the first application of ML to ESMs (Chevallier et al., 1998). One domain of those applications is the representation of all physical parameterizations at once or a superparameterization for a speed-up (Brenowitz & Bretherton, 2018; Gentine et al., 2018; Kochkov et al., 2024; Rasp et al., 2018; Watt-Meyer et al., 2024; Yuval et al., 2021; Yuval & O'Gorman, 2020). Another domain focuses on learning a single parameterization, such as radiation or convection, for a better representation of this process or speed-up by using short high-resolution and higher fidelity simulations at the same resolution (Bolton & Zanna, 2019; Espinosa et al., 2022; Grundner et al., 2022; Heuer et al., 2024; O'Gorman & Dwyer, 2018).

The development of an ML-based radiation scheme can be approached as an emulation or parameterization. The parameterization approach aims to improve the radiation scheme by learning from another more accurate radiation model than what is present in the climate model, for example, a wide-band or line-by-line model. The first ML-based radiation parameterization was presented in Chevallier et al. (1998), which was trained on a more accurate radiation scheme. They developed parameterizations for longwave radiation based on a wide-band model and a line-by-line model and trained a fully connected neural network (NN) for the clear-sky component and $2 \times M$ NNs for the cloudy-sky component, where M is the number of cloudy layers. Although this was useful at that time, ESMs have evolved, and this multi-network approach is not applicable anymore because the speed-up depends on the number of cloud-layers and does not provide speed-up with more than 60 layers in the atmosphere (Morcrette et al., 2007).

The emulation approach has the goal to speed up the radiation scheme by emulating the existing parameterization while preserving substantial accuracy. A fast radiation scheme has the advantage being called more often than traditional parameterizations. Thereby, interactions with clouds can be better represented, which may indirectly improve the overall accuracy of simulations. The emulation of radiation can be addressed in different ways by dividing the radiation parameterizations into two tasks. The first part deals with calculating cloud and gas optics, and the second part approximates the radiative transfer equations. Some efforts are focusing on gas optics only (Ukkonen et al., 2020; Veerman et al., 2021). The argument to only emulate gas optics is that the overall radiation parameterization would be more robust because the radiative transfer approximation is not changed, but the speed-up potential would be smaller compared to emulating the full radiation parameterization. The machine-learned gas optics module was successfully tested online (Ukkonen & Hogan, 2023). Pal et al. (2019) emulated only a part of radiation, including gas optics but not cloud and aerosol optics.

Most of the ML-based radiation parameterizations emulate full radiation, including cloud and gas optics as well as radiative transfer equations, because it has more potential to speed up the simulation, thus allowing either to call

radiation more often or to increase the horizontal resolution of the climate model, or both. First attempts to perform this emulation were based on fully connected NNs, which were tested online in CAM2 and GFS (V. M. Krasnopolsky et al., 2005; V. Krasnopolsky, 2012; V. M. Krasnopolsky et al., 2008). About a decade later, the same approach was used in a modern ESM, for example, for numerical weather prediction in WRF (Roh & Song, 2020; Song & Roh, 2021) and a 6-month simulation in GFSv16 (Belochitski & Krasnopolsky, 2021). Recently, there have been approaches using more advanced deep learning architectures, such as U-Net, Bidirectional Long Short-Term Memory (BiLSTM), transformer, and neural operator, to emulate full radiation (Lagerquist et al., 2021, 2023; Ukkonen, 2022; Yao et al., 2023). Some of these studies compared different architectures and found that bidirectional recurrent NNs performed better than fully connected NNs because recurrent NNs can better handle the autocorrelation in the vertical profile. Despite good overall offline performance, the remaining question is why the NNs perform well and how they use specific inputs, that is, which inputs are important. This is a very relevant question to verify reliability and physical consistency of the ML-based emulator.

In this study, we build on the findings from previous studies and develop an ML-based alternative to emulate the radiation scheme RTE+RRTMGP (Pincus et al., 2019) used in the atmosphere component of the ICOsahedral Non-hydrostatic (ICON-A) model (Giorgetta et al., 2018). With speed-up and accuracy in mind, we design NNs that are as small as possible but also sufficiently complex and expressive for good performance. The speed-up allows for more frequent radiation calls, implicitly improving the cloud-radiation interactions. It is not naturally given that an ML-based emulator learns the underlying physics of the radiation processes. Therefore, we also focus on the interpretation of the NNs and explain what they learned physically.

The paper is structured as follows. We first introduce the data used and how we pre-process and select training data. In Section 3, we explain the NN architectures and their training process. Then, we analyze the predicted heating rates and fluxes in Section 4. It is followed by an interpretation using Shapley values in Section 5.

2. Data

We develop an ML-based radiation emulator for the atmosphere component of the ICOsahedral Non-hydrostatic (ICON-A) model (Giorgetta et al., 2018) and use explainability methods to interpret the prediction post-hoc. We use a historical Atmospheric Model Intercomparison Project (AMIP)-like setup (Eyring et al., 2016) with a coupled land model. The land model reacts to temperature changes but does not have an interactive carbon cycle. The AMIP setup includes prescribed sea surface temperature and sea ice concentration. Concentrations of the well-mixed greenhouse gases are prescribed as annual global mean mole fractions. Ozone is prescribed using monthly mean historical values. The prognostic atmospheric variables are initialized from the Integrated Forecasting System analysis files. ICON is a flexible, state-of-the-art model using a modern and accurate radiation scheme. Our ICON setup uses a triangular grid with a resolution of R2B5, where R2 means that every edge of the icosahedron is divided into 2 parts, creating smaller triangles, and B5 describes 5 subsequent edge bisections. An R2B5 grid corresponds to a horizontal resolution of 80 km. The vertical dimension has 47 levels using sigma coordinates. These levels span 80 km in the atmosphere. More details on the horizontal and vertical grids are given in Section 2 of Giorgetta et al. (2018). Subgrid-scale processes are parameterized, which include cloud cover, radiation, vertical diffusion, cumulus convection, stratiform clouds, orographic drag, and non-orographc gravity wave drag. The radiation scheme used here is RTE+RRTMGP (Pincus et al., 2019) where RRTMGP (Rapid Radiative Transfer Model for GCM application—Parallel) defines the radiative transfer problem based on optical properties and RTE (Radiative Transfer for Energetics) approximates a solution for the radiative transfer problem. The radiation scheme follows a correlated-k scheme to represent spectral variations and two-stream approximation, which can be described as upward and downward fluxes. Moreover, longwave (terrestrial) and shortwave (solar) radiation are treated separately because they cover different ranges of the radiative spectrum. Additionally, the separation has practical reasons because shortwave radiation is only calculated during the day and scattering is neglected for longwave radiation.

In ICON, the parameterizations are easily interchangeable which is convenient when comparing different parameterizations (traditional vs. ML-based). The triangular grid has the advantage that the grid cells are almost equally sized everywhere while a regular latitude-longitude grid has a decreasing grid size polewards. A regular grid has more grid points near the poles that cover a smaller area leading to oversampling in the zonal direction.

Because of the triangular grid, there is no oversampling of grid points in the polar region with ICON, which is helpful for ML-based approaches.

We run ICON-A for the year 1979 and save 5 hourly instantaneous output for one day every 2 weeks to get a data set that is as diverse as possible. The first output day saves the output starting at 00:00, the second day at 01:00, the third day at 02:00, and so on. Lagerquist et al. (2023) used a fixed interval of 6 h, which resulted in four equally spaced peaks in the spatial error distribution (see their Figure 7). Therefore, the odd output interval of 5 h is chosen on purpose to cover the diurnal cycle and more solar zenith angles with different local conditions. This is similar to Bertoli et al. (2025) and could also be achieved by randomly sampling solar zenith angles for a given state (Ukkonen, 2022). The time step of the physics parameterizations is 6 min including radiation. Usually, the radiation time step is 1−2 h. We chose a shorter radiation time step because we want to call the ML-based radiation emulator more often and more aligned with cloud cover, and therefore get the same distribution of atmospheric states as in simulation with high frequent radiation calls. The simulation data are always saved right before and after the radiation call to save the exact input/output of the traditional parameterization in order to capture the correct causality for our emulation. For training, we use the first 10 days of every month, for validation, the center 10 days, and for testing, we use the last 10 days. Although not every month is represented, every season is represented in each subset. That way, we reduce any type of autocorrelation and save storage space. For training speed and to increase variability in the training set, we do not use every cell for each time step, which results in 546k training samples.

2.1. Variables

The input (predictor) and output (target) variables are column-wise values of the model's radiation scheme, and are summarized in Table 1. We divided the training process into two separate components: one focused on shortwave (SW) radiation and the other on longwave (LW) radiation. This division aligns with how these components are treated separately in the original radiation scheme, with the SW component excluded during nighttime. In order to reduce the error from predicting intermediate variables such as the vertically resolved upward and downward flux, we only predict variables that are needed to couple the ML-based emulator to ICON, which involves SW and LW heating rates. Alternatively, we could predict upward and downward flux profiles and construct heating rates, but that may lead to larger errors in the upper layers (see SI for more details). Predicting flux profiles directly is not shown here; however, (Bertoli et al., 2025) reported that doing so can lead to stability issues when coupled to a model such as ICON, requiring additional scaling and smoothing of the upper layers of the fluxes to ensure stable online performance.

Additionally, we predict downward surface fluxes. The total shortwave downward flux $F_{\downarrow,surf,SW}$ can be partitioned into near-infrared (NIR), visible (vis), and photo-synthetically active radiation (PAR), which can be partitioned further into a direct and diffuse component. These partial fluxes and also the $F_{\downarrow,surf,SW}$ are important for coupling the emulator to ICON and its land model component. We also predict the upward flux at the top of the atmosphere, which is not needed to couple the emulator to the model but which is a variable that is needed for model tuning and is also interesting to check for energy consistency.

Unlike other ML-based radiation emulations, we omit the solar zenith angle as a direct input. In our study, the solar zenith angle is indirectly included in the incoming flux at the top of the atmosphere $F_{\downarrow,TOA,SW}$, which is the solar constant weighted by the Earth-Sun distance and solar zenith angle. We also neglect changes in greenhouse gas concentration (in particular CO_2) in our input, since our focus was solely on learning the radiation scheme from 1 year of data. During this period, GHG concentrations were fixed as a single annual global mean value. In addition, our approach omits aerosols in the input as we focus on the interpretation of an ML-based radiation emulation. Additionally, we focus on the impact of clouds and cloud-related variables because they are the largest contributor to the overall uncertainty (Forster et al., 2021). Aerosols affect radiation both directly and indirectly, with the indirect effect occurring through aerosol-cloud interactions. Since this indirect effect tends to be larger, it reinforces the importance of accurately representing cloud-radiation interactions (Forster et al., 2021).

2.2. Normalization

Normalization is essential for ML. One reason is to bring input variables to the same scale preventing the dominance of larger variables, such as temperature with a magnitude of 10^2 K while having smaller variables

Table 1
Input and Output Variables for Each Network and Training Phase

Variable	Unit	SW HR	SW flux	LW HR	LW flux
Input					
$F_{\downarrow,TOA,SW}$	W/m^2	✓	✓	–	–
α	–	✓	✓	–	–
$\alpha_{NIR,dir}$	–	–	✓	–	–
$\alpha_{NIR,dif}$	–	–	✓	–	–
$\alpha_{vis,dir}$	–	–	✓	–	–
$\alpha_{vis,dif}$	–	–	✓	–	–
$\vec{q_i}$	kg/kg	✓	✓	✓	✓
$\vec{q_l}$	kg/kg	✓	✓	✓	✓
$\vec{q}_{H2O} = \vec{q_v} + \vec{q_l} + \vec{q_i}$	kg/kg	✓	✓	✓	✓
$\overrightarrow{O_3}$	kg/kg	✓	✓	✓	✓
$\vec{\rho}$	kg/m^3	✓	✓	✓	✓
$\vec{cl}$	–	✓	✓	✓	✓
$\vec{T}$	K	✓	✓	✓	✓
T_{surf}	K	–	–	✓	✓
Output					
$\partial\vec{T}_{SW}/\partial t$	K/d	✓	–	–	–
$\partial\vec{T}_{LW}/\partial t$	K/d	–	–	✓	–
$F_{\downarrow,surf,SW}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,NIR,dir}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,NIR,dif}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,vis,dir}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,vis,dif}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,PAR,dir}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,SW,PAR,dif}$	W/m^2	–	✓	–	–
$F_{\uparrow,TOA,SW}$	W/m^2	–	✓	–	–
$F_{\downarrow,surf,LW}$	W/m^2	–	–	–	✓
$F_{\uparrow,TOA,LW}$	W/m^2	–	–	–	✓

Note. The network learns heating rates in the first training phase, denoted as HR. Boundary fluxes are learned in the second phase, denoted as Flux. F stands for upward ($\uparrow$) or downward ($\downarrow$) flux at the surface (surf) or top of the atmosphere (TOA), α is surface albedo, q_i is cloud ice, q_l is cloud liquid, q_v is specific humidity, O_3 is ozone concentration, ρ is density, cl is cloud area fraction, T is the atmospheric temperature profile, and $\partial T/\partial t$ is heating rate (HR). The variable used for each network and training phase are indicated by ✓. The vector sign indicates that a variable is defined on all vertical levels.

such as water vapor with a magnitude of 10^{-4}. Another reason for normalization is the context of the variables regarding their physical meaning. For example, the SW flux cannot exceed the incoming flux at the top of the atmosphere. Therefore, normalizing by a parameter that changes based on the context provides consistency across different data distributions (Beucler et al., 2024; Connolly et al., 2025; Shamekh et al., 2023). The following explanation provides more detail on this normalization process for each variable.

Cloud ice and liquid concentrations are normalized using level-wise total water concentration (q_{H2O}), where $q_{\mathrm{H2O}} = q_v + q_l + q_i$. This approach places greater focus on cloud-containing levels, and especially where ice and liquid concentrations are larger or comparable to water vapor concentrations. We find that this is especially

helpful in reducing the error of shortwave heating rate in the upper troposphere. There, the impact of cloud ice on heating rate can become larger than that of water vapor but the total concentration is still smaller than in lower layers as the density decreases. If normalizing the cloud-related variables only by their global mean values, the effect that cloud ice can become more important is not as strong.

Furthermore, $F_{\downarrow,TOA,SW}$ is normalized using the solar constant 1360 W/m^2. Shortwave fluxes are normalized using incoming shortwave fluxes $F_{\downarrow,TOA,SW}$. Longwave fluxes are normalized by σT_{surf}^4. Albedo and cloud fraction values naturally range between 0 and 1 and thus do not require normalization. Heating rates are not normalized because the majority of values lie between −10 and 10. All other variables are normalized using Z-score normalization

$$x_{\text{norm}} = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ is the mean and σ the standard deviation of the variable distribution. The mean and standard deviation are computed from the data of one time step using all cells and levels.

3. Method

We use PyTorch to develop our ML-based radiation emulation (Ansel et al., 2024). The training of the networks for SW and LW radiation is separated by data availability as SW radiation is calculated only during the day. However, the architecture of the networks is the same for SW and LW. Additionally, we differentiate between heating rates and fluxes, as heating rates are defined as vertical array variables containing all levels, whereas fluxes are scalar variables defined at a single level. Moreover, fluxes are defined at half levels, which is the upper and lower boundaries of a vertical grid cell, whereas heating rates are defined at full levels located at the cell center. We predict heating rates and fluxes using a single NN but split the training process into two phases. In the first phase, we optimize the prediction of heating rates (HR_{SW}, HR_{LW}). In the second phase, we learn predicting the boundary fluxes (FLUX_{SW}, FLUX_{LW}).

3.1. Energy Consistency

During training of the ML schemes, we enforce energy consistency, which is an inherent property of the physics-based radiation scheme. Ensuring this consistency is crucial for applying the ML schemes in climate simulations and for maintaining online stability. An unphysical energy source or sink can cause spurious local temperature changes, which in turn may trigger unrealistic responses in circulation and cloud distribution. Over time, these effects can accumulate to unphysical values, and potentially lead to a model crash. Therefore, we assess here the statistics of the imbalance between radiative energy changes in atmospheric columns and the accompanying divergence of radiative net fluxes at the atmospheric boundaries. This imbalance can arise because the ML scheme predicts heating rates and fluxes separately. The radiative balance is defined as follows:

$$(F_{\downarrow,TOA} - F_{\uparrow,TOA}) - (F_{\downarrow,surf} - F_{\uparrow,surf}) = \int_{surf}^{TOA} \frac{\partial T}{\partial t} c_{p,air} \rho dz \approx \sum_{l=0}^{n_{lev}} (F_{net,l+1/2} - F_{net,l-1/2}), \tag{2}$$

where l is defined at the layer center and $l \pm 1/2$ at the layer boundaries. The incoming flux at the top of the atmosphere $F_{\downarrow,TOA}$ defined by the solar constant, eccentricity and solar zenith angle for SW radiation and is zero for LW radiation. The fluxes $F_{\uparrow,TOA}$ and $F_{\downarrow,surf}$ are calculated by the NNs. The upward flux at the surface $F_{\uparrow,surf}$ is $\alpha F_{\downarrow,surf}$ for SW radiation and $\varepsilon \sigma T_{surf}^4$ for LW radiation where ε is emissivity of the surface, and σ is the Stefan-Boltzmann constant. We approximate the vertical integral of radiative energy as the sum of the vertical net flux divergence. Here, we construct the net flux divergence from the heating rates predicted by the NN, heat capacity $c_{p,air}$, density ρ, and vertical thickness of a layer dz. In the physics-based scheme, all terms of the sum over the net flux divergence except the boundary fluxes cancel each other. Therefore, the energy consistency is an inherent property.

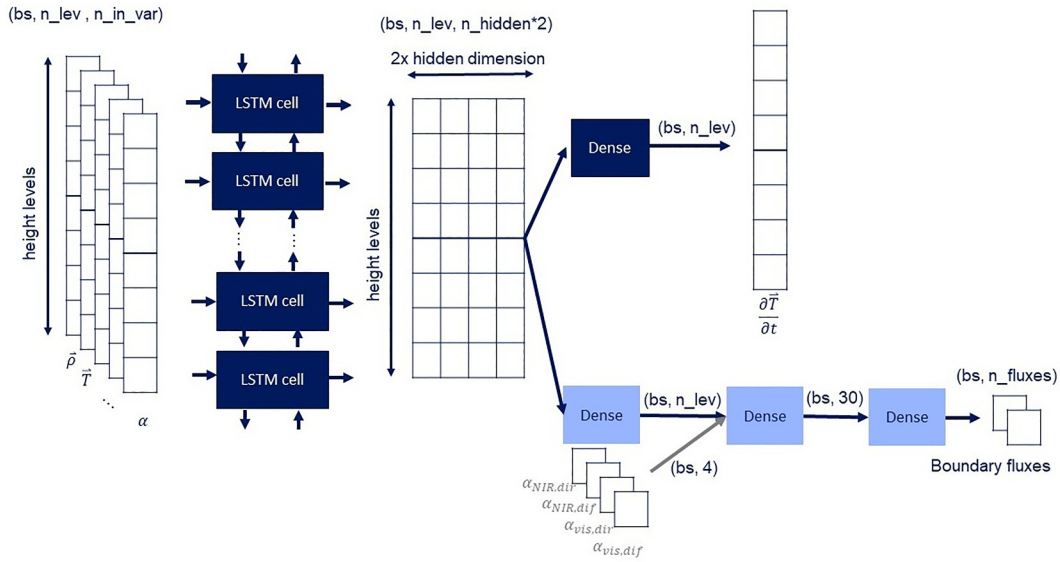

Figure 1. Schematic of the neural network architectures to emulate radiative heating rates and boundary fluxes. On the left, example input variables are density, atmospheric temperature, and albedo. Profile variables span all height levels, such as atmospheric temperature $\vec{T}$. Scalar variables, such as surface albedo α, are defined on one level and are expanded to match the height. In the first training phase, the LSTM cells predict one height level at a time, scanning the input by height level in both directions. A dense layer transforms the learned features of every height level to a heating rate ($\partial\vec{T}/\partial t$). In the second training phase, we freeze the LSTM weights, predicting the boundary fluxes using the Bidirectional Long Short-Term Memory output, and for SW, we add the albedos to compute partial fluxes. The size of the input and output of every layer is given in brackets, where *bs* stands for batch size, *n_nlev* is the number of vertical levels, *n_in_var* is the number of input variables, *n_hidden* is the number of nodes in an LSTM cell, *n_fluxes* is the number of boundary fluxes.

3.2. First Training Phase: Heating Rates

The radiation scheme we aim to emulate computes two column-wise streams of radiation throughout the atmosphere—the upward and downward fluxes—and then derives heating rates from the divergence of these flux profiles (Pincus et al., 2019). Given that radiation processes involve non-locality, with fluxes influenced by conditions in distant atmospheric layers, we chose a BiLSTM network. This non-locality can arise from various sources, such as clouds or moisture anomalies in distant layers. This model is well-suited to handle the bidirectional nature of the radiation streams and the complex dependencies across different layers. In a BiLSTM, each LSTM cell looks at one element of a sequence at a time. Here, the sequence corresponds to the levels of the atmospheric profile. The term *bidirectional* means that the network analyzes the vertical sequence (i.e., atmospheric layers) from the top of the atmosphere to the surface and the other way around, just like upward and downward fluxes in the radiation scheme. Bidirectional architectures have been found to perform better than a multi-layer perceptron (Ukkonen, 2022). The architecture choice is motivated by Yao et al. (2023), who compared various advanced architectures for radiative transfer problems and found that BiLSTM architectures were among the best performing models. Note that here, however, we only learn heating rates using a BiLSTM with significantly less trainable parameters (10 times less). The only parameter that controls the number of trainable parameters is the hidden dimension of the BiLSTM which we set to 96. The dense layer uses the hidden dimension as input and has one output feature. The total number of trainable parameters is 82.4k for SW and 81.6k for LW.

Figure 1 shows the architecture, where the dark blue boxes represent the layers trained during the first training phase. The BiLSTM takes a two-dimensional array as input, where the first dimension corresponds to the vertical dimension, and the second dimension, also known as channel, represents the physical properties of the current atmospheric state at each level. To match the size of the other variables in the vertical dimension, we expand scalar variables into a vertical array by repeating their value. Then, they are stacked with all other variables to create the 2D input array. Each LSTM cell processes all variables at the vertical level it is scanning as well as all scalar variables. The downward stream of the BiLSTM uses latent features from all levels above, whereas the

upward stream uses information from the levels below. The latent features contain information from all cells that the BiLSTM looked at before. This feature is known as memory, and informs the current cell if there was something important such as a cloud above or below which has a strong effect on the state of the cell. This bidirectional aspect represent the upward and downward direction of radiative fluxes, which is similar to the two-stream approximation in traditional radiation schemes. In other words, each LSTM cell learns to estimate the amount of radiation reflected, transmitted and absorbed by the atmosphere above or below. Then, the network returns a set of learned features for each level, with the length determined by the hidden dimension parameter, which controls the number of trainable parameters. Next, a dense layer combines the learned features at each level to compute heating rates. Note, the dense layer works only on the last dimension and has only one output feature, which is the heating rate at the current level. The dense layer shares the weight for all levels. For the shortwave network, we use a Rectified Linear Unit (*ReLU*) activation for the output to ensure that the prediction remains positive. Longwave heating rates are typically negative (indicating cooling), but they can also be positive when the surface is warmer than the air above, leading to atmospheric heating. To accommodate this variability, the longwave network does not use an output activation function, allowing it to handle both positive and negative values effectively.

To accurately model the large variability in our data and make reliable predictions for cloudy pixels, which are more difficult compared to clear-sky pixels, we construct a tailored loss function using multiple components as follows:

$$\mathcal{L}_{HR} = \text{MSE} + \text{MAE} + \min\left(10^{-8} * 10^{\frac{e-e_s}{n_e}}, 10^{-1}\right) * energy, \tag{3}$$

where the mean squared error (MSE) governs the loss during the early stage of training. However, as the MSE tends to diminish significantly due to its squared operation, the optimization process shifts its focus on the mean absolute error (MAE). The last term enforces energy consistency by minimizing the difference between the left- and right-hand side of Equation 2. This term is introduced after epoch e_s, which we defined as the epoch where MSE and MAE almost converged. Then, this term increases by a factor of 10 every n_e epochs, which we set as 10, whereas e is the current epoch. However, the maximum weight of this term is 10^{-1} so that it will not be much larger than the other terms. Model data provides the boundary flux terms in the energy term, allowing adjustments to the heating rate to maintain energy conservation.

We use the Adam optimizer with a learning rate of $5 * 10^{-3}$, along with a learning rate scheduler that reduces the learning rate by a factor of 2 when a plateau is reached. The plateau is reached when the minimum of the validation loss does not decrease by 0.01% for 20 subsequent epochs. Additionally, we employ early stopping with patience of 150 epochs to avoid overfitting (Goodfellow et al., 2016).

3.3. Second Training Phase: Boundary Fluxes

For the fluxes, we want to leverage what the BiLSTM has learned already. Therefore, we use the BiLSTM output in the second training phase and add three dense layers (see Figure 1 light blue). The first dense layer has an input size depending on the hidden dimension and one output feature. In other words, it combines the BiLSTM output to one feature per height level. For SW, we include the partial albedos in the input feature vector. The idea is that the first dense layer extracts sufficient spectral and vertical information, which is then combined with the partial albedos to predict the fluxes. After the first dense layer, we apply a `tanh` activation, followed by another dense layer and a second `tanh` activation. This dense layer has a hidden dimension of 32. The last dense layer depends on the number of output variables and is 8 for SW and 2 for LW. The output is limited between 0 and 1 for SW and 0 and 2 for LW which is due to normalization. The total incoming flux at the top of the atmosphere normalizes the SW fluxes, whereas surface emission limits the LW fluxes. The normalized LW fluxes can be larger than one if the cell above the surface is warmer than the surface itself. The three dense layers add in total 2.1k trainable parameters to the SW NN and 1.8k to the LW NN.

The loss function is the same as before (Equation 3) but all components in the energy term come from the NN. We choose the optimization and early stopping configuration as for the first training phase but start training with a learning rate of $1 * 10^{-3}$ and use the AdamW optimizer.

Table 2
Bulk Statistics for Heating Rate Results

	MAE [K/d]	Bias [K/d]	R^2	RMSE [K/d]
SW HR–total	0.045 (2.77%)	0.004 (0.38%)	0.98	0.154 (12.50%)
SW HR–clear	0.036 (1.90%)	0.005 (0.63%)	0.99	0.090 (6.47%)
SW HR–cloudy	0.047 (3.13%)	0.002 (0.30%)	0.98	0.166 (14.26%)
LW HR–total	0.060 (4.50%)	0.008 (0.60%)	0.99	0.214 (16.86%)
LW HR–clear	0.038 (7.00%)	0.007 (1.09%)	0.98	0.130 (18.03%)
LW HR–cloudy	0.069 (4.87%)	0.008 (0.60%)	0.99	0.230 (17.12%)

Note. MAE is mean absolute error and R^2 is coefficient of determination. RMSE is root mean squared error. The percentage values in brackets denote the relative values of MAE, bias, and RMSE.

4. Results

In this section, we evaluate the performance of all four components, HR_{SW}, HR_{LW}, $FLUX_{SW}$ and $FLUX_{LW}$, using the test set of the ICON-A simulation described above. This is an offline evaluation and comparison to the output of the traditional radiation parameterization.

4.1. Heating Rates

We begin by evaluating the predictions of the machine-learned heating rates, as summarized in Table 2. The overall MAE for both shortwave (SW) and longwave (LW) heating rates is 0.045 K/d and 0.060 K/d with biases of 0.004 K/d and 0.008 K/d, respectively. Although the longwave radiation calculation neglects scattering, it is not easier to compute than shortwave radiation because it has a source of radiation in every layer of the atmosphere itself. The coefficient of determination R^2 is 0.98 for SW and 0.98 for LW, where 0 indicates that the mean network prediction matches the mean value of the data distribution, which means that the sample-by-sample comparison could be bad. The closer the value is to 1, the better the prediction accuracy in a sample-by-sample comparison.

Figure 2 shows the vertical profiles of MAE and biases, averaged globally and over all time steps of the test set, as well as the coefficient of determination (R^2) for both longwave and shortwave heating rates. The prediction of the SW HR and LW HR components are virtually bias-free in the troposphere and stratosphere. For SW heating rates, the pronounced peak and spread in MAE in the upper stratosphere result from the significantly larger heating rates in that region, induced by ozone absorption. For LW heating rates, the MAE and its spread are very small in the stratosphere due to an overall reduced variability in heating rates. The spread in the troposphere primarily results from the presence of clouds. When evaluating clear-sky and cloudy-sky samples separately, the results show a reduced error and error spread for clear-sky samples in the troposphere (see Figures S2 and S3 in Supporting Information S1). The R^2 is very close to one for all levels for both, SW and LW heating rates. Nevertheless, the R^2 is slightly smaller in the troposphere than in stratosphere, which is also visible in the vertically resolved R^2. The R^2 has the MSE in the nominator and the deviation from the mean in the denominator. A larger variability in states is usually hard to capture for a model. The cloud variability is larger in the troposphere compared to the stratosphere, resulting in a larger MSE and therefore a smaller R^2 in the troposphere. If the variability in states is very small, the states are usually close to their mean value which means the denominator of R^2 gets very small. Despite a small MSE, the R^2 can be smaller in those cases, which can be seen at the upper troposphere and lower stratosphere at around 10–12 km. This region is cloud-free with ozone effects beginning at higher levels.

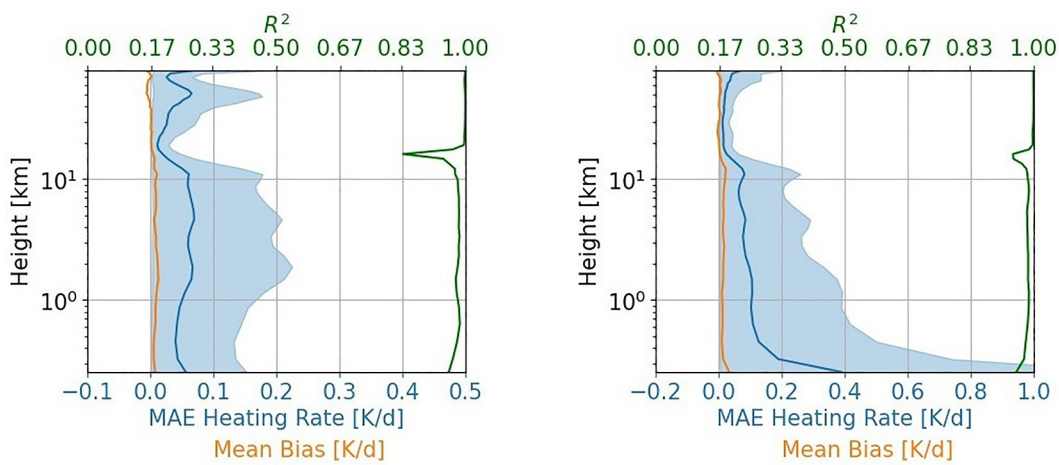

Figure 2. Global and time mean vertical profiles of heating rates. Mean absolute error, bias, and R^2 are shown for shortwave heating rates (left) and longwave heating rates (right). The shaded area shows 90% of error spread.

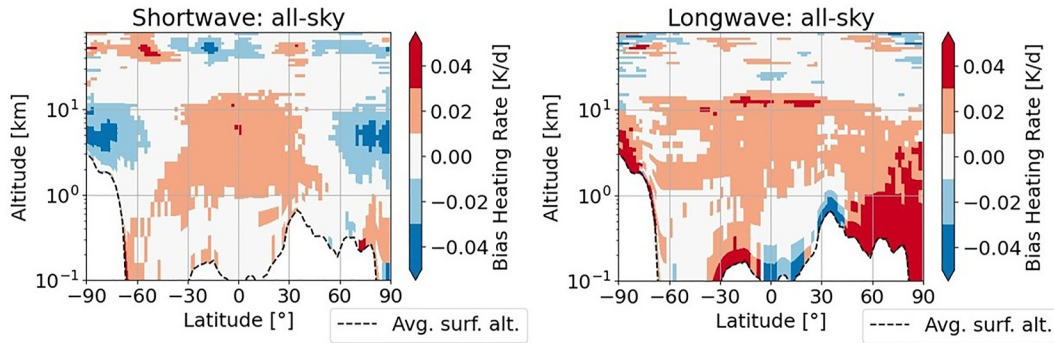

Figure 3. Zonal- and time-mean machine learned biases for (left) shortwave heating rates, and (right) longwave heating rates.

Figure 3 presents the bias of the heating rates in a height-latitude projection, covering both SW and LW heating rates. In the troposphere, the heating rate shows a small positive bias in the tropics and small negative bias in high latitudes. The LW heating rates are overall unbiased except for a small negative bias near the surface in the southern hemisphere and a small positive bias near the surface in the northern hemisphere. The bias is an important measure that does not guarantee online stability, but is a prerequisite.

For comparison with other studies, we also present the root mean squared error (RMSE) (Table 2). Hogan and Matricardi (2022) developed a tool for generating fast gas-optics models and report an RMSE of less than 0.18 K/d for clear-sky samples. Czarnecki et al. (2023) use an approach based on a linear weighted sum of optimally chosen frequencies and report an RMSE of 0.2 K/d for clear-sky longwave heating rates while we can reduce the RMSE to 0.13 K/d. A similar ML-based study is Lagerquist et al. (2023) using a U-Net variant and also covering 80 km of the vertical profile. They report in their Tables 8 and 9 an RMSE of 0.14 K/d for shortwave and 0.22 K/d for longwave heating rates, whereas having $10^{7.52}$ (approx. 33 million) and $10^{7.28}$ (approx. 19 million) trainable parameters. Ukkonen (2022) report an MAE of 0.07 K/d and an RMSE of 0.16 K/d for shortwave heating rates (their Figure 6) using a bidirectional NN with only 5,698 trainable parameters and a model top of 10 Pa. In comparison, Yao et al. (2023) report an RMSE of 0.032 K/d for shortwave heating rates and 0.139 K/d for longwave heating rates (their Table 3), using a BiLSTM with 1.12 million trainable parameters and a model top at 30 km. We can get a similar RMSE for heating rates of 0.154 K/d for shortwave and 0.214 K/d for longwave heating rates and an MAE of 0.045 K/d and 0.060 K/d, respectively, while using only a fraction of trainable parameters (80k).

4.2. Fluxes

The SW flux component predicts in total eight scalar SW fluxes and the LW flux component predicts two scalar fluxes. Table 3 summarizes the performance statistics. The upward flux at the top of the atmosphere $F_{\uparrow,TOA,SW}$ was predicted well with an error of 5.7 $\mathrm{W/m^2}$. The downward fluxes at the surface are in general predicted worse, where $F_{\downarrow,surf,SW}$ has an error of around 30 $\mathrm{W/m^2}$. The partial fluxes exhibit a smaller MAE of around 9 $\mathrm{W/m^2}$ for diffuse fluxes and 15 $\mathrm{W/m^2}$ for direct fluxes, but direct fluxes are on average larger than diffuse fluxes. The bias remains minimal, ranging from −0.6 to 0.4 $\mathrm{W/m^2}$. The NIR and visible fluxes approximately add up to the total SW downward flux (see Figure S4 in Supporting Information S1). The R^2 of >0.76 is generally high, and we observe that direct fluxes usually have higher R^2 values of >0.83. However, the R^2 values for SW fluxes are smaller than for LW fluxes, where the R^2 exceeds 0.99. The LW fluxes have a smaller MAE of 2 $\mathrm{W/m^2}$ and bias of $-0.29 - 0.17$ $\mathrm{W/m^2}$. For further analysis, we focus on the SW and LW downward flux at the surface and refer to the Supporting Information S1 for the other fluxes.

The MAE errors are larger in the tropics, see Figure 4. However, the map plot shows no clear spatial pattern, indicating that these errors are distributed relatively evenly across the globe. This is an important detail to note, as other studies, such as in Figures 7e–7f of Lagerquist et al. (2023), show peaks in the MAE at regular intervals, corresponding to their regular time step sampling. The larger errors in the tropics can be explained mainly by the frequent presence of clouds. The bias, Figure 4 right, appears somewhat erratic but is overall slightly negative.

Table 3
Bulk Statistics for all Fluxes

Variable	MAE $[W/m^2]$	Bias $[W/m^2]$	R^2
$F_{\uparrow,TOA,SW}$	5.70	−0.34	0.99
$F_{\downarrow,surf,SW}$	28.95	0.22	0.88
$F_{\downarrow,surf,SW,vis,dir}$	13.51	−0.57	0.85
$F_{\downarrow,surf,SW,vis,dif}$	8.21	0.03	0.81
$F_{\downarrow,surf,SW,NIR,dir}$	16.08	−0.52	0.83
$F_{\downarrow,surf,SW,NIR,dif}$	8.55	0.42	0.76
$F_{\downarrow,surf,SW,PAR,dir}$	14.71	−0.62	0.84
$F_{\downarrow,surf,SW,PAR,dif}$	8.57	0.21	0.78
$F_{\uparrow,TOA,LW}$	2.06	−0.29	0.99
$F_{\downarrow,surf,LW}$	1.78	0.17	0.99

For LW, the MAE is very small everywhere, Figure 5, but slightly larger in elevated areas such as the Andes and the Tibetan plateau and the bias is very small.

4.3. Energy Consistency

Taking the difference of the left and right side of Equation 2, we expect a mean of 0 W/m^2 if the training of the heating rates and the fluxes can approximate the energy consistency on average. The histograms of differences, computed separately for the SW and LW radiation, are shown in Figure 6. The mean for SW radiation is 0.59 W/m^2 and for LW radiation −0.07 W/m^2. The values are within ±0.5 W/m^2 which is acceptable.

When heating rate profiles and boundary fluxes are trained separately using distinct NNs (not shown but tested in a previous version), their predictions can become inconsistent, particularly in terms of energy balance. To address this, we train both components jointly with an energy constraint, ensuring that the predicted fluxes and heating rates are physically consistent. Compared to separate NNs for fluxes and heating rates, the presented approach also improves efficiency: the flux component now contains only a fraction of the trainable parameters and leverages shared representations learned by the BiLSTM. As a result, the spread in energy imbalance is reduced by a factor of two, the R^2 scores improve, and biases—especially in stratospheric shortwave heating rates—are significantly reduced. Although the MAE for total downward shortwave surface flux is slightly higher, this may reflect compensation for residual energy inconsistencies. Crucially, the bias in total shortwave boundary fluxes is reduced by an order of magnitude.

5. Interpretation

Neural networks do not necessarily learn the underlying physical relationships. Instead, they might rely on spurious links, which could lead to false heating rates and fluxes when applying the network to states that only slightly deviate from the training distribution. Therefore, we now focus on interpreting the predictions of the different networks. Here, our interest lies in understanding and assessing the extent of physically meaningful relationships within the networks. To achieve this, we employ a Shapley Additive exPlanations (SHAP) analysis (Lundberg & Lee, 2017), a method used for interpreting complex ML models by attributing predictions to input features. For the calculation of Shapley values, we use the captum package (Kokhlikyan et al., 2020). Here, we assess the strength of the contribution of specific inputs to specific outputs by comparing the mean absolute Shapley values using a subset of the data. Specifically, we use the test set as background data set and a random subset that corresponds to 1% of the background data set.

5.1. Shortwave Radiation

The top panel of Figure 7 shows the mean absolute Shapley values for the SW heating rates, predicted by a BiLSTM. Looking at the air density ρ (Figure 7f), the large Shapley values are present in the troposphere and lower stratosphere. Air density decreases exponentially with height. So, there is almost no impact of density on

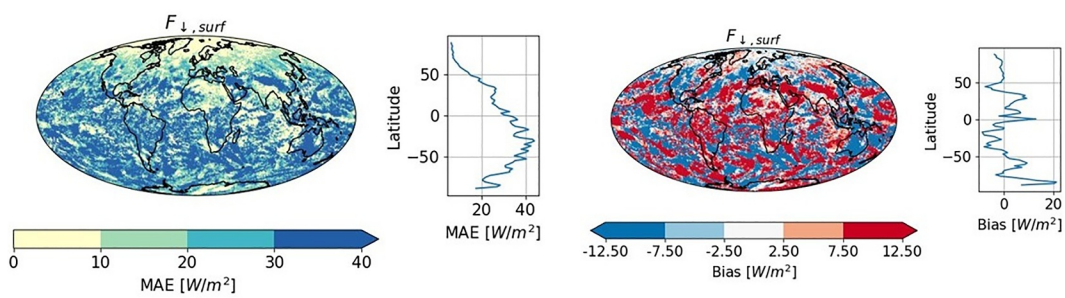

Figure 4. Time-averaged maps of shortwave downward flux at the surface. Mean absolute error (left) and (right) bias are shown. Right panels show zonal-mean values.

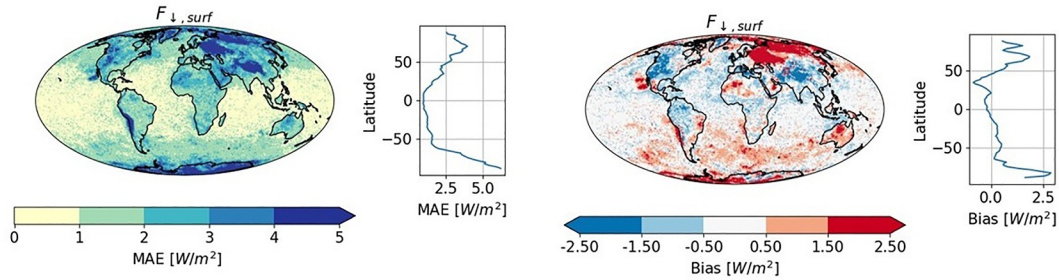

Figure 5. Same as Figure 4 but for longwave downward flux at the surface.

the heating rates in the stratosphere and mesosphere. Consequently, the network learned a sensible relation, as it directly links density and the amount of SW radiation absorbed and emitted. The temperature input for the SW HR output has non-negligible values that are primarily concentrated around the diagonal, indicating that the model uses temperature at each level to predict the heating rate at the same level, demonstrating a local dependency (Figure 7g). The BiLSTM primarily relies on local atmospheric variables to predict the heating rates, which are locally affected by absorption and emission of matter that is locally available (Figures 7a, 7b, and 7e).

The cloud fraction has the strongest contribution in the troposphere and affects the heating rate at the location of the cloud. However, it also exhibits strong non-local effects on all levels, particularly on lower levels below the cloud layer for SW, due to cloud shading. The non-local effects of clouds are consistent with our physical understanding, as clouds block or reflect SW radiation from the top, thereby reducing heating in the lower layers. Additionally, there is a moderate contribution from reflected radiation in the troposphere to the cloudless stratosphere at an approximate height of 30 km, which leads to heating in the stratosphere. This non-local contribution in the stratosphere is smaller than the local contribution in the troposphere potentially due to the following reasons: only a fraction of radiation gets reflected, there is less matter to heat and also the contribution of incoming radiation is the strongest in the stratosphere to mesosphere (see $F_{in,TOA}$ in Figure 7o). The upper stratosphere to mesosphere is cloud-free, and therefore, there is no impact on any level. Similar effects, local, non-local as well as affected layers, can be found for the cloud liquid q_l, and cloud ice q_i variables (Figures 7a and 7b). Unlike cloud variables, the contribution of ozone is concentrated in the stratosphere and mesosphere (Figure 7d). Ozone mixing ratio is highest in the stratosphere at 15−32 km and is the dominating factor that influences the shortwave heating rate and therefore the vertical temperature profile in the stratosphere (Wallace & Hobbs, 2006). The contribution of surface albedo α is strongest closer to the surface, associated with reflected radiation (Figure 7o).

The middle panel of Figure 7 shows the mean absolute Shapley values for the shortwave fluxes. In general, the fluxes $F_{\uparrow,TOA,SW}$ and $F_{\downarrow,surf,SW}$ have higher Shapley values than the partial fluxes, which are fractions of $F_{\downarrow,surf,SW}$.

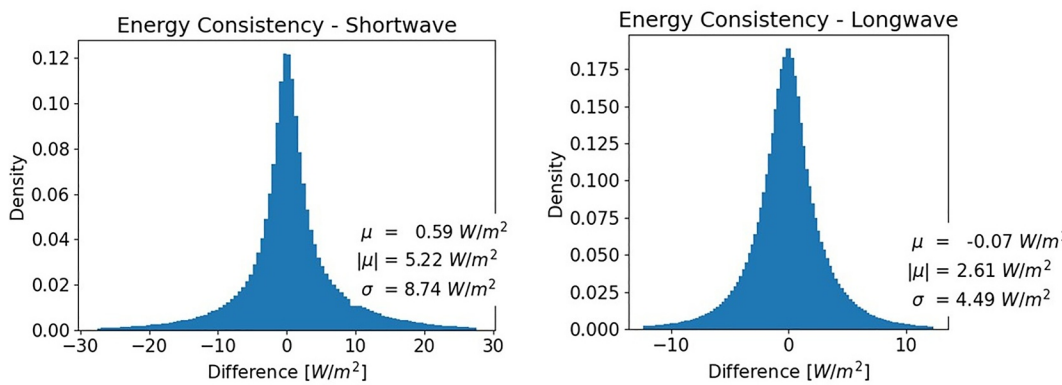

Figure 6. Energy balance check for combined neural networks (NNs) for SW radiation (left) and LW radiation (right). The histogram shows the difference between boundary fluxes and the vertical integral of radiative energy, both predicted by the NNs.

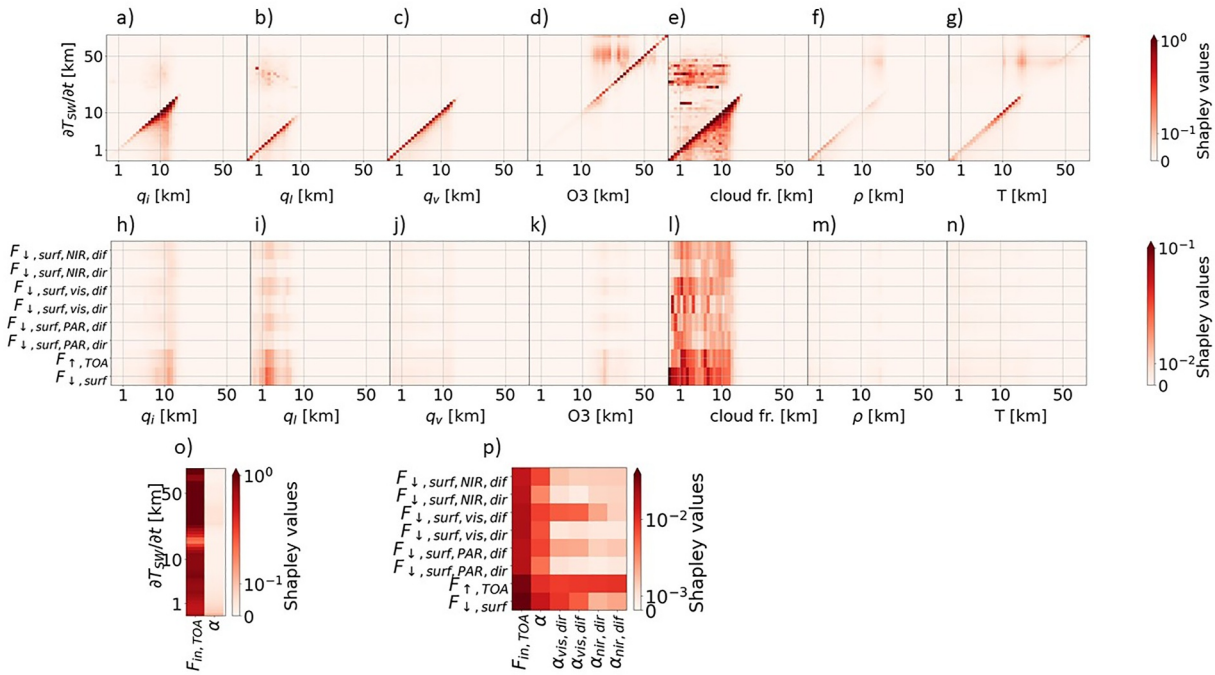

Figure 7. Mean absolute Shapley values for the neural network used for predicting SW heating rates and SW boundary fluxes. The *x*-axis represents the input variables, whereas the *y*-axis represents the predicted output, indicating how each layer of the input affects the corresponding layers of the output. The height scale is in model levels. 1, 10, and 50 km are marked for reference. Panels (a–g and o) show the input variables for the SW heating rate, whereas panels (h–n and p) show the input variables for SW boundary fluxes.

The albedo has a strong effect on $F_{\uparrow,TOA,SW}$ because it sets a lower limit to how much SW radiative flux can go out at the top of the atmosphere. Overall, input variables show a greater influence where they have larger values. For example, cloud cover is largest in the troposphere, and is associated with a strong effect on shortwave fluxes. Interestingly, almost all variables influence diffuse fluxes to a greater extent than direct fluxes. The stronger effect for diffuse fluxes can be attributed to the scattering of radiation in the presence of clouds, which contributes to the diffuse component.

The SW fluxes include both broadband fluxes at the TOA and surface, and the partial fluxes specific to certain bands (NIR, vis, PAR). In principle, the BiLSTM output at the top and bottom levels should retain sufficient vertical and spectral information to predict the corresponding boundary fluxes. However, our SHAP analysis (Figures 7h, 7i, and 7l) reveals that the model relies heavily on the nonlocal information from across the column when predicting these fluxes, in contrast to the heating rate predictions (Figures 7a–7g), which are dominated by local input features from most variables. This suggests that, in practice, the BiLSTM latent states at the boundaries do not encapsulate all necessary context for accurate flux prediction, likely due to the partial forgetting and compression inherent to the recurrent network but also heating rate prediction requirement.

To test this directly, we implemented at alternative version of the model that used only the top and bottom BiLSTM latent vectors to predict TOA and surface fluxes respectively. This variant resulted in higher biases (on the order of 20 W/m^2) and worse energy consistency, despite slightly improved accuracy for some partial SW flux component. These results reinforce the SHAP-based conclusion that explicitly using the full-column latent information leads to more reliable and physically consistent flux estimation in our case. However, we note that other studies have successfully predicted boundary fluxes when their approach was predicting flux profiles with a biLSTM (Ukkonen, 2022; Yao et al., 2023). As mentioned above and discussed in Supporting Information S1, we did not investigate this approach.

5.2. Longwave Radiation

Figure 8 displays the mean absolute Shapley values for the BiLSTM used to compute the LW heating rate. One of the strongest contributions comes from the surface temperature (Figure 8o), which is strongest directly above the

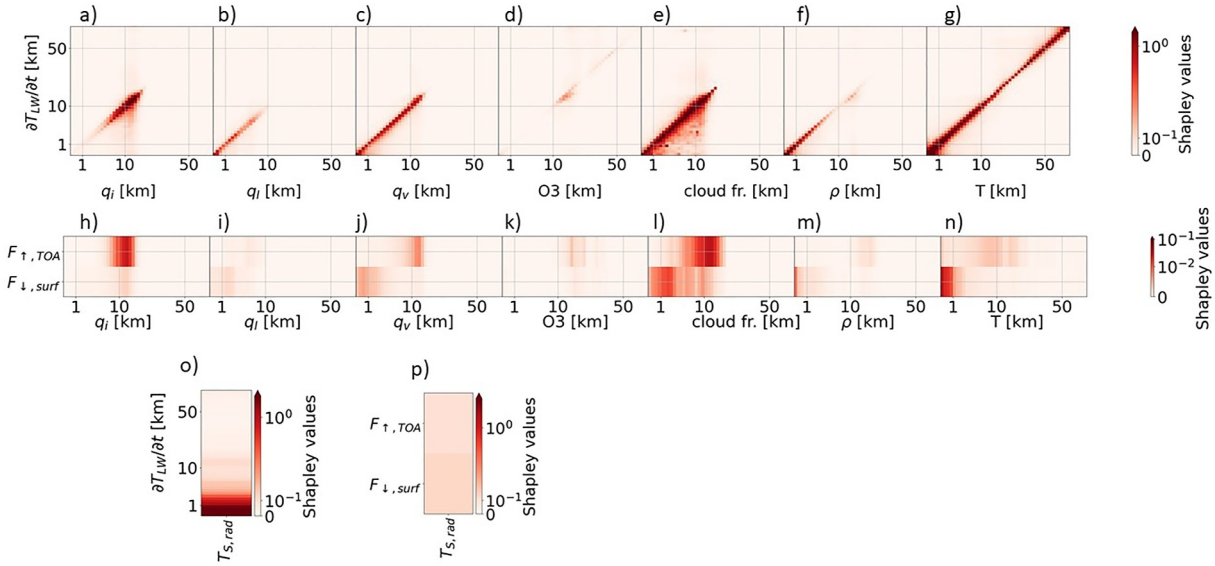

Figure 8. Similar to Figure 7 but for longwave radiation.

surface and decreases with height. Another significant contribution comes from the local temperature, which primarily exhibits a local impact. The impact of temperature on longwave heating rate is strongest at the same level (diagonal in Figure 8g) and also affects neighboring layers due to emission. The contribution from air density is strongest in lower layers because there is more matter radiating and absorbing in the longwave spectrum (Figure 8f). As density decreases with height, there is a smaller contribution to longwave radiation above the troposphere. Cloud-related variables—namely cloud fraction, cloud liquid q_l, and cloud ice q_i—contribute similarly to longwave heating rates as they do to shortwave heating rates (Figures 8a and 8b). Their effect is strongest locally, concentrated within the troposphere, and closely associated with convective processes. A notable difference, however, is that cloud-related variables exhibit slightly weaker and diffuse non-local effects. This is a physically meaningful effect, as scattering does not occur in longwave radiation. Instead, the effect is primarily driven by the absorption and emission of radiation, leading to diffuse local impacts. Moreover, the effect of ozone is much smaller on longwave heating rates and mostly local.

$F_{\downarrow,surf,LW}$ is more influenced by lower levels because they are closer to the surface, whereas $F_{\uparrow,TOA,LW}$ is more influenced by higher levels of the variables. For example, cloud fraction of the lower to middle troposphere strongly influences $F_{\downarrow,surf,LW}$, whereas cloud fraction up to the upper troposphere influences $F_{\uparrow,TOA,LW}$ (Figure 8l). Furthermore, $F_{\downarrow,surf,LW}$ is influenced by low tropospheric water vapor and cloud liquid, whereas $F_{\uparrow,TOA,LW}$ gets more impact from high ice clouds. This can be associated with locality, meaning the largest contribution comes from closer emission points. There is almost no contribution to $F_{\uparrow,TOA,LW}$ from the stratosphere and mesosphere except from ozone because the air density is very small and thereby also the emitted radiation.

The training process and architecture design did not include physical constraints, except energy conservation. However, the explainable AI analysis using Shapley values revealed physically meaningful relations between input and output for all networks. For instance, it showed the non-local cloud dependence of SW heating rate. Additionally, it demonstrated the local temperature dependence of LW heating rate. The BiLSTM has an important feature that is close to the physical scheme: the bidirectional scanning of the atmospheric column mimics the upward and downward fluxes of the radiation scheme.

5.3. Comparison to Multilayer Perceptron

For comparison, we conducted the same analysis using a multilayer perceptron (MLP). Unlike the BiLSTM, which efficiently leverages spatial structure and shared weights, the MLP requires more trainable parameters to achieve comparable performance. Specifically, the MLP consists of four hidden layers with 256 nodes each, totaling approximately 300,000 trainable parameters. Apart from the architecture, the training procedure—including the two-phase training strategy—was kept identical to that of the BiLSTM. In the first training

Table 4
Bulk Statistics for Heating Rate Results With the MLP

	MAE [K/d]	Bias [K/d]	R^2	RMSE [K/d]
SW HR–total	0.59 (19%)	0.32 (2.8%)	0.33	0.93 (48%)
SW HR–clear	0.52 (14%)	0.31 (2.1%)	0.31	0.67 (25%)
SW HR–cloudy	0.61 (22%)	0.31 (3.2%)	0.37	0.99 (55%)
LW HR–total	0.44 (31%)	0.003 (1.1%)	0.72	0.89 (64%)
LW HR–clear	0.29 (62%)	0.004 (0.7%)	0.72	0.49 (89%)
LW HR–cloudy	0.52 (33%)	−0.001 (1.4%)	0.72	1.02 (68%)

Note. MAE is mean absolute error and R^2 is coefficient of determination. RMSE is root mean squared error. The percentage values in brackets denote the relative values of MAE, bias, and RMSE.

phase, the learning rate is set to $5 * 10^{-4}$. The bulk statistics for heating rates are shown in Table 4. The corresponding plots as in Figures 2–6 and statistics for the boundary fluxes are provided in Supporting Information S1.

The MAE for the heating rate profiles ranges from 0.29 to 0.61 K/d for the MLP, which is roughly an order of magnitude larger than that of the BiLSTM (0.036−0.069 K/d). For the longwave heating rates, the average bias of the MLP (−0.001−0.004 K/d) is slightly smaller than that of the BiLSTM (0.007−0.008 K/d). However, the MLP's vertical bias profile is noticeably noisier (see in Supporting Information S1) compared to the BiLSTM (see Figure 2). For the shortwave heating rates, the MLP exhibits a much larger bias of 0.31−0.32 K/d compared to only 0.002−0.005 K/d for the BiLSTM. Although the MLP achieves an R^2 value above zero, indicating some predictive skill, its performance remains inferior to that of the BiLSTM across all metrics.

Overall, the MAE is comparable with similar MLP architectures, where Ukkonen (2022) report an MAE of 0.49 K/d for shortwave heating rates and Roh and Song (2020) report an RMSE of 0.92−1.03 K/d for longwave heating rates and 0.40−0.47 K/d for shortwave heating rates. Yao et al. (2023) reports an RMSE of 0.189 K/d for shortwave heating rates and 0.394 K/d for longwave heating rates, which is better, but their NN has twice as many trainable parameters and their model top is 30 km.

Figure 9 shows the mean absolute Shapley values for shortwave radiation using an MLP. The MLP captures some local relationships, particularly for specific humidity, cloud liquid water, and cloud ice (Figures 9a–9c). However, for variables such as ozone, density, and temperature, the MLP relies on non-physical or non-causal associations to predict heating rates. For example, it learns to use stratospheric ozone to predict tropospheric temperature tendencies (Figure 9d) or lower tropospheric density to predict heating rates in the upper stratosphere (Figure 9f).

The mean absolute Shapley values for longwave radiation are shown in Figure 10. As for the MLP applied to shortwave radiation, the longwave MLP identifies the importance of certain local features such as cloud liquid water influencing the longwave heating rate at the same vertical level (Figure 10c). However, the MLP also attributes strong non-local influence to temperature, with significant contributions from levels above and below

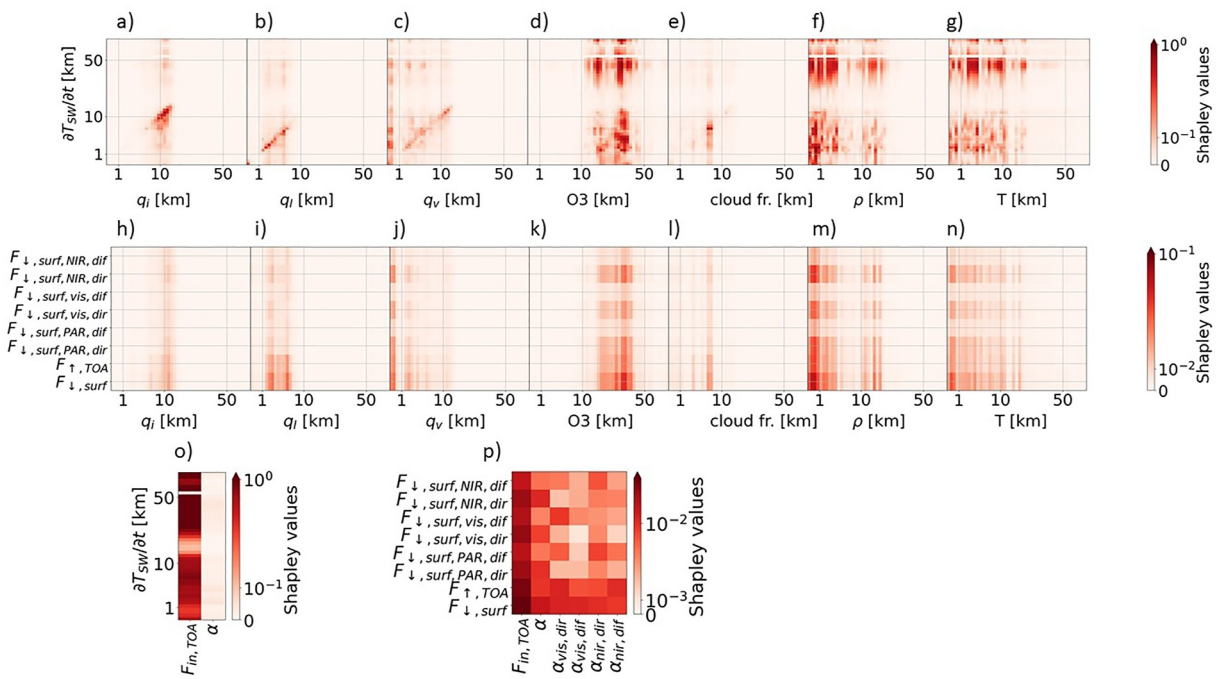

Figure 9. Similar to Figure 7 but using a multilayer perceptron for shortwave radiation.

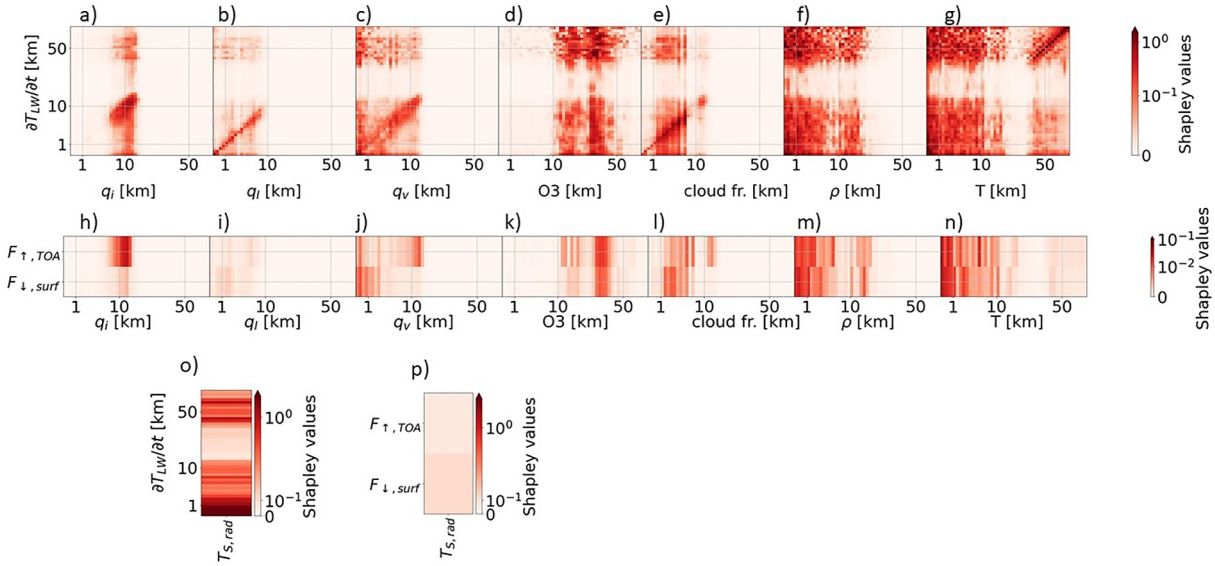

Figure 10. Similar to Figure 8 but using a multilayer perceptron for longwave radiation.

the target level (Figure 10g). This contrasts with the BiLSTM, which predominantly relies on local temperature information for predicting longwave heating rates.

The SHAP plots show that the MLP learns some important things, such as stratospheric ozone is important, or the density in the troposphere is more important than the density in the stratosphere, where the values are larger. However, the MLP fails to attribute it to the correct location for the heating rates. For instance, stratospheric ozone affects heating on all levels. Additionally, some levels appear completely irrelevant, leading to the checkerboard pattern. This suggests that physics-inspired networks, such as BiLSTMs, are able to capture important aspects of the underlying physics.

6. Conclusion and Discussion

Radiation is one of the most computationally expensive components in ESMs, despite several simplifications built into radiation parameterization and its application in ESMs. Machine learning can potentially help to speed up the calculation related to radiation—a key energy transfer in the climate system—while retaining accuracy. There have been attempts to emulate radiation using ML for different applications, but so far none for RTE+RRTMGP tailored to ICON. Additionally, the interpretation of the ML-based radiation emulation has often been missing. Here, we develop two NNs to emulate shortwave and longwave heating rates and surface fluxes. We use Bidirectional Long Short-Term Memory (BiLSTMs) to compute vertically resolved heating rates and a fully connected NN that computes boundary fluxes from the BiLSTM output.

Our ML-based model accurately emulates heating rates. The shortwave heating rates have an MAE of 0.045 K/d (2.77%) and a bias of 0.004 K/d (0.38%). The longwave heating rates have an MAE of 0.060 K/d (4.50%) and a bias of 0.008 K/d (0.60%). Both networks perform better on clear sky conditions than under cloudy sky conditions, emphasizing the need for further research on handling clouds with ML-based emulation. This is a subgrid process, as coarse resolutions do not resolve clouds, and clouds are not homogeneously distributed horizontally.

Using SHAP, we found that the networks learned relationships consistent with established physical principles. The BiLSTM predicting shortwave heating rates learned that locally absorbed and non-locally reflected radiation by clouds is significant, whereas the BiLSTM model for longwave heating rates identified the temperature profile as the most important contributor, given that the atmosphere itself is a source of longwave radiation. Additionally, the local cloud effect due to absorption and emission extends non-locally to influence adjacent regions in the atmosphere. In contrast, an MLP cannot account for such spatial dependencies and instead relies on correlations that may not reflect the underlying physics, highlighting the advantage of BiLSTMs for radiative transfer problems.

In this study, we focus on developing an accurate and interpretable data-driven architecture for implementation into the coarse-resolution version of the ICON model, providing a framework to overcome the "black box" approach in previous ML-based radiation developments. We neglected greenhouse gases and aerosols in this study, as we used only 1 year of training data, and they are prescribed by global annual mean values. This limitation is planned to be addressed in future work targeting long-term projections. We show that the NNs have good offline accuracy, and our interpretability analysis shows that the networks learned physically meaningful input-output connections. Additionally, we show that the NNs are statistically energy consistent, enforcing it during training. These connections and approximate energy consistency hold promise for our ML-based emulators to also perform well online when coupled to a model. The analysis of online performance will be presented in a future study. This study is paving the way for trustworthy physically consistent ML-based radiation calculations in a state-of-the-art ESM such as ICON, which may allow for more frequent radiation calls, and thereby an improved representation of cloud-radiation interactions.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The code can be found on GitHub and is archived on Zenodo (Hafner, 2025b). The software code for the ICON model is available from https://icon-model.org. Sample data can be found at Hafner (2025a).

Acknowledgments
KH and VE were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (Reference No. EY 22/2-1). FIS, PG, and VE additionally acknowledge funding by the European Research Council (ERC) Synergy Grant "Understanding and Modeling the Earth System with Machine Learning" (USMILE) under the Horizon 2020 Research and Innovation program (Grant Agreement No. 855187). KH was also supported by a fellowship of the German Academic Exchange Service (DAAD). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179. PG, RP, and SS were supported by the US National Science Foundation through the Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625). SS acknowledges support provided by Schmidt Sciences, LLC. We thank Peter Ukkonen and one anonymous reviewer for their helpful comments and suggestions, which improved our manuscript. Open Access funding enabled and organized by Projekt DEAL.

References

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., et al. (2024). PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (Vol. 2, pp. 929–947). ACM. https://doi.org/10.1145/3620665.3640366

Belochitski, A., & Krasnopolsky, V. (2021). Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art general circulation model. *Geoscientific Model Development*, *14*(12), 7425–7437. https://doi.org/10.5194/gmd-14-7425-2021

Bertoli, G., Mohebi, S., Ozdemir, F., Jucker, J., Rüdisühli, S., Perez-Cruz, F., et al. (2025). Revisiting machine learning approaches for short- and longwave radiation inference in weather and climate models. *Journal of Advances in Modeling Earth Systems*, *17*(9), e2025MS004956. https://doi.org/10.1029/2025MS004956

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., et al. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Bolton, T., & Zanna, L. (2019). Applications of deep learning to Ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018MS001472

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018GL078510

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., et al. (Eds.), Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change (pp. 147–286). https://doi.org/10.1017/9781009157896.003

Chevallier, F., Chéruy, F., Scott, N. A., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology and Climatology*, *37*(11), 1385–1397. https://doi.org/10.1175/1520-0450(1998)037⟨1385:ANNAFA⟩2.0.CO;2

Connolly, A., Cheng, Y., Walters, R., Wang, R., Yu, R., & Gentine, P. (2025). Deep learning turbulence closures generalize best with physics-based methods. https://doi.org/10.22541/essoar.173869578.80400701/v1

Czarnecki, P., Polvani, L., & Pincus, R. (2023). Sparse, empirically optimized quadrature for broadband spectral integration. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003819. https://doi.org/10.1029/2023MS003819

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., et al. (2021). The earth's energy budget, climate feedbacks and climate sensitivity. In *Climate change 2021 – The physical science basis: Working group i contribution to the sixth assessment report of the intergovernmental panel on climate change* (pp. 923–1054). Cambridge University Press.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018GL078202

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., et al. (2018). ICON-A, the atmosphere component of the ICON Earth system model: I. Model description. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1613–1637. https://doi.org/10.1029/2017MS001242

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., & Eyring, V. (2022). Deep learning based cloud cover parameterization for ICON. *Journal of Advances in Modeling Earth Systems*, *14*(12), e2021MS002959. https://doi.org/10.1029/2021MS002959

Hafner, K. (2025a). Interpretable machine learning-based radiation emulation for ICON [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.15199085

Hafner, K. (2025b). Interpretable machine learning-based radiation emulation for ICON [Software]. *Zenodo*. https://doi.org/10.5281/ZENODO.15199158

Heuer, H., Schwabe, M., Gentine, P., Giorgetta, M. A., & Eyring, V. (2024). Interpretable multiscale machine learning-based parameterizations of convection for ICON. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2024MS004398. https://doi.org/10.1029/2024MS004398

Hogan, R. J., & Matricardi, M. (2020). Evaluating and improving the treatment of gases in radiation schemes: The correlated k-distribution model intercomparison project (ckdmip). *Geoscientific Model Development*, *13*(12), 6501–6521. https://doi.org/10.5194/gmd-13-6501-2020

Hogan, R. J., & Matricardi, M. (2022). A tool for generating fast k-Distribution gas-optics models for weather and climate applications. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003033. https://doi.org/10.1029/2022MS003033

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., et al. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv:2009.07896*. https://doi.org/10.48550/arXiv.2009.07896

Krasnopolsky, V. (2012). Accurate and fast neural network emulations of long and short wave radiation for the NCEP global forecast system model. Retrieved from https://repository.library.noaa.gov/view/noaa/6951

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2008). Decadal climate simulations using accurate and fast neural network emulation of full, longwave and shortwave, radiation. *Monthly Weather Review*, *136*(10), 3683–3695. https://doi.org/10.1175/2008MWR2385.1

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370–1383. https://doi.org/10.1175/MWR2923.1

Lagerquist, R., Turner, D., Ebert-Uphoff, I., Stewart, J., & Hagerty, V. (2021). Using deep learning to emulate and accelerate a radiative transfer model. *Journal of Atmospheric and Oceanic Technology*, *38*(10), 1673–1696. https://doi.org/10.1175/JTECH-D-21-0007.1

Lagerquist, R., Turner, D. D., Ebert-Uphoff, I., & Stewart, J. Q. (2023). Estimating full longwave and shortwave radiative transfer with neural networks of varying complexity. *Journal of Atmospheric and Oceanic Technology*, *40*(11), 1407–1432. https://doi.org/10.1175/JTECH-D-23-0012.1

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research*, *102*(D14), 16663–16682. https://doi.org/10.1029/97JD00237

Morcrette, J.-J., Mozdzynski, G., & Leutbecher, M. (2007). A reduced radiation grid for the ecmwf integrated forecasting system (no. 538). *ECMWF*. https://doi.org/10.21957/ulnklieu

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018MS001351

Pal, A., Mahajan, S., & Norman, M. R. (2019). Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. *Geophysical Research Letters*, *46*(11), 6069–6079. https://doi.org/10.1029/2018GL081646

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3074–3089. https://doi.org/10.1029/2019MS001621

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Roh, S., & Song, H.-J. (2020). Evaluation of neural network emulations for radiation parameterization in cloud resolving model. *Geophysical Research Letters*, *47*(21), e2020GL089444. https://doi.org/10.1029/2020GL089444

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, *120*(20), e2216158120. https://doi.org/10.1073/pnas.2216158120

Song, H.-J., & Roh, S. (2021). Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002609. https://doi.org/10.1029/2021MS002609

Ukkonen, P. (2022). Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002875. https://doi.org/10.1029/2021MS002875

Ukkonen, P., & Hogan, R. J. (2023). Implementation of a machine-learned gas optics parameterization in the ecmwf integrated forecasting system: Rrtmgp-nn 2.0. *Geoscientific Model Development*, *16*(11), 3241–3261. https://doi.org/10.5194/gmd-16-3241-2023

Ukkonen, P., Pincus, R., Hogan, R. J., Pagh Nielsen, K., & Kaas, E. (2020). Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002226. https://doi.org/10.1029/2020MS002226

Veerman, M. A., Pincus, R., Stoffer, R., van Leeuwen, C. M., Podareanu, D., & van Heerwaarden, C. C. (2021). Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200095. https://doi.org/10.1098/rsta.2020.0095

Wallace, J. M., & Hobbs, P. V. (2006). *Atmospheric science: An introductory survey* (2nd ed. ed.), Vol. 92). Academic Press.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (2024). Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *Journal of Advances in Modeling Earth Systems*, *16*(2), e2023MS003668. https://doi.org/10.1029/2023MS003668

Yao, Y., Zhong, X., Zheng, Y., & Wang, Z. (2023). A physics-incorporated deep learning framework for parameterization of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003445. https://doi.org/10.1029/2022MS003445

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. https://doi.org/10.1029/2020GL091363