# Interpreting and Improving the Radiation Parameterization for ICON with Machine Learning

Doctoral Dissertation of
Katharina Hafner


2025

# Interpreting and Improving the Radiation Parameterization for ICON with Machine Learning

Am Institut für Umweltphysik
vom Fachbereich für Physik und Elektrotechnik
der Universität Bremen

*zur Erlangung des akademisches Grades*
**Doktor der Naturwissenschaften (Dr. rer. nat.)**
*genehmigte Dissertation*

von

Katharina Hafner

| | |
|---|---|
| Erster Gutachter: | Prof. Dr. Veronika Eyring |
| Zweiter Gutachter: | Prof. Dr. Pierre Gentine |
| Eingereicht am: | 10. Oktober 2025 |
| Tag des Promotionskolloquiums: | 23. Januar 2026 |


BY

# Abstract

With progressing climate change, we need robust climate projections supporting the development of mitigation and adaptation strategies to inform policy makers. Radiation is an important part of the climate system as it regulates the energy balance of the Earth but cannot be resolved in the spectral dimension in climate models and thus needs to be parameterized. Machine Learning (ML) has been shown to be useful to improve the performance of physical parameterizations in climate models. The main focus of this thesis is to develop an ML-based radiation scheme and to investigate possible improvements of the radiation parameterization in the ICOsahedral Nonhydrostatic (ICON) model.

In the first part of this thesis, I develop an ML-based radiation emulation for the ICON model. A Bidirectional Long Short-Term Memory (BiLSTM) network predicts radiative heating rates and boundary fluxes at the surface and top of the atmosphere based on the current atmospheric state including boundary conditions such as incoming solar radiation. The bidirectional part of the BiLSTM scans the vertical direction of an atmospheric column, similarly to the upward and downward fluxes in physics-based radiation schemes. To analyze what the Neural Network (NN) has learned, I used SHapley Additive exPlanations (SHAP) to identify which input variables determine the radiative predictions of the BiLSTM. The analysis of the Shapley values reveals that the BiLSTM learned relationships related to known physical laws: a cloud reflects incoming shortwave radiation which has a strong local effect on the heating rate, but also affects all layers below leading to less heating due to less downwelling shortwave radiation below the cloud. Additionally, the layers above the cloud are affected as well, as the reflected radiation can interact with ozone. For longwave radiation, the ambient temperature plays a significant role because it determines how much longwave radiation is emitted by the gases and clouds that are locally available. This shows that the ML-based radiation emulator learns relationships that are consistent with physical processes in radiative transfer models, which is a crucial step for its operational use in an Earth System Model (ESM).

The second part of this thesis focuses on the coupling of the proposed ML-based radiation emulator to the ICON model to perform hybrid simulations and test stability. The hybrid model ICON-$A_{ML}$ is compared with two reference setups using low and high frequent radiation updates (ICON-$A_{LF}$ , ICON-$A_{HF}$). The comparison of 10 ensemble members of each setup with a length of one year showed that ICON-$A_{ML}$ has no statistically significant deviations compared to the reference simulations. A long-term simulation of 10 years demonstrates that ICON-$A_{ML}$ shows no sign of accumulated biases compared to the references ICON-$A_{HF}$ and ICON-$A_{LF}$. This indicates that ICON-$A_{ML}$ is stable in longer simulations while also being four times faster than the original radiation parameterization. Testing with increased prescribed

sea surface temperatures showed that ICON-$A_{ML}$ can generalize to a +4 K warmer climate. This indicates that the ML-based radiation emulator is fit-for-purpose for its use in ICON simulations.

The final study of this thesis aims to improve cloud-radiation interactions in climate models utilizing high-resolution simulations. Global Storm Resolving Models (GSRMs) with a horizontal resolution of 5 km resolve cloud systems better than coarse-scale models with a horizontal resolution of around 100 km. Coarse-graining high-resolution data allows to implicitly learn subgrid-scale effects related to the horizontal and vertical distribution of clouds. I develop an approach to specifically target the improvement of the cloud radiative effects on heating rates. This approach showed 4-10 times smaller errors compared to the coarse-scale physics-based radiation scheme that was evaluated on coarse-grained data.

This thesis shows that the presented ML-based radiation emulator is interpretable and that it effectively mimics physical laws. The developed ML-based radiation scheme is fast and stable when coupled to a model like ICON. Additionally, it demonstrates that high-resolution simulations can be used to improve cloud radiative effects on heating rates by implicitly learning subgrid-scale effects paving the way for reducing longstanding biases in cloud radiative effects in machine learning enhanced ESMs.

# Integrated Author's References

Parts of this thesis (text, figures, and tables) are already published in the following peer-reviewed studies, studies that are submitted and under review. More details on this are given in Section 1.3 and at the beginning of the corresponding chapters.

## Publications as first author and co-author

**Hafner, K.**, Iglesias-Suarez, F., Shamekh, S., Gentine, P., Giorgetta, M. A., Pincus, R., & Eyring, V. (2025a). Interpretable Machine Learning-Based Radiation Emulation for ICON. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(4), e2024JH000501. https://doi.org/10.1029/2024JH000501

**Hafner, K.**, Iglesiaz-Suarez, F., Shamekh, S., Gentine, P., Pincus, R., Giorgetta, M., & Eyring, V. (2025b). Stable Machine Learning based Radiation Emulation for ICON. *Under Review for Journal of Advances in Modeling Earth Systems*. https://doi.org/10.22541/essoar.174708082.27787580/v1

**Hafner, K.**, Shamekh, S., Bertoli, G., Lauer, A., Pincus, R., Savre, J., & Eyring, V. (2025c). Representing Subgrid-Scale Cloud Effects in a Radiation Parameterization using Machine Learning: MLe-radiation v1.0. *Submitted to Geoscientific Model Development, arXiv preprint*. https://doi.org/10.48550/arXiv.2510.05963

# Contents

# 1. Introduction

## 1.1. Motivation

Climate change is one of the most pressing issues of our time with broad consequences for the environment, such as increased frequency and severity of extreme events that put humans at risk (Seneviratne et al. 2021). To address this global challenge, 193 countries agreed to the Paris agreement of 2015 limiting the increase of global mean surface temperature to a maximum of 2 $K$ compared to pre-industrial time (UNFCCC 2015). By combining observations and model simulations, it has already been shown that humans are the cause of climate change (Eyring et al. 2021). The development of accurate climate models is important to inform policy makers and prepare mitigation and adaptation strategies. The importance of climate research, especially climate modelling, was acknowledged with the Noble prize in physics in 2021 for Manabe, Parisi and Hasselmann who laid the foundation for climate modelling and understanding of complex systems (Manabe et al. 2021). The first model was a single column radiative convective equilibrium model (Manabe and Wetherald 1967). Since then, climate models have been further developed and now cover the whole globe and more processes and components are included in a model such as land, ocean, biogeochemical processes. These complex models are also called Earth System Models (ESMs). The Coupled Model Intercomparison Project (CMIP) Phase 6 (CMIP6) (Eyring et al. 2016), provides a protocol that allows to compare ESMs with each other under the same experiment setup. One important metric in climate modelling is the Equilibrium Climate Sensitivity (ECS), which is used to quantify the global temperature increase for doubling of the atmospheric $CO_2$ concentration. However, the inter-model spread of ECS in CMIP6 did not decrease compared to CMIP5 indicating that models still have remaining uncertainties. A large portion of remaining spread can be related to different representations of clouds (Zelinka et al. 2020). Additionally, ESMs have long standing biases such as the shortwave cloud radiative effect of local deviations of up to 30 $W/m^2$ (Lauer et al. 2023). Therefore, ESMs need to be improved especially in terms of cloud-radiation interactions.

Radiation is driving the climate system, heating and cooling the atmosphere. Clouds reflect, absorb and transmit radiation. The absorption spectrum of the atmosphere is very complex, which requires accurate calculations. State-of-the-art radiative transfer models are based on a *line-by-line* approach that requires performing integrals over $10^5$ *lines* (very narrow bands in the wavelength spectrum) (Hogan and Matricardi 2020). The computational cost of these models is so large that it is impossible to include a line-by-line model in an ESM with simulations spanning multiple decades. Therefore, different approximations have been proposed to design

radiation parameterizations that can be run at a reduced computational cost in ESMs (Hogan and Bozzo 2018; Pincus et al. 2019), and line-by-line models are only used to validate these efficient radiative transfer codes. Despite simplifications and approximations, radiation is still one of the most expensive components in ESMs. One way to reduce the computational burden is to calculate radiation less often than other processes. However, the state variables and clouds are kept constant for periods of $1 - 2\ h$ while other processes are updated every $10\ min$. Therefore, the radiative temperature tendencies that update the temperature in the atmosphere are based on a state and cloud distribution that does not represent the current state. This leads to uncertainties in the cloud-radiation interactions.

One possibility to reduce the computational cost while retaining accuracy is Machine Learning (ML). A faster radiation scheme would allow to calculate radiative heating rates more often and improve the cloud-radiation interactions. In the last decade, the development of Graphics Processing Units (GPUs), efficient ML algorithms and increased data availability boosted research in the field of Artificial Intelligence (AI). Applications include image recognition, natural language processing and speech recognition, as well as applications in physics like AI weather models (Bi et al. 2023; Lam et al. 2023) that even beat state-of-the-art physics based models (Rasp et al. 2023). However, these models can not run at climate time scales yet because they have no information about the boundary conditions that lead to a changing climate i.e., changing Greenhouse Gas (GHG) concentrations. A different approach is to replace single components of a model such as a parameterization which represents the physics of a subgrid-scale process with a faster machine learned alternative, or by learning from high-resolution simulation to include subgrid-scale effects (Gentine et al. 2018; Rasp et al. 2018; Yuval and O'Gorman 2020).

The aim of this dissertation is to develop an improved ML-based radiation scheme to speed up the computation and calculate radiative tendencies as often as other physical processes to ultimately improve cloud-radiation interactions. This has been targeted by other studies already (Song and Roh 2021; Ukkonen 2022; Yao et al. 2023). But, instead of climate projections, most applications focused on Numerical Weather Prediction (NWP), and thus the troposphere and lower stratosphere. However, for climate projections the stratosphere is also important, which is difficult to model with existing ML-based radiation emulators. Another implication of the NWP focus of ML-based radiation schemes is that they were often, if at all, tested only for a few days online, i.e., implemented in an ESM such as the ICOsahedral Nonhydrostatic (ICON) model.

Climate models usually have a horizontal resolution of $100 - 200\ km$ (Chen et al. 2021) where the subgrid-scale structure of i.e., clouds is not known. The development of Global Storm Resolving Models (GSRMs) with a horizontal resolution of $1 - 10\ km$ accelerated in the past few years. While it is still a huge effort to run high resolution climate projections, these models provide an excellent source of training data for the development of ML-based parameterizations for coarse-scale climate models. Coarse-graining high resolution data to a target resolution allows to implicitly learn subgrid-scale effects.

## 1.2. Key Scientific Questions

The aim of this thesis is to replace the radiation parameterization in the ICON model with an ML-based scheme that is fast, stable and improved by using high resolution simulations. Therefore, the thesis focuses on three key scientific questions:

1. Can an ML-based emulator learn to represent atmospheric radiative transfer based on physical laws?
2. Can the ML-based emulator be coupled to a complex ESM like ICON and produce stable simulations?
3. Can we improve the representation of cloud-radiation interactions by using high-resolution data for training?

## 1.3. Structure

The thesis contains parts that were published in peer-reviewed journals or are currently under review. A list of publications that contributed to the results of this thesis where I was first-author can be found on page vii. Whenever I use parts of these studies in this thesis, the chapters and sections will be prefaced with a more detailed description. Additionally, I keep the pronoun "we" to maintain readability and consistency as well as to acknowledge the contributions of my co-authors. However, I declare that all content from these publications including text, tables and figures, originates from me unless stated otherwise.

This thesis is structured as follows: Chapter 2 introduces the scientific background starting with a theoretical introduction of radiative transfer in Section 2.1, and a general introduction to climate models (Section 2.2) followed by a description of atmospheric radiation in Section 2.3. Then, a few modelling concepts that are important for radiation are presented in Section 2.4. These concepts include the plane parallel approximation (Section 2.4.1), the correlated k-distribution method (Section 2.4.2) and the treatment of clouds (Section 2.4.3). All of these are important for a radiation parameterization (Pincus et al. 2019), which is described in Section 2.4.4. In Section 2.5, I introduce the concept of machine learning including basics (Section 2.5.1) and general terms of training procedure (Section 2.5.2). Chapter 2 concludes with a literature review and current status of ML-based radiation approaches, which is also part of the literature review presented in Hafner et al. (2025a) and Hafner et al. (2025b). Chapter 3 describes the ICON model including simulations (Section 3.1) and methods used in this thesis, specifically the neural network architecture and interpretability methods (Section 3.3).

Chapters 4–6 contain the results of this thesis, where the key science questions from Section 1.2 are answered step by step. First, I focus on addressing key science question 1 in Chapter 4, which is about the development of the ML-based radiation emulator and interpretation of the ML-based scheme. These results have been published in the *Journal of Geophysical Research: Machine Learning and Computation* (Hafner et al. 2025a). Chapter 5 addresses key

science question 2 by presenting simulations where the ML-based emulator is implemented in the ICON model. These results were presented in Hafner et al. (2025b) which is currently under review for the *Journal of Advances in Modeling Earth Systems*. Chapter 6 explores the use of high resolution data to improve radiative heating rates, which has been submitted to the journal *Geoscientific Model Development* (Hafner et al. 2025c). The results of the thesis are summarized in Chapter 7, where I also present an outlook on possible future work.

# 2. Scientific Background

Atmospheric radiation determines how the atmosphere heats when the sun is shining during the day and cools during the night. However, it becomes more complicated with clouds as they reflect, absorb and transmit radiation. Considering aerosols and GHGs further increases complexity. First, I present the theoretical background of these processes. Before diving deeper into atmospheric radiation, I give a general introduction into ESMs, followed by a description of modelling concepts that are relevant for radiation parameterizations. Then, a general introduction into ML is followed by an introduction of certain integral studies that pioneered the development of ML-based radiation schemes in the last decades, and the current status of the field.

## 2.1. Theory of Radiation and Radiative Transfer

In general, electromagnetic radiation consists of propagating electromagnetic waves. In terms of the wave-particle dualism (Compton 1923; Young 1804), a wave can be considered as a photon. Thermal radiation is electromagnetic radiation that is emitted by every body above a temperature of absolute zero. A system can be in local thermal equilibrium which means that the entropy of a system has reached its maximum. This equilibrium can also be characterized by a constant temperature and pressure of the system. In local thermal equilibrium, a body emits as much radiation as it absorbs, also known as Kirchhoff's law

$$\epsilon = \alpha_a, \tag{2.1}$$

with the emissivity $\epsilon$ and absorptivity $\alpha_a$, which is limited between 0 and 1. A blackbody is a body that absorbs all incoming radiation ($\alpha_a = \epsilon = 1$).

The emitted amount of radiation can be described as radiance or irradiance, where radiance means the emitted intensity per solid angle and irradiance means the total flux in all directions. The emitted irradiance $B$ of a black body can described by Planck's law

$$B(\lambda, T) = \frac{2hc^2}{\lambda^5\left(exp\left(\frac{hc}{k_B T \lambda}\right) - 1\right)}, \tag{2.2}$$

where $h$ is the Planck constant, $c$ is the speed of light, $\lambda$ is the wavelength, $k_B$ is the Boltzmann constant and $T$ is temperature. With $f = \lambda/c$ Planck's law can be rewritten in terms of frequency $f$. When integrating over the solid angle and the wavelength of Planck's law, the total irradiance $F_b$ of a black body with temperature $T$ can be derived as

$$F_b = \pi \int B(\lambda, T) d\lambda = \sigma T^4, \tag{2.3}$$

where $\sigma$ is the Stefan-Boltzmann constant. This equation is also known as the Stefan-Boltzmann law, which can be used to determine, e.g., the "surface" temperature of the Sun for which the black body assumption is a good approximation. The Stefan-Boltzmann law can be multiplied by emissivity $\epsilon$ for gray bodies ($\epsilon < 1$). Emissivity can also be defined as the amount of emitted radiation divided by the corresponding black body radiation for a given wavelength.

The explained concepts of absorption and emission of black and gray bodies are relevant to further describe the propagation of electromagnetic radiation through a medium such as the atmosphere. This transfer of electromagnetic radiation requires the definition of source and sink terms in a medium or along a path $ds$. The main source term is emission while the main sink term is absorption. An additional source and sink term of electromagnetic radiation is scattering. Scattering towards the path - also called *in-scattering* - is a source term while scattering away from the path or *out-scattering* is a sink term. For in-scattering, the change in monochromatic intensity $I_\lambda$ along a path $s$ is described as

$$\left(\frac{dI_\lambda}{ds}\right)_{in-scat} = \alpha_s \frac{1}{4\pi} \int_0^{4\pi} I_\lambda(\Omega')P(\Omega, \Omega')d\Omega, \tag{2.4}$$

where $\alpha_s$ is the scattering coefficient, $\Omega'$ the incoming and $\Omega$ the scattering solid angle, and $P$ is the phase function describing the probability of scattering given the solid angles $\Omega'$ and $\Omega$. The phase function highly depends on the size of the particle and the wavelength $\lambda$ of the photon that scatters on the particle. Usually, a beam of radiation is considered and not just a single photon which then yields a distribution of scattered photons. The scattered distribution is representative for different types of scattering regimes that are classified by the size parameter. The size parameter is defined as

$$x = \frac{2\pi r}{\lambda}, \tag{2.5}$$

with the wavelength $\lambda$ of incoming radiation and scattering particle size $r$, where the scattering particle is assumed to be spherical. $x << 1$ defines Rayleigh scattering with a scattering efficiency that is proportional to $\lambda^{-4}$. The Mie scattering regime is defined as $x \approx 1$ with a scattering efficiency proportional to $\lambda^{-1}$. For Mie scattering with $x > 1$, forward scattering is preferred. For $x >> 1$, geometric optics describes the behavior of radiation and the interacting particle.

The scattering sink or *out-scattering* is usually described along with absorption as extinction

$$\left(\frac{dI_\lambda}{ds}\right)_{ext} = -(\alpha_s + \alpha_a)I_\lambda, \tag{2.6}$$

with the absorptivity $\alpha_a$. The absorption depends on the characteristics of a particle. Therefore, different absorption mechanisms exist. A photon can excite an electron in an atom or molecule to a higher energy level which allows the electron to change orbitals. The energy that a photon needs to cause this corresponds to a specific wavelength that depends on the atom and the energy levels, which is known as a spectral line. If an electron is already excited i.e.,

in a higher energy state, it can emit a photon in any direction and transfer to a lower energy state. Therefore, the spectral line can be observed due to an emitted or absorbed photon. Another absorption mechanism is the excitation of a molecule into higher vibrational and rotational modes. Vibration means that the atoms move relative to each other. Rotation means that the molecule rotates around an axis. A molecule can be described by a number called degree of freedom, which depends on the number of atoms in that molecule. A larger degree of freedom means that the molecule can absorb more energy. Therefore, methane (CH4) can absorb more energy than water (H2O) because it has more atoms and therefore more degrees of freedom. Usually, the excitation of higher energy levels either of electrons, rotational or vibrational modes, is a discrete process, absorbing photons of a specific wavelength. However, other processes can influence absorption and measurements show that these discrete lines are broadened. There are different mechanisms that lead to line broadening e.g., natural, pressure and Doppler broadening. Natural broadening is explained by Heisenberg's uncertainty principle. This relates to the uncertainty of energy and time which means that the time at which the absorbed line is re-emitted is uncertain. Pressure broadening is associated with collisions of nearby particles which is more dominant in the lower atmosphere due to larger pressure (Strong and Plass 1950). Doppler broadening is associated with the movement of the absorber to or away from the observer which is more dominant in the upper atmosphere above 50 $km$ where less particles are present (Wallace and Hobbs 2006).

The emission of matter is another source term and can be described by Planck's law (Equation (2.2)) and emissivity $\epsilon$ as

$$\left(\frac{dI_\lambda}{ds}\right)_{emiss} = \epsilon B(\lambda, T). \tag{2.7}$$

Combining all source (Equations (2.4) and (2.7)) and sink (Equation (2.6)) terms yields the radiative transfer equation:

$$\begin{aligned} \frac{dI_\lambda}{ds} &= \left(\frac{dI_\lambda}{ds}\right)_{emiss} &&+\left(\frac{dI_\lambda}{ds}\right)_{ext} &&+\left(\frac{dI_\lambda}{ds}\right)_{in-scat} \\ &= \epsilon B(\lambda, T) &&-(\alpha_s + \alpha_a) I_\lambda &&+\alpha_s \frac{1}{4\pi} \int_0^{4\pi} I_\lambda(\Omega') P(\Omega, \Omega') d\Omega. \end{aligned} \tag{2.8}$$

Solving the radiative transfer equation for the atmosphere is analytically not possible. One of the reasons is multiple scattering. A photon can be scattered more than once along a path in the atmosphere. Calculating multiple scattering would require the knowledge of the scattering phase function which depends on the shape of the scattering particle and thereby also the atmospheric composition. Though, a beam or rather a spectrum of radiation is considered for radiative transfer and not just a single photon. Additionally, a radiative transfer solution would require the integration over the absorption spectrum which is complex due the different absorption mechanisms explained above. However, radiation is an important physical process in ESMs which will be explained in the next section.

## 2.2. Climate Models

Earth System Models (ESMs), or more general climate models, help us to understand and model the Earth and its climate. In the 60s and 70s, the development of atmospheric General Circulation Models (GCMs) gained momentum and the first coupled atmosphere ocean GCM was presented (Manabe and Bryan 1969). These early models were used to study the drivers of atmospheric circulation (Phillips 1956) and how the surface temperature changes when doubling $CO_2$ (Manabe and Wetherald 1967). Since then, model development has advanced massively and now includes more components of the Earth system such as chemical and biogeochemical processes, like the carbon cycle. Additionally, since the 70s many satellite and ground-based observations have become available to evaluate ESMs. Today, ESMs can be used to study how and why the climate changes by changing boundary conditions like the concentration of GHGs. To address growing concerns about the changing climate, the Intergovernmental Panel on Climate Change (IPCC) was formed and is now regularly reporting on the changing climate (IPCC 2021).

A climate model is based on the *primitive equations* that describe the large-scale changes of the atmospheric state using the six *prognostic variables*: horizontal and vertical velocity, humidity, temperature, pressure and density. The primitive equations include prognostic equations describing the time-dependent change of the state variables and diagnostic equations defining the instantaneous change between the variables (Wallace and Hobbs 2006). Fundamental physical laws behind the primitive equations include conservation laws such as conservation of mass, momentum, energy and moisture.

The primitive equations describe the large-scale dynamics which is calculated by the dynamical core in a climate model. For the numerical calculation of the dynamics, the Earth is discretized into grid boxes of finite volume. The size of these boxes depends on the application: for climate modelling the horizontal size or resolution is in the order of $100 - 200\ km$ and for NWP the horizontal resolution is in the order of $10\ km$. Depending on the resolution some processes can not be resolved as they happen on smaller scales. The small scale physical processes need to be approximated which is called parameterization (Gettelman and Rood 2016).

By now, there exist many climate models and all represent physical processes slightly differently. CMIP (Eyring et al. 2016) provides a federated experimental protocol for ESMs, i.e., models with interactive components that represent atmosphere, land, ocean, and cryosphere. Models that participate in CMIP provide model output in a common format that allows to more easily compare and evaluate different climate models to help us understand the climate system.

Sometimes, it is useful to focus on one part of the climate system e.g., the atmosphere. The Atmospheric Model Intercomparison Project (AMIP) provides a protocol within CMIP that focuses on the atmosphere with prescribed ocean parameters i.e., sea ice concentrations (SICs) and sea surface temperatures (SSTs).

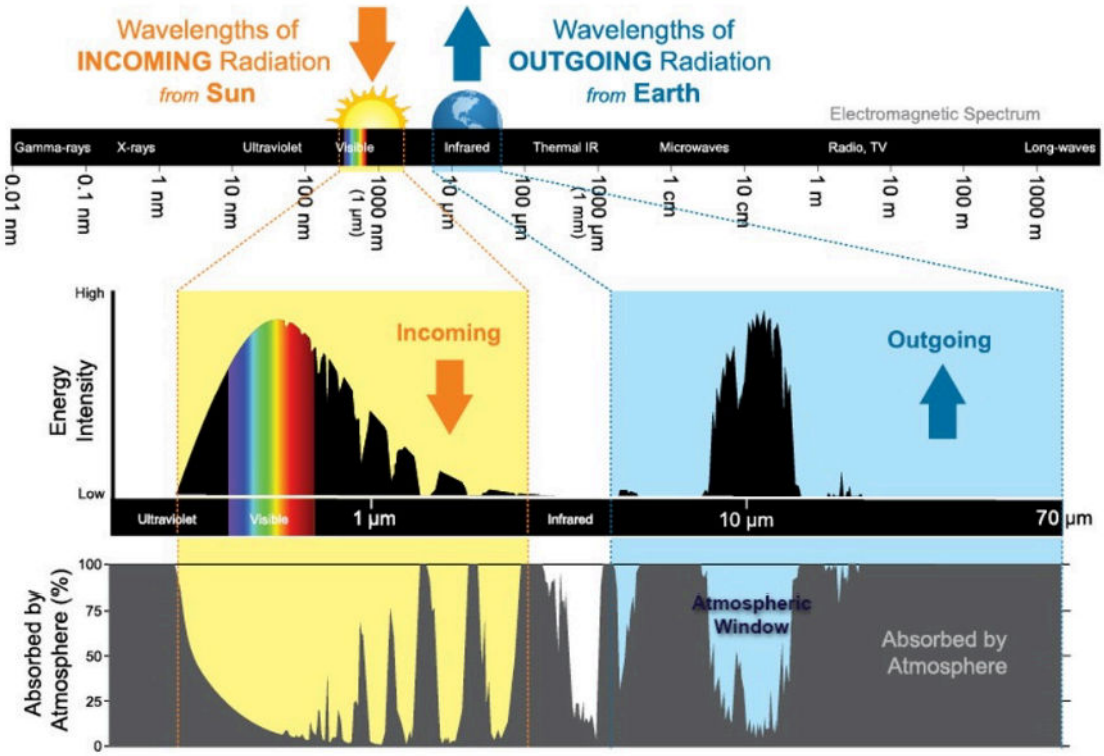


**Figure 2.1.:** The incoming radiation of the Sun is in the visible to near infrared range while the outgoing longwave radiation is in the infrared range. Depending on the wavelength, radiation is absorbed in the atmosphere by clouds, GHGs and aerosols. Modified from (National Oceanic and Atmospheric Administration 2023).

## 2.3. Atmospheric Radiation

Atmospheric radiation is divided into longwave and shortwave radiation because the assumption that the overlap is negligible is valid in the Earth's atmosphere, see Figure 2.1. Shortwave radiation originates from the Sun, describing incoming radiation and spanning short wavelengths below 4 $\mu m$. The shortwave spectrum includes ultra violet, visible and near-infrared radiation. In contrast, longwave radiation is emitted by the surface of the Earth and the atmosphere itself such as GHGs and clouds. Longwave radiation has wavelengths larger than 4 $\mu m$ and spans the infrared range of the electromagnetic spectrum. The spectrum has two windows with small absorption for both spectra which can be used for observations. The visible window in the shortwave spectrum can be used for ground based measurements while the atmospheric window in the longwave spectrum can be used for satellite observations.

The global energy balance can be simply derived from a geometrical point of view. Due to the distance between Sun and Earth, the radiation from the Sun can be approximated as parallel beams illuminating the Earth. The intensity is described by the solar constant $S$ which is in fact not a constant but varies according to the 11-year solar cycle. The solar constant is used to define the incoming shortwave radiation at the top of the atmosphere. In ESMs, the solar constant is also scaled by the distance between the Earth and Sun which changes throughout the year due to the elliptic orbit. The effective illuminated area corresponds to

a circle with the radius of the Earth $r_E$. A part of the radiation from the Sun is reflected by the Earth which is described by the planetary albedo $\alpha$. The emission of the Earth can be approximated by the Stefan-Boltzmann law Equation (2.3). The Earth emits in all directions, so the emission is scaled by the surface area of the Earth. Then, the radiation that the Earth receives and emits can be set into balance

$$S\ \pi r_E^2\ (1-\alpha) = 4\pi r_E^2\ \sigma T_{Earth}^4. \tag{2.9}$$

This equation can be rearranged to calculate the global mean surface temperature of the Earth. Considering the Earth as a black body, the solar constant as 1360 $W/m^2$ and an albedo of 0.3, this would yield a surface temperature of 255 $K$ or $-18°C$. However, this rather simple derivation does not reflect reality because it neglects the greenhouse effect.

The greenhouse effect describes the effect of GHGs which act like an absorbing layer around the Earth. The simple model in Equation (2.9) can be extended by a factor $\beta$ accounting for the amount of radiation that is absorbed by the GHGs instead of emitted to space. The factor can be added to the right hand side of Equation (2.9) yielding

$$S\ \pi r_E^2\ (1-\alpha) = 4\pi r_E^2\ \sigma T_{Earth}^4\ \beta. \tag{2.10}$$

With $\beta = 0.62$, the average surface temperature of the Earth is 287 $K$ or $14°C$. A more detailed look at the absorbed radiation in the atmosphere is shown in Figure 2.1. The figure shows the relative amount of absorbed radiation depending on the wavelength which can be very complex due the presence of GHGs.

A more detailed look at the global energy budget including the involved processes is shown in Figure 2.2. The total incoming and outgoing radiation is approximately in balance with a small imbalance of 0.7 $W/m^2$ of more incoming radiation leading to a net warming of the Earth system. Depending on local atmospheric and surface conditions, incoming radiation is absorbed or reflected. Besides GHGs and aerosols, clouds play an important role as they reflect incoming radiation but also absorb and emit longwave radiation.

Clouds can have a positive and negative feedback on the climate e.g., high optically thin clouds absorb a large portion of outgoing longwave radiation while only reflecting a small amount of incoming shortwave radiation. Therefore, high, optically thin clouds have a net warming effect and therefore a positive climate effect on the mean surface temperature. In contrast, optically thick low clouds that are usually not much colder (around 10 $K$) than the surface beneath them reflect a large portion of incoming solar radiation and emit outgoing longwave radiation due to their relatively high temperature compared to high, optically thin clouds. Therefore, optically thick low clouds are leading to cooling and thus have a negative climate effect on the mean surface temperature (Wallace and Hobbs 2006). The total cloud feedback on climate is uncertain as both effects depend on the areal coverage which can change due to climate change, and clouds strongly contribute to the overall uncertainty of climate models (Forster et al. 2021b). Comparing multi-model means of Coupled Model Intercomparison Project Phase 6 (CMIP6) models with observations for 20-year annual averages, the

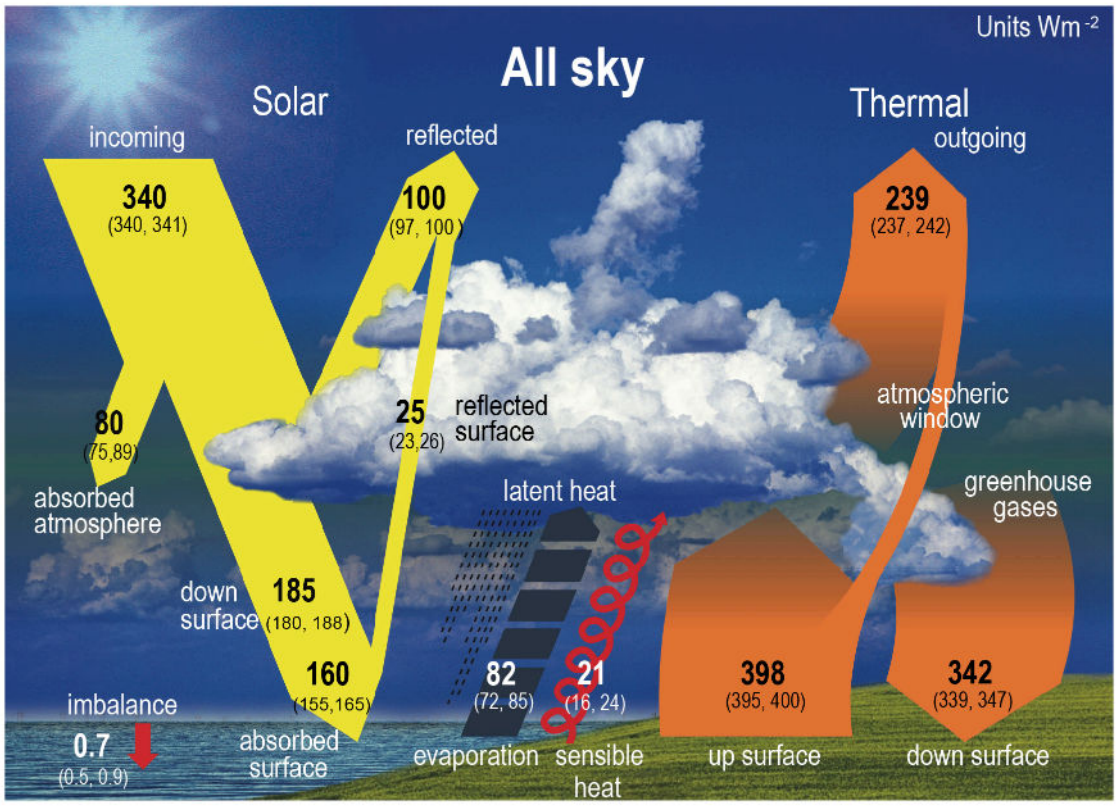


**Figure 2.2.:** The energy budget of incoming and outgoing radiation is approximately in balance. Modified from Figure 7.2 in IPCC, 2021: Chapter 7 (Forster et al. 2021a)

shortwave cloud radiative effect has a bias of 1.490 $W/m^2$ and a global root mean squared distance of 7.468 $W/m^2$ with local deviations up to 30 $W/m^2$ (Lauer et al. 2023). The longwave cloud radiative effect has a bias of $-2.837$ $W/m^2$ and a global root mean squared distance of 4.689 $W/m^2$ with local deviations of $\pm 10$ $W/m^2$. While some biases and uncertainties have been reduced over time, i.e., the bias in total cloud cover was reduced in CMIP6 compared to Coupled Model Intercomparison Project Phase 5 (CMIP5) from $-7\%$ to $-1\%$ (Lauer et al. 2023), other biases and uncertainties remain, i.e., the underestimation of clouds in stratocumulus regions which is related to large biases in the shortwave cloud radiative effect (Lauer et al. 2023).

For longwave radiation, scattering is usually neglected because the wavelength is larger than the typical particle size in the atmosphere. Therefore, with $\alpha_s = 0$ and Equation (2.1), Equation (2.8) reduces to

$$\frac{dI_\lambda}{ds} = \left(\frac{dI_\lambda}{ds}\right)_{emiss} + \left(\frac{dI_\lambda}{ds}\right)_{ext} = \alpha_a(B(\lambda, T) - I_\lambda), \tag{2.11}$$

which is also known as *Schwarzschild equation*. The Schwarzschild equation explains that the change in monochromatic intensity $I_\lambda$ for longwave radiation depends only on the absorption and emission along the path $s$.

The absorption component of Equation (2.6) can also be expressed as

$$\frac{dI_\lambda}{ds} = -I_\lambda \rho r k_\lambda, \tag{2.12}$$

with air density $\rho$, mass mixing ratio of the absorbing gas $r$ and the mass absorption coefficient $k_\lambda$. Substituting the path $s$ with the height $z$ with $ds = sec(\theta)dz$ with zenith angle $\theta$ and integrating from the top of the atmosphere ($z_\infty$) to a given height $z$ yields

$$\begin{aligned} I_\lambda &= I_{\lambda\infty} e^{-\tau_\lambda sec(\theta)} = I_{\lambda\infty} T_\lambda \\ \tau &= \int_z^\infty k_\lambda \rho r dz \\ T_\lambda &= e^{-\tau_\lambda sec(\theta)}, \end{aligned} \tag{2.13}$$

where $T_\lambda$ is the *transmissivity* and $\tau_\lambda$ is the *optical depth* or *optical thickness* of a layer. This set of equations is known as the *Beer-Bouguer-Lambert law* which describes the exponential reduction of radiative intensity when passing through a medium (Wallace and Hobbs 2006). Without scattering, the monochromatic absorptivity can be inferred from the transmissivity

$$\alpha_{a,\lambda} = 1 - T_\lambda. \tag{2.14}$$

These equations are the theoretical building blocks to understand radiative transfer in the atmosphere. The next section focuses on how radiative transfer is calculated in ESMs.

## 2.4. From Modelling Concepts Towards Radiation Parameterizations in ESMs

Accurate calculations of radiative transfer can be done using line-by-line radiative transfer models that calculate the transfer for almost every line (wavelength or wavenumber). It is necessary to consider so many lines because of the complex absorption spectrum (Figure 2.1). This requires immense computing time which is not feasible when performing climate projections for multiple decades. In practice, modelling radiative transfer in ESMs involves the use of several simplifications and approximations. The separation of longwave and shortwave radiation and neglecting of scattering for longwave radiation were already discussed. Here, a few more concepts are presented to make radiative transfer calculations feasible for decade long climate simulations. The final representation that approximates radiative transfer in ESMs is called radiation parameterization.

### 2.4.1. Plane Parallel Approximation

The plane parallel approximation is widely used in radiation parameterizations. It means that the layers in an atmospheric column are parallel, and the variables like temperature and flux density are height-dependent and horizontally homogeneous distributed in a column. This also means that 3D cloud radiative effects are neglected (Wallace and Hobbs 2006). 3D

cloud radiative effects describe horizontal effects such as the horizontal emission or reflection of radiation by clouds. Considering these effects requires another parameterization but is usually not included in climate models (Hogan and Shonk 2013). Then, the commonly used two-stream approximation is used in ESMs to calculate radiative fluxes in each layer. Here, only upward and downward radiative fluxes are calculated (Meador and Weaver 1980).

### 2.4.2. Correlated k-distribution

To calculate the flux density in a grid cell, it is necessary to integrate over the wavelengths because the absorption coefficient $k$ depends on wavelength, pressure, and temperature. The integrand includes the Planck function (Equation (2.2)) and transmissivity (Equation (2.13)). To approximate the integral, the spectrum is divided into bands meaning a smaller range of wavelength compared to the full spectrum of longwave and shortwave radiation. The width of these bands is defined such that the Planck function can be assumed to be constant while the transmissivity strongly depends on the wavelength due to the absorption coefficient. The correlated k-distribution (CKD) method (Fu and Liou 1992) makes the assumption that only the distribution of absorption coefficients is important (k-distribution). Therefore, the method introduces a transformation into the so-called $g$ space which describes the cumulative probability distribution of the absorption coefficients $k$. In other words, the $g$ space describes a reordering of the absorption coefficient such that $g$ describes a smooth function. Then, the smooth function is discretized for integration. The discretization points are called $g$ points and vary per band as the k-distribution varies per band.

The essence of the CKD method describes the transformation into $g$ space for an easier integration for a band

$$T_{\tilde{\nu}} = \int_{\Delta\nu} exp\Big[-\int_{z_1}^{z_2} k(\nu,p,T)\rho dz\Big]\frac{d\nu}{\Delta\nu} = \int_0^1 exp\Big[-\int_{z_1}^{z_2} k(g,p,T)\rho dz\Big]dg \tag{2.15}$$

with the wavenumber $\nu$, pressure $p$, temperature $T$, density $\rho$ in a layer between height $z_1$ and $z_2$ (Fu and Liou 1992). The absorption coefficient depends on $g$ but also on pressure and temperature. The term *correlated* in CKD (Fu and Liou 1992) denotes the assumption that the absorption coefficient at different pressure and temperature points can be calculated from a reference state because the coefficients are correlated.

### 2.4.3. Clouds

In climate models, the common horizontal grid resolution is about $100 - 200\ km$ (Chen et al. 2021). Obviously, clouds are smaller than that and are considered sub-grid scale. However, radiation schemes usually assume that physical quantities are horizontally homogeneous distributed. There are different methods to include the sub-grid scale cloud statistics into radiation parameterizations. One method is called Independent Column Approach (ICA) (Cahalan et al. 1994) where a column is divided into $n$ sub-columns that are either clear-sky or full cloudy. The disadvantage of this method is that the radiative transfer has be calculated

for all $n$ sub-columns which is usually in the order of 100 (Hogan and Matricardi 2020). A fast alternative of this approach, called Monte Carlo Independent Column Approach (McICA) (Pincus et al. 2003), applies the cloud statistics to the $g$ points instead. For a number of $g$ points, the assumption is that these points see cloudy conditions and for the rest the assumption is clear-sky.

### 2.4.4. RTE+RRTMGP

The radiation parameterization Radiative Transfer for Energetics + Rapid Radiative Transfer Model for GCMs - Parallel (RTE+RRTMGP) (Pincus et al. 2019) provides a flexible framework for computing radiation in climate and weather models. It divides the radiation into two distinct problems which are defining gas optics based on Rapid Radiative Transfer Model for GCMs - Parallel (RRTMGP) (Mlawer et al. 1997a) and solving the Radiative Transfer for Energetics (RTE).

RRTMGP specifies only the gas optics based on volume mixing ratios for GHGs. Cloud and aerosol optics are specified outside of RRTMGP, but are essential for the following radiative transfer calculation. RTE then combines the gas optics with cloud and aerosol optics to get the optical properties for calculating radiative fluxes. There is no explicit treatment of clouds or sub-grid scale cloudiness as the RTE solver only sees optical properties (Pincus et al. 2019). Therefore, cloud related radiative effects have to be approximated via e.g., McICA externally in the ESM or NWP model.

RTE+RRTMGP is based on CKD and choosing the $g$-points is a tuning problem. RTE+RRTMGP provides a set of 16 $g$-points per band which leads to a total number of g-points of 256 for shortwave radiation and 244 for longwave radiation. Reducing the number of $g$-points would increase computational efficiency but might also increase errors. However, an optimized selection of fewer g-points can keep errors in the same range depending on the application, e.g., longwave clear-sky calculations (Czarnecki et al. 2023).

The modern radiation parameterization RTE+RRTMGP is used here as a baseline to train the ML-based radiation scheme. It is used in ESMs such as the ICON model and is optimized to run efficiently on Central Processing Units (CPUs) and GPUs.

## 2.5. Machine Learning

In this section, ML basics including typical terms for the training procedure are described. ML gained more recognition in the recent decades especially with the rise of new hardware like GPUs that enabled the efficient use of large datasets. Besides recent developments in natural language processing, image generation and speech recognition, ML received more attention in the science community i.e., boosting the development of parameterizations for climate models (Eyring et al. 2024).

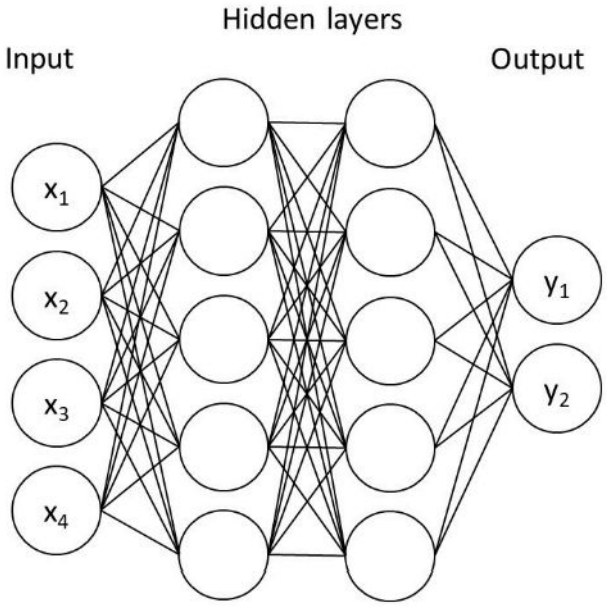


**Figure 2.3.:** Sketch of a neural network with two hidden layers. The input $x$ has four nodes, the hidden layers have 5 nodes each and the output $y$ has two nodes.

### 2.5.1. Basics

Supervised machine learning describes a statistical method to approximate a function $f$ with Neural Networks (NNs) between an input $\vec{x}$ and output $\vec{y}$ given some trainable parameters $\theta$

$$\vec{y} \approx f(\vec{x}; \theta). \tag{2.16}$$

The analytical form of the function $f$ is either unknown or complex, therefore NNs are used. NNs are graphs that are defined by nodes and edges that connect the input and output. The basic NNs is called Feedforward Neural Network (FNN) which uses hidden layers connecting input and output (Figure 2.3). A hidden layer is defined by a number of nodes where every node $x_i$ from the previous layer e.g., the input layer, is connected to every node $x'_j$ from the hidden layer. The functional form is

$$x'_j = w_{i,j} * x_i + b_j, \tag{2.17}$$

where the weights $w_{i,j}$ and biases $b_j$ are trainable parameters. The number of nodes in a layer defines the width of a NN while the number of layers defines the depth, therefore often the term Deep Neural Network (DNN) is used.

The universal approximation theorem says that every function $f$ can be approximated by a FNN with one hidden layer if a sufficient number of nodes is given (Hornik et al. 1989). For most real-world cases, this is not feasible. Therefore, complexity is added to NNs. One way is to add more layers. But, only adding more layers is not enough as it is still a linear combination that can be reduced to the form of Equation (2.17). A non-linear activation function $\sigma$ is

added after every layer. Typical activation functions are Rectified Linear Unit (ReLU), *tangent hyperbolic* and *sigmoid*. The functional form of the NN in Figure 2.3 then becomes

$$\begin{aligned} \vec{x_{h1}} &= \sigma_1(\, \boldsymbol{W_1}(\ \vec{x}\ ) + \vec{b_1}) \\ \vec{x_{h2}} &= \sigma_2(\, \boldsymbol{W_2}(\, \vec{x_{h1}}) + \vec{b_2}) \\ \vec{y} &= \sigma_3(\, \boldsymbol{W_3}(\, \vec{x_{h2}}) + \vec{b_3}). \end{aligned} \tag{2.18}$$

When these NNs contain many layers, they are referred to as deep. Hence, sometimes the term *deep learning* is used.

### 2.5.2. Training

After defining a NN architecture as in Equation (2.18), the next step is to define the training process i.e., how to find the optimal weights $\boldsymbol{W}$ and biases $\vec{b}$. The first step is to define an objective function aiming to optimize the trainable parameters. It is often called cost or loss function $\mathcal{L}$ (Goodfellow et al. 2016). Depending on the definition, the goal is to minimize or maximize the cost. A typical loss function is the Mean Squared Error (MSE) and the objective is to minimize the MSE between expected output $\vec{y}_{True}$ and predicted output $\vec{y}_{Pred}$.

To update the trainable parameters $\theta$, the partial derivative of the loss function $\partial\mathcal{L}/\partial\theta$ is computed for every trainable parameter in the network. An efficient algorithm for this purpose is called *backpropagation* (Goodfellow et al. 2016) where the gradients are first calculated for the last layer and then via chain rule propagated through the NN. Then, the gradients are used to update the trainable parameters for the next iteration which is called *gradient descent*

$$\theta' = \theta - \alpha \frac{\partial \mathcal{L}}{\partial \theta}, \tag{2.19}$$

with the learning rate $\alpha$ (Goodfellow et al. 2016). The learning rate is a tunable hyperparameter for optimal performance (Goodfellow et al. 2016). If the learning rate is too large, the optimal set of trainable parameters will not be reached because the updated parameters will oscillate around the optimal parameters. In contrast, a small learning rate could lead to only a small reduction of the loss or the model will get stuck in a local minimum. A common technique to avoid this is to use a learning rate schedule that changes the learning rate and helps approaching an optimal set of trainable parameters.

The dataset is divided into a training, validation and test set. The training set usually holds the largest fraction of all samples and is used to update the trainable parameters. The validation set is used to tune hyperparameters and monitor how the NN would perform on unseen data during training. The final evaluation is then done on the test set.

To decide when a training is finished, the training and validation losses are monitored. If both losses converge, then the training is finished. If the validation loss starts to diverge from the training loss, the NN starts to overfit to the training data. The consequence is that the trained NN will perform worse on unseen data e.g., the test set.

To avoid overfitting, there are techniques that are summarized under the term regularization, where some are explicit and some are implicit. An implicit option is to use *early stopping*,

it basically stops the training when the validation loss starts increasing. An explicit option is to use L1 or L2 regularization (Goodfellow et al. 2016), where the absolute or squared norm, respectively, of the weights is added to the loss with a scaling factor $\lambda$. The idea of penalizing the weights is that they should be closer to zero so that the NNs are not too sensitive to small changes. L2 regularization is often called *weight decay* (Goodfellow et al. 2016).

The trainable parameters are usually updated in *batches* which is a collection of samples. Then, the gradients are averaged over the samples in a batch leading to more stable parameter updates and faster convergence. This way of updating the trainable parameters is also called *Stochastic Gradient Descent (SGD)*. Another method to optimize the trainable parameters is the optimizer Adam (Kingma and Ba 2017), which uses SGD, adaptive learning rates and momentum. Momentum means that the optimizer makes use of past gradients and adaptive learning rate describes the idea that the learning rate is adapted throughout the training, often separately for different parameters (Goodfellow et al. 2016). Meanwhile, there exist many different optimizers and different versions of Adam such as AdamW (Loshchilov and Hutter 2017). AdamW decouples the weight decay from the loss function as it is not equivalent to L2 regularization for adaptive learning rates (Loshchilov and Hutter 2017).

## 2.6. Previous Work on ML-based Radiation

The development of ML-based radiation schemes is an active field of research because it has the potential to significantly reduce computational cost while maintaining accuracy. The first developments of such schemes started over two decades ago. Nevertheless, there is still no ML-based radiation scheme used operationally in ESMs. In this section, a literature review of ML-based radiation schemes is presented.

When referring to ML-based parameterizations, often the terms *offline* and *online* are used. *Offline* means the ML-based parameterization has been trained or evaluated on model output, whereas *online* means the ML-based parameterization is coupled to an ESM like ICON. The usual workflow is to train and evaluate an ML-based parameterization first offline. If the performance is good, which is defined by different metrics, the ML-based parameterization can be implemented into a model and tested online.

### 2.6.1. Offline Approaches

This section has already been published in Hafner et al. (2025a). As mentioned in Section 1.3, I created all text, that is presented from this publication.

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 for the clear-sky component and $2 \times M$ NNs for the cloudy-sky component, where $M$ is the number of cloudy layers. While 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 of 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 and 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 neural networks, which were tested online in CAM2 and GFS (Krasnopolsky et al. 2005; Krasnopolsky et al. 2008; Krasnopolsky 2012). About a decade later, the same approach was used in a modern ESM, for example, for numerical weather prediction in Weather Research and Forcasting model (WRF) (Roh and Song 2020; Song and Roh 2021) and a 6-month simulation in GFSv16 (Belochitski and 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; Lagerquist et al. 2023; Ukkonen 2022; Yao et al. 2023). Some of these studies compared different architectures and found that bidirectional recurrent neural networks performed better than fully connected neural networks because recurrent neural networks can better handle the autocorrelation in the vertical profile. Despite good overall offline performance, the remaining question is why the neural networks 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.

### 2.6.2. Online Approaches

This section has already been published in Hafner et al. (2025b). As mentioned in Section 1.3, I created all text, that is presented from this publication.

Only a few studies focused on emulating one component of radiation and testing it online as the speed-up potential is smaller. Emulating gas optics (Ukkonen and Hogan 2023) or an ML-based radiation approach without emulating cloud optics (Pal et al. 2019) has been shown successful in one-year long hybrid atmosphere-only simulations. The advantage is that it is more flexible and can be used with other modules but the speed-up potential is limited.

Emulating full radiation with a fully connected neural network has been successful for decade long simulations with the Community Atmosphere Model 2 (CAM-2) (Krasnopolsky et al. 2005; Krasnopolsky et al. 2008). The same approach was used in GCMs like GFSv16 (Belochitski and Krasnopolsky 2021) in a continuous simulation of 7 months and WRF (Roh and Song 2020) for numerical weather prediction in a 6-hour simulation. More complex neural networks for emulating full radiation (Lagerquist et al. 2023; Ukkonen 2022; Yao et al. 2023) have been tested online at most in a 3-day numerical weather prediction context (Zhong et al. 2023).

### 2.6.3. Further Work

This section has already been published in Hafner et al. (2025c). As mentioned in Section 1.3, I created all text, that is presented from this publication.

As more high-resolution GSRM data become available, they offer increasing opportunities to enhance coarse-scale ESMs with machine learning. High-resolution simulations have been applied to nudge coarse-scale models toward fine-scale states (Bretherton et al. 2022), to learn subgrid tendencies directly (Busecke et al. 2025; Heuer et al. 2024), to infer subgrid effects from coarse-scale states (Grundner et al. 2022; Shamekh et al. 2023) and to learn all physics parameterizations (Watt-Meyer et al. 2024). Beyond these applications, high-resolution model simulations enable new strategies for representing fractional cloudiness and cloud overlap in coarse-scale ESMs using ML. Specifically, GSRM output can be coarse-grained to the resolution of the target ESM, and an NN can then be trained to learn the subgrid-scale distribution of clouds from the underlying statistics. For instance, Henn et al. (2024) predicted coarse-grained cloud fields to reduce radiative biases; however, the cloud overlap assumption in the radiation scheme remains unchanged, limiting improvements. Leveraging ML to represent subgrid-scale cloud effects on radiation thus provide a promising pathway toward accurate radiation schemes in ESMs for climate projections.

A separation between clear-sky and cloudy radiative fluxes was previously proposed by Chevallier et al. (1998). Moreover, Meyer et al. (2022) focused on learning 3D cloud radiative effects. However, the explicit separation of cloud radiative impacts from all-sky radiation, especially including subgrid-scale effects, remains largely unexplored.

# 3. Data and Methods

In this chapter, I will describe the ICON model, the climate model for which I developed an ML-based radiation emulator. The ICON model is used to generate training data but also to test the ML-based emulator online. All types of simulations will be described here. Also, the specific ML methods are described in more detail.

## 3.1. The ICON model

The ICOsahedral Nonhydrostatic (ICON)-Earth System Model (ESM) is the numerical weather prediction and climate model mainly developed by the German Weather Service (DWD) and the Max Planck Institute for Meteorology (MPI-M). It is designed to run globally or regionally on different spatial and temporal scales. ICON can be used with a coupled ocean, land and atmosphere model (Jungclaus et al. 2022). The focus of this thesis is the atmospheric component or ICON-A as described in Giorgetta et al. (2018).

As described in Section 2.2, an ESM is based on the primitive equations. ICON uses a non-hydrostatic dynamical core that solves the primitive equations numerically. The numerical method is based on finite volume which means that the atmosphere is discretized into grid boxes of finite volume. Everything that cannot be resolved in a grid box, e.g., convection and turbulence, is parameterized which is also referred to as *physics parameterizations*.

The ICON grid is defined on an icosahedron, a twenty-sided object, where every side is a triangle, see Figure 3.1. To refine the grid to higher resolutions, every triangle is divided into finer equally-sized triangles. The refinement is defined by two parameters $n$ and $k$. Every edge can be divided into $n$ parts and further refined by $k$ subsequent bisections. The total number of cells $n_c$ is given as:

$$\begin{aligned} n^2 &= \sum_{i=1}^{n}(2i-1) && \text{, cells per triangle for } n \text{ edge divisions} \\ 4^k & && \text{, cells per triangle for } k \text{ bisections} \\ n_c &= 20n^2 4^k && \text{, total cells.} \end{aligned} \tag{3.1}$$

Figure 3.1 shows how this refinement process looks like. The grid-refinement can be global or regional depending on the application i.e., global modelling or regional modelling in the *limited area mode*.

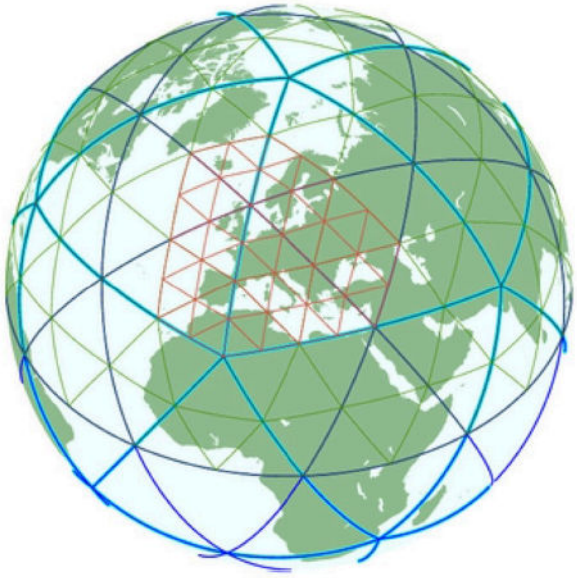

**Figure 3.1.:** The horizontal ICON grid is based on an icosahedron. Each triangle can be refined by subsequent bisections or edge divisions. Reproduced without changes from Figure 1b) in Jungclaus et al. (2022) in line with creative commons license CC BY-NC-ND 4.0.

The ICON grid size is then given as R$n$B$k$. The effective horizontal grid resolution $\bar{\Delta x}$ can be derived as

$$\Delta\bar{x} = \sqrt{\frac{A_E}{n_c}} = \sqrt{\frac{\pi}{5}}\frac{r_E}{2^k n} \tag{3.2}$$

using the surface area of the Earth $A_E$ which can be calculated using the Earth's radius $r_E$. A common resolution for climate simulations with ICON is R2B5 which corresponds to an effective grid resolution of 78.91 $km$. This resolution is used in this thesis for the coarse scale simulations. A common high-resolution grid is R2B9 with an effective horizontal grid resolution of 4.93 $km$ which is used for fine scale simulations.

Since ICON uses nonhydrostatic equations, it does not guarantee that the pressure is decreasing monotonically with height (Prill et al. 2024). Therefore, the vertical grid is defined as terrain following coordinates (Figure 3.2). The layers can be described by *half levels* and *full levels* where the half levels are defined as the layer boundaries and full levels as layer centers. For example, the radiative fluxes are defined on half levels as they describe the transfer of radiation between layers. Though, tendencies or heating rates are defined on full levels as they described the change of a property i.e., temperature, in the layer.

As mentioned above, the ICON model can be used for NWP or climate projection. The difference is that NWP depends mainly on the initial conditions ("What is the temperature in Bremen in one hour knowing the temperature at the moment?") and climate projection depends on the boundary conditions ("How do the $CO_2$ concentrations change over the next 20 years?"). Another difference within the ICON model is the use of different physics parameterizations. The atmospheric component of the climate projection mode is also known as ICON-A (Giorgetta et al. 2018) which is based on ECHAM physics. ECHAM is developed

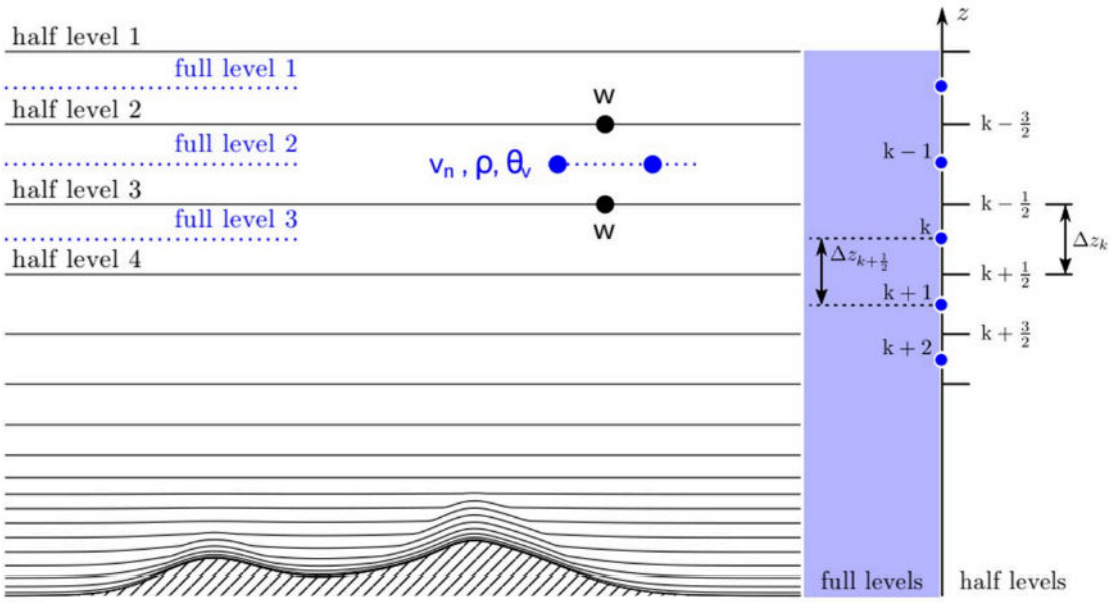


**Figure 3.2.:** Illustration of ICON's vertical levels. With num_lev layers, there are num_lev +1 so-called *half levels*. The half levels $k - 1/2, k + 1/2$ enclose layer $k$ at the centers of which are the corresponding full levels $k$, for $k = 1, ...,$ num_lev. Layer 1 is at the top of the atmosphere and layer $n$ at the bottom of the atmosphere. Half level num_lev +1 coincides with the Earth's surface. Reproduced without changes, including the caption, from (Prill et al. 2024) in line with creative commons license CC BY-NC-ND 4.0.

by MPI-M based on a model from European Center for Medium Range Weather Forcasts (ECMWF). Therefore, the name mixes EC from ECMWF and HAM for Hamburg (ECHAM) (Stevens et al. 2013). The NWP branch is mainly developed by DWD. The differences are summarized in Table 3.1.

In the ICON model, it is easy to change the parameterization for different physical processes. Therefore, Table 3.1 has multiple options for some processes. For NWP, radiation can be parameterized with RRTM (Barker et al. 2003; Mlawer et al. 1997a) or ecRad (Hogan and Bozzo 2018) and for climate projection with PSrad (Pincus and Stevens 2013) or RTE+RRTMGP (Pincus et al. 2019) which is the successor of PSrad. Both latter schemes have the option to run on GPUs. For cloud cover, the *all-or-nothing* scheme (0/1) is used or a diagnostic probability density function based on the turbulent variability of water (Prill et al. 2024). ICON-A uses the Sundqvist scheme which is based on relative humidity (Sundqvist et al. 1989). In both modes, the operational microphysics scheme is a single-moment scheme. For ICON-A, it is based on Lohmann and Roeckner (1996) and for NWP, it is based on Doms et al. (2011) and Seifert (2008) with the option use a double-moment scheme (Seifert and Beheng 2005) or warm spectral bin microphysics (Khain and Sednev 1996; Khain et al. 2004). The vertical diffusion or turbulence parameterization is based on prognostic turbulent kinetic energy (Raschendorfer 2001) for NWP and total turbulent energy (Mauritsen et al. 2007) for ICON-A which additionally considers turbulent potential energy. The convection parameterizations are all based on mass flux schemes (Bechtold et al. 2008; Nordeng 1994; Tiedtke 1989). The subgrid scale orography drag

**Table 3.1.:** Overview of the parameterizations used in ICON for NWP and climate projection (Giorgetta et al. 2018; Giorgetta et al. 2022; Prill et al. 2024).

| **Process** | **NWP** | **Climate Projection/ ICON-A** |
|---|---|---|
| Radiation | RRTM (Barker et al. 2003; Mlawer et al. 1997a), ecRad (Hogan and Bozzo 2018) | PSrad (Pincus and Stevens 2013) (old configuration), RTE+RRTMGP (Pincus et al. 2019) |
| Cloud cover | 0/1, Diagnostic probability density function (Prill et al. 2024) | Sundqvist et al. (1989) |
| Microphysics | Single-moment scheme (Doms et al. 2011; Seifert 2008) Double-moment scheme (Seifert and Beheng 2005) Warm Spectral Bin Microphysics (Khain and Sednev 1996; Khain et al. 2004) | Lohmann and Roeckner (1996) |
| Vertical Diffusion | Prognostic turbulent kinetic energy (Raschendorfer 2001) | Total turbulent energy (Mauritsen et al. 2007) |
| Cumulus Convection | Bechtold et al. (2008), Tiedtke (1989) | Nordeng (1994) |
| Sub-grid scale orography drag | Lott and Miller (1997) | Lott (1999) |
| Non-orographic gravity wave drag | Orr et al. (2010) | Hines (1997) |
| Land | TERRA (Schulz et al. 2016) | JSBACH4-lite (Reick et al. 2021) |

parameterization accounts for orography induced gravity waves (Lott and Miller 1997; Lott 1999). In contrast, the non-orographic gravity wave drag parameterization determines effects induced by imbalances in the atmospheric flow i.e., convection, jet streams and fronts (Hines 1997; Orr et al. 2010). The land schemes are different because of the coupling to the vertical diffusion scheme. Therefore, the land model TERRA (Schulz et al. 2016) is used for NWP and a simplified version of the land model JSBACH (Reick et al. 2021) is used in ICON-A (Giorgetta et al. 2018). The land model is used for heat exchange with the land surface, whereas SSTs are prescribed.

ICON-A as in Giorgetta et al. (2018) was further developed to run on higher horizontal resolutions and on GPUs. The setup was presented in Giorgetta et al. (2022) to mainly study the Quasi-Biennial Oscillation (QBO). Therefore, the simulations are referred to as Quasi-Biennial Oscillation In a Changing Climate (QUBICC). The horizontal grid is R2B9 which corresponds to an effective horizontal resolution of 4.93 $km$ (Equation (3.2)). The aim of a higher resolution simulation is that more physical processes are resolved requiring less physics parameterization. The parameterizations are as for the climate projection simulations in Table 3.1 unless otherwise stated in the following. The radiation scheme is RTE+RRTMGP. Due to the high horizontal resolution, the cloud cover scheme is *all-or-nothing*. The cloud microphysics scheme is a single moment scheme (Doms et al. 2011). The parameterizations

for convection, subgrid scale orography drag and non-orographic gravity wave drag are not used because these processes are considered as resolved. ICON-A is run with 47 vertical levels that cover 83 $km$ of the atmosphere. QUBICC covers the same height with 191 vertical levels. GHGs including CO2, CH4, N2O, O2, CFC-11, CFC-12, are well mixed and prescribed with annual global mean volume mixing ratios. Ozone is spatially resolved with monthly mean values. Interactive aerosols are not included for QUBICC as it is still an active area of research (Weiss et al. 2025). Therefore, the ICON-A simulations in this thesis also do not include interactive nor radiatively active aerosols for better comparability.

ICON uses different time steps describing the updates of the dynamical core, slow and fast physics. The fast physics time step corresponds to the model time step $\Delta t$. The time step of the dynamical core $\Delta\tau$ is defined by the model time step and a number of sub time steps `ndyn_substep` for the dynamical core which is usually 5. The slow physics time step depends on the process and is usually multiples of the model time step. For the radiation parameterization, it is usually $1-2$ $h$ (Giorgetta et al. 2018). The maximum possible time step is determined by the fastest wave that should be resolved, which is the sound wave, and the effective grid resolution. It can be approximated as

$$\Delta t = 9 * 10^{-3} \overline{\Delta x} \frac{s}{m}, \tag{3.3}$$

with the effective grid resolution $\overline{\Delta x}$ from Equation (3.2) (Prill et al. 2024). For ICON-A with an R2B5 grid, the maximum time step is 12 minutes and for QUBICC with an R2B9 grid, it is 45 seconds.

The ICON model is constantly under development. The versions used in this thesis are v2.6.4 for ICON-A and icon-2024.10 for QUBICC. In Giorgetta et al. (2018), the radiation scheme PSrad (Pincus and Stevens 2013) was used which was later replaced with the successor RTE+RRTMGP (Pincus et al. 2019) as it was designed to run also on GPUs. Both schemes are available for ICON-A in v2.6.4 and for newer version only RTE+RRTMGP. Hence, the radiation scheme RTE+RRTMGP was used in all simulations.

## 3.2. The Long Short-Term Memory architecture

Learning a complex process such as a parameterization in the ICON model, requires careful design of the NN architecture. Bigger NNs, i.e., more layers and more nodes, can learn more complex relationships between input and output variables. Another option to make NNs more complex is to use more complex architectures. Depending on the problem, there are different suitable architectures, e.g., convolutions for image recognition, graph neural networks for unstructured data, LSTMs for sequence data, and many more. There exist different versions for LSTMs. Here, the version that is used in PyTorch (Ansel et al. 2024) will be explained which is based on Sak et al. (2014). In an LSTM architecture, the LSTM cell looks at one element of the sequence at a time. Within the LSTM cell, the input but also some information from the previous sequence steps are processed in a specific way as follows (Figure 3.3). The LSTM cell is designed in gates that all have a specific task. The gates are defined by a set of equations:

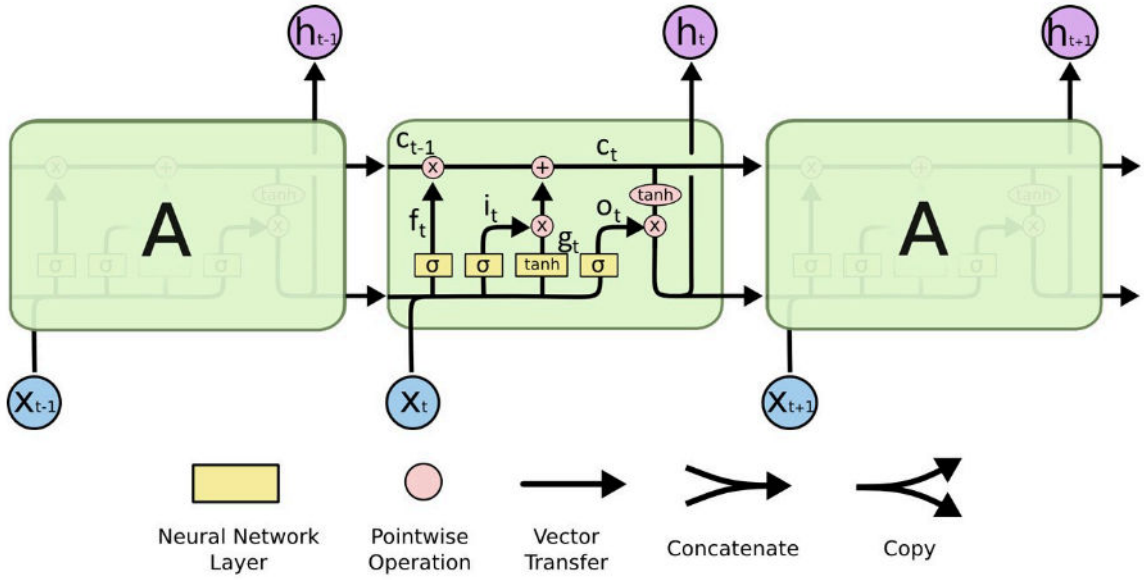


**Figure 3.3.:** Structure of a Long Short-Term Memory (LSTM) cell. The cell at sequence step $t$ gets input from the LSTM cell of the previous step $t-1$ and forwards its output to the LSTM cell that processes the next sequence step $t+1$. Here, A stand for an LSTM cell. Modified with permission from Olah (2015).

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \circ c_{t-1} + i_t \circ g_t \\
h_t &= o_t \circ tanh(c_t),
\end{aligned}
\tag{3.4}
$$

where all weights $W$ and biases $b$ are trainable parameters, similar to the FNN. $x_t$ is the input, or the element at sequence step $t$. $c_t$ is the cell state. $h_{t-1}$ is the hidden state from the previous element in the sequence and $h_t$ is the hidden state from current sequence step $t$ which is also the output of the LSTM cell. To calculate the output of the LSTM cell, the input $x_t$, cell state $c_t$ and also the previous hidden state $h_{t-1}$ have to be passed through the gates. First, the $x_t$ and $h_{t-1}$ are passed through the input gate $i_t$ to extract important information from the input. Then, the forget gate $f_t$ decides which part should be discarded. $g_t$ is the cell gate, an intermediate step to calculate the new cell state $c_t$. The cell state which is the core of the memory, is determined by the input, forget, and cell gate, and the cell state of the previous sequence step $c_{t-1}$. The output gate $o_t$ determines the output of the LSTM cell and defines the new hidden state by combining information of output gate $o_t$ and cell state $c_t$. This process is repeated for every element in the sequence.

### 3.2.1. Bidirectional LSTM

The LSTM architecture from Section 3.2 can be extended to a BiLSTM which processes a sequence in two directions and then combines the result from both directions. In the field of natural language processing, this would mean that a sentence is parsed from first to last word and also from last to first word. For variables in the atmosphere, this would mean the the BiLSTM scans a vertical profile from the top of the atmosphere to the surface and the other way round.

## 3.3. Interpretability: SHAP

To enhance the interpretability and trustworthiness of emulation results of ML models, interpretability methods were developed in the last decade. Now, many interpretability methods are available and used to analyze how an NN makes a prediction. For example, PyTorch provides a full package for model interpretability called captum (Kokhlikyan et al. 2020). One of the methods is called SHapley Additive exPlanations (SHAP) based on Lundberg and Lee (2017). SHAP can identify if and how a certain input feature contributes to a certain prediction based on a trained NN. The contribution of each input feature is calculated by multiplying the gradient of the NN with the difference of the input feature with respect to a baseline distribution. The baseline distribution describes the average input sample leading to the average NN prediction. This calculated contribution is known as the Shapely value, which approximates each input features' contribution to the prediction. Then, the sum of the Shapley values yields an approximation of the NN prediction. This gives insights on how an NN makes a prediction.

# 4. Interpretable Machine Learning-based Radiation Emulation for ICON

As the radiation parameterization is one of the most expensive components in ESMs in terms of computing time, ML can be used to create a faster alternative. ML-based radiation schemes have been developed in the past (Chevallier et al. 1998; Krasnopolsky et al. 2005; Lagerquist et al. 2023; Song and Roh 2021; Yao et al. 2023) but the target application was mostly NWP, where a model top of 36 $km$ is sufficient. However, climate projections require high model tops to include stratospheric effects. Additionally, ML-based radiation schemes often lack interpretability raising the demand for explainability. Therefore, the key science question 1 is "Can an ML-based emulator learn to represent atmospheric radiative transfer based on physical laws?".

This chapter focuses on the development of an ML-based radiation emulator for the ICON model. The underlying relationships that are learned by the ML-based emulator are analyzed with interpretability methods like SHAP. This study has been presented in Hafner et al. (2025a) which is published in the *Journal of Geophysical Research: Machine Learning and Computation*. As indicated in Section 1.3, I produced all content that is presented from this publication including text, figures, tables and code[1].

This chapter is structured as follows: First, I present the data used in this study (Section 4.1) including variables and normalization. The data is based on the ICON model described in Section 3.1. In Section 4.2, the model architecture and physical constraints for the training are described. The results for the ML-based radiation emulator are presented in Section 4.3. The emulator itself is interpreted in Section 4.4. This chapter finalizes with a summary in Section 4.5.

## 4.1. 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

[1] https://github.com/EyringMLClimateGroup/hafner24jgrml_MLradiationemulation_offline (last accessed: 19.08.2025) preserved at Hafner (2025b)

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 (IFS) 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-orographic 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.

### 4.1.1. Variables

The input (predictor) and output (target) variables are column-wise values of the model's radiation scheme, and are summarized in Table 4.1. 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 Appendix A 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 like 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. 2021b). 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. 2021b).

**Table 4.1.:** Input and output variables for each network and training phase. 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 (↑) or downward (↓) flux at the surface (surf) or top of the atmosphere (TOA), $\alpha$ is surface albedo, $q_i$ is cloud ice, $q_l$ is cloud liquid, $q_v$ is specific humidity, $O_3$ is ozone concentration, $\rho$ 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. Adapted with permission from Hafner et al. (2025a).

| Variable | Unit | SW HR | SW Flux | LW HR | LW Flux |
|---|---|---|---|---|---|
| Input | | | | | |
| $F_{\downarrow,\,TOA,\,SW}$ | $W/m^2$ | ✓ | ✓ | - | - |
| $\alpha$ | - | ✓ | ✓ | - | - |
| $\alpha_{NIR,\,dir}$ | - | - | ✓ | - | - |
| $\alpha_{NIR,\,dif}$ | - | - | ✓ | - | - |
| $\alpha_{vis,\,dir}$ | - | - | ✓ | - | - |
| $\alpha_{vis,\,dif}$ | - | - | ✓ | - | - |
| $\vec{q}_i$ | $kg/kg$ | ✓ | ✓ | ✓ | ✓ |
| $\vec{q}_l$ | $kg/kg$ | ✓ | ✓ | ✓ | ✓ |
| $\vec{q}_{H2O} = \vec{q}_v + \vec{q}_l + \vec{q}_i$ | $kg/kg$ | ✓ | ✓ | ✓ | ✓ |
| $\vec{O}_3$ | $kg/kg$ | ✓ | ✓ | ✓ | ✓ |
| $\vec{\rho}$ | $kg/m^3$ | ✓ | ✓ | ✓ | ✓ |
| $\vec{cl}$ | - | ✓ | ✓ | ✓ | ✓ |
| $\vec{T}$ | $K$ | ✓ | ✓ | ✓ | ✓ |
| $T_{surf}$ | $K$ | - | - | ✓ | ✓ |
| Output | | | | | |
| $\partial\vec{T}_{SW}/\partial t$ | $K/d$ | ✓ | - | - | - |
| $\partial\vec{T}_{LW}/\partial t$ | $K/d$ | - | - | ✓ | - |
| $F_{\downarrow,\,surf,\,SW}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,NIR,\,dir}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,NIR,\,dif}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,vis,\,dir}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,vis,\,dif}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,PAR,\,dir}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,SW,\,PAR,\,dif}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\uparrow,\,TOA,\,SW}$ | $W/m^2$ | - | ✓ | - | - |
| $F_{\downarrow,\,surf,\,LW}$ | $W/m^2$ | - | - | - | ✓ |
| $F_{\uparrow,\,TOA,\,LW}$ | $W/m^2$ | - | - | - | ✓ |

### 4.1.2. Normalization

Normalization is essential for machine learning. 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 like 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 $q_{H2O} = q_v + q_l + q_i$. 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 $\sigma T_{surf}^4$. 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

$$x_{norm} = \frac{x - \mu}{\sigma}, \tag{4.1}$$

where $\mu$ is the mean and $\sigma$ 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.

## 4.2. 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, while heating rates are defined at full levels located at the cell center. We predict heating rates and fluxes using a single neural network but split the training process into two phases. In the first phase, we optimize the prediction of heating

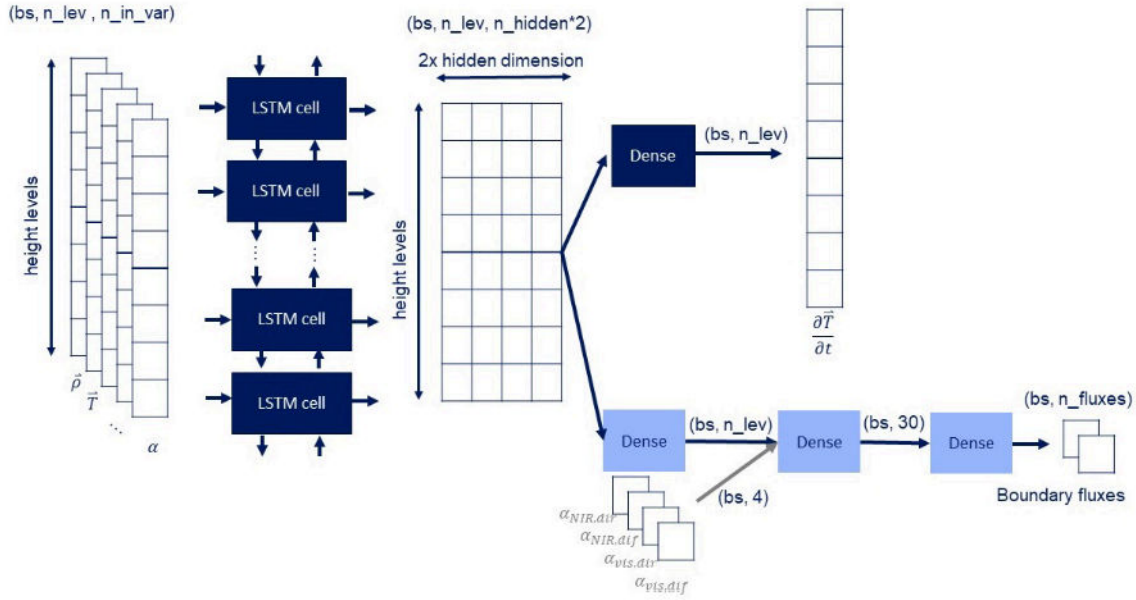


**Figure 4.1.:** Schematic of the neural network architectures to emulate radiative heating rates and boundary fluxes. On the left, example input variables are density, atmospheric temperature and albedo. Profile variables span all height levels, such as atmospheric temperature $\vec{T}$. Scalar variables, such as surface albedo $\alpha$, 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 BiLSTM 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* stand 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 notes in an LSTM cell, *n_fluxes* is the number of boundary fluxes. Adapted with permission from Hafner et al. (2025a).

rates ($\mathrm{HR}_{SW}$, $\mathrm{HR}_{LW}$). In the second phase, we learn predicting the boundary fluxes ($\mathrm{FLUX}_{SW}$, $\mathrm{FLUX}_{LW}$).

### 4.2.1. 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:

$$(F_{\downarrow,TOA} - F_{\uparrow,TOA}) - (F_{\downarrow,surf} - F_{\uparrow,surf}) = \int_{surf}^{TOA} \frac{\partial T}{\partial t} c_{p,air} \rho dz \approx \sum_{l=0}^{n_{lev}} (F_{net,l+1/2} - F_{net,l-1/2}), \quad (4.2)$$

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 neural networks. 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 $\varepsilon$ is emissivity of the surface, and $\sigma$ 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 neural network, heat capacity $c_{p,air}$, density $\rho$, and vertical thickness of a layer. 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.

### 4.2.2. 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 Bidirectional Long Short-Term Memory (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 82.4$k$ for SW and 81.6$k$ for LW.

Figure 4.1 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:

$$\mathcal{L}_{HR} = MSE + MAE + min(10^{-8} * 10^{\frac{e-e_s}{n_e}}, 10^{-1}) * energy, \tag{4.3}$$

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 (4.2). 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).

### 4.2.3. Second Training Phase: Boundary Fluxes

For the fluxes, we want to leverage what the BiLSTM has learned already. Therefore, we used the BiLSTM output in the second training phase and added three dense layers (see Figure 4.1 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, while surface emission limits the LW fluxes. The normalized LW fluxes can be larger than 1 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 (4.3)) 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.

## 4.3. Results

In this section, we evaluate the performance of all four components, $\mathrm{HR}_{SW}$, $\mathrm{HR}_{LW}$, $\mathrm{FLUX}_{SW}$ and $\mathrm{FLUX}_{LW}$, 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.

### 4.3.1. Heating Rates

We begin by evaluating the predictions of the machine-learned heating rates, as summarized in Table 4.2. The overall MAE for both SW and 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.

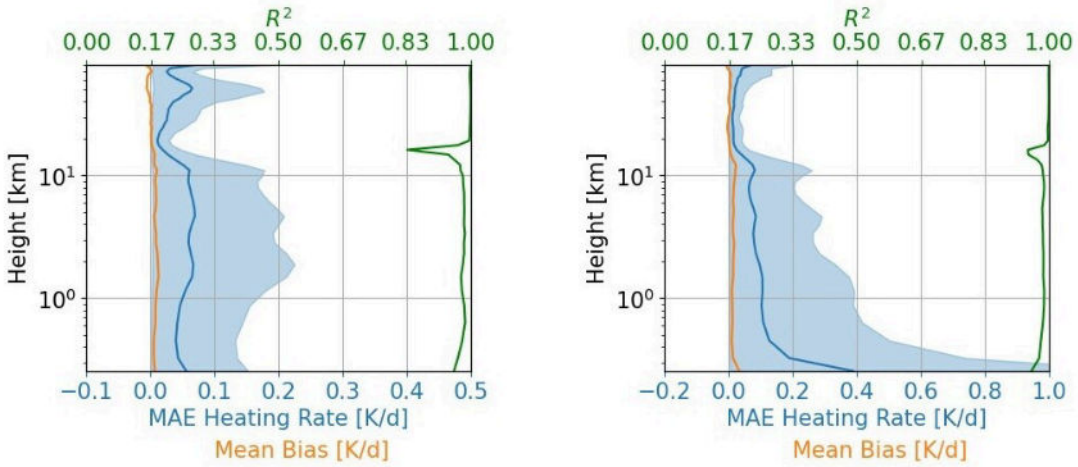


**Figure 4.2.:** Global and time mean vertical profiles of heating rates. Mean Absolute Error (MAE), bias, and $R^2$ are shown for shortwave heating rates (left) and longwave heating rates (right). The shaded area shows 90% of error spread. Adapted with permission from Hafner et al. (2025a).

**Table 4.2.:** Bulk statistics for heating rate results. MAE is mean absolute error and $R^2$ is coefficient of determination. Root Mean Squared Error (RMSE) is root mean squared error. The percentage values in brackets denote the relative values of MAE, bias and RMSE. Adapted with permission from Hafner et al. (2025a).

| | 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%) |

Figure 4.2 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 A.2 and A.3). 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

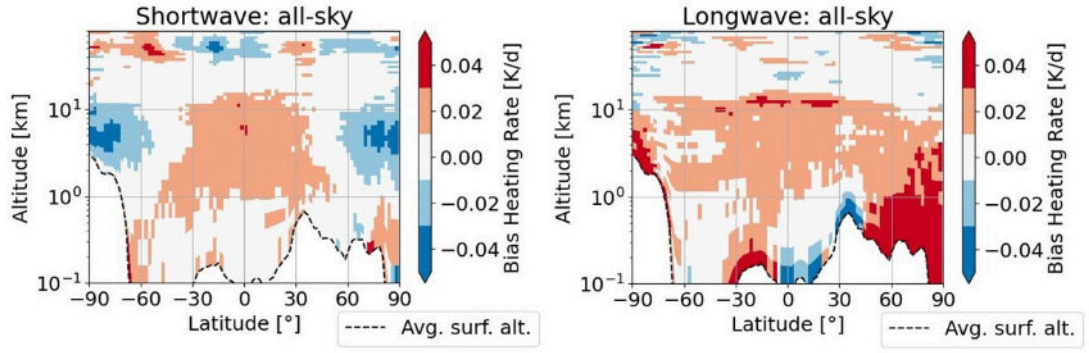


**Figure 4.3.:** Zonal- and time-mean machine-learned biases for (left) shortwave heating rates, and (right) longwave heating rates. Adapted with permission from Hafner et al. (2025a).

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 (UTLS) at around 10-12 km. This region is cloud-free with ozone effects beginning at higher levels.

Figure 4.3 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 4.2). 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 5698 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

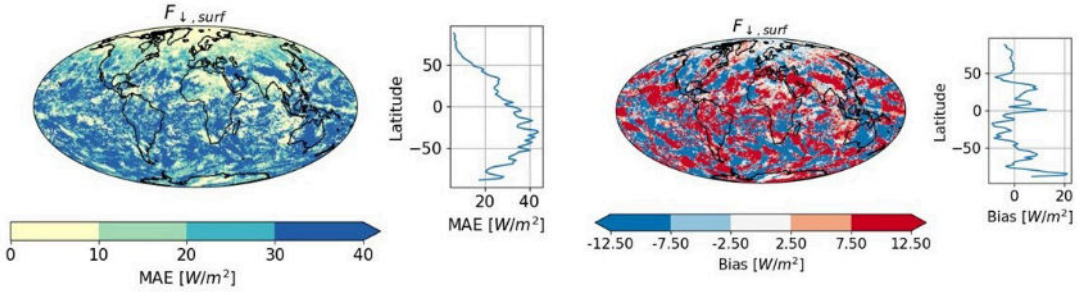


**Figure 4.4.:** Time-averaged maps of shortwave downward flux at the surface. (left) Mean Absolute Error (MAE), and (right) bias are shown. Right panels show zonal-mean values. Adapted with permission from Hafner et al. (2025a).

**Table 4.3.:** Bulk statistics for all fluxes. Adapted with permission from Hafner et al. (2025a).

| 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 |

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).

### 4.3.2. Fluxes

The SW flux component predicts in total eight scalar SW fluxes and the LW flux component predicts two scalar fluxes. Table 4.3 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 $W/m^2$. The downward fluxes at the surface are in general predicted worse, where $F_{\downarrow, surf, SW}$ has an error of around 30 $W/m^2$. The partial fluxes exhibit a smaller MAE of around 9 $W/m^2$ for diffuse fluxes and 15 $W/m^2$ for direct fluxes, but direct fluxes are on average larger than diffuse fluxes. The bias remains minimal, ranging from $-0.6$ to $0.4 W/m^2$. The NIR and visible fluxes approximately add up to the total SW downward flux (see Figure A.8). 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 $R^2$ vales 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 $W/m^2$ and bias of $-0.29 - 0.17$ $W/m^2$. For further analysis, we focus on the SW and LW downward flux at the surface and refer to the supplementary material (Figures A.4–A.7) for the other fluxes.

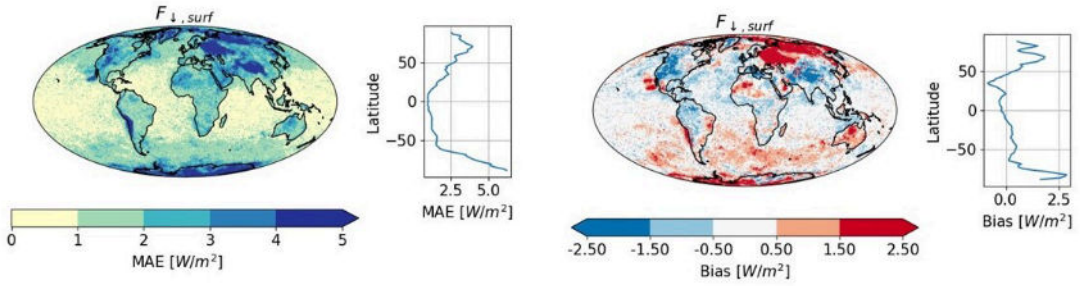


**Figure 4.5.:** Similar to Figure 4.4 but for longwave downward flux at the surface. Adapted with permission from Hafner et al. (2025a).

The MAE errors are larger in the tropics, see Figure 4.4. 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 Figure 7 e-f 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 4.4 right, appears somewhat erratic but is overall slightly negative.

For LW, the MAE is very small everywhere, Figure 4.5, but slightly larger in elevated areas like the Andes and the Tibetan plateau and the bias is very small.

### 4.3.3. Energy Consistency

Taking the difference of the left and right side of Equation (4.2), 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 4.6. The mean for SW radiation is 0.59 $W/m^2$ and for LW radiation $-0.07$ $W/m^2$. The values are within $\pm 0.5$ $W/m^2$ which is acceptable.

When heating rate profiles and boundary fluxes are trained separately using distinct neural networks (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. While 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.

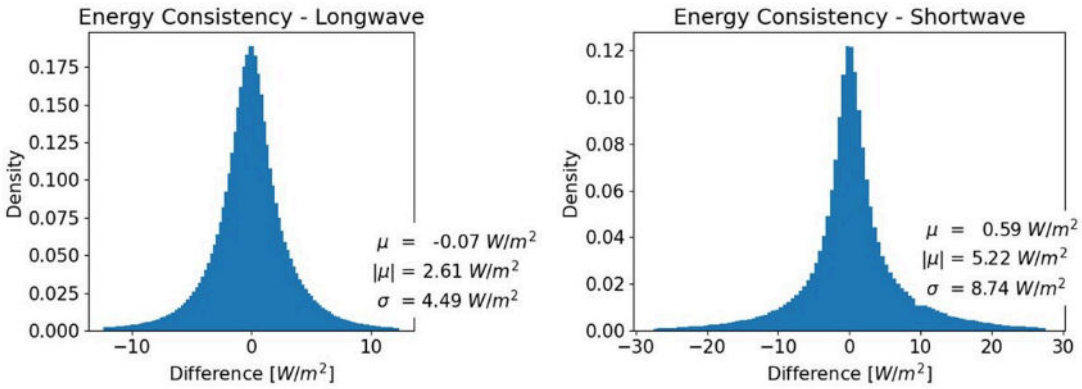


**Figure 4.6.:** Energy balance check for combined neural networks for SW radiation (left) and LW radiation (right). The histogram shows the difference between boundary fluxes and the vertical integral of radiative energy, both predicted by the neural networks. Adapted with permission from Hafner et al. (2025a).

## 4.4. 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 and Lee 2017), a method used for interpreting complex machine learning 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 dataset and a random subset that corresponds to 1% of the background dataset.

### 4.4.1. Shortwave Radiation

The top panel of Figure 4.7 shows the mean absolute Shapley values for the SW heating rates, predicted by a BiLSTM. Looking at the air density $\rho$ (Figure 4.7 f), 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 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 4.7 g). The BiLSTM primarily relies on local

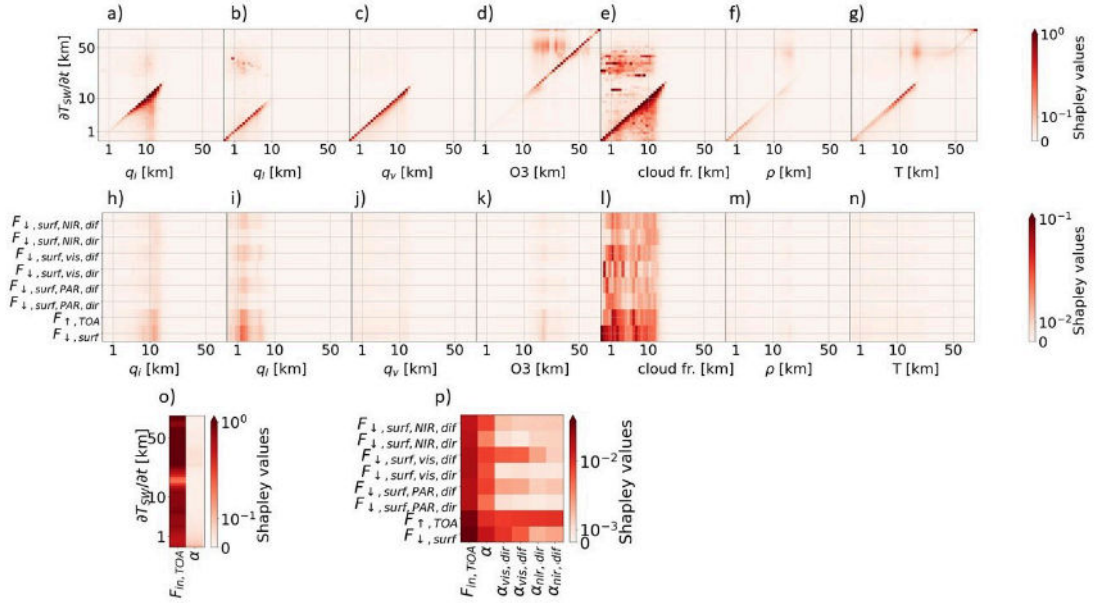


**Figure 4.7.:** Mean absolute Shapley values for the NN used for predicting SW heating rates and SW boundary fluxes. The x-axis represents the input variables, whereas the y-axis represents the predicted output, indicating how each layer of the input affects the corresponding layers of the output. The 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. Adapted with permission from Hafner et al. (2025a).

atmospheric variables to predict the heating rates, which are locally affected by absorption and emission of matter that is locally available (Figure 4.7 a, b, e).

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 4.7 o). 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 (Figure 4.7. a,b). Unlike cloud variables, the contribution of ozone is concentrated in the stratosphere and mesosphere (Figure 4.7.d). 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 and Hobbs

2006). The contribution of surface albedo $\alpha$ is strongest closer to the surface, associated with reflected radiation (Figure 4.7 o).

The middle panel of Figure 4.7 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 $F_{\downarrow,\, surf,\, SW}$. 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 (Figure 4.7 h, i, l) 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 (Figure 4.7 a-g), 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 the SI, we did not investigate this approach.

### 4.4.2. Longwave Radiation

Figure 4.8 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 4.8 o), which is strongest directly above the 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 4.8 g) 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 4.8 f). As density decreases with height, there is

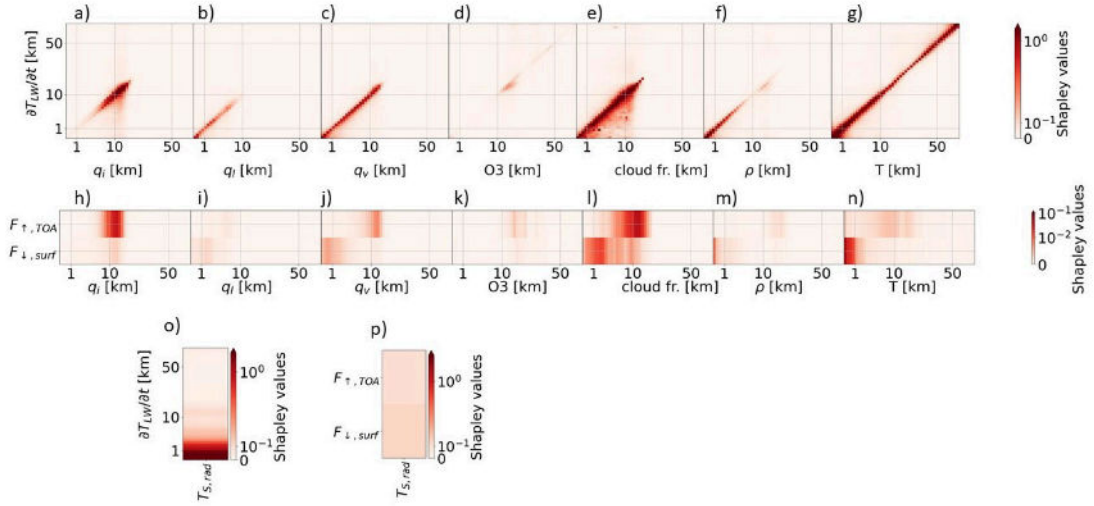


**Figure 4.8.:** Similar to Figure 4.7 but for longwave radiation. Adapted with permission from Hafner et al. (2025a).

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 (Figure 4.8 a, b). 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 4.8 l). 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.

**Table 4.4.:** Bulk statistics for heating rate results with the MLP. 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. Adapted with permission from Hafner et al. (2025a).

| | 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%) |

### 4.4.3. Comparison to Multilayer Perceptron

For comparison, we conducted the same analysis using a Multi Layer 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 phase, the learning rate is set to $5 * 10^{-4}$. The bulk statistics for heating rates are shown in Table 4.4. The corresponding plots as in Figures 4.2–4.6 and statistics for the boundary fluxes are provided in the appendix Figures A.9–A.12 and Table A.1.

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 Figure A.9) compared to the BiLSTM (see Figure 4.2). 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 4.9 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 (Figure 4.9 a-c). However, for variables such as ozone, density, and temperature, the MLP relies on non-physical or non-causal associations to predict heating

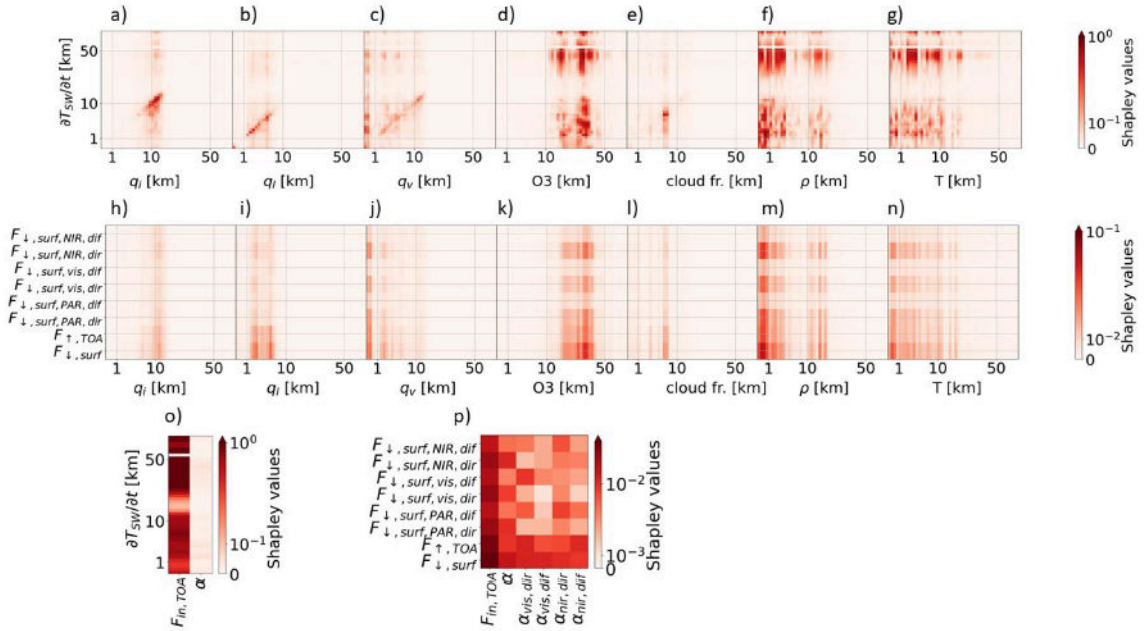


**Figure 4.9.:** Similar to Figure 4.7 but using a multilayer perceptron for shortwave radiation. Adapted with permission from Hafner et al. (2025a).

rates. For example, it learns to use stratospheric ozone to predict tropospheric temperature tendencies (Figure 4.9 d) or lower tropospheric density to predict heating rates in the upper stratosphere (Figure 4.9 f).

The mean absolute Shapley values for longwave radiation are shown in Figure 4.10. 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 4.10 c). However, the MLP also attributes strong non-local influence to temperature, with significant contributions from levels above and below the target level (Figure 4.10 g). 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.

## 4.5. Conclusion and Discussion

Radiation is one of the most computationally expensive components in Earth system models (ESMs), despite several simplifications built into radiation parameterization and its application in ESMs. Machine Learning (ML) can potentially help to speed up the calculation related

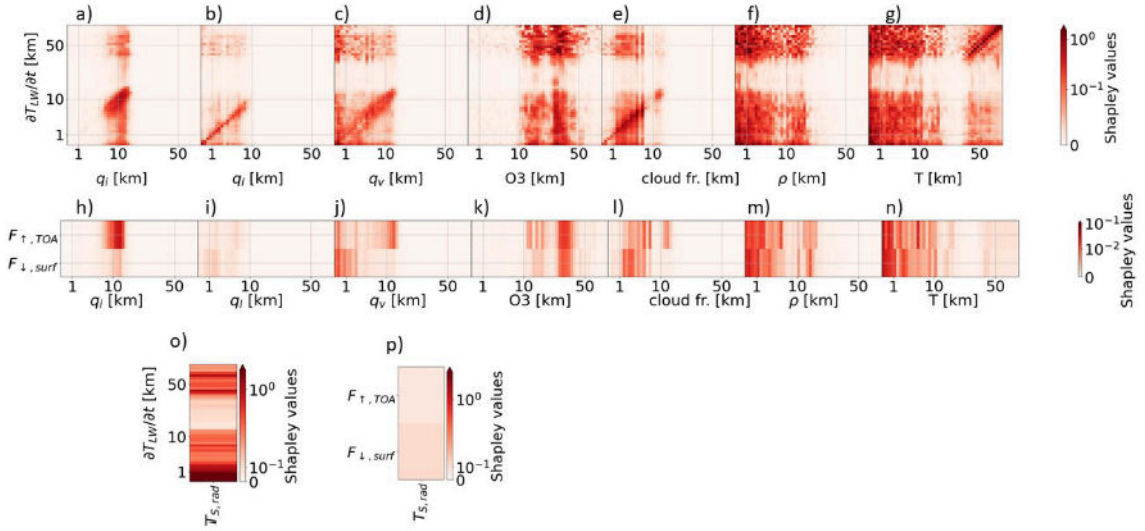


**Figure 4.10.:** Similar to Figure 4.8 but using a multilayer perceptron for longwave radiation. Adapted with permission from Hafner et al. (2025a).

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 neural networks 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 neural network that computes boundary fluxes from the BiLSTM output.

Our ML-based model accurately emulates heating rates. The shortwave heating rates have a mean absolute error of 0.045 K/d (2.77%) and a bias of 0.004 K/d (0.38%). The longwave heating rates have a mean absolute error 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 Shapley Additive exPlanations (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 neural networks have good offline accuracy, and our interpretability analysis shows that the networks learned physically meaningful input-output connections. Additionally, we show that the neural networks 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.

# 5. Stable Machine Learning-based Radiation Emulation for ICON

Most ML-based radiation schemes that are based on advanced architectures like BiLSTMs have been tested online for only a few days in the context of NWP, if tested at all, as already mentioned in Section 2.6. Therefore, key science question 2 is formulated as "Can an ML-based emulator learn to represent atmospheric radiative transfer based on physical laws?" which is discussed here.

The focus of this chapter is the online coupling of the ML-based radiation emulator to the ICON model. The results of this study are presented in Hafner et al. (2025b) which is currently under review for *Journal of Advances in Modeling Earth Systems*. As mentioned in Section 1.3, I produced all content in this thesis that is presented from that publication including text, figures and tables.

This chapter starts with the technical aspect of coupling an ML-based scheme to a Fortran based ESM (Section 5.1). Then, the experimental setup for the hybrid ICON simulation and two reference simulations is explained (Section 5.2). The analysis of the online results is presented in Section 5.3. A test of how well the ML-based radiation emulator performs in a warmer climate is presented in Section 5.4. This chapter concludes with a summary and discussion (Section 5.5).

## 5.1. Neural Networks and Online Coupling

We couple an ML-based radiation emulator to ICON-A that emulates the full radiation parameterization RTE+RRTMGP (Pincus et al. 2019) accurately and in a physical meaningful way (Hafner et al. 2025a). The full radiation scheme includes cloud and gas optics as well as the radiative transfer equations. The emulator is separated into SW and LW radiation and returns heating rate profiles (HR) and fluxes at the surface and the top of the atmosphere. Note that shortwave and longwave radiation are also treated separately in the original radiation scheme. Additionally, shortwave radiation is only available during daytime and longwave neglects scattering.

As in Hafner et al. (2025a), the inputs are vertical profiles of cloud ice $q_i$, cloud liquid $q_l$, water vapor $q_v$, ozone $O_3$, density $\rho$, cloud fraction $cl$, and temperature $T$. For the SW emulator, we additionally use the scalar inputs incoming flux at the top of the atmosphere $F_{\downarrow,TOA,SW}$, albedo $\alpha$ and partial albedos for near infra-red (NIR) and visible (vis) range for direct and diffuse radiation $\alpha_{NIR,dir}$, $\alpha_{NIR,dif}$, $\alpha_{vis,dir}$, and $\alpha_{vis,dif}$. The partial albedos are

only used to predict the partial SW surface fluxes. For the LW emulator, we additionally use the surface temperature $T_{surf}$.

The outputs of the emulator are heating rates and boundary fluxes. The boundary fluxes include the downward flux at the surface and the upward flux at the top of the atmosphere. Additionally, we predict partial SW fluxes including photosynthetically active radiation (PAR), visible (vis) and near-infrared (NIR) each with a direct (dir) and diffuse (dif) component. All surface fluxes, including the partial SW fluxes, are needed to couple to the (land) surface scheme while the upward flux at the top of the atmosphere is used to calibrate ICON-A.

In the traditional parameterization, the heating rate profiles are calculated from upward and downward fluxes in an atmospheric column. The emulator consists of a BiLSTM. An LSTM processes a sequence element by element, retaining information from earlier elements; the bidirectional variant also incorporates information from preceding and subsequent elements. In other words, the BiLSTM scans the input in upward and downward direction, remembers what it has seen above and below, to learn important features from the atmospheric state—similarly to the physics based scheme. These features are then used to infer the heating rate and boundary fluxes. The NNs have 84.5k trainable parameters for the SW emulator and 83.4k trainable parameters for the LW emulator. The emulator accurately predicts LW and SW heating rates with a MAE of 0.060 $K/d$ and 0.045 $K/d$ respectively. The LW fluxes have a MAE of 1.9 $W/m^2$ and bias of $-0.06$ $W/m^2$ and the SW fluxes have an MAE of 13 $W/m^2$ and bias of $-0.15$ $W/m^2$. More details about the neural networks, their offline performance and interpretability can be found in Hafner et al. (2025a).

The neural networks are implemented in Python using PyTorch (Ansel et al. 2024), compiled and saved. For coupling, we use FTorch (Atkinson et al. 2025). The saved neural networks can be called from Fortran using FTorch. FTorch uses the Torch C++ interface and avoids Python calls.

## 5.2. Experimental Setup

We have in total three setups for the simulations analyzed in this study, two reference simulations and one hybrid ML simulation, where hybrid is referring to a simulation that combines the traditional physics based simulation with ML. They all have the same settings except for the radiation parameterization, which is completely replaced with our newly developed ML-based radiation emulator for the hybrid ML simulation. We conducted our simulations starting from IFS initial conditions for January 1st 1979. Aerosol interactions are omitted in this setting and are subject to further work. Radiation is called less often than other physics parameterizations, i.e. every $1-2$ $h$ vs. every 6 $min$. The simulations have a length of one year and their differences are summarized in Table 5.1. Here, time per simulated day describes only the time that is spent for the radiation parameterization. The time needed to compute the rest stays the same in all setups.

**Table 5.1.:** Summary of the differences between the simulations used for the evaluation of online radiation emulator performance. Adapted with permission from Hafner et al. (2025b).

| | ICON-$A_{HF}$ | ICON-$A_{LF}$ | ICON-$A_{ML}$ |
|---|---|---|---|
| Rad. scheme | RTE+RRTMGP (Pincus et al. 2019) | | ML-based (Hafner et al. 2025a) |
| Rad. time step | 6 $min$ | 60 $min$ | 6 $min$ |
| Time per sim. day [s] for radiation | 74.5 | 7.5 | 19.2 |
| Time per sim. day [s] total | 86.4 | 15.4 | 28.7 |

The first reference simulation has a high frequency radiation call (ICON-$A_{HF}$), with a time step of 6 $min$ shared for all physics parameterizations. ICON-$A_{HF}$ is used to train and develop the ML-based radiation emulator, and it is the target ('ground truth') for this analysis. Frequent radiation calls, in principle, provide a better representation of interactions with clouds compared to less frequent calls.

The second reference simulation follows the common approach of low frequency ( 60 $min$, ICON-$A_{LF}$) calls to the radiation scheme, trading-off computational burden and accuracy. We use ICON-$A_{LF}$ as a baseline compared to the computationally costly ICON-$A_{HF}$. With a smaller temporal frequency—updating radiative fluxes only once per hour— ICON-$A_{LF}$ introduces errors. The main contribution to these errors results from the cloud radiative effects, which cannot be adjusted in the time steps between radiation calls.

The third simulation, referred to as the hybrid ML simulation (ICON-$A_{ML}$), uses a radiation time step of 6 $min$. ICON-$A_{ML}$ is computationally more efficient than ICON-$A_{HF}$, and potentially captures cloud-radiation interactions better compared to ICON-$A_{LF}$.

For each of the three setups, we make 10 ensemble runs each. The ensembles are generated by initializing the model with 10 distinct initial conditions, while the boundary conditions are kept the same. To minimize the influence of internal variability on inter-setup differences, we primarily compare ensemble means. Additionally, to assess the potential accumulating biases, we also run one 10-year simulation for each setup to examine whether ICON-$A_{ML}$exhibits long-term drift.

### 5.2.1. Comparison of Computational Performance

To provide a baseline, we first compare wall-clock times at the same radiation calling frequency. Under these conditions, the ML-based radiation emulator delivers a speedup of approximately 4× relative to RTE+RRTMGP without any additional code optimization. In a standalone prototype test (Python implementation) on the Levante system at the German Climate Computing Center (DKRZ), the emulator ran about 20× faster on GPU than on CPU, underscoring the potential for further acceleration. The achieved speed-up though depends on several factors, including hardware (CPU vs. GPU), the number of columns evaluated in parallel (batch size), and implementation details. Because the ICON configuration used in

this study does not run on GPUs, we do not report end-to-end ICON-on-GPU benchmarks; we simply note the larger speedup potential on GPUs suggested by the standalone tests.

Comparing the speed-up with other studies requires careful interpretation because the run-time depends many different factors, for instance, factors such as hardware, compiler, reference radiation scheme, and offline versus online configuration. Pal et al. (2019) reports a speed-up of 8-10 on CPU in comparison with the older radiation code Rapid Radiative Transfer Model (RRTM) for GCMs (RRTMG) (Iacono et al. 2008). Song and Roh (2021) also compared to RRTMG and reports a speed-up of 60. Lagerquist et al. (2023) reports a speed-up of 90-7510 but compared to RRTM (Mlawer et al. 1997b), which is slower than RRTMG. Notably, code optimization accelerates the computation by a factor of 12 (Ukkonen and Hogan 2024). Furthermore, the switch from double to single precision already halves the computing time (Cotronei and Slawig 2020). Compared to optimized radiation schemes ML-based emulators are not necessarily faster Bertoli et al. (2025), Ukkonen and Chantry (2025). One of the issues reported is the communication between CPU and GPU. In contrast, emulators exhibit higher energy efficient when evaluated on GPUs (Ukkonen and Chantry 2025). In conclusion, a computational performance increase of 1-10 is considered limited when compared to state of the art radiation schemes such as RTE+RRTMGP or ecRad (Hogan and Bozzo 2018), since further optimized codes improve the speed by $O(10)$. However, ML-based radiation emulators can potentially still be useful for GPU-based ESMs. In contrast to conventional radiation schemes, emulators possess the property of differentiability, making them suitable for use in fully differentiable ESMs (Klöwer et al. 2024; Kochkov et al. 2024). However, their stability must be evaluated prior to their implementation, which will be analyzed in the following sections.

## 5.3. Online Results

To evaluate the performance of ICON-$A_{ML}$, we first assess the output variables of the emulator itself by analyzing the differences to ICON-$A_{LF}$ and ICON-$A_{HF}$. We expect slight deviations in ICON-$A_{ML}$ compared to ICON-$A_{HF}$ because the residual error of the ML-based emulator, while small, is nonzero (Hafner et al. 2025a) (see above). Small deviations in prognostic variables, such as the temperature profile, can lead to a different simulated climate state. Over time, large deviations can accumulate and climate drift may result in a nonphysical state, leading the model to crash. To identify whether the differences are systematic or statistical, we analyze the ensemble runs: systematic differences are indicated when ensemble means of selected variables show consistent deviations, whereas the ensemble standard deviation reflects statistical differences associated with internal variability. It is important to note that the errors in ICON-$A_{ML}$ differ in nature from those in ICON-$A_{LF}$. Errors in the ML-based emulator depend on the quality and coverage of the training data as well as the capacity of the ML tool in emulating the data, whereas errors from low-frequency radiation calls arise

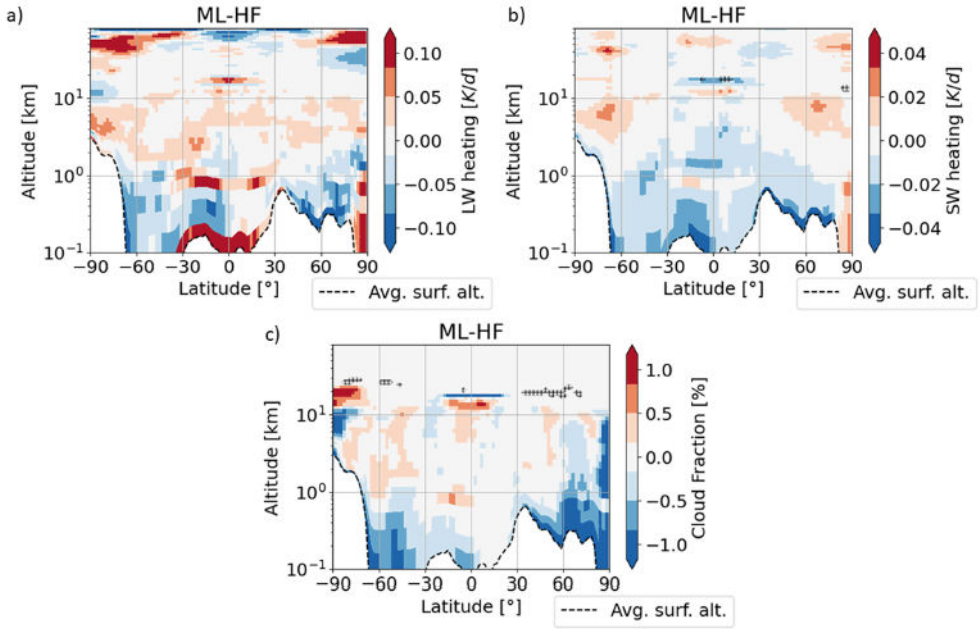


**Figure 5.1.:** Ensemble mean for annual and zonal mean difference of ICON-$A_{HF}$ and ICON-$A_{ML}$ for a) LW heating rate, b) SW heating rate and c) cloud fraction. Hatching indicates where the difference is three times larger then the standard deviation of annual zonal-means of ICON-$A_{HF}$. Adapted with permission from Hafner et al. (2025b).

from temporal sampling of the diurnal cloud cycle. The latter biases can often be reduced by tuning or by increasing the radiation call frequency (Yin and Porporato 2017).

To assess long-term stability, we perform a single, long-duration simulation for each setup. This provides evidence as to whether the ICON-$A_{ML}$configuration remains stable and free of long-term drift. For the ensemble runs, we subsequently examine the radiative energy balance at the top of the atmosphere (TOA), followed by an evaluation of fields—such as cloud cover—that are directly or indirectly influenced by radiation.

### 5.3.1. Heating Rates and Surface Fluxes

The ensemble mean of annual zonal-mean differences in SW and LW heating rate of the ensemble means of ICON-$A_{ML}$ and ICON-$A_{HF}$ are in general very small, not significant (Figure 5.1 a, b) and comparable to offline biases (Hafner et al. 2025a). A difference is evident in the tropical upper troposphere, and the same pattern is present in the cloud-fraction field (Figure 5.1 c). Clouds are slightly lower in this region, shifting cloud-radiation interaction and thereby increasing the heating rate difference compared to ICON-$A_{HF}$. Compared to the offline analysis (Hafner et al. (2025a), Figure 3), the annual, zonal-mean SW heating rates differences are smaller in the upper stratosphere and mesosphere. These differences are not present in the hybrid simulation ICON-$A_{ML}$. A potential reason is that the differences mixed away with the atmospheric circulation (Pincus and Stevens 2009). Moreover, we observe larger deviations near the surface in the tropics for SW and LW heating rates, although of opposite sign, whereas cloud fraction shows only a minor difference. The enhanced heating could lead

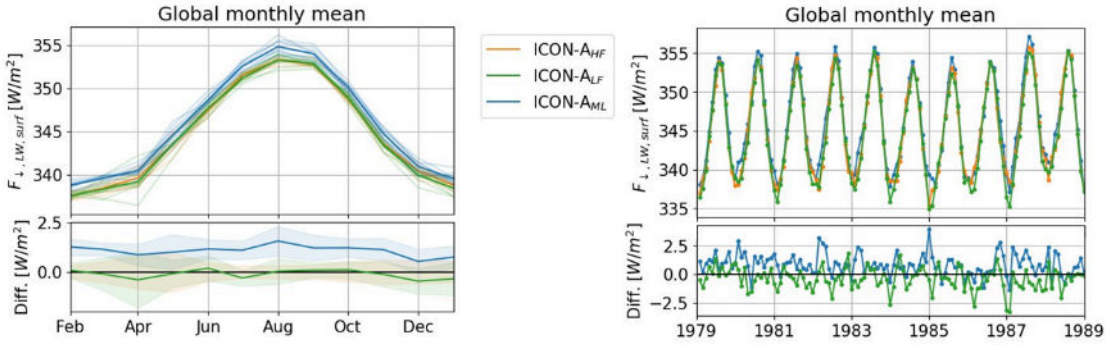


**Figure 5.2.:** Global monthly mean of downward LW flux at the surface $F_{\downarrow,LW,surf}$. Bottom panels show the difference to the 'ground truth' ICON-A$_{\mathrm{HF}}$. The left plot shows the ensemble means (bold lines) and the corresponding single runs (transparent lines). The lower panel shows mean difference to ICON-A$_{\mathrm{HF}}$ including $1\sigma$ spread. The right plot shows the single 10-year simulation for each setup. Adapted with permission from Hafner et al. (2025b).

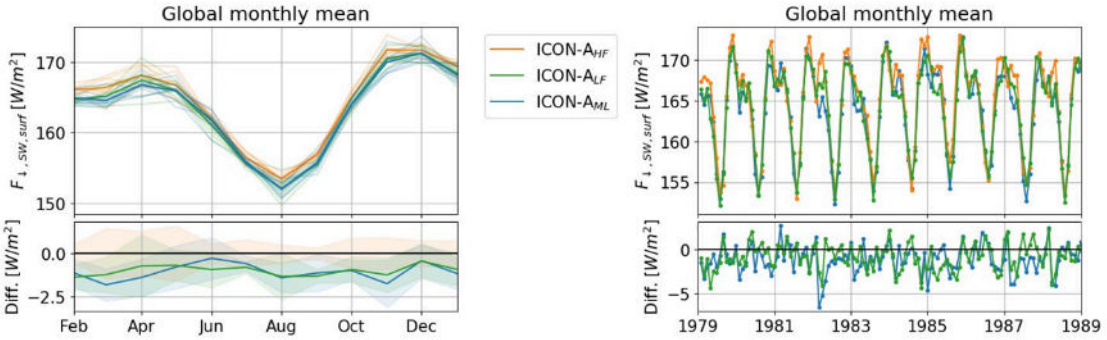


**Figure 5.3.:** Global monthly mean of downward SW flux at the surface $F_{\downarrow,SW,surf}$. Similar to Figure 5.2. Adapted with permission from Hafner et al. (2025b).

to enhanced circulation and therefore not directly correlate with differences in cloud fraction. Nevertheless, the differences are non-significant.

Figure 5.2 shows the temporal evolution for the LW downward flux at the surface, which is an important variable because it is coupled to the surface scheme. Specifically, in our simulations, it is coupled to the land surface only as sea surface temperature and sea ice concentration are prescribed. All three experimental setups have the same seasonal cycle and similar spread between the ensemble members. But, ICON-A$_{\mathrm{ML}}$ is on average 1 $W/m^2$ larger than the ICON-A$_{\mathrm{HF}}$ and ICON-A$_{\mathrm{LF}}$ but the difference does not increase over time. The long simulations (Figure 5.2 right) complement this result. The differences between the three runs does not diverge although there is a bias of about 1 $W/m^2$ indicating that ICON-A$_{\mathrm{ML}}$ runs in equilibrium that slightly differs from ICON-A$_{\mathrm{HF}}$.

The corresponding results for downward SW flux at the surface (Figure 5.3) show similar performance. Here, the average difference in the ensemble runs is of opposite sign of about $-1$ $W/m^2$. Interestingly, ICON-A$_{\mathrm{LF}}$ experiences the same deviation from ICON-A$_{\mathrm{HF}}$ as ICON-A$_{\mathrm{ML}}$. But, the differences of ICON-A$_{\mathrm{LF}}$ and ICON-A$_{\mathrm{ML}}$ are in range of 1 $\sigma$. The

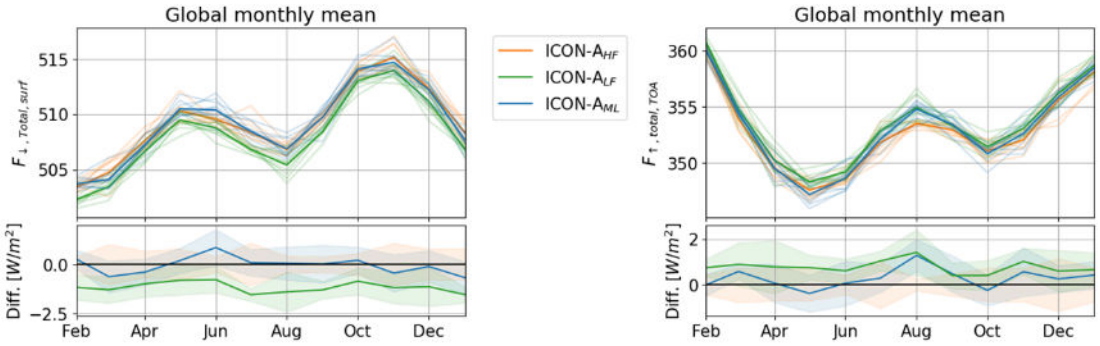


**Figure 5.4.:** (left) Total (SW+LW) downward flux at the surface and (right) total upward flux at the top of the atmosphere. Similar to Figure 5.2. Adapted with permission from Hafner et al. (2025b).

long simulations show that the differences do not build up over time for both ICON-A$_{\mathrm{ML}}$ and ICON-A$_{\mathrm{LF}}$.

While $F_{\downarrow,SW,surf}$ and $F_{\downarrow,LW,surf}$ from ICON-A$_{\mathrm{ML}}$ have a bias of $\pm 1\ W/m^2$, the bias cancels for the total (SW+LW) flux (Figure 5.4). However, ICON-A$_{\mathrm{LF}}$ shows a bias of $-1\ W/m^2$ which comes from $F_{\downarrow,SW,surf}$, as $F_{\downarrow,LW,surf}$ is unbiased. We observe similar results for $F_{\uparrow,total,TOA}$, where ICON-A$_{\mathrm{ML}}$ is closer to ICON-A$_{\mathrm{HF}}$.

### 5.3.2. Energy Balance at the Top of the Atmosphere

To evaluate our simulations, we compare the energy balance at the top of the atmosphere, as ICON-A$_{\mathrm{ML}}$ is expected to closely follow the original simulation used for training. Although the common method for evaluation would involve comparing the energy balance with observations, as a future step, we plan to calibrate the radiation balance of the atmosphere against the observed state of the Earth's climate system. Currently, the radiation balance has not been calibrated, and the ICON model itself exhibits biases when compared to observations. The net radiation $F_{net,\,TOA}$ at the top is described by

$$F_{net,\,TOA} = F_{\downarrow,\,SW,\,TOA} - F_{\uparrow,\,SW,\,TOA} - F_{\uparrow,\,LW,\,TOA}, \tag{5.1}$$

where $F_{\downarrow}$ is the incoming SW radiation and $F_{\uparrow,\,SW/LW,\,TOA}$ is the outgoing SW/LW radiation at the top of the atmosphere. The incoming solar radiation is prescribed and is the same in all runs.

Figure 5.5 shows the annual-mean net radiation at the top of the atmosphere for ICON-A$_{\mathrm{HF}}$ (Figure 5.5 a), ICON-A$_{\mathrm{ML}}$ (Figure 5.5 b), and their difference (Figure 5.5 c). For the absolute net radiation, red means more incoming radiation (warming) and blue means more outgoing radiation (cooling). By visual inspection, both runs exhibit very similar patterns and absolute magnitudes. To better quantify the differences, the bottom panel displays the difference between ICON-A$_{\mathrm{HF}}$ and ICON-A$_{\mathrm{ML}}$ with red indicating larger net radiation in the ICON-A$_{\mathrm{HF}}$ run. The differences are generally minimal, though notable discrepancies are

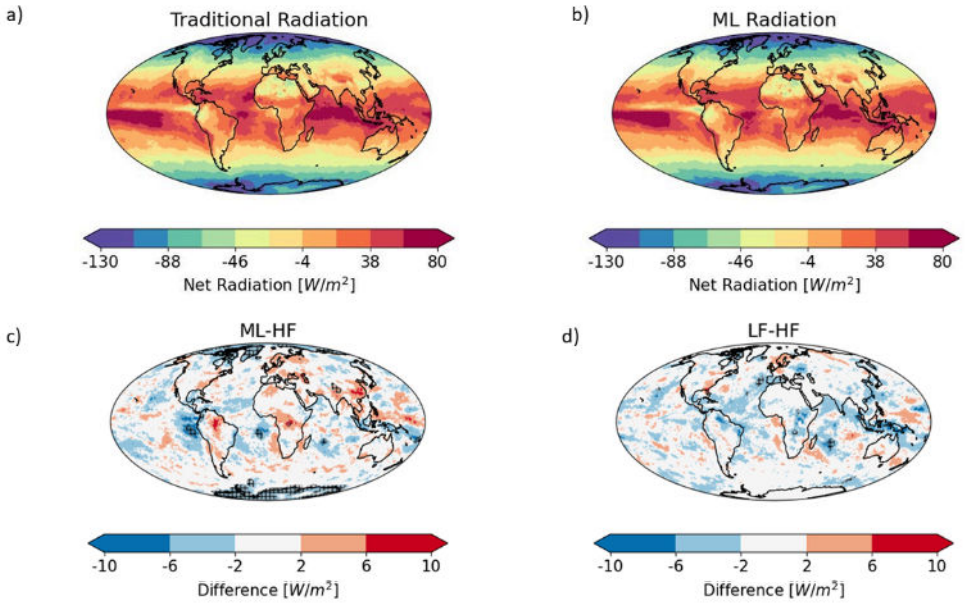


**Figure 5.5.:** Top of the atmosphere net radiation for a) ICON-$A_{HF}$ and b) ICON-$A_{ML}$. The bottom panel shows the difference compared to ICON-$A_{HF}$. The scaled difference between c) ICON-$A_{ML}$ and ICON-$A_{HF}$ and d) ICON-$A_{LF}$ and ICON-$A_{HF}$. Net radiation and differences were calculated from ensemble means. Adapted with permission from Hafner et al. (2025b).

observed, particularly over the Atlantic and Pacific Oceans, where the ICON-$A_{ML}$ exhibits increased cloud coverage. More clouds result in a greater absorption of radiation and a smaller net radiative effect in those regions. Comparing the normalized differences of ICON-$A_{ML}$ and ICON-$A_{LF}$ to ICON-$A_{HF}$, we see that the differences are both mostly in the tropics. The total bias of ICON-$A_{LF}$ is $-0.78\ W/m^2$ while the bias of ICON-$A_{ML}$ is $-0.28\ W/m^2$. Therefore, ICON-$A_{ML}$shows a slightly improved energy balance, which can be partially attributed to compensating biases. However, we find that ICON-$A_{ML}$ has in some regions still larger deviations as indicated by an RMSE of $2.24\ W/m^2$. Moreover, these deviations are comparable to ICON-$A_{LF}$ with an RMSE of $2.10\ W/m^2$.

On another note, all simulations are untuned and the energy balance could potentially be improved by tuning. The bias of the energy balance at TOA of ICON-$A_{HF}$ is $-12.31\ W/m^2$ with a standard deviation $\sigma$ of $0.23\ W/m^2$ between the ensemble runs. So, ICON-$A_{ML}$ differs by $1.2\sigma$, which is statistically non-significant. However, ICON-$A_{LF}$ shows significant deviations of $3.4\sigma$.

### 5.3.3. Total Cloud Cover

We observe that cloud radiative effects contribute to differences between ICON-$A_{ML}$ and ICON-$A_{HF}$. Therefore, we explore total cloud cover, as clouds influence and are influenced by radiation, due to climate feedbacks.

The left plot in Figure 5.6 illustrates the time series of the global, monthly mean total cloud cover for all three simulation setups, which exhibit a consistent temporal evolution. ICON-$A_{ML}$

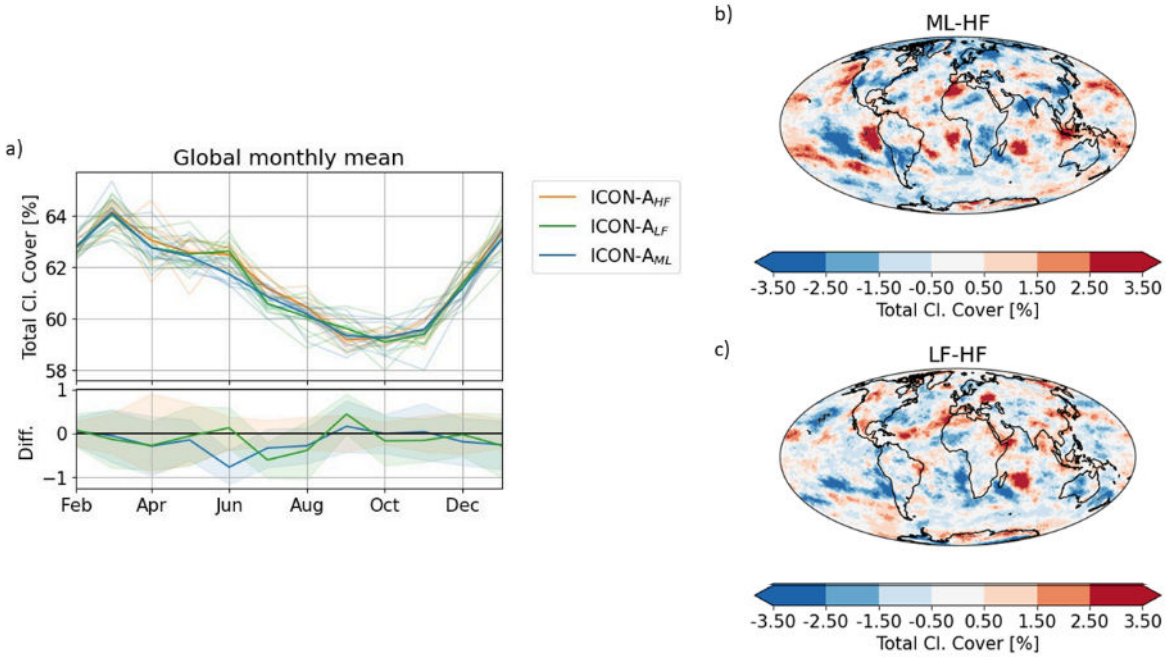


**Figure 5.6.:** (left) Global daily mean of total cloud cover for traditional high-frequent, traditional infrequent and hybrid run. (right) Annual ensemble mean difference of total cloud cover. Adapted with permission from Hafner et al. (2025b).

exhibits differences that are similar to those of ICON-$A_{LF}$ when compared to ICON-$A_{HF}$. It is important to note that the differences appear to be almost unbiased and experience similar variability. Here, the biases of ICON-$A_{ML}$ and ICON-$A_{LF}$ are smaller than 0.2%. However, differences in the spatial distribution of cloud cover are evident (Figure 5.6 b), which can significantly impact other fields, such as surface temperature. The red areas mean that the ICON-$A_{ML}$ is cloudier. It is notable that there are more clouds in regions of low clouds, such as over the Atlantic and Pacific ocean (Muhlbauer et al. 2014). Figure 5.1 c) also shows more clouds at around 1 $km$ height and $-15°$ latitude.

### 5.3.4. Clouds and Radiative Surface Fluxes

In Figure 5.1, we have demonstrated that changes in cloud distribution correlate heating rates. In this section, we examine in greater detail the correlation between changes in clouds and surface fluxes. The temporal evolution of globally averaged total downward SW and LW fluxes were discussed earlier (Section 5.3.1). Here, we focus on the spatial differences in downward SW and LW flux at the surface.

The total downward SW flux is slightly biased by $-1.17\ W/m^2$ for ICON-$A_{ML}$ (Figure 5.7 a). But, ICON-$A_{LF}$ is also biased by $-1.10\ W/m^2$ (Figure 5.8 b) . To analyze the origin, we look at the spatial distribution of differences. The pattern closely resembles the differences in total cloud cover shown in Figure 5.6, but with the opposite sign. Physically, this means more clouds reflect more radiation and therefore less SW radiation reaches the surface. To further verify this link, we show a 2D histogram of the local monthly mean differences in

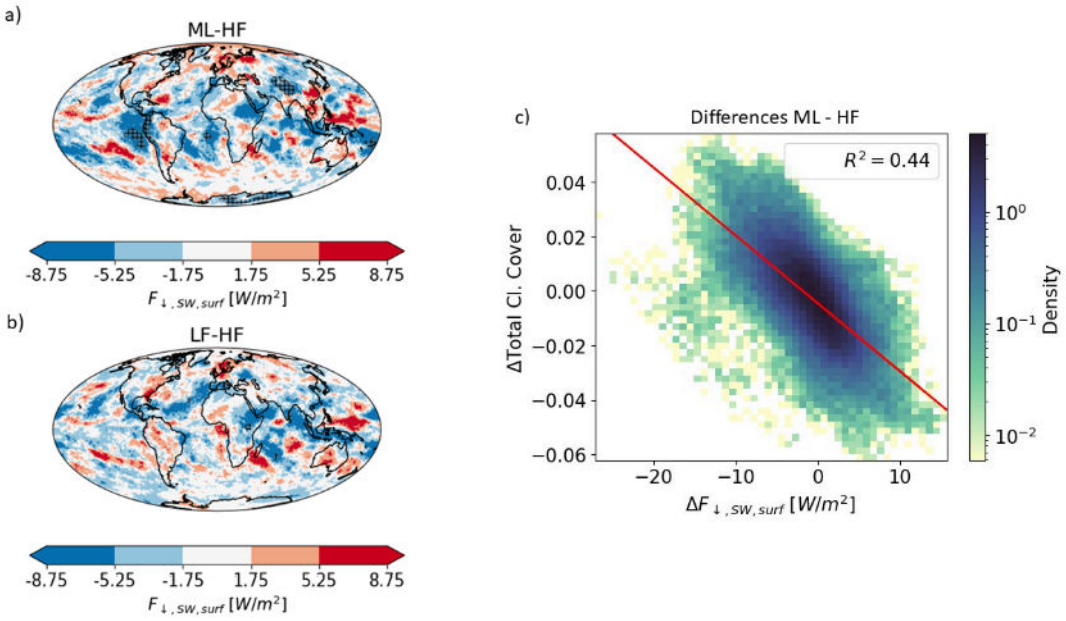


**Figure 5.7.:** Difference in downward SW flux at the surface $F_{\downarrow, SW, surf}$ of ensemble means of a) ICON-$A_{ML}$ and ICON-$A_{HF}$ and b) ICON-$A_{LF}$ and ICON-$A_{HF}$. Hatching indicates significant differences compared to ICON-$A_{HF}$. c) 2D histogram of monthly mean differences of $F_{\downarrow, SW, surf}$ vs total cloud cover. The red line is a linear fit and $R^2$ is coefficient of determination. Adapted with permission from Hafner et al. (2025b).

cloud cover and $F_{\downarrow, SW, surf}$ (Figure 5.7 c). Here, local differences refer to differences in the same grid cells. The red line is a linear fit to the data shown in the histogram. The coefficient of determination $R^2$ is 0.44, which means the fit describes the link between the difference of cloud cover and $F_{\downarrow, SW, surf}$ well. If the $R^2$ would be zero, one could not infer the difference in $F_{\downarrow, SW, surf}$ from the difference in total cloud cover. We see this link consistently with all direct SW flux components which are shown in the supplementary information (Figures B.1–B.6 and Table B.1). This link is also known in literature as more low clouds reflect more direct SW radiation (Forster et al. 2021b).

The downward LW flux at the surface $F_{\downarrow, LW, surf}$ is also biased by 1.12 $W/m^2$ for ICON-$A_{ML}$ (Figure 5.8 a) as already seen in Figure 5.2. Whereas ICON-$A_{LF}$ is rather unbiased with a bias of $-0.06$ $W/m^2$ (Figure 5.8 b). However, the biases are compensating for ICON-$A_{ML}$ but not for ICON-$A_{LF}$. Comparing the differences in total cloud cover and $F_{\downarrow, LW, surf}$, we observe similar behavior as for $F_{\downarrow, SW, surf}$ but with opposite sign. Here, more clouds correlate with higher $F_{\downarrow, LW, surf}$. Significant differences in $F_{\downarrow, LW, surf}$ are most prominent in the eastern Pacific. It is worth to note that significant changes in $F_{\downarrow, LW, surf}$ correlate with non-significant changes in total cloud cover.

This analysis shows that the differences between ICON-$A_{HF}$ and ICON-$A_{ML}$ are not just errors but rather the physically consistent responses to different cloud distributions. What exactly caused the different distribution is difficult to diagnose due to cloud feedbacks. Also, small changes can have significant effects in complex non-linear models such as ICON. As we are working monthly mean output, it is hard to isolate grid cells related to deep or shallow

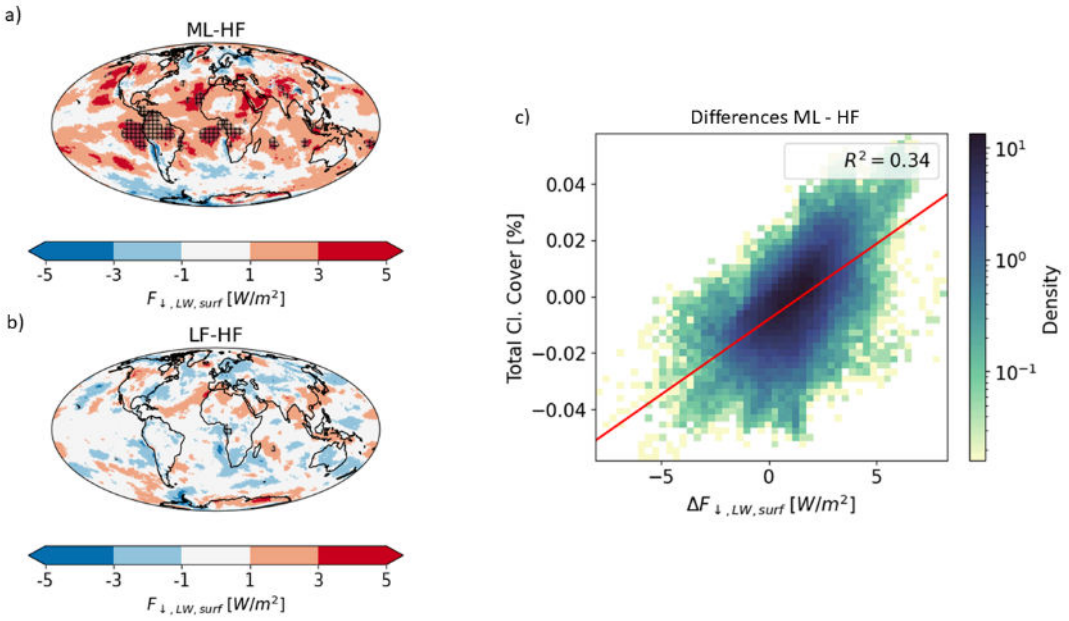


**Figure 5.8.:** Difference in downward LW flux at the surface $F_{\downarrow, LW, surf}$ of ensemble means. Similar to Figure 5.8. Adapted with permission from Hafner et al. (2025b).

convection. Therefore, we analyzed different regions where certain cloud types are most relevant to find potential causes of the slightly different cloud distribution. We looked at three stratocumulus regions of southeast Pacific ($10-30°$S, $75-90°$W), southeast Atlantic ($10-30°$S, $10°$W-$10°$E), and northeast Pacific ($15-35°$N, $120-140°$W) as defined in (Bock and Lauer 2024). We combined these three regions and calculated the annual and domain mean. From this, we calculated the ensemble mean and standard deviation and compared certain variables such as LW heating rate $\partial T_{LW}/\partial t$, cloud fraction, and cloud liquid $q_l$ as shown in Figure 5.9. First of all, the heating rates for all three setups match well (Figure 5.9 a). The cloud fraction in the stratocumulus region shows two peaks, one in the lower and one in the upper troposphere (Figure 5.9 b), where the upper peak can be associated with cloud ice. Cloud liquid is overall small and we observe slight differences near the surface (Figure 5.9 c). Despite the good agreement between the three setups, we find that LW cooling is slightly stronger at 1 km for ML-based emulator. This would lower temperatures; by the Clausius-Clapeyron relation, the saturation vapor pressure decreases, so water vapor reaches saturation more quickly (given our fixed sea surface temperature) and clouds form more easily. Indeed, we observe an increase in cloud fraction and cloud liquid at the same height for ML. Below a height of 500 m, LW cooling is slightly weaker. This would increase the temperature, making ICON-$A_{ML}$ slightly warmer than ICON-$A_{LF}$ and ICON-$A_{HF}$. However, we do not observe a different temperature profile for ICON-$A_{ML}$. Additionally, we looked at the same variables over the tropical pacific, which can be associated with deep convection. There, the profiles of LW heating rate, cloud fraction and cloud liquid match well and small differences as for the stratocumulus region were not observed (see Figure B.7). On another note, the changed clouds

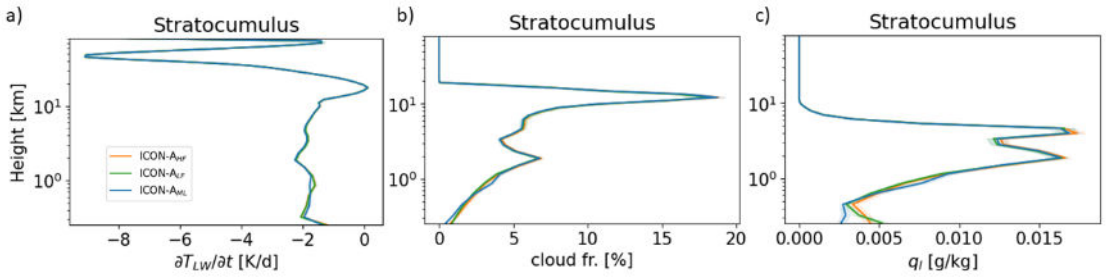


**Figure 5.9.:** Ensemble mean and standard deviation of a) LW heating rate, b) cloud fraction, and c) cloud liquid in the stratocumulus region. Adapted with permission from Hafner et al. (2025b).

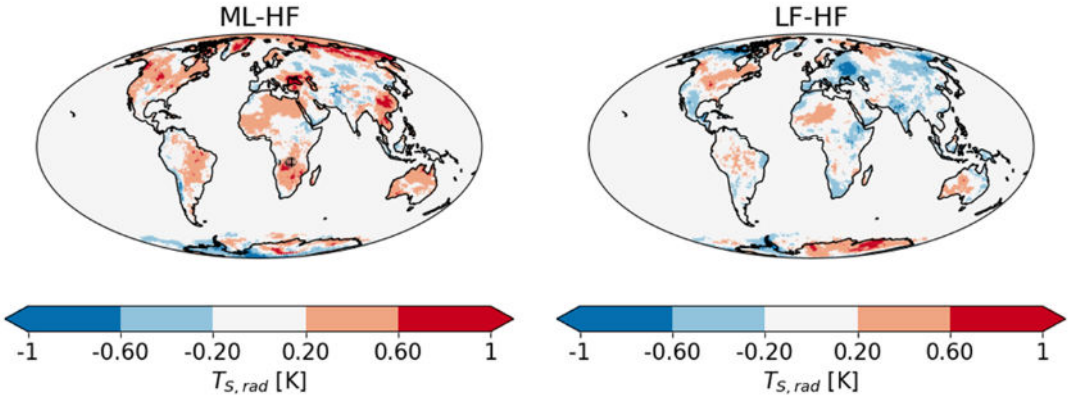


**Figure 5.10.:** Ensemble mean difference of surface temperature of ICON-$A_{ML}$ and ICON-$A_{LF}$ compared to ICON-$A_{HF}$. Adapted with permission from Hafner et al. (2025b).

can affect the circulation (Dinh et al. 2023) and the atmospheric circulation can mix positive and negative errors such that differences may not persist over time (Pincus and Stevens 2009).

### 5.3.5. Surface Temperature

The land surface is sensitive to downward radiative fluxes as they warm the surface and the surface also emits and reflects radiation leading feeding the upward radiative flux in the atmosphere. One parameter to assess the coupling between radiative fluxes and the surface scheme is the surface temperature. Here, we focus on the land temperature as the sea surface temperature is prescribed. The difference between ICON-$A_{ML}$ and ICON-$A_{HF}$ is that the hybrid run is slightly warmer with a bias of 0.05 $K$ and MAE of 0.09 $K$ (Figure 5.10 left). For reference and in order to classify the scale, we also evaluated the difference between ICON-$A_{LF}$ and ICON-$A_{HF}$. Here, the bias is $-0.02$ $K$ and MAE is 0.07 $K$ (Figure 5.10 right). So, the difference induced by using the ML-based radiation emulator on the land surface is similar in scale, but of opposite sign.

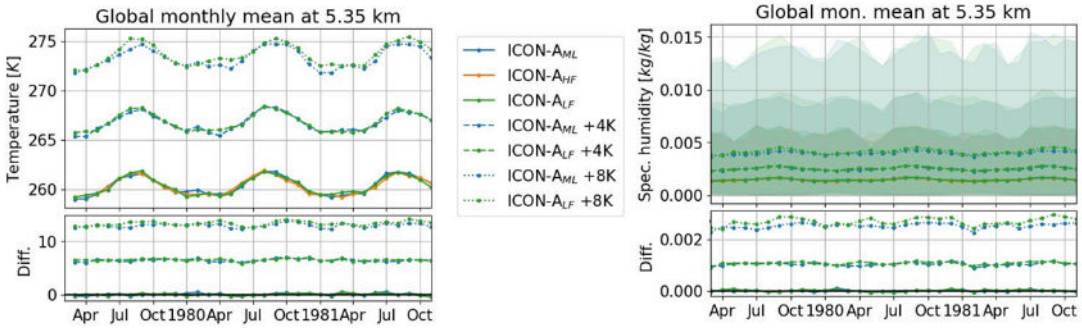


**Figure 5.11.:** Temperature (left) and specific humidity (right) in the troposphere in a +4K (dashed) and +8K (dotted) climate compared to current climate. Specific humidity additionally shows the local minimum and maximum values of the monthly means. Adapted with permission from Hafner et al. (2025b).

## 5.4. Generalizability to Warmer Climates

To test the generalizability, we make simulations as before but add 4 $K$ to the prescribed SSTs. Here, we only compare the warmer simulations with ICON-$A_{LF}$to reduce resource cost. In a warmer climate, the troposphere warms and the air can hold more moisture. Therefore, we analyze the temperature and specific humidity in the troposphere (Figure 5.11) and stratosphere (Fig 5.12). As expected, the temperature is larger than in the reference climate in the troposphere at around 5 $km$ (Figure 5.11 left). Also, the ICON-$A_{ML}$ and ICON-$A_{LF}$ have the same seasonal cycle and deviation from the reference indicating that the ML-based radiation scheme is generalizing well to a warmer surface. Similar for specific humidity, it rises with increasing temperature (Figure 5.11 right). In the stratosphere, the temperature difference is not that large and ICON-$A_{ML}$ and ICON-$A_{LF}$ have a similar temporal pattern (Figure 5.12 left). For specific humidity, it takes about a year to react to the strong SST changes in the stratosphere. For both, ICON-$A_{ML}$ and ICON-$A_{LF}$, the global mean values of specific humidity are still in or at least close to the min-max range of the reference climate. So, the tested warmer climate is not too far from what the ML-based scheme has seen during training which is why it generalizes well.

To test the limits of the ML-based scheme, we do the same as before but add 8 $K$ to the SSTs. The temperature in the troposphere increases twice as much as for the SST +4 $K$ case, similarly for specific humidity (Figure 5.11). Both runs, ICON-$A_{ML}$ and ICON-$A_{LF}$ are close to each other. Therefore, we conclude that the ML-based scheme can also generalize well to SST +8 $K$ in the troposphere. But, in the stratosphere, there is a difference between ICON-$A_{ML}$ and ICON-$A_{LF}$ that increases over time. While the temperature of ICON-$A_{ML}$ stays in a similar range as the reference climate and the SST +4 $K$ case, the temperature of ICON-$A_{LF}$ decreases. The reason becomes evident when looking at specific humidity in the stratosphere (Figure 5.12). After 10 months, specific humidity increases drastically and then keeps increasing with a smaller slope. The min-max values are completely out of range of the reference. The ML-based scheme cannot react to those differences and fails to generalize.

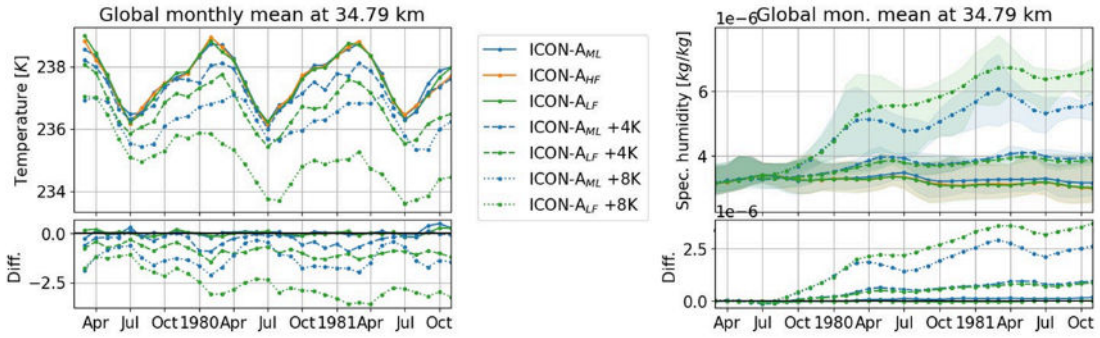


**Figure 5.12.:** Same as Figure 5.11 but for stratosphere. Adapted with permission from Hafner et al. (2025b).

In addition to the temporal evolution of the atmospheric temperature, we looked at the warming pattern for the +4K and +8K runs (Figure 5.13). We compare the annual mean temperature of the lowest model level of the +4K and +8K runs with the unperturbed simulation ICON-A$_{HF}$. For the warmer simulations, we only used the last year of the 3-year long simulation. For ICON-A$_{LF}$+4K and ICON-A$_{ML}$+4K, we observe strong warming over land (Figure 5.13 a and b). However, ICON-A$_{ML}$+4K exhibits stronger warming over land in the northern hemisphere (Figure 5.13 b). We observe similar behavior for the +8K experiments, where the air over land exhibits increased warming (Figure 5.13 c and d). Again, ICON-A$_{ML}$+8K shows more warming over land in the northern hemisphere (Figure 5.13 d).

## 5.5. Conclusion and Discussion

The radiation parameterization is one of the most expensive components of Earth System Models. Therefore, reducing the computational cost is one of the key priorities. There have been many attempts to replace the radiation parameterization with a machine learning based emulator. But, so far only rather simple architectures have been tested online and the longest coupled run for a state-of-the-art general circulation model (GCM) was 7 months (Belochitski and Krasnopolsky 2021).

In this study, we used a BiLSTM architecture to emulate shortwave and longwave heating rates and boundary fluxes, indirectly incorporating both upward and downward radiative fluxes within the bidirectional architecture. These networks were developed and evaluated offline, demonstrating accuracy and physical consistency, as detailed in (Hafner et al. 2025a).

The neural networks emulate the radiation parameterization RTE+RRTMGP (Pincus et al. 2019) used in the ICON-A model (Giorgetta et al. 2018). We compared the hybrid simulation ICON-A$_{ML}$ with simulations using high frequency radiation calls (ICON-A$_{HF}$, 'ground truth') and low frequency radiation calls (ICON-A$_{LF}$, baseline). The ML-based radiation scheme in ICON-A$_{ML}$, is four times faster than ICON-A$_{HF}$ enabling more frequent radiation calls. We demonstrated that ICON-A$_{ML}$ runs stable for at least 10 years and would likely continue to do so over a longer period. The next step is to include GHGs and aerosol interaction with

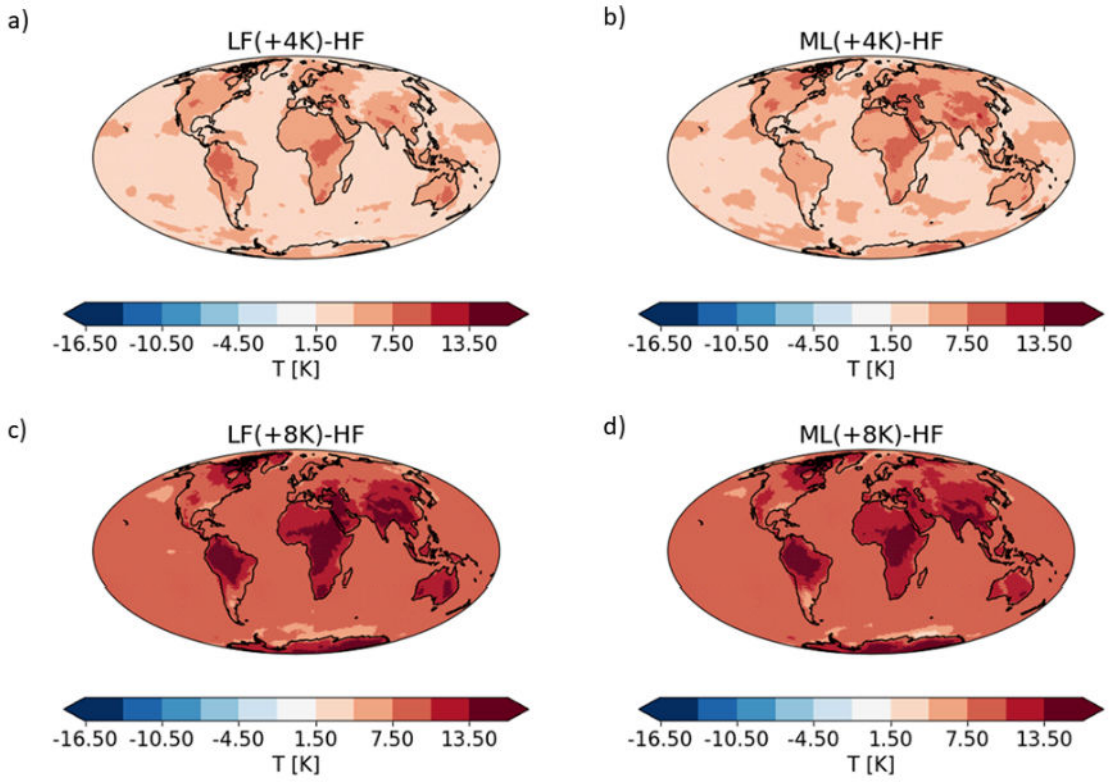


**Figure 5.13.:** Warming pattern for a) ICON-$A_{LF}$+4K, b) ICON-$A_{ML}$+4K, c) ICON-$A_{LF}$+8K, d) ICON-$A_{ML}$+8K. The annual mean air temperature of the last year in the lowest model level is compared with the unperturbed simulation ICON-$A_{HF}$. Adapted with permission from Hafner et al. (2025b).

radiation and compare the hybrid run to observations. However, when including aerosols, the larger contribution to effective radiative forcing is through aerosol-cloud interactions while a smaller fraction is due to direct aerosol-radiation interactions (Forster et al. 2021b). Therefore, it is important to get the cloud-radiation interactions correct.

Nevertheless, small differences between ICON-$A_{ML}$ and ICON-$A_{HF}$ persist, likely due to interactions between radiative heating, circulation, and cloud distribution. Such differences are expected, given that a GCM like ICON-A is a highly non-linear system where small changes can lead to different outcomes. It is, however, a notable success that ICON-$A_{ML}$ reaches a new equilibrium very close to the reference simulations, indicating that the ML-based scheme operates in a physically consistent manner. For the energy balance at the top of the atmosphere, we even see a slight improvement of the bias compared to ICON-$A_{LF}$. This outcome also open doors for further research to explore the feedback mechanisms that contribute to new equilibrium state. Additionally, experiments with coupled ocean would be an interesting next step to analyze feedbacks between the atmosphere and the ocean induced by radiative surface fluxes.

A test with increased SSTs showed that the ML-based radiation scheme can in fact generalize to a warmer climate where the SSTs are increased by 4 $K$. Even in an extreme case of SST +8 $K$,

the ML-based scheme can generalize well in the troposphere. But once we leave the 'known' range, the ML-based scheme starts to struggle as seen in the stratosphere. This motivates further development of climate-invariant formulations (Beucler et al. 2024) or training sets that cover a wide range of different climates.

# 6. Representing Subgrid-Scale Cloud Effects in a Radiation Parameterization using Machine Learning

For coarse-scale models, which are often used for climate projections, clouds are not resolved on horizontal scales of $100 - 200\ km$ (Chen et al. 2021). To incorporate fractional cloudiness, models often use statistical schemes i.e., McICA (Pincus et al. 2003) for clouds or cloud effect on radiation where g-points are randomly assigned as cloudy or clear depending on cloud fraction. This simplification introduces uncertainties in cloud-radiation interactions due to randomness, which can be large. Random errors can be as large as $100\ W/m^2$, but average out with the flow within an ESM (Pincus et al. 2003). Some models use the *all-or-nothing* cloud cover scheme which is a good approximation for high-resolution simulations (Giorgetta et al. 2022; Hohenegger et al. 2023). With more high-resolution model data available, there are new possibilities of incorporating fractional cloudiness in coarse ESMs. Global Storm Resolving Models (GSRMs) can be used to generate data and a NN can learn the subgrid distribution of clouds from the underlying statistics. As a result, coarse-grained data has subgrid-scale effects implicitly included in the variables. This motivates the key science question 3: "Can we improve the representation of cloud-radiation interactions by using high-resolution data for training?". First, a baseline needs to be defined to assess the improvement of cloud radiative impact on heating rates when comparing to the *ground truth*, which is the coarse-grained high-resolution simulation. The baseline is the physics-based radiation scheme RTE+RRTMGP (Pincus et al. 2019), which is used also used in ICON. This naturally poses the question how a coarse-scale radiation scheme would perform on coarse-grained data.

This chapter focuses on the improvement of the cloud radiative impact on heating rates and comparison to a coarse-scale reference scheme. This study has been presented in Hafner et al. (2025c), which has been submitted to the journal *Geoscientific Model Development*. As indicated in Section 1.3, I produced all content that is presented from this publication including text, figures, tables and code[1].

This chapter is structured as follows: In Section 6.1, we describe the method of learning the cloud radiative impact from high-resolution simulation data. Then, we described the specifics about the coarse- and high-resolution simulations and compare selected variables (Section 6.2). In Section 6.3, we present the results of the coarse-scale and the ML-enhanced

[1] https://github.com/EyringMLClimateGroup/hafner25GMD_MLe_radiation (last accessed: 07.10.2025 and archived at (Hafner 2025a))

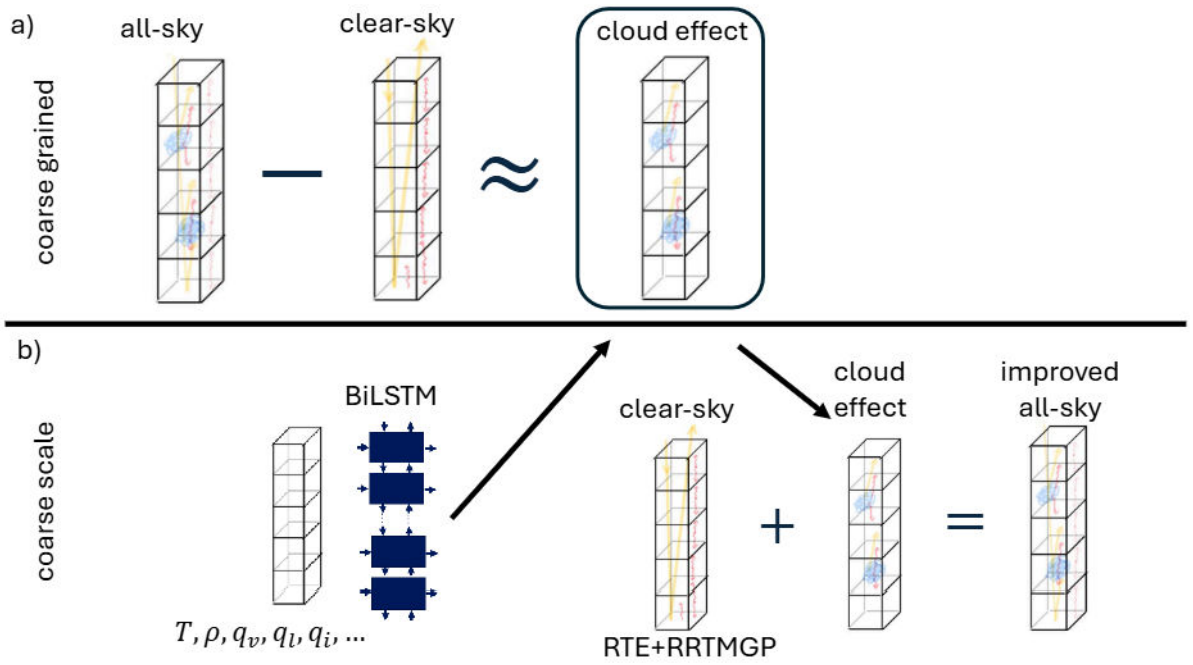


**Figure 6.1.:** Sketch of constructing the cloud radiative impact on heating rates. Radiation schemes calculate fluxes for the same scene once with and once without clouds resulting in all-sky and clear-sky fluxes. The corresponding heating rates can be inferred from the fluxes and the residual yields an approximation of the cloud radiative impact on heating rates for all layers in a column. Adapted with permission from Hafner et al. (2025c).

radiation scheme evaluated on coarse-grained data. This chapter finalizes with a summary and discussion in Section 6.4.

## 6.1. Learning the Cloud Radiative Impact on Heating Rates

We define the cloud radiative impact on heating rates in a column as the residual between the all-sky and clear-sky heating rates, where clear-sky represents the same atmospheric conditions as all-sky but without clouds (Figure 6.1 a). In ESMs such as the ICON model (Giorgetta et al. 2018; Giorgetta et al. 2022), radiation parameterizations like RTE+RRTMGP (Pincus et al. 2019) first calculate gas optical properties and then add cloud optical properties. Then, these combined properties are used to calculate all-sky radiative fluxes, and clear-sky fluxes can be obtained omitting the cloud optical properties. Heating rates are then derived from the flux divergence for both all-sky and clear-sky conditions. The residual of these heating rates, which is the Cloud Radiative Impact (CRI) on the heating rates, serves as the training target for the ML-based radiation parameterization:

$$\frac{\partial T_{CRI}}{\partial t} = \frac{\partial T_{all-sky}}{\partial t} - \frac{\partial T_{clear-sky}}{\partial t}. \tag{6.1}$$

The heating rates $\frac{\partial T_k}{\partial t}$ in a layer $k$ are calculated from the net flux $F_{Net}$ at the layer boundaries $k \pm \frac{1}{2}$

$$\frac{\partial T_k}{\partial t} = \frac{F_{Net,k+1/2} - F_{Net,k-1/2}}{c_v m_{air}}, \tag{6.2}$$

where $m_{air}$ is the mass of moist air per area, and specific heat at constant volume $c_v$ scales with the tracer mixing ratios.

The central idea is to train a neural network (NN) that learns only the cloud impact on heating rates (see Figure 6.1). Cloud-radiation interactions are subject to large uncertainties, since the subgrid-scale horizontal and vertical distributions of clouds are not resolved in coarse-scale ESMs. In the hybrid ML-physics radiation parameterization (Figure 6.1 b), the NN predicts only the cloud radiative impact, while the clear-sky component is still computed by the original radiation parameterization. This design ensures that the ML-enhanced radiation scheme retains sensitivity to changes in GHG and aerosols through the clear-sky part, potentially improving generalization across different climates. The ML-cloud component can respond to GHG and aerosols changes only indirectly, for example through modifications of the cloud distribution. However, this hybrid approach does not capture secondary effects arising from reflected radiation.

At first glance, a linear decomposition of the clear-sky heating rate and the cloud radiative impacts on heating rates may seem counterintuitive, given the inherently nonlinear interaction among specific humidity, clouds, and trace gases such as ozone (Bony et al. 2015). Nevertheless, we adopt a linear decomposition framework in this work, as illustrated on Figure 6.1. Within this framework, the NN is tasked with learning the nonlinear relationships based on the prevailing atmospheric conditions and cloud-related variables (e.g., cloud liquid water and cloud ice).

### 6.1.1. Method

We use a BiLSTM based on Hafner et al. (2025a) to learn the cloud radiative impact in an atmospheric column. Bidirectional architectures are particularly well-suited for radiative transfer problems (Bertoli et al. 2025; Ukkonen and Chantry 2025; Ukkonen 2022; Yao et al. 2023). Unlike their common usage for temporal sequences, BiLSTMs for radiation scan the vertical dimension in both directions, resembling upward and downward fluxes. The NN consists of one BiLSTM layer with $tanh$ activation and a hidden dimension of 96, with two LSTMs scanning the input vertical profiles in upward and downward directions. The combined output of the LSTMs is then processed by a dense layer that predicts the heating rate at each level. The training is split between SW and LW radiation as shortwave temperature tendencies are only calculated for sun-lit areas and longwave temperature tendencies are always computed, totaling in 82k trainable parameters per NN.

The input variables are vertical profiles of mass mixing ratios of specific humidity $q_v$, cloud liquid $q_l$, cloud ice $q_i$, snow $q_s$ and ozone $O_3$, plus the cloud fraction $cl$, air density $\rho$, and temperature $T$. For SW, we additionally use surface albedo $\alpha$ and incoming solar flux at the

top of the atmosphere $F_{\downarrow,TOA,SW}$, which is the solar constant weighted by the solar zenith angle and change in Earth-Sun distance to account for daily and seasonal variations. For LW, we additionally use the surface temperature $T_{surf}$ as input. As mentioned above, the output is the cloud radiative effect on heating rates derived from all-sky and clear-sky heating rates (Figure 6.1).

Normalization is important for faster convergence of the training (LeCun et al. 2012) and generalization (Beucler et al. 2024). $O_3$, $\rho$, $T$, and $T_{surf}$ are normalized using their mean values $\mu$ and standard deviation $\sigma$, also known as z-score normalization ($x_{norm} = \frac{x-\mu}{\sigma}$). The water related variables $q_l$, $q_i$, $q_s$ are normalized by the ambient total (radiativly active) water ($q_v + q_l + q_i + q_s$). Here, radiatively active means used in the radiation scheme. $q_v$ uses z-score normalization providing information that relates to the absolute mass mixing ratios. $cl$ and $\alpha$ are not normalized as they already vary between 0 and 1. $F_{\downarrow,TOA,SW}$ is normalized by 1360 $\frac{W}{m^2}$, which is close to the solar constant. The cloud radiative impact on heating rates is only converted to K/d, which is on the order of one.

The loss we minimize during training consists of the sum ofMAE and MSE. We use the Adam optimizer (Kingma and Ba 2017), and set a learning rate of $10^{-3}$, which is reduced by a factor of 2 when the validation loss does not decrease by 0.1% for 5 epochs. To avoid overfitting, we use early stopping, which stops the training if the validation loss does not decrease for 10 epochs.

## 6.2. Data

ICON is a weather and climate model permitting simulations across different resolutions, from a few to hundreds of kilometers. For global and long-term applications, the atmospheric component ICON-A (Giorgetta et al. 2018) often runs at 80 km horizontal resolution, and 47 vertical levels covering the altitude range $0 - 83$ km, with parameterizations for radiation, cloud microphysics, turbulence, convection and gravity waves. The ICON model has the option to be run as a GSRM (Giorgetta et al. 2022; Hohenegger et al. 2023; Stevens et al. 2019). The high-resolution simulations used here follow the QUBICC protocol from Giorgetta et al. (2022). The QUBICC simulations have 5 km horizontal resolution, and the vertical dimension spans 83 km on 191 levels. The high-resolution allows the model to run without a convection scheme and gravity wave parameterization, as these processes are starting to be resolved.

We performed QUBICC simulations that cover a total of 40 days evenly distributed across four months: November 2004, January, April, July 2005. The simulations have a physics time step of 40 s and a radiation time step of 6 min. These simulations are run with prescribed sea surface temperatures, sea ice concentrations, greenhouse gas concentrations but no aerosols. The outputs are saved every 192 min. The uneven output interval was chosen to cover a large variability of different solar zenith angles at different locations. The first 6 days of each 10-day period is used for training, the next 2 days for validation and the last 2 days for testing. Using QUBICC data to train our ML model for ICON-A requires coarse-graining the high-

resolution QUBICC dataset. All variables are horizontally and vertically coarse-grained from high-resolution simulations as in Grundner et al. (2022). We discarded a few coarse-grained cells, e.g., if the surface height of the coarse-grained cell deviated by more than 0.5 m from the coarse-scale surface height. Then, we randomly sampled 35k and 5k grid points per time step for the training and validation set. For the test set, we randomly selected 75k cells per time step for LW and 35k cells for SW. This yields 2.3 million training samples, 260k validation samples, 1.9 million test samples for shortwave and 4.2 million test samples for longwave radiation.

In order to evaluate the high-resolution data, we also made coarse-scale ICON-A simulations for the same periods and with a similar configuration at 80 km horizontal resolution to compare differences in various distributions. The ICON-A simulations are based on the version 2.6.4 described in Giorgetta et al. (2018), while the QUBICC simulations are based on the version icon-2024.10 (ICON partnership (DWD and MPI-M and DKRZ and KIT and C2SM) 2024). To make the coarse-scale simulation more comparable, we used the same microphysics scheme (Doms et al. 2011) and no aerosols. See Appendix for a comparison with the default microphysics scheme. Both simulations use the radiation scheme RTE+RRTMGP (Pincus et al. 2019). The physics and radiation time step is 6 min for the ICON-A simulation.

The remaining differences between the coarse-scale ICON-A and high-resolution QUBICC simulation include the horizontal and vertical resolution, the higher temporal resolution of physical process in QUBICC, which is required due to the high spatial resolution. Moreover, QUBICC intends to resolve gravity waves and (deep) convection. Additionally, the radiation scheme RTE+RRTMGP (Pincus et al. 2019) in QUBICC uses snow mixing ratio as an input. Other differences could be due to differences between the code versions and tuning, as we did not retune ICON-A.

### 6.2.1. Comparison of input and output variables

If an ML-based scheme is trained on high-resolution simulations like QUBICC and the goal is to couple it to a coarse-scale model like ICON-A, one needs to ensure that the distributions of input and output variables match. Otherwise, the ML-based scheme could be faced with out-of-distribution samples, which can lead to errors that quickly build-up while the model is integrated (Rasp 2020). Therefore, we analyze systematic differences between the coarse-scale ICON-A simulations and the high-resolution QUBICC simulations. This analysis is conducted by comparing the spatial and temporal means and spread in the input and output used and produced by the radiation parameterization. Specifically, we focus on $q_v$, $q_l$, $q_i$, $cl$, total cloud cover, $q_s$, as well as longwave and shortwave heating rates. For the comparison, we use all samples in the test set. The samples of both simulations cover the same time period which is November 2004, January, April, and July 2005. Comparing two simulations with different grids enables us to uncover systematic differences, with a focus on identifying larger variations.

Distributions of water related input variables are shown in Figure 6.2. The distributions of specific humidity look similar in the coarse-grained and coarse-scale data set, where the

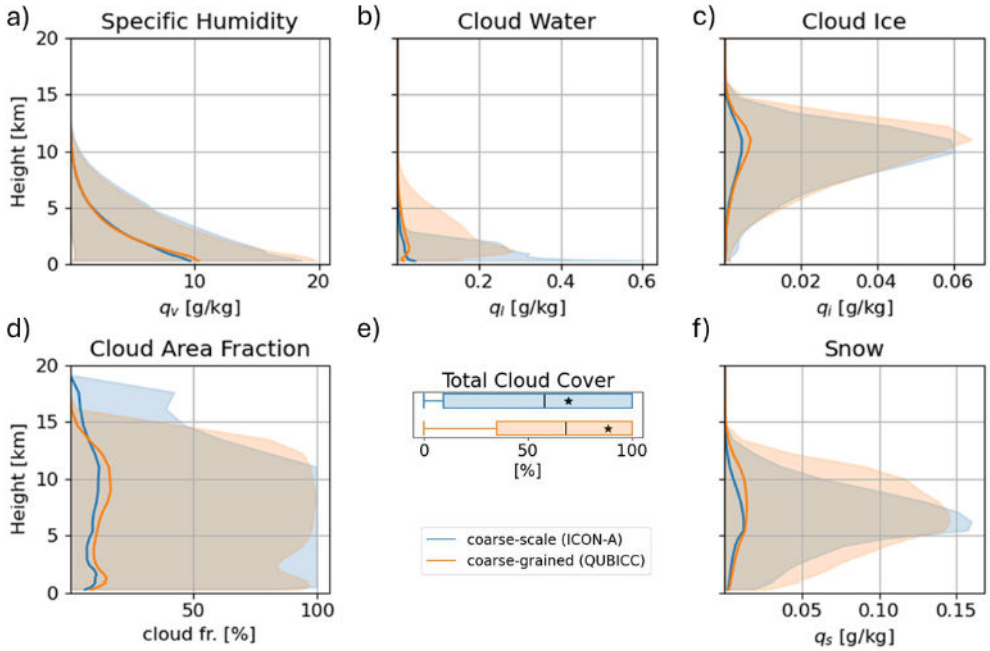


**Figure 6.2.:** Distributions of water related input variables for ICON-A and coarse-grained QUBICC data. The bold line shows the mean and the shaded area shows 95% of the spread between the 2.5% percentile and the 97.5% percentile. The boxplot is limited by the minimum and maximum values. The box edges are defined at the 25% and 75% percentile of the distribution. The black line illustrates the mean of the distribution and the star is the median. Adapted with permission from Hafner et al. (2025c).

maximum differences between the means is 0.9 g/kg (Figure 6.2 a). The spread in humidity values also overlaps for both simulations. Cloud water has higher values below 3 km for the coarse-scale simulation while cloud water is more evenly distributed throughout the troposphere for the QUBICC simulations (Figure 6.2 b). Here, the maximum difference of the mean values is 0.02 g/kg. The distributions of cloud ice have similar shapes, but the coarse-grained distribution peaks higher by 0.002 g/kg (Figure 6.2 c). The mean cloud area fraction is on average larger by 3 % for the coarse-grained data set below 14 km (Figure 6.2 d). Snow peaks between 5 – 10 km (Figure 6.2 f). However, the vertical distribution is wider in the coarse-grained simulation. Nevertheless, snow was not used as input for radiation in ICON-A.

For comparability, the coarse-grained heating rates from QUBICC had to be rescaled by a factor $c_v/c_p$ to account for differences in the two model versions used, where $c_p$ is specific heat at constant pressure.

For SW all-sky heating rates, the mean profiles match very well and the mean difference is 0.18 K/d (Figure 6.3 a). There are only small differences in the spread of heating rates in the troposphere. The heating rate is decomposed into a clear-sky heating rate—which is calculated from the clear-sky fluxes—and the cloud radiative impact on heating rates. Their distributions are also shown in Figure 6.3 b-c. The SW clear-sky heating rate has a mean difference of 0.11 K/d for the mean profiles. For the SW cloud radiative impact, the mean

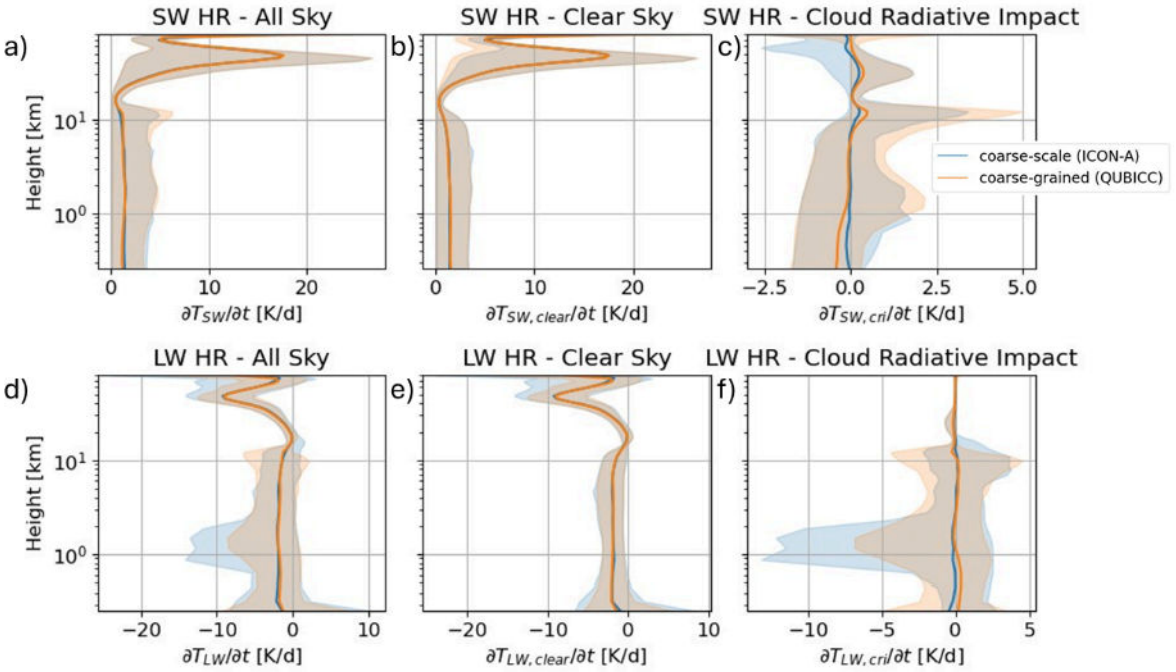


**Figure 6.3.:** The distribution of shortwave (top row) and longwave (bottom row) heating rates in coarse-scale and scaled coarse-grained data. The bold line shows the mean and the shaded area shows 95% of the spread, which is defined as the spread between the 2.5% percentile and the 97.5% percentile. The left column shows all-sky heating rates as it is used in the ICON model. The middle column shows clear-sky heating rate computed from clear-sky fluxes, which is a diagnostic output in the ICON model. The right column shows the cloud radiative impact on the heating rate which was computed by subtracting the clear-sky heating rate from the all-sky heating rate. Adapted with permission from Hafner et al. (2025c).

difference is also 0.11 K/d. Here, we expect small differences due to mostly resolved clouds in the coarse-grained dataset. However, there is no clear bias between coarse-scale and coarse-grained cloud impacts.

For the LW all-sky heating rates, the mean profiles match well and have a mean difference of 0.18 K/d (Figure 6.3 d). However, the spread in heating rates is slightly different, which may be due to the different spread in cloud water at 1 km (Figure 6.2 b). The LW clear-sky heating rates look very similar in their mean values and their spread where the mean difference is 0.14 K/d (Figure 6.3 e). The LW cloud impact is concentrated in the troposphere (Figure 6.3 f). The mean impact is almost the same between coarse-scale and coarse-grained simulations with a mean difference of 0.09 K/d. However, there is a difference in spread, which can be expected, because the coarse-grained data set implicitly includes subgrid-scale cloud effects.

In the unscaled comparison, the clear-sky heating rates show the same bias as the all-sky heating rates for both SW and LW (Figure C.2). However, this bias does not directly translate to the cloud radiative impact on heating rates because adding the cloud impact is a highly non-linear process (Bony et al. 2015). Additionally, this indicates that the mean cloud effect is similar for a (quasi)-hydrostatic and a non-hydrostatic model.

## 6.3. Results

For the coarse-scale radiation reference, we use the Python front-end (hereafter pyRTE) (Pincus et al. 2025) of the radiation scheme RTE+RRTMGP (Pincus et al. 2019), which is used in all our simulations. Using pyRTE, we replicate the implementation of radiation in QUBICC as closely as possible. For subgrid-scale variability, we employ McICA (Pincus et al. 2003) together with maximum-random cloud overlap (Räisänen et al. 2004). The procedure is as follows: first we calculate gas optical properties, assign them to the atmospheric state and calculate clear-sky fluxes. Next, we compute cloud optical properties, apply McICA with maximum-random overlap to represent subgrid-scale variability, and add cloud optical properties to the gas optical properties. Because QUBICC also includes snow in its radiation parameterization, we additionally calculate snow optical properties and combine them with the other optical properties, before computing all-sky fluxes. Then, heating rates are obtained applying Equation (6.2). We calculate heating rates for all samples in the test dataset and compare them to the coarse-grained heating rates. The same procedure is applied to the ML-enhanced heating rates, which are compared with pyRTE. As evaluation metric, we use MAE, bias, and the coefficient of determination ($R^2$). The results are presented in Figure 6.4 with the column-averaged metrics summarized in Table C.1. We restrict the presentation to the lowest 20 km of the atmosphere, as cloud impacts are most relevant in the troposphere (see Figure 6.3). Nevertheless, the NN predicts the cloud impact on the entire atmospheric column, and the results for the full column can be found in the appendix C.3.

The first column of Figure 6.4 shows clear-sky heating rates, for which the cloud distribution plays no role. Accordingly, only small and statistically insignificant differences are expected, arising mainly from variability in water vapor. For pyRTE, the assumption of horizontally homogeneous input parameters introduces a small error of 0.367 K/d (SW) and 0.571 K/d (LW). For comparison with other studies, Table C.1 also reports the RMSE. For clear-sky heating rates from pyRTE, the RMSE is 0.443 K/d (SW) and 0.688 K/d (LW) compared with the coarse-grained QUBICC rates. Hogan and Matricardi (2022) developed a fast tool for computing gas-optical properties and reports an RMSE of 0.18 K/d. Although the error source differs (gas optical properties vs. spatial resolution), the magnitudes are comparable.

For completeness, we also evaluated the ML-model on clear-sky samples. However, it is not intended for clear-sky scenes as the cloud impact is zero. Therefore, the corresponding panels are grayed out. The MAE is 0.049 K/d for SW and 0.028 K/d for LW.

The second column of Figure 6.4 shows results for fully cloudy samples (total cloud cover of 100%). For pyRTE, the MAE peaks near 10 km, exceeding 5 K/d for both SW and LW. The corresponding $R^2$ are low, with average values of 0.83 (SW) and 0.66 (LW), compared to 0.98 for the ML-enhanced scheme. $R^2$ is weighted by variability, and an $R^2$ of zero indicates that the error is as large as the variability itself. Since, McICA within pyRTE is supposed to produce unbiased noise, the $R^2$ is therefore a less informative metric. In contrast, the bias reveals that pyRTE systematically struggles to represent the cloud impact near 10 km, particularly for SW.

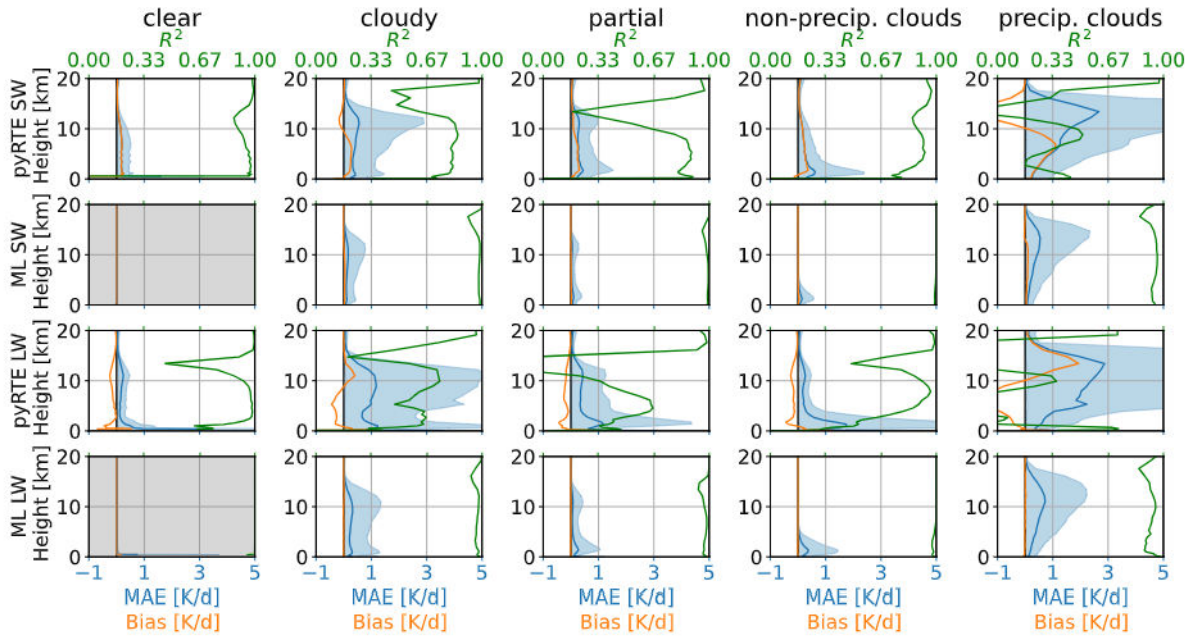


**Figure 6.4.:** Comparison of the pyRTE and the hybrid ML-based radiation scheme on coarse-grained QUBICC data. Results are shown for the shortwave (top rows) and longwave (bottom rows) spectral range. The results are shown separately for clear-sky samples (no clouds, left column), fully cloudy sky samples (middle column) and samples with partial cloudiness (right column). The shown metrics are coefficient of determination $R^2$ (green), bias (orange) and MAE (blue) with 95% of the spread, which is defined as the spread between the 2.5% percentile and the 97.5% percentiles. The bias and MAE share the x-axis. The ML-clear-sky panels are gray because the clear-sky fluxes are not calculated by the ML model, but shown as reference. Adapted with permission from Hafner et al. (2025c).

The ML-enhanced scheme produces nearly unbiased heating rates, with MAEs of 0.106 K/d (SW) and 0.127 K/d (LW), representing errors 4-6 times smaller than those from pyRTE.

The third column of Figure 6.4 shows results for partially cloudy scenes, defined here as total cloud cover between 10-90%. In these cases, both schemes exhibit smaller errors than in fully cloudy scenes, consistent with the weaker overall cloud radiative impact. For pyRTE, the bias is substantially smaller than fully cloudy conditions but still show a pronounced peak near 1 km for LW and a double peak at 1 and 10 km for SW. In contrast, the ML-enhanced samples produces nearly unbiased heating rates, with an MAE of 0.082 K/d (SW) and 0.068 K/d (LW), representing errors that are 5-10 times smaller than those for pyRTE.

To interpret the double-peak error observed in partially cloudy samples, we divided the samples into precipitating and non-precipitating clouds, as a rough proxy for deep and shallow convection. Non-precipitating (warm) clouds were identified using thresholds of maximum 0.01 mm/h, total cloud cover larger than 10%, and vertically integrated ice water path of less than 1 $\frac{g}{m^2}$. For reference, drizzle has a precipitation rate of 0.2 mm/d (Wood 2012). Precipitating clouds were identified by total cloud cover > 10% and precipitation more than 3 mm/h. For reference, Zhao et al. (2024) reports mean precipitation of 3.5 mm/h for deep convective cores over the tropical ocean. For non-precipitating clouds, both pyRTE and the ML-enhanced scheme show an MAE peak near 1 km for SW and LW. However, the ML-

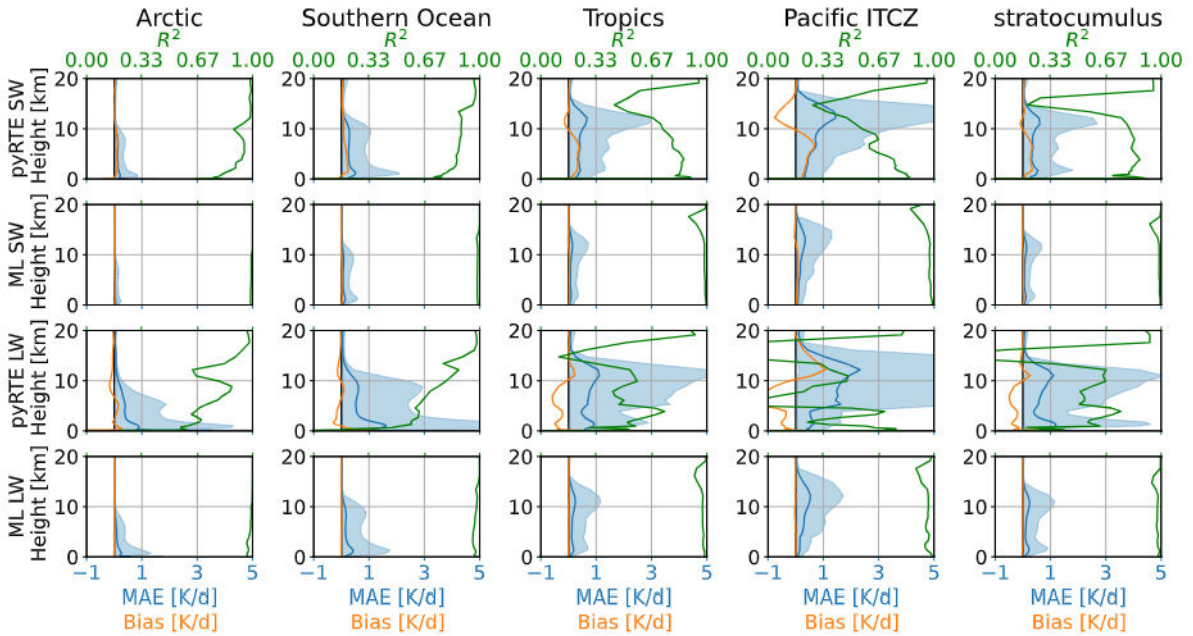


**Figure 6.5.:** As Figure 6.4, but for the selected regions shown in Figure 5 of Bock and Lauer (2024). Adapted with permission from Hafner et al. (2025c).

enhanced scheme achieves substantially smaller errors of 0.080 K/d for SW and 0.069 K/d for LW, which are 6-10 times lower than those of pyRTE. For the precipitating clouds, pyRTE exhibits an MAE peak at 10 – 12 km, while the ML-enhanced scheme shows enhanced error in the upper troposphere but without distinct peak. Instead, the ML-enhanced scheme shows a broader peak in MAE between 12 – 14 km. On average, however, the ML-enhanced error remain about 4-5 times smaller than those from pyRTE.

We further evaluated the performance of the ML-enhanced scheme across five selected regions characterized by different predominant cloud regimes, following the classification of Bock and Lauer (2024). The results are shown in Figure 6.5 and summarized in Table C.2. In the arctic region (70-90°N), errors remain confined below 10 km consistent with the lower tropopause height in this region. For pyRTE, the SW MAE is 0.215 K/d, even smaller than in the clear-sky conditions, although the spread in MAE is slightly larger. For LW, the MAE is 0.614 K/d, exceeding the clear-sky values. In contrast, the ML-enhanced scheme achieves errors that are 4-8 times smaller than those of pyRTE.

In the Southern Ocean (30-65°S), pyRTE exhibits large errors of 0.417 K/d for SW and 0.760 K/d for LW, with a characteristic double peak in the MAE at 1-2 and 10 km. The ML-enhanced scheme also reproduces this double peak structure in the MAE spread but reduces the MAE by a factor of 4-7 relative to pyRTE. Over the tropical ocean (30°N-30°S), pyRTE shows large errors in the upper troposphere, likely associated with deep convection. In this region, the ML-enhanced scheme again reduces the MAE by a factor of 5-9. A subregion within the tropical region, the Pacific Intertropical Convergence Zone (ITCZ) region (0-12°N, 135°E-85°W), shows similar behavior but with even larger errors at higher altitudes.

The stratocumulus region is represented by three sub-regions: the Southeast Pacific (10-30°N, 75-97°W), Southeast Atlantic (10-30°S, 10°W-10°E), and Northeast Pacific (15-35°N, 120-140°W). In these regions, both models show two peaks in the MAE: one in the lower troposphere at $1 - 2\,\text{km}$ and another in the upper troposphere at $10 - 12\,\text{km}$. Notably, the upper tropospheric peak is larger in both SW and LW for both models. Nevertheless, the ML-enhanced scheme achieves errors 5-9 times smaller than those from pyRTE.

## 6.4. Conclusion

ESMs struggle to represent subgrid-scale cloudiness and commonly rely on statistical schemes, such as McICA, to account for subgrid-scale cloud radiative effects (Pincus et al. 2003). Although random, unbiased errors can be mitigated by large-scale atmospheric mixing (Pincus et al. 2003), the column-by-column error can be large. ML algorithms trained on high-resolution, global storm-resolving simulations now provide an opportunity to represent fractional cloudiness in radiative transfer more accurately. To bridge scales, the high-resolution model output is coarse-grained to the target resolution, such that the coarse-grained variables implicitly retain subgrid-scale effects. Then, the coarse-grained variables implicitly include subgrid-scale effects.

We developed a hybrid physics-ML radiation parameterization, where the physics-based component computes clear-sky fluxes, while the ML component predicts cloud impact, implicitly accounting for subgrid-scale variability. This ML-enhanced framework offers a more robust and generalizable radiation scheme: the physics-based parameterization retains its responsiveness to changes in GHGs and aerosols, thereby mitigating potential out-of-distribution issues in climate projections. The ML component is implemented as a BiLSTM neural network, which has previously demonstrated strong performance in radiation applications (Bertoli et al. 2025; Hafner et al. 2025a; Ukkonen and Chantry 2025; Ukkonen 2022; Yao et al. 2023). For training, we use data from high-resolution QUBICC simulations with a horizontal resolution of $\approx 5\,\text{km}$ and 191 vertical layers expanding up to 83 km. These fields are coarse-grained to $\approx 80\,\text{km}$ and 47 vertical layers, matching the target resolution for a coarse-scale ESM. For comparison and to assess systematic differences between high-resolution and coarse-scale models, we additionally perform a coarse-scale ICON-A simulation. The distributions of the relevant input variables are found to be comparable between the coarse-scale and coarse-grained simulations.

We find that a coarse-scale radiation scheme such as pyRTE performs well for clear-sky samples, but exhibits large errors in cloudy conditions, reflecting its inability to represent subgrid-scale distributions. In contrast, the ML-enhanced radiation scheme consistently outperforms pyRTE, reducing errors from unresolved clouds in the radiative transfer calculations by a factor of 4-10. Although the ML-enhanced radiation scheme does not explicitly resolve subgrid-scale distributions, it learns how specific combinations of grid-scale mean states map to heating rates that implicitly include subgrid effects. pyRTE showed substantial biases of

$1 - 3\,\mathrm{K/d}$ in the upper troposphere at $10 - 15\,\mathrm{km}$ for precipitating clouds, highlighting the strong influence of subgrid-scale cloud ice on heating rates. In general, both pyRTE and the ML-enhanced scheme produce larger errors in cloudier conditions, but the ML-enhanced scheme consistently yields smaller errors. These results emphasize the need for a more explicit treatment of subgrid-scale clouds, particularly in the upper troposphere.

Therefore, we conclude that high-resolution model data combined with ML can improve the representation of cloud-radiation interactions in coarse-scale radiation parameterizations. Nevertheless, the presented approach has caveats. High-resolution simulations at 5 km horizontal resolution cannot resolve shallow convection directly, leaving associated cloud radiative effects on heating rates—particularly within the planetary boundary layer—unresolved (Stevens et al. 2019). Using finer horizontal resolutions could help reduce these uncertainties. In addition, aerosols and heterogeneous GHG concentrations are typically absent from current high-resolution models. If future simulations include substantial variability in GHG and aerosol concentrations, these could be incorporated as additional NN inputs to capture secondary effects of reflected radiation.

One of the next steps is the online implementation of the ML-enhanced radiation scheme in a coarse-resolution model such as ICON-A. Although the online stability remains to be tested, the comparison with the coarse-scale model and results from previous stable hybrid simulations (Hafner et al. 2025b) are promising, suggesting potential improvement for climate projections. An additional advantage of the presented scheme is that clear-sky fluxes can be computed less frequently, while the cloud radiative impact can be updated every time step. This provides a pathway to both reducing computational costs and improving the representation of cloud-radiation interactions.

# 7. Conclusion and Outlook

## 7.1. Summary

Radiation is one of the most important processes that need to be represented in ESMs as it is driving the climate, heating and cooling the Earth system. It requires precise knowledge about the absorption spectrum for all gases, aerosols and condensed water particles present in the atmosphere and the integration over many wavelengths to accurately calculate radiative transfer in the atmosphere, which can be calculated by *line-by-line* models. Radiation parameterizations have been developed to reduce the computational costs of a line-by-line model (Hogan and Bozzo 2018; Pincus et al. 2019) by making several simplifications, e.g., using a reduced number of integration points in wavelength spectrum. However, the radiation parameterization is still a computationally expensive component in ESMs. As a consequence, the radiation parameterization is calculated less frequently than other parameterizations, every $1-2\ h$ as opposed to fast processes computed every $10\ min$. This leads to the issue that the radiation scheme does not "see" fast changes in the state variables and cloud distributions, and the resulting heating tendencies cannot reflect the effects of these changes. Additionally, ESMs have long standing biases such as the shortwave cloud radiative effect with local deviations of up to $30\ W/m^2$ (Lauer et al. 2023). A faster radiation scheme that is called as often as other physical processes and thus reacting to changes in clouds, will not improve these biases but would allow for faster, more rigorous testing of other components, e.g., a different microphysics scheme, and diagnosing where these biases come from. Therefore, a faster radiation scheme has the potential to find sources of inconsistencies faster and ultimately reduce these biases. Considering the above mentioned limitations, benefits can be expected by leveraging ML to improve radiative transfer parameterizations. Therefore, this thesis answered the following key scientific questions:

1. Can an ML-based emulator learn to represent atmospheric radiative transfer based on physical laws?
2. Can the ML-based emulator be coupled to a complex ESM like ICON and produce stable simulations?
3. Can we improve the representation of cloud-radiation interactions by using high-resolution data for training?

Even though the development of ML-based radiation schemes already started two decades ago (Chevallier et al. 1998; Krasnopolsky et al. 2005), the development of more advanced

computing hardware like GPUs and the accelerated progression in ML-based algorithms sparked interest in the development of ML-based radiation schemes in the past few years, as shown by recent research activities (Song and Roh 2021; Ukkonen 2022; Yao et al. 2023). With the increased availability of approaches for ML-based radiation schemes, one may wonder whether an NN based architecture can learn physical relationships. This question, formulated in the key science question 1 "Can an ML-based emulator learn to represent atmospheric radiative transfer based on physical laws?", was addressed in Chapter 4 and presented in Hafner et al. (2025a).

In this first study, I developed an ML-based radiation scheme to replace the RTE+RRTMGP radiation parameterization used in the German weather and climate community model ICON. The goal of the ML-based scheme is to emulate the physics-based radiation scheme accurately but at lower computational cost. Bidirectional NN architectures for ML-based radiation schemes have been shown to be superior when compared to other architectures like feed forward NNs or convolutional NNs (Ukkonen 2022; Yao et al. 2023). Therefore, the BiLSTM network architecture was chosen, which is also physically motivated through the bidirectional aspect scanning in the vertical dimension resembling the upward and downward fluxes of radiation in the atmosphere - just like the physics-based radiation schemes. A physics-based constraint was used in the loss function during training ensuring that the ML-based radiation emulator produces energetically consistent boundary fluxes and heating rates. The resulting emulator accurately learns radiative heating rates with a mean absolute error below 0.06 $K/d$ and boundary fluxes at the surface and top of the atmosphere with a mean absolute error of 2 $W/m^2$ for longwave radiation but up to 30 $W/m^2$ for shortwave radiation. The emulator is in itself energetically consistent within 0.5 $W/m^2$. Overall, these results are comparable to other ML-based approaches (Lagerquist et al. 2023; Ukkonen 2022; Yao et al. 2023). The ML-based emulator was then interpreted using SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). The analysis showed that the emulator's predictions rely mostly on local features. Cloud fraction only affects shortwave heating rates where clouds are available in the troposphere. The strongest effect is local: a cloud affects the radiative heating rate where the cloud is located. This effect however also propagates to lower layers, although with a smaller strength, which is sensible as a cloud reflects incoming radiation leading to less heating in the lower layers. Additionally, reflected radiation leads to heating in the ozone layer. For longwave radiation, the biggest impact comes from the temperature profile. Gases, aerosols or clouds present in every layer are a source of longwave radiation whose magnitude strongly depends on the temperature in each cell. The Shapley values show strong local effects of temperature on longwave heating rates that also carry over to the neighboring layers which is in agreement with the atmosphere as a source of longwave radiation. I conclude this study by providing a concrete answer to the key science question 1: "Yes, the ML-based radiation emulator indeed learns relationships that are related to known physical laws such as cloud reflection and longwave emission of the atmosphere".

Provided that most previous ML-based radiation approaches focus on NWP applications instead of climate projections, only the lower part of the atmosphere, up to 36 $km$, is generally

considered. Additionally, online tests, when available, are only performed over a few days (Song and Roh 2021; Zhong et al. 2023), which is sufficient for NWP applications. This motivated the key science question 2, "Can the ML-based emulator be coupled to a complex ESM like ICON and produce stable simulations?", discussed in detail in Chapter 5 based on Hafner et al. (2025b).

To address this question, the ML-based emulator is coupled to the atmospheric component of ICON. However, good offline performance does not guarantee good online performance (Brenowitz et al. 2020; Wang et al. 2022). Additionally, ICON - similarly to other state-of-the-art ESMs - is written in Fortran, and ML-based schemes are usually trained in Python. The latter issue has been addressed with recently developed coupling bridges like *FTorch* (Atkinson et al. 2025) that allow to execute a trained ML scheme in Fortran. However, the former issue is more difficult to address and requires testing. Therefore, I used *FTorch* to implement the ML-based radiation emulator presented in Chapter 4 and Hafner et al. (2025a) into the ICON model for simulations with a horizontal resolution of 80 $km$. The ICON model was run in three different experimental setups: ICON-$A_{ML}$, ICON-$A_{HF}$ and ICON-$A_{LF}$. In ICON-$A_{ML}$ the ML-based radiation emulator was used and the radiation time step was set to the physics time step of 6 $min$. The other two setups are used for reference with the original radiation scheme RTE+RRTMGP (Pincus et al. 2019). The ICON-$A_{HF}$ setup also uses a radiation time step of 6 $min$ and is labeled as *ground truth* whereas the ICON-$A_{LF}$ setup uses a radiation time step of 1 $h$, which is labeled as *baseline*. In terms of performance, the ML-based radiation emulator is four times faster than the original radiation parameterization. In a complex ESM like ICON, small differences in a few parameters can lead to large changes such that two similar simulations are not comparable. Therefore, 10 simulations with different initial conditions - or *ensemble* runs with 10 ensemble members - with a length of one year were created for each setup. The ensemble mean and standard deviation was then used to compare the three setups. The results presented in Chapter 5 and Hafner et al. (2025b) show that the setups align within their standard deviation. Whenever there are deviations, they are related to changes in the cloud distribution, e.g., if there are more clouds then there is less downward shortwave radiation at the surface. In other words, the ICON-$A_{ML}$ simulations react physically consistent to changes in clouds. Regarding the top of the atmosphere energy balance, both ICON-$A_{ML}$ and ICON-$A_{LF}$, have a small bias (below $1 W/m^2$) with respect to ICON-$A_{HF}$, but ICON-$A_{ML}$ is even less biased than ICON-$A_{LF}$. Moreover, as some biases need time to build up, long-term stability was tested by comparing single runs of each setup over 10 years. The ICON-$A_{ML}$ setup remained stable without showing signs of deviations. The generalizability of the ML-based scheme was tested by increasing the prescribed SSTs by 4 $K$ and 8 $K$. The ICON-$A_{ML}$ simulations showed good agreement with ICON-$A_{LF}$ for the +4 $K$ experiment. In the +8 $K$ experiment, the ICON-$A_{ML}$ simulation still showed agreement with the ICON-$A_{LF}$ simulation in the troposphere, while in the stratosphere the runs started to diverge. This behavior was shown to be due to high water vapor content in the stratosphere, far outside the training range of the emulator. To answer the key science question 2: "Yes, we can couple the ML-based radiation emulator to a complex model like ICON and produce stable

simulations. It is four times faster than the original parameterization while having statistically non-significant differences from the reference, except for a limited number of areas that can be related to physical effects. The ML-based emulator can even generalize to a +4 $K$ warmer climate, or even warmer, as long as the atmospheric state is not too far out of the training range."

So far, the ML-based radiation emulator increased the performance of the radiation scheme by allowing more frequent calls and therefore more frequent updates of the cloud-radiation interactions. However, the improvement due to more frequent radiation calls was limited and the reduction in computing time was less than desired. One possible issue is that the radiation scheme assumes that clouds are horizontally homogeneous, which is not accurate at a horizontal resolution of 80 $km$, and the subgrid-scale distribution of clouds could have a bigger impact. The assumption of horizontal homogeneity can however be relaxed by learning from a high-resolution dataset with horizontal grid spacing of 5 $km$. In such a GSRM, clouds are better resolved motivating the key science question 3 "Can we improve the representation of cloud-radiation interactions by using high-resolution data for training?" discussed in Chapter 6 and based on (Hafner et al. 2025c).

For the final study, an ML-enhanced radiation scheme is developed to specifically target the cloud radiative impact on heating rates. The cloud radiative impact is defined as the difference between all-sky and clear-sky heating rate. Here, the approach is called ML-enhanced because the physics-based radiation scheme computes clear-sky heating rates, while the uncertain cloud impact is predicted by a neural network. The NN is trained on horizontally and vertically coarse-grained high-resolution simulations, where the high-resolution dataset is created using the ICON model based on the QUBICC protocol presented in Giorgetta et al. (2022). The high-resolution data is coarse-grained to a horizontal resolution of 80 $km$ and vertically from 191 to 47 levels. This way the atmospheric state, water related variables and corresponding radiative heating rates implicitly include the effects of the subgrid-scale distribution of clouds. The coarse-grained input and output variables of the radiation parameterization are compared to their coarse-scale counterpart to ensure consistency between scales. The physics-based radiation scheme RTE+RRTMGP (Pincus et al. 2019), which is also used in ICON, is evaluated on coarse-grained data, which serves as reference for a coarse-scale radiation scheme. The coarse-scale reference is compared to the ML-enhanced radiation scheme depending on different cloud related categories and regions. While the coarse-scale reference has small errors on clear-sky scenes, the error significantly increases when clouds are involved because the subgrid-distribution of clouds is unknown. The ML-enhanced radiation scheme consistently performs better with errors that are 4-10 times smaller and unbiased compared to the coarse-scale reference. This ML-enhanced approach has potentially better generalization skills since the machine-learned part focuses only on clouds, while the clear component can still react to changes in GHGs and aerosols. Hence, this demonstrates the importance of using high-resolution simulations as a training baseline to learn subgrid-scale effects. To answer the key science question 3, I therefore conclude that "Yes, cloud-radiation interactions can be improved using high-resolution data for training, because this type of data implicitly includes

subgrid-scale effects. Therefore, the NN is trained to improve the cloud radiative impact on heating rates and reduced errors by a factor of 4-10 compared to a coarse-scale radiation scheme."

## 7.2. Context and Outlook

This thesis presents an ML-based radiation scheme that accurately and efficiently emulates the radiation parameterization RTE+RRTMGP for the ICON model. The interpretability of the ML-based emulator increases trustworthiness while also being four times faster than the original scheme. However, additional code optimization efforts could also speed-up traditional radiation parameterizations (Ukkonen and Hogan 2024). But still ESMs that are optimized for GPUs could benefit from ML-based radiation schemes because ML-based algorithms are highly optimized for the use on GPUs and they are also more energy efficient (Ukkonen and Chantry 2025).

The faster radiation scheme allows to test potentially improved representations of various physical processes within the ICON model such as different cloud cover or microphysics schemes, that could reduce biases like the shortwave cloud radiative effects at the top of the atmosphere. One limitation of the presented ML-based radiation emulator is that the effect of changing GHGs and aerosols was not yet considered due to technical reasons. Also, the ML-enhanced radiation approach of learning only the cloud radiative impact on heating rates as presented in the third study still needs to be tested online.

To reduce long standing systematic biases, the development of ESMs for climate projections currently follows two directions. One direction is increasing the horizontal resolution such that more processes are resolved and fewer parameterizations would be required, which would potentially have fewer uncertainties but GSRMs are computationally expensive. The other direction is improving coarse-scale models e.g., by using ML, potentially with additional high-resolution data as done in this thesis. Coarse-scale models are computationally much cheaper, allowing for testing multiple different scenarios and many ensemble members. High-resolution data can be used to quantify uncertainty that comes from the subgrid-scale distributions of the variables of interest i.e., clouds in the context of this thesis. GSRMs make fewer assumptions about the subgrid-scale processes as many are directly simulated. This allows for learning the underlying statistics that can be used to improve coarse-scale models. Another option is that high-resolution data can be coarse-grained to different horizontal resolutions allowing for the development of ML-based parameterizations that are scale independent or learn how to use the scale information to make accurate predictions. A better source of training data, i.e., a line-by-line model, could also improve an ML-based radiation scheme, which would allow to cover a wider range of values in trace gases, which are globally and annually prescribed in ICON.

In conclusion, this thesis showed that ML-based parameterizations can accurately and efficiently replace otherwise computationally expensive radiation parameterizations. The pre-

sented work complements ongoing work on ML-based radiation schemes by showing interpretability and stability in long-term simulations. Additionally, the presented advancements in terms of improving radiative heating rates with high-resolutions simulation show promising results laying the foundation for further research in terms of uncertainty quantification and the development of resolution- or model-agnostic parameterizations.

# A. Supplementary Materials for Chapter 4

The Appendix A was published as supplementary material for (Hafner et al. 2025a). I produced all content from this paper including text, figures and tables.

## Why predict heating rates directly instead of flux profiles?

Most of the existing emulators were targeted for NWP applications. Therefore, the top is usually defined at 10 $Pa$ or 36 $km$. Here, the top is defined at 83 $km$. To get from fluxes to heating rates you divide by the difference in pressure which gets really small in the top. ICON calculates using double precision but the outputs were saved at single precision. We made the same calculation to get from fluxes to heating rates offline and compared to the heating rate calculated within ICON. The results for a few columns for longwave heating rates are shown in Figure A.1. For reference, we added a line at 36 $km$ which is the usual top in NWP. The difference clearly increases with height as the difference in pressure gets smaller. For this reason, Bertoli et al. (2025) had to add extra modifications for heating rates above 20 and 45 km for LW and SW. We would like to emphasize that the differences in Figure A.1 are only due to differences in single and double precision. Therefore, the approach to predict fluxes seems to be insufficient when emulating radiation for an atmosphere up to 80 $km$ as NN errors are usually larger. The only other study that we know of that emulated radiation up to 80 $km$ is Lagerquist et al. (2023) and they also predicted heating rates directly.

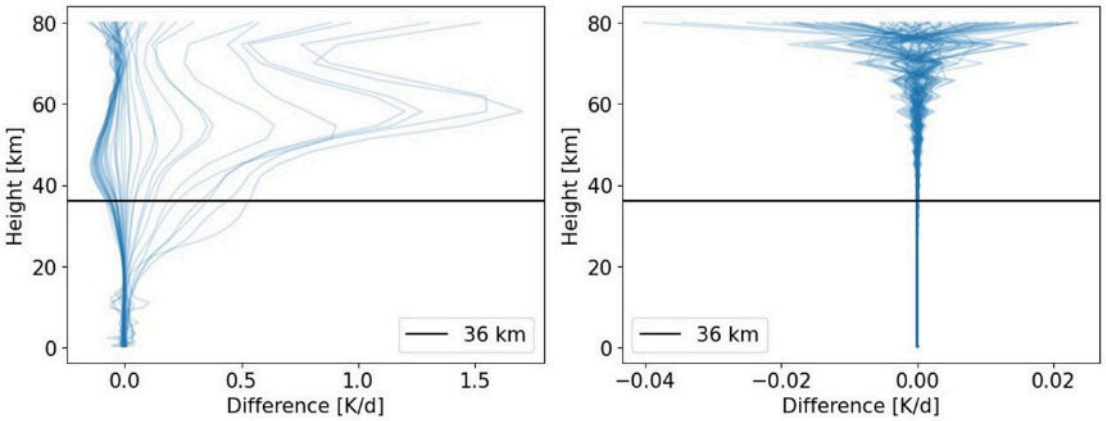


**Figure A.1.:** Difference between SW (left) LW (right) heating rates calculated offline and within ICON. Adapted with permission from Hafner et al. (2025a).

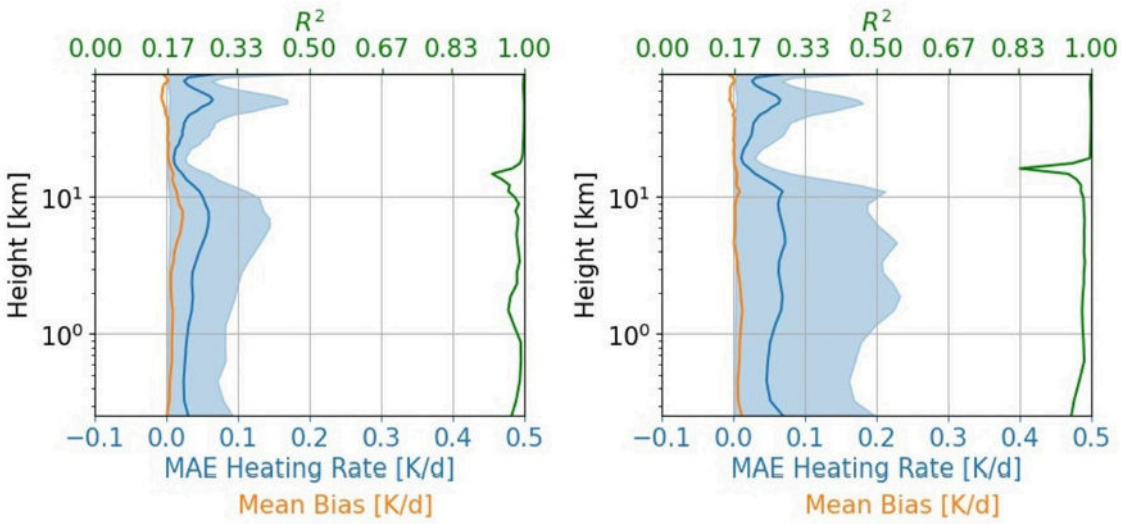


**Figure A.2.:** Vertically resolved mean absolute error (MAE), mean bias and coefficient of determination $R^2$ for shortwave heating rate for clear-sky samples (left) and cloudy sky samples (right). Adapted with permission from Hafner et al. (2025a).

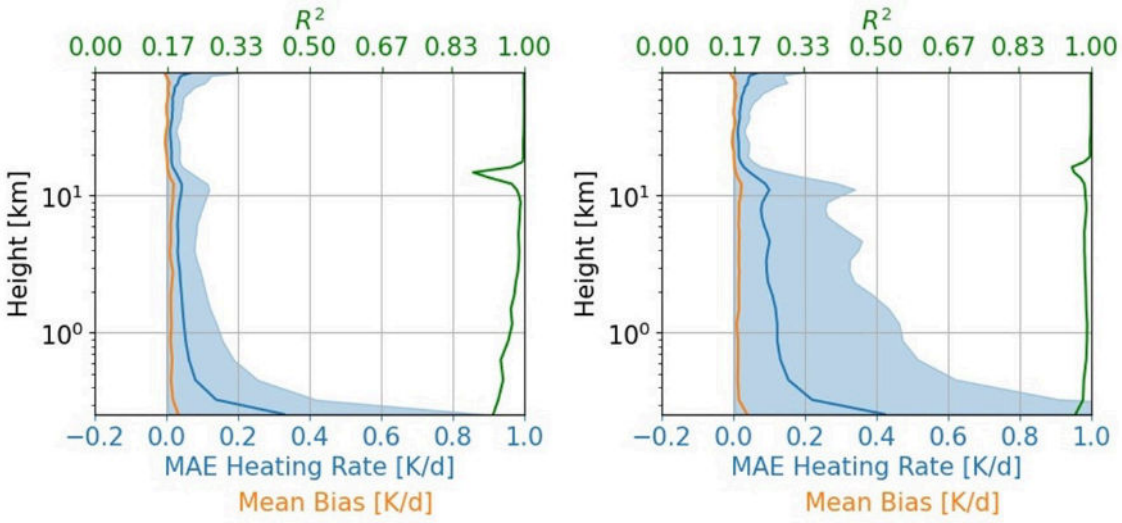


**Figure A.3.:** Vertically resolved mean absolute error (MAE), mean bias and coefficient of determination $R^2$ for longwave heating rate for clear-sky samples (left) and cloudy sky samples (right). Adapted with permission from Hafner et al. (2025a).

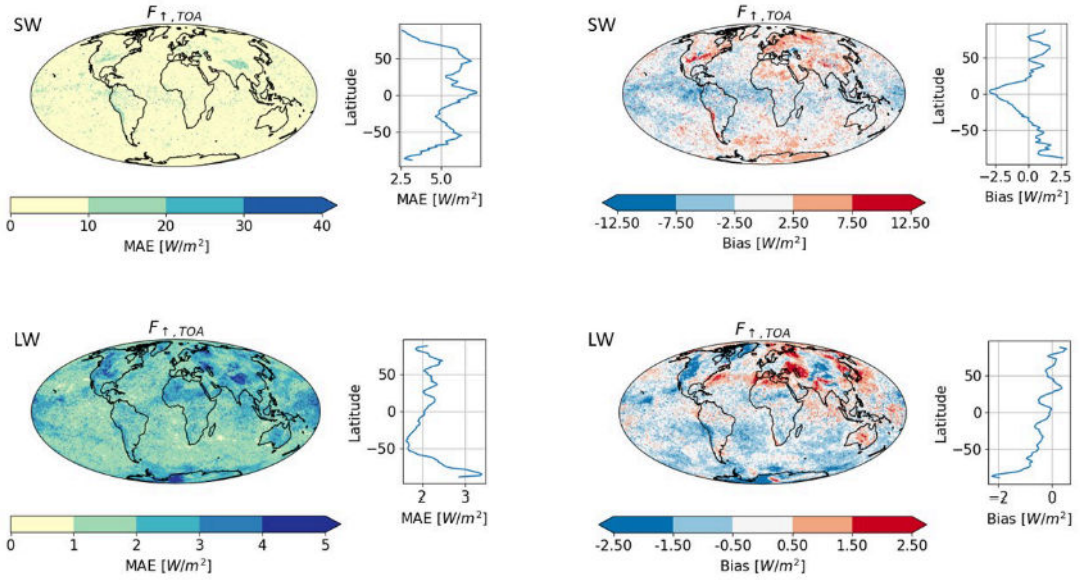


**Figure A.4.:** Mean absolute error (MAE) and bias for predicted upward fluxes at the top of the atmosphere for shortwave (top panel) and longwave (bottom panel) radiation. Adapted with permission from Hafner et al. (2025a).

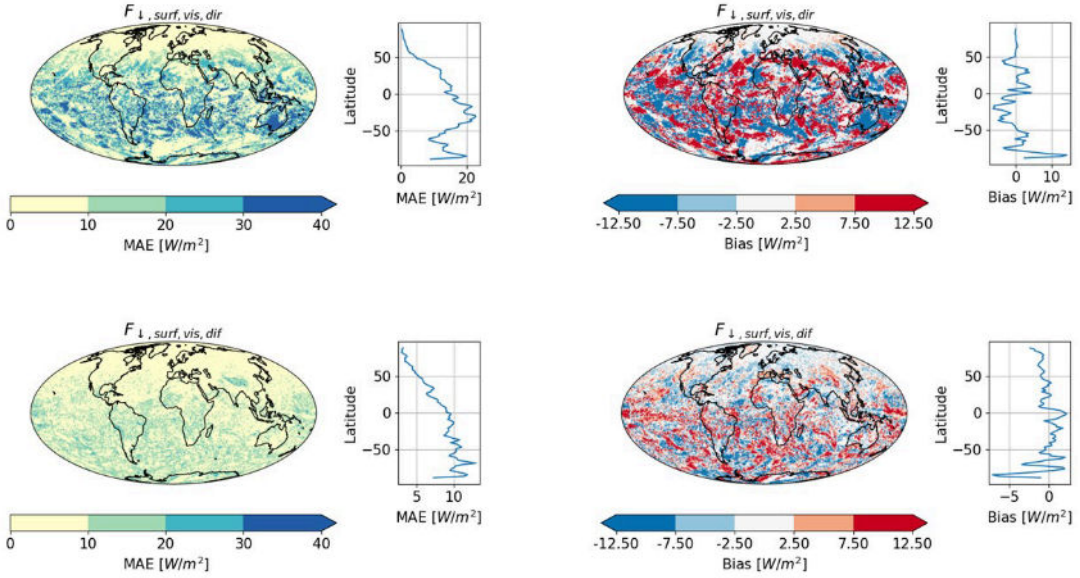


**Figure A.5.:** Mean absolute error (MAE) and bias for predicted downward fluxes at the surface for direct (top panel) and diffuse (bottom panel) visible shortwave radiation. Adapted with permission from Hafner et al. (2025a).

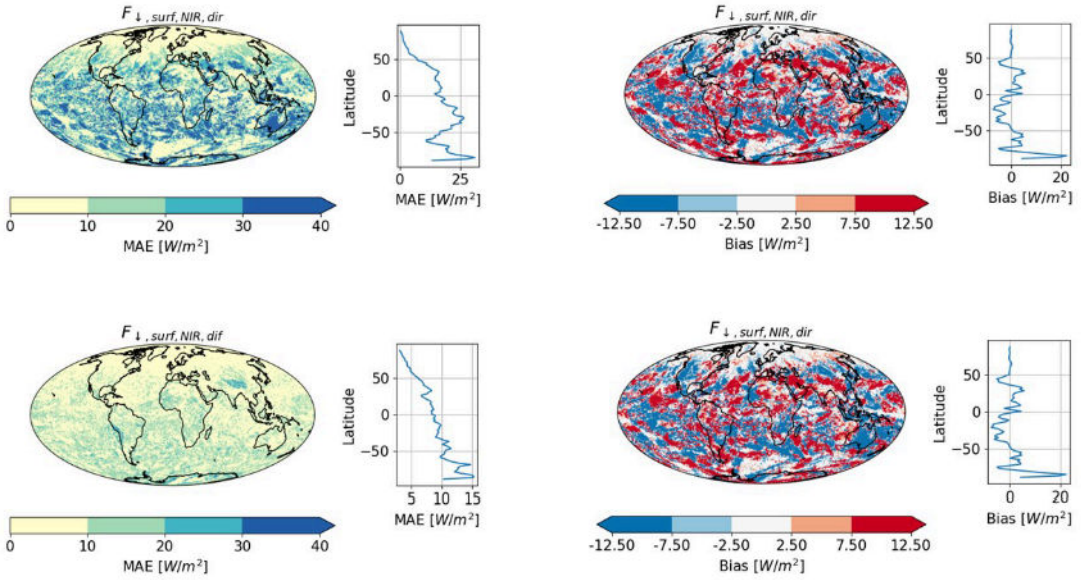


**Figure A.6.:** Mean absolute error (MAE) and bias for predicted downward fluxes at the surface for direct (top panel) and diffuse (bottom panel) near-infrared shortwave radiation. Adapted with permission from Hafner et al. (2025a).

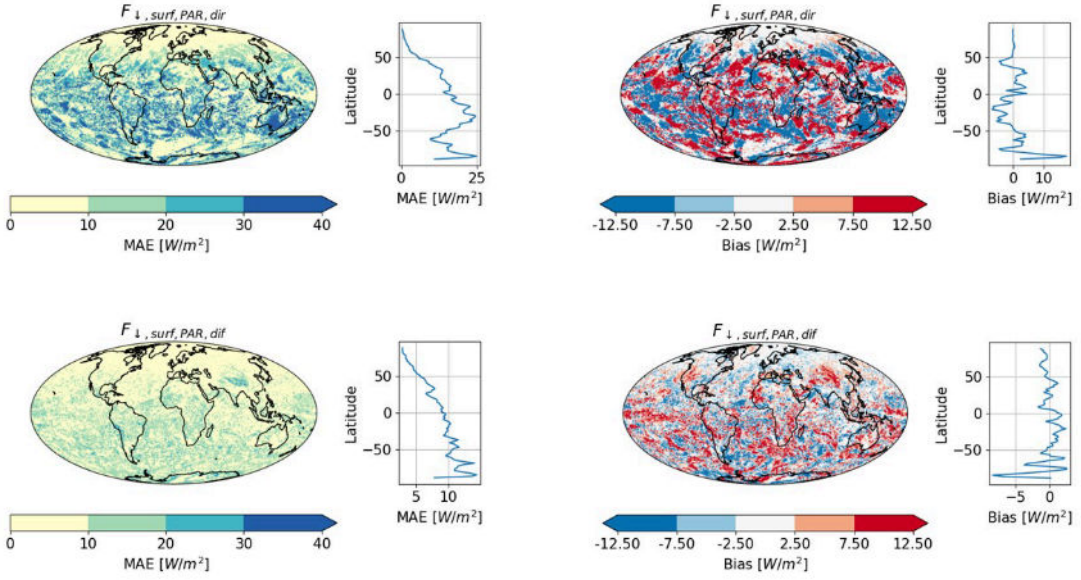


**Figure A.7.:** Mean absolute error (MAE) and bias for predicted downward fluxes at the surface for direct (top panel) and diffuse (bottom panel) photosynthetically active shortwave radiation. Adapted with permission from Hafner et al. (2025a).

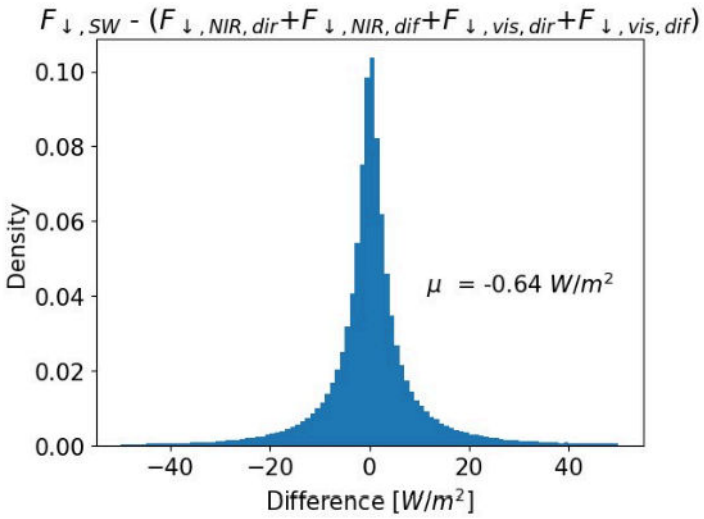


**Figure A.8.:** Difference between shortwave downward flux at the surface and the partial shortwave fluxes. Adapted with permission from Hafner et al. (2025a).

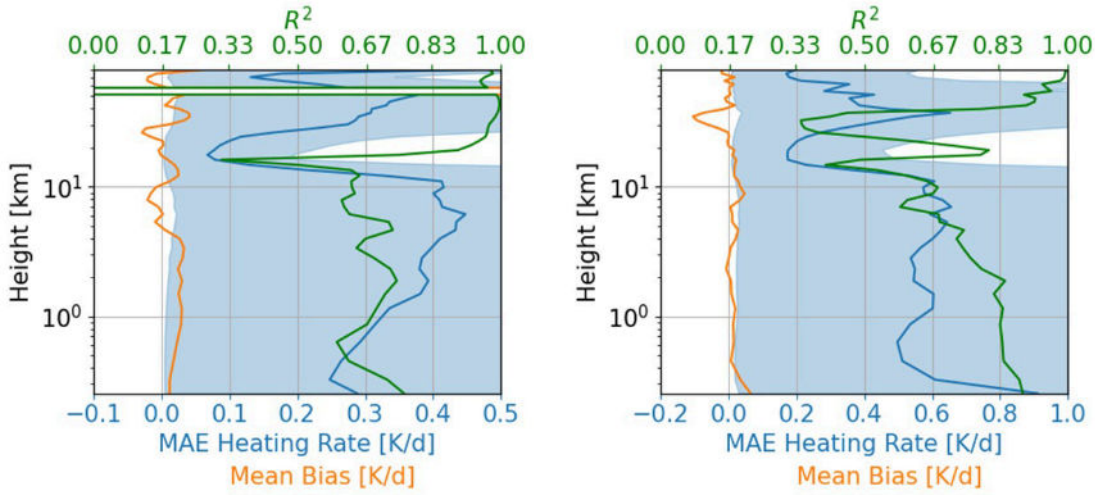


**Figure A.9.:** As Figure 2 but for the MLP: Global and time mean vertical profiles of heating rates. Mean absolute error (MAE), bias, and $R^2$ are shown for shortwave heating rates (left) and longwave heating rates (right). The shaded area shows 90% of error spread. Adapted with permission from Hafner et al. (2025a).

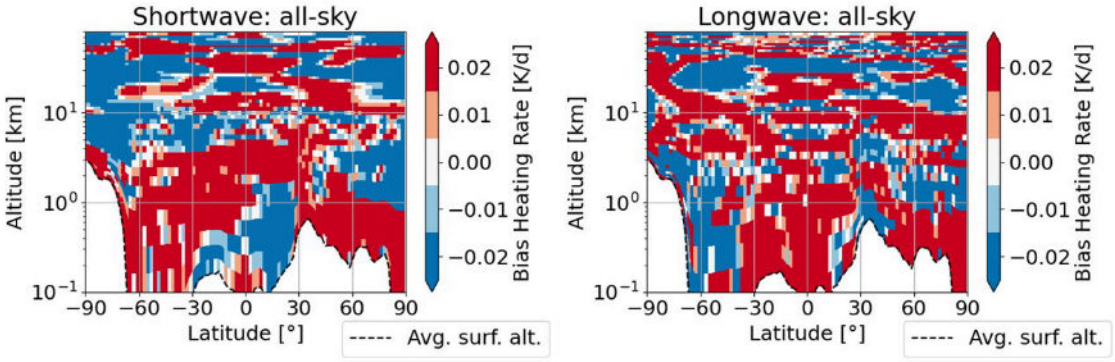


**Figure A.10.:** As Figure 3 but for the MLP: Zonal- and time-mean machine learned biases for (left) shortwave heating rates, and (right) longwave heating rates. Adapted with permission from Hafner et al. (2025a).

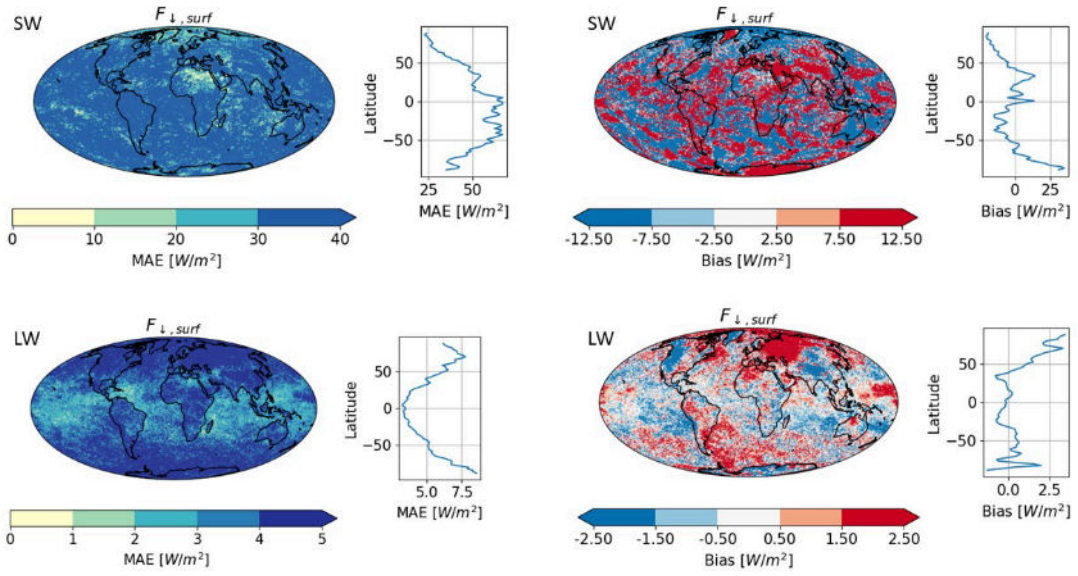


**Figure A.11.:** As Figure 4 and 5 but for the MLP: Time-average maps of shortwave (top panel) and longwave (bottom panel) downward flux at the surface. (left) Mean absolute error (MAE), and (right) bias are shown. Right panels show zonal-mean values. Adapted with permission from Hafner et al. (2025a).

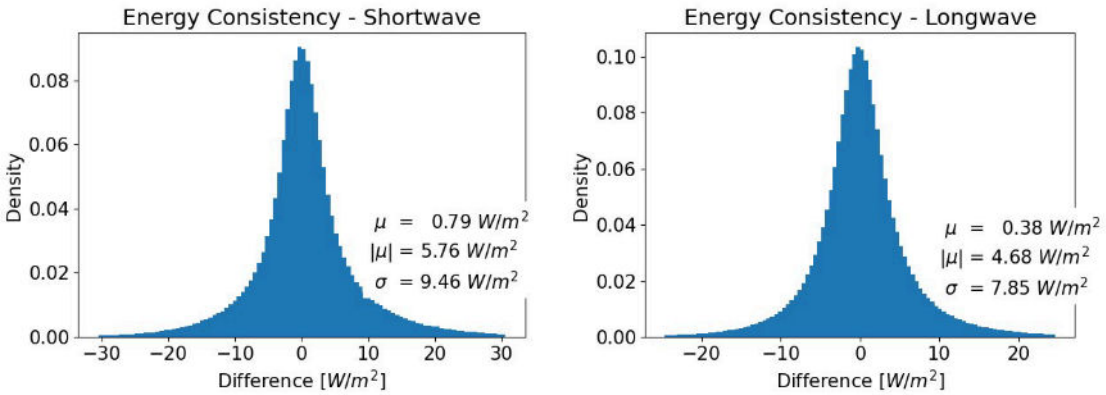


**Figure A.12.:** As Figure 6 but for the MLP: Energy balance check for combined neural networks for SW radiation (left) and LW radiation (right). The histogram shows the difference between boundary fluxes and the vertical integral of radiative energy, both predicted by the neural networks. Adapted with permission from Hafner et al. (2025a).

**Table A.1.:** Bulk statistics for all fluxes. Adapted with permission from Hafner et al. (2025a).

| Variable | MAE $[W/m^2]$ | Bias $[W/m^2]$ | $R^2$ |
|---|---|---|---|
| $F_{\uparrow,\, TOA,\, SW}$ | 28 | 2.1 | 0.92 |
| $F_{\downarrow,\, surf,\, SW}$ | 54 | -3.0 | 0.84 |
| $F_{\downarrow,\, surf,\, SW,\, vis\ dir}$ | 25 | -3.2 | 0.78 |
| $F_{\downarrow,\, surf,\, SW,\, vis\ dif}$ | 18 | 2.1 | 0.63 |
| $F_{\downarrow,\, surf,\, SW,\, nir\ dir}$ | 29 | -3.1 | 0.76 |
| $F_{\downarrow,\, surf,\, SW,\, nir\ dif}$ | 18 | 2.9 | 0.53 |
| $F_{\downarrow,\, surf,\, SW,\, par\ dir}$ | 27 | -3.3 | 0.77 |
| $F_{\downarrow,\, surf,\, SW,\, par\ dif}$ | 19 | 2.6 | 0.57 |
| $F_{\uparrow,\, TOA,\, LW}$ | 7.70 | -0.40 | 0.95 |
| $F_{\downarrow,\, surf,\, LW}$ | 4.63 | 0.22 | 0.99 |

# B. Supplementary Materials for Chapter 5

The Appendix B was published as supplementary material for (Hafner et al. 2025b). I produced all content from this paper including text, figures and tables.

**Table B.1.:** Summary of differences in boundary fluxes and total cloud cover. The annual, ensemble mean values of ICON-$A_{ML}$ and ICON-$A_{LF}$ are compared to those of ICON-$A_{HF}$. Adapted with permission from Hafner et al. (2025b).

| | MAE | RMSE | Bias |
|---|---|---|---|
| ML $F_{\uparrow,SW,TOA}$ $[W/m^2]$ | 2.51 | 3.38 | 1.28 |
| LF $F_{\uparrow,SW,TOA}$ $[W/m^2]$ | 2.18 | 2.94 | 0.91 |
| ML $F_{\downarrow,SW,surf}$ $[W/m^2]$ | 3.03 | 4.04 | -1.18 |
| LF $F_{\downarrow,SW,surf}$ $[W/m^2]$ | 2.69 | 3.53 | -1.10 |
| ML $F_{\downarrow,surf,PAR,dir}$ $[W/m^2]$ | 1.64 | 2.20 | -0.70 |
| LF $F_{\downarrow,surf,PAR,dir}$ $[W/m^2]$ | 1.46 | 1.92 | -0.49 |
| ML $F_{\downarrow,surf,PAR,dif}$ $[W/m^2]$ | 0.76 | 1.05 | 0.39 |
| LF $F_{\downarrow,surf,PAR,dif}$ $[W/m^2]$ | 0.46 | 0.59 | -0.04 |
| ML $F_{\downarrow,surf,NIR,dir}$ $[W/m^2]$ | 1.80 | 2.40 | -0.87 |
| LF $F_{\downarrow,surf,NIR,dir}$ $[W/m^2]$ | 1.55 | 2.02 | -0.48 |
| ML $F_{\downarrow,surf,NIR,dif}$ $[W/m^2]$ | 0.82 | 1.16 | 0.37 |
| LF $F_{\downarrow,surf,NIR,dif}$ $[W/m^2]$ | 0.44 | 0.57 | -0.08 |
| ML $F_{\downarrow,surf,vis,dif}$ $[W/m^2]$ | 0.66 | 0.88 | 0.34 |
| LF $F_{\downarrow,surf,vis,dif}$ $[W/m^2]$ | 0.43 | 0.56 | -0.04 |
| ML $F_{\uparrow,LW,TOA}$ $[W/m^2]$ | 1.67 | 2.22 | -0.99 |
| LF $F_{\uparrow,LW,TOA}$ $[W/m^2]$ | 1.14 | 1.52 | -0.13 |
| ML $F_{\downarrow,LW,surf}$ $[W/m^2]$ | 1.53 | 1.88 | 1.13 |
| LF $F_{\downarrow,LW,surf}$ $[W/m^2]$ | 0.88 | 1.13 | -0.06 |
| ML Total Cl. Cover [%] | 1.08 | 1.39 | -0.21 |
| LF Total Cl. Cover [%] | 0.93 | 1.21 | -0.15 |

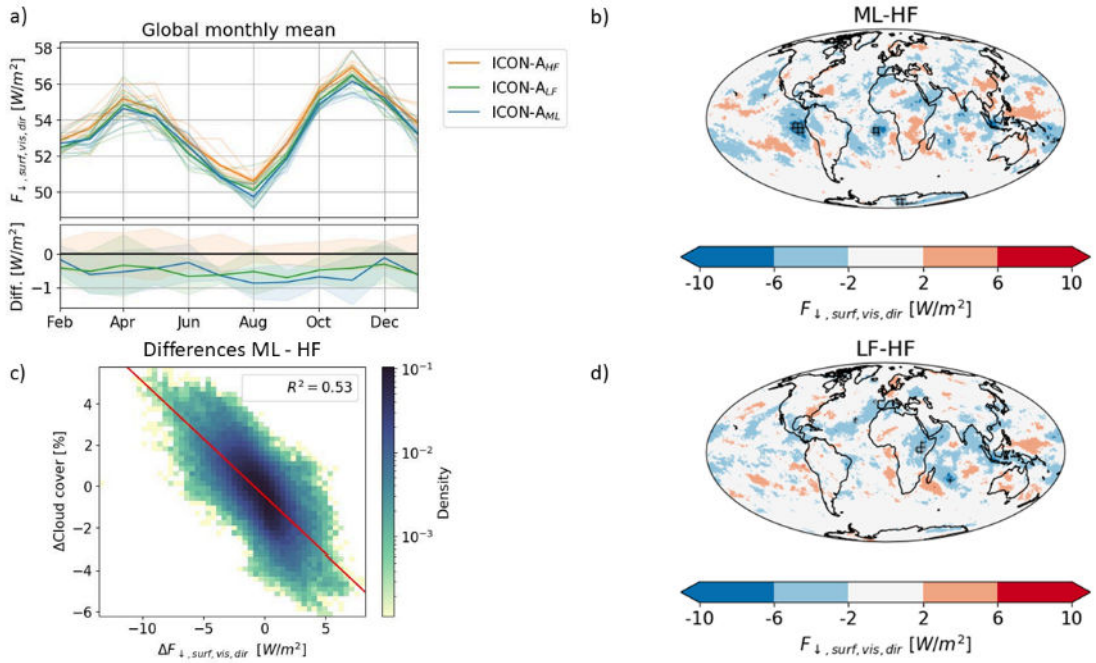


**Figure B.1.:** a) Global monthly mean of direct visible component of SW downward flux at the surface $F_{\downarrow, SW, surf, vis, dir}$ ICON-A$_{HF}$, ICON-A$_{LF}$ and ICON-A$_{ML}$ run. Map of annual, ensemble-mean difference between b) ICON-A$_{HF}$ and ICON-A$_{ML}$ run and d) ICON-A$_{HF}$ and ICON-A$_{LF}$ run. c) 2D Histogram of local monthly mean differences of total cloud cover vs $F_{\downarrow, SW, surf, vis, dir}$. The red line is a linear fit and $R^2$ is coefficient of determination. Adapted with permission from Hafner et al. (2025b).

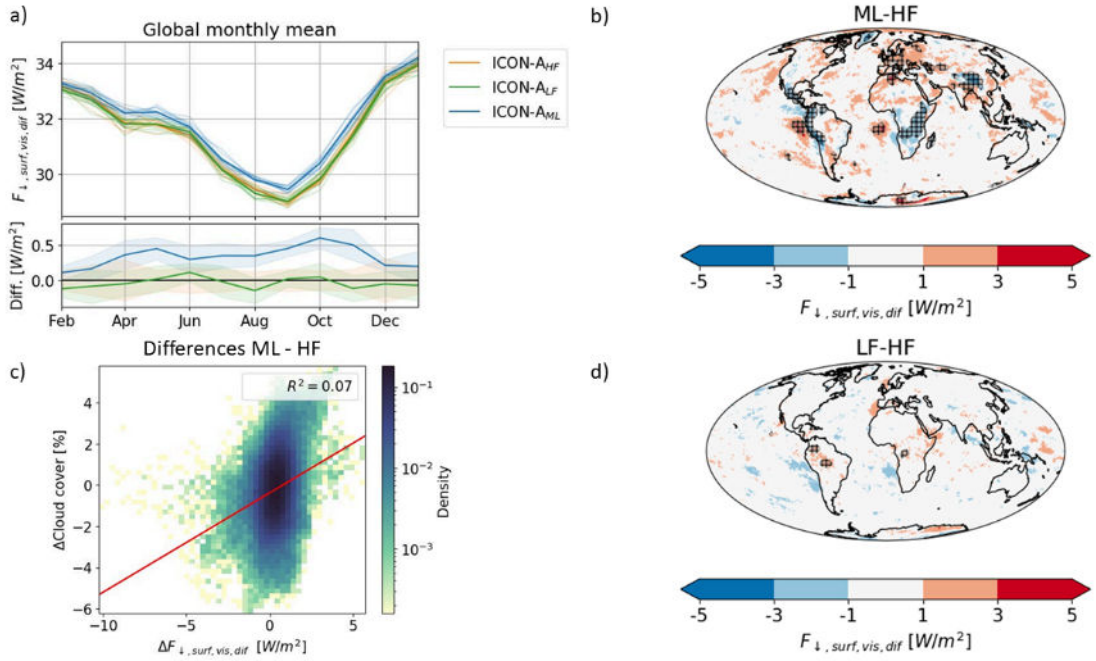


**Figure B.2.:** Same as Figure B.1 but for the diffuse visible component $F_{\downarrow, SW, surf, vis, dif}$. Adapted with permission from Hafner et al. (2025b).

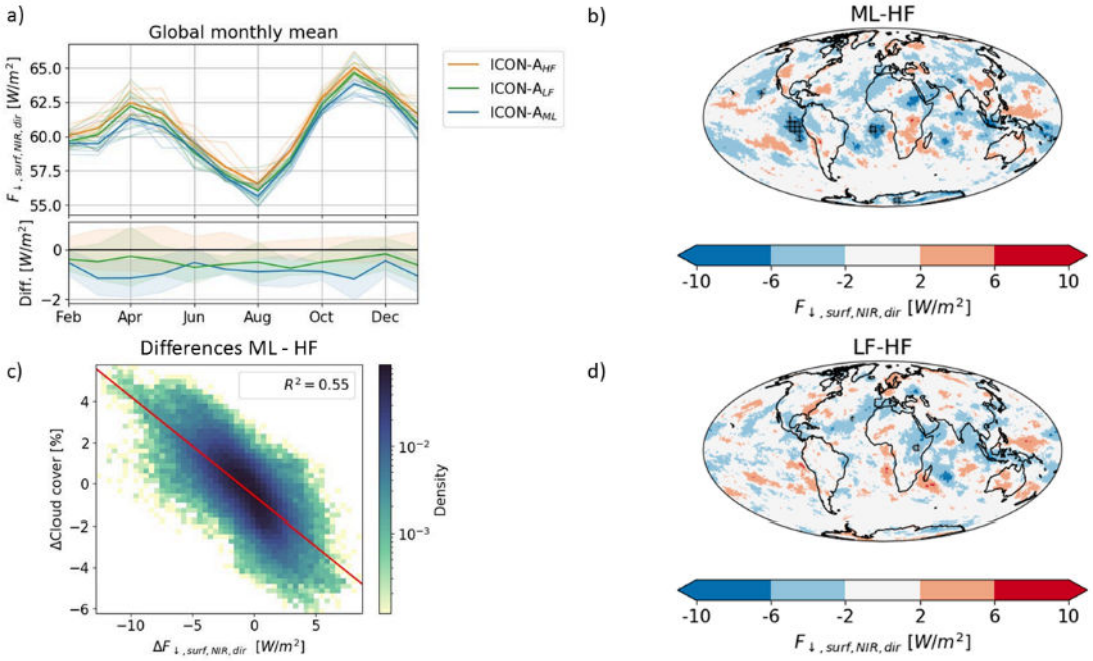


**Figure B.3.:** Same as Figure B.1 but for the direct near-infrared component $F_{\downarrow, SW, surf, NIR, dir}$. Adapted with permission from Hafner et al. (2025b).

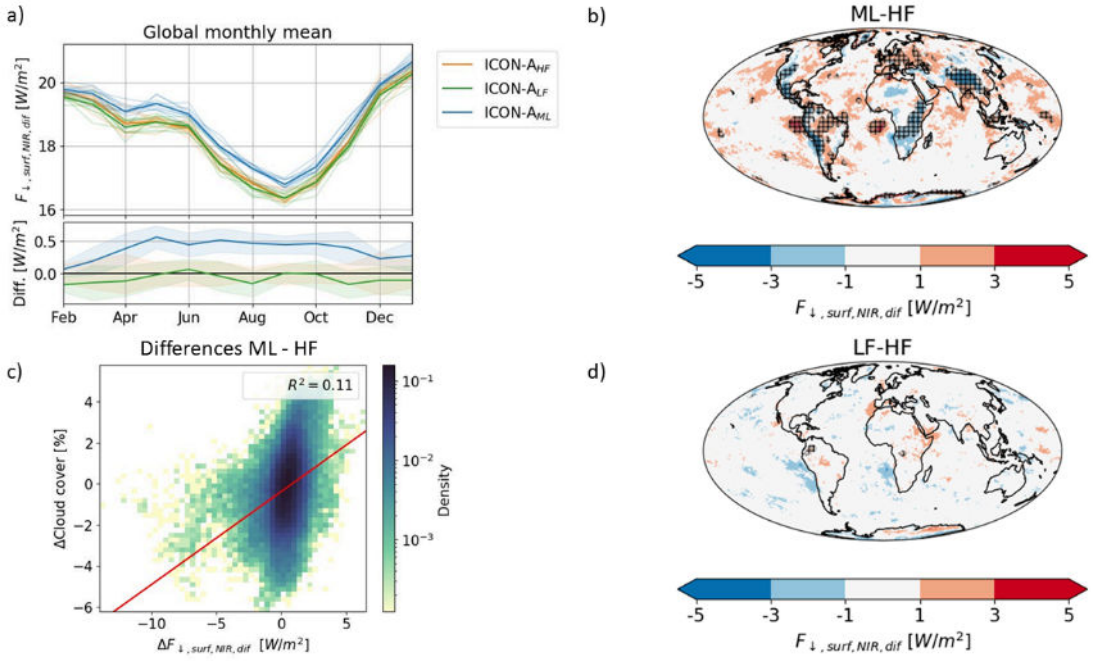


**Figure B.4.:** Same as Figure B.1 but for the diffuse near-infrared component $F_{\downarrow, SW, surf, NIR, dif}$. Adapted with permission from Hafner et al. (2025b).

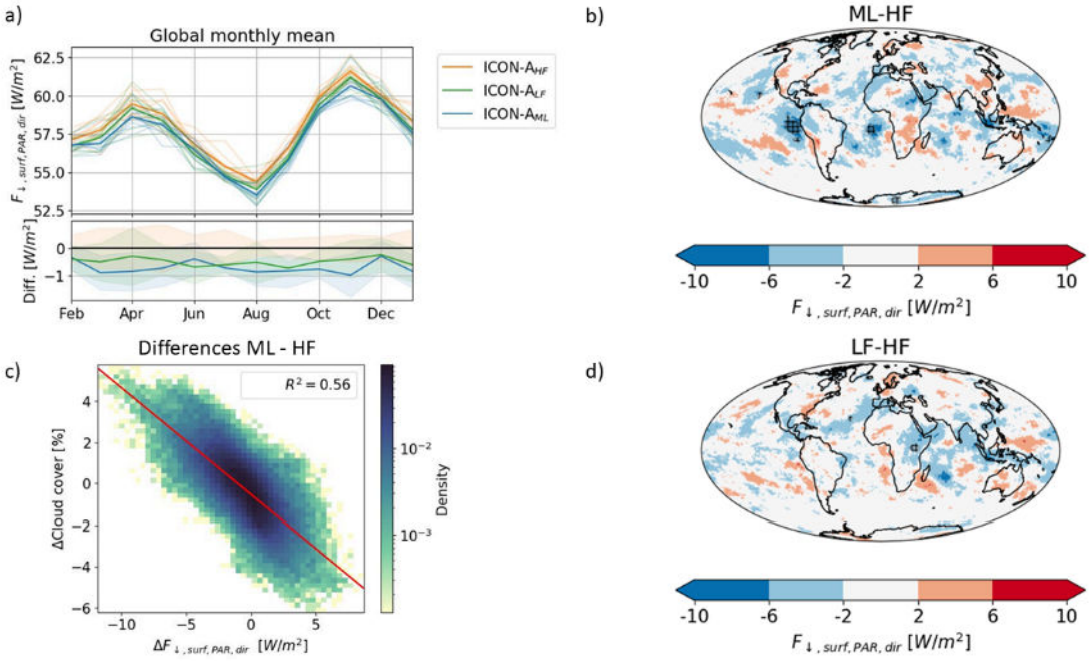


**Figure B.5.:** Same as Figure B.1 but for the direct photosynthetically active component $F_{\downarrow, SW, surf, PAR, dir}$. Adapted with permission from Hafner et al. (2025b).

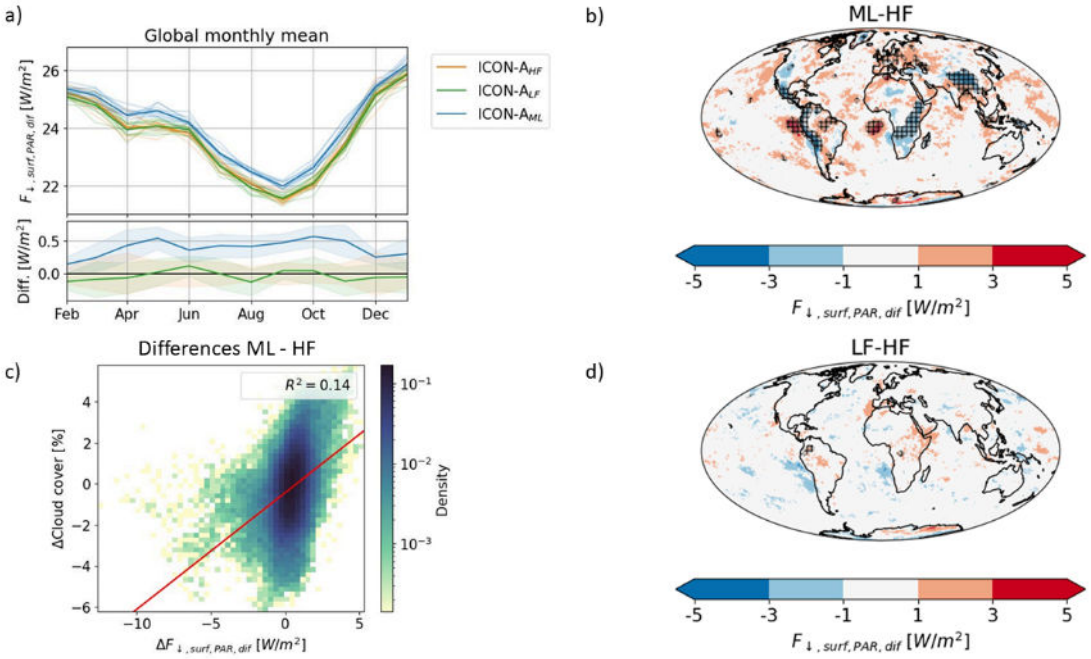


**Figure B.6.:** Same as Figure B.1 but for the diffuse photosynthetically active component $F_{\downarrow, SW, surf, PAR, dif}$. Adapted with permission from Hafner et al. (2025b).

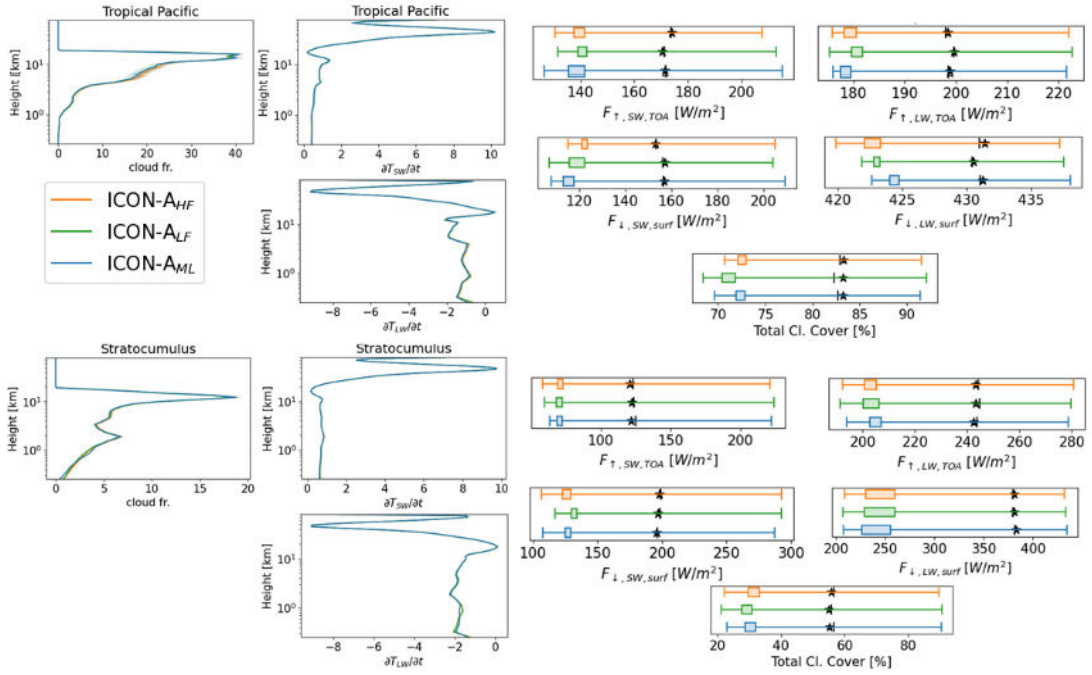


**Figure B.7.:** Additional variable distributions in the tropical pcific and stratocumulus region. Adapted with permission from Hafner et al. (2025b).

# C. Supplementary Materials for Chapter 6

The Appendix C was published as supplementary material for (Hafner et al. 2025c). I produced all content from this paper including text, figures and tables.

## Default microphysics scheme

One major difference between the ICON-A and QUBICC versions is the microphysics scheme. ICON-A uses a modified version of the Lohmann and Roeckner (1996) scheme and the QUBICC simulation uses the graupel scheme described in Doms et al. (2011). While both schemes are single-moment schemes, the latter treats precipitating tracers like snow, rain and graupel as prognostic variables while the former only diagnoses snow and rain. Moreover, it is known that cloud ice is too large in the upper troposphere in ICON-A (Doktorowski et al. 2024), which we also see when comparing cloud ice in ICON-A and QUBICC (Figure C.1).

## Calculation of heating rates

When comparing the unscaled heating rates between coarse-scale and coarse-grained simulations, we find huge biases of up to 10 K/d between the mean heating rates, especially in the stratosphere. We found that there is a difference in heating rate calculation between the code version. Usually, the heating rate is calculated from flux divergence, pressure difference and constants:

$$\frac{\partial T_k}{\partial t} = \frac{g}{c_p} \frac{F_{Net,k+1/2} - F_{Net,k-1/2}}{P_{k+1/2} - P_{k-1/2}}, \tag{C.1}$$

where $g$ is gravitational acceleration, $c_p$ specific heat at constant pressure, $F_{Net}$ is the difference between downward and upward flux, $P$ is pressure. $k$ is defined at the center of a layer (also full levels) while $k \pm \frac{1}{2}$ is defined at the layer boundaries (also half levels). This form of converting fluxes to heating rates is usually found in hydrostatic models but does not work in ICON because pressure is a diagnostic variable. Instead, the density is kept constant and specific heat at constant volume $c_v$ needs to be used for the conversion (Zängl et al. 2014). This transforms Equation C.1 to Equation 6.2. In the ICON-A version, they use $c_p$ which is valid for quasi-hydrostatic models because then the hydrostatic pressure holds $dP/g = \rho dz$. Additionally, the specific heat is scaled only by water vapor, while all tracers are included in the code version used for QUBICC. Therefore, we scaled the coarse-grained heating rates in Figure 6.3 with the ratio $c_v/c_p$, which is on average 0.7. The unscaled heating rates are shown in Figure C.2.

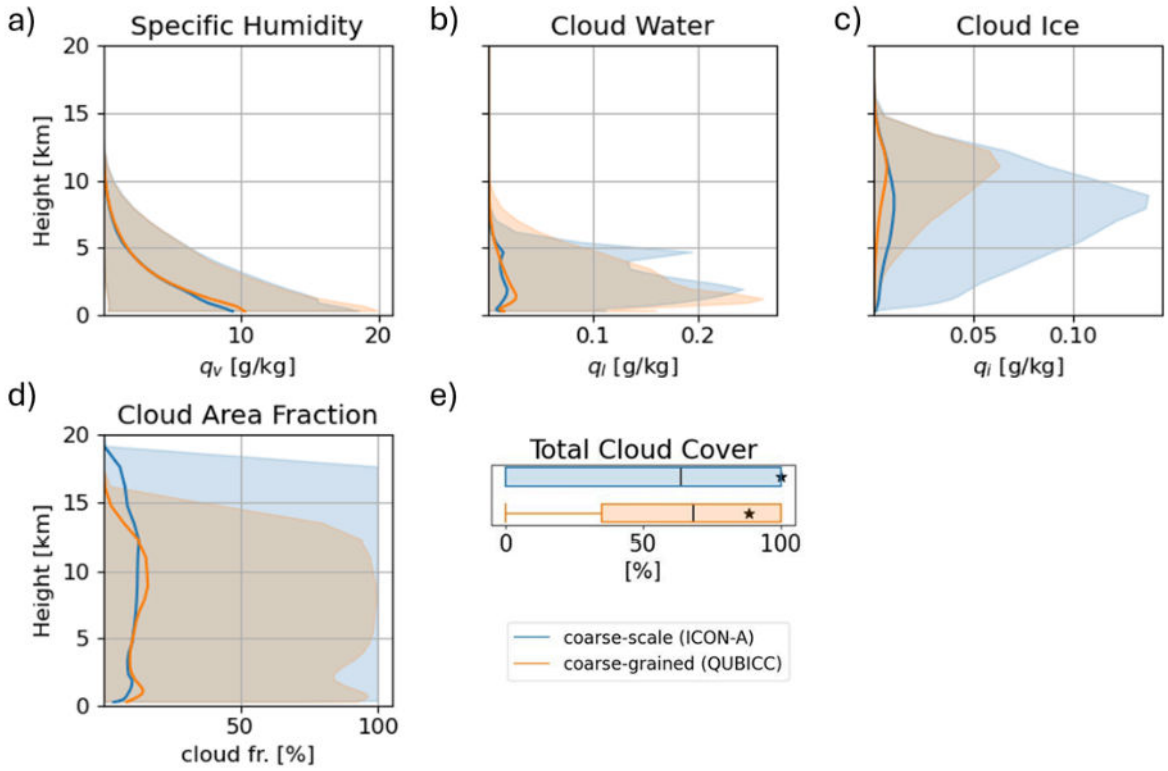


**Figure C.1.:** As Figure 6.2 but for the default microphysics scheme based on Lohmann and Roeckner (1996).

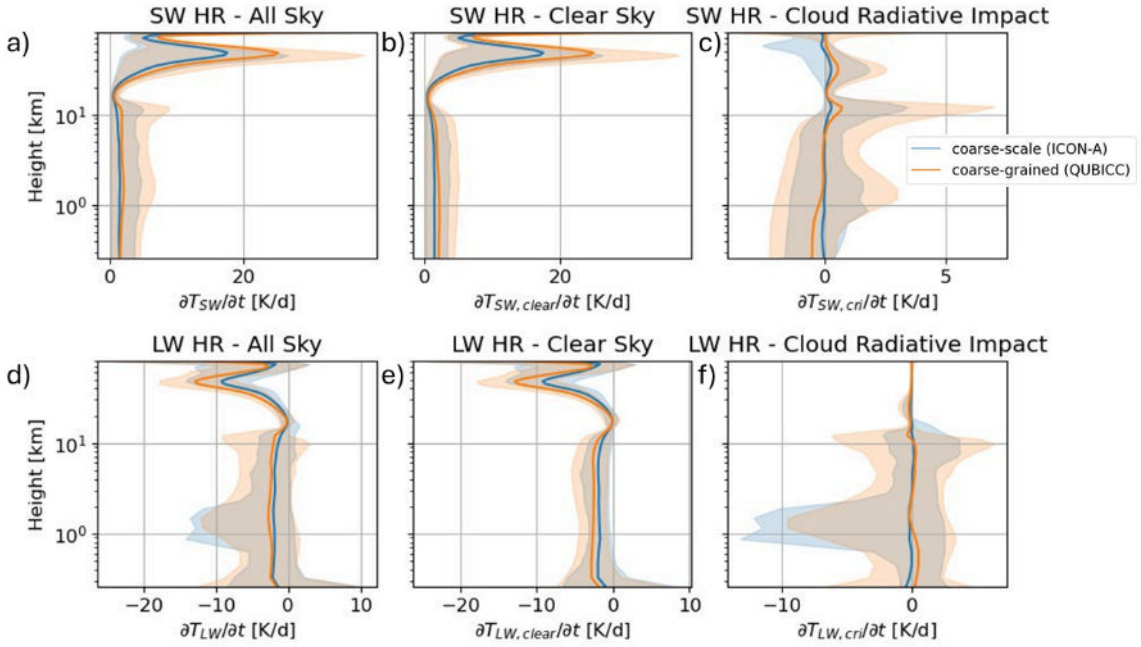


**Figure C.2.:** As Figure 6.3 but here the coarse grained heating rates are not scaled.

**Table C.1.:** Bulk statistics for heating rate results of the coarse-scale ML-based radiation emulator on coarse-grained QUBICC data. MAE is mean absolute error and $R^2$ is coefficient of determination. RMSE is root mean squared error. The percentage values in parentheses denote the relative values of MAE, bias and RMSE.

| | MAE $[K/d]$ | Bias $[K/d]$ | $R^2$ | RMSE $[K/d]$ |
|---|---|---|---|---|
| pyRTE | | | | |
| SW clear | 0.367 (8.47 %) | 0.234 (4.91 %) | 0.91 | 0.443 (10.56 %) |
| SW cloudy | 0.445 (16.73 %) | 0.204 (4.22 %) | 0.83 | 0.789 (41.54 %) |
| SW partial | 0.470 (12.24 %) | 0.273 (4.65 %) | 0.82 | 0.683 (23.55 %) |
| SW non-precip. clouds | 0.493 (12.11 %) | 0.292 (4.79 %) | 0.87 | 0.711 (19.98 %) |
| SW precip. clouds | 0.778 (32.62 %) | 0.244 (17.28 %) | 0.59 | 1.250 (58.35 %) |
| LW clear | 0.564 (23.56 %) | -0.349 (-9.85 %) | 0.83 | 0.677 (30.09 %) |
| LW cloudy | 0.862 (41.08 %) | -0.296 (-8.35 %) | 0.67 | 1.478 (81.55 %) |
| LW partial | 0.694 (25.83 %) | -0.328 (-7.74 %) | 0.56 | 1.112 (47.24 %) |
| LW non-precip. clouds | 0.725 (48.35 %) | -0.294 (-27.53 %) | 0.70 | 1.230 (73.66 %) |
| LW precip. clouds | 1.109 (77.76 %) | -0.337 (-23.85 %) | 0.34 | 1.578 (130.42 %) |
| ML-enhanced | | | | |
| SW clear | 0.049 (0.49 %) | -0.000 (-0.00 %) | 0.99 | 0.074 (0.72 %) |
| SW cloudy | 0.106 (4.46 %) | 0.012 (0.98 %) | 0.98 | 0.214 (11.30 %) |
| SW partial | 0.082 (2.00 %) | 0.005 (0.23 %) | 0.99 | 0.150 (5.00 %) |
| SW non-precip. clouds | 0.080 (1.57 %) | 0.006 (0.24 %) | 0.99 | 0.144 (3.50 %) |
| SW precip. clouds | 0.188 (9.16 %) | 0.024 (2.87 %) | 0.96 | 0.341 (17.71 %) |
| LW clear | 0.028 (2.58 %) | 0.003 (0.52 %) | 1.00 | 0.053 (5.50 %) |
| LW cloudy | 0.127 (7.42 %) | -0.001 (0.00 %) | 0.98 | 0.275 (17.70 %) |
| LW partial | 0.068 (3.33 %) | -0.001 (0.02 %) | 0.99 | 0.158 (8.70 %) |
| LW non-precip. clouds | 0.069 (4.43 %) | -0.002 (0.43 %) | 0.99 | 0.160 (9.41 %) |
| LW precip. clouds | 0.197 (19.37 %) | -0.004 (-0.99 %) | 0.96 | 0.319 (36.18 %) |

**Table C.2.:** Bulk statistics for results of cloud radiative effect on heating rates. MAE is the mean absolute error and $R^2$ is the coefficient of determination. RMSE is the root mean squared error. The percentage values in parentheses denote the relative values of MAE, bias and RMSE.

| | MAE $[K/d]$ | Bias $[K/d]$ | $R^2$ | RMSE $[K/d]$ |
|---|---|---|---|---|
| pyRTE | | | | |
| SW Arctic | 0.215 (9.55 %) | 0.114 (3.73 %) | 0.92 | 0.326 (17.99 %) |
| SW Southern Ocean | 0.417 (13.34 %) | 0.218 (3.32 %) | 0.87 | 0.665 (27.68 %) |
| SW Tropics | 0.535 (13.86 %) | 0.290 (4.39 %) | 0.80 | 0.817 (29.42 %) |
| SW Pacific ITCZ | 0.638 (18.33 %) | 0.271 (6.27 %) | 0.76 | 0.964 (33.67 %) |
| SW stratocumulus | 0.529 (13.66 %) | 0.289 (4.38 %) | 0.80 | 0.825 (28.97 %) |
| LW Arctic | 0.634 (30.75 %) | -0.254 (-14.89 %) | 0.77 | 1.046 (55.94 %) |
| LW Southern Ocean | 0.787 (30.60 %) | -0.318 (-8.64 %) | 0.69 | 1.307 (56.70 %) |
| LW Tropics | 0.755 (30.96 %) | -0.326 (-4.21 %) | 0.59 | 1.192 (65.09 %) |
| LW Pacific ITCZ | 0.944 (50.80 %) | -0.311 (-2.19 %) | 0.49 | 1.326 (99.80 %) |
| LW stratocumulus | 0.742 (31.38 %) | -0.327 (-6.13 %) | 0.58 | 1.234 (62.76 %) |
| ML-enhanced | | | | |
| SW Arctic | 0.056 (2.17 %) | 0.006 (0.58 %) | 0.99 | 0.098 (4.84 %) |
| SW Southern Ocean | 0.089 (2.98 %) | 0.005 (0.34 %) | 0.99 | 0.174 (7.30 %) |
| SW Tropics | 0.097 (2.61 %) | 0.009 (0.47 %) | 0.98 | 0.187 (7.32 %) |
| SW Pacific ITCZ | 0.142 (4.50 %) | 0.019 (1.17 %) | 0.98 | 0.247 (9.73 %) |
| SW stratocumulus | 0.095 (2.48 %) | 0.008 (0.41 %) | 0.99 | 0.179 (6.38 %) |
| LW Arctic | 0.071 (4.96 %) | -0.000 (0.01 %) | 0.99 | 0.176 (11.91 %) |
| LW Southern Ocean | 0.103 (4.86 %) | -0.002 (-0.07 %) | 0.99 | 0.231 (11.55 %) |
| LW Tropics | 0.079 (4.67 %) | -0.002 (-0.02 %) | 0.99 | 0.172 (14.14 %) |
| LW Pacific ITCZ | 0.132 (10.55 %) | -0.002 (-0.46 %) | 0.98 | 0.219 (23.47 %) |
| LW stratocumulus | 0.080 (4.75 %) | -0.001 (0.11 %) | 0.99 | 0.175 (11.59 %) |

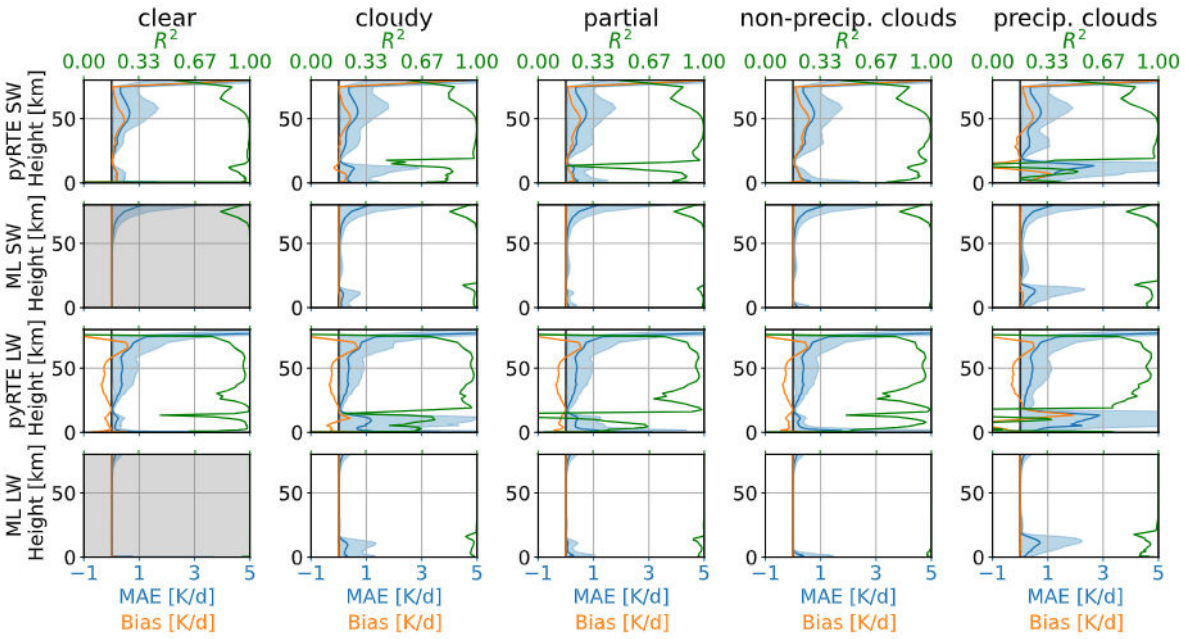


**Figure C.3.:** As Figure 6.4 but for the full column.

# List of Abbreviations

| | | |
|---|---|---|
| **AI** | Artificial Intelligence | 2 |
| **AMIP** | Atmospheric Model Intercomparison Project | 8 |
| **BiLSTM** | Bidirectional Long Short-Term Memory | 18 |
| **CPU** | Central Processing Unit | 14 |
| **CKD** | correlated k-distribution | 13 |
| **CMIP** | Coupled Model Intercomparison Project | 1 |
| **CMIP5** | Coupled Model Intercomparison Project Phase 5 | 11 |
| **CMIP6** | Coupled Model Intercomparison Project Phase 6 | 10 |
| **CRI** | Cloud Radiative Impact | 68 |
| **DNN** | Deep Neural Network | 15 |
| **DWD** | German Weather Service | 21 |
| **ECHAM** | EC from ECMWF and HAM for Hamburg | 23 |
| **ECMWF** | European Center for Medium Range Weather Forcasts | 23 |
| **ECS** | Equilibrium Climate Sensitivity | 1 |
| **ESM** | Earth System Model | 1 |
| **FNN** | Feedforward Neural Network | 15 |
| **GCM** | General Circulation Model | 8 |
| **GHG** | Greenhouse Gas | 2 |
| **GPU** | Graphics Processing Unit | 2 |
| **GSRM** | Global Storm Resolving Model | 2 |
| **HR** | Heating Rate | 32 |
| **ICA** | Independent Column Approach | 13 |
| **ICON** | ICOsahedral Nonhydrostatic | 2 |
| **IPCC** | Intergovernmental Panel on Climate Change | 8 |
| **ITCZ** | Intertropical Convergence Zone | 76 |
| **LSTM** | Long Short-Term Memory | 26 |
| **LW** | Longwave | 31 |

| | | |
|---|---|---|
| **MAE** | Mean Absolute Error | 36 |
| **McICA** | Monte Carlo Independent Column Approach | 14 |
| **ML** | Machine Learning | 2 |
| **MLP** | Multi Layer Perceptron | 46 |
| **MPI-M** | Max Planck Institute for Meteorology | 21 |
| **MSE** | Mean Squared Error | 16 |
| **NN** | Neural Network | 15 |
| **NWP** | Numerical Weather Prediction | 2 |
| **QUBICC** | Quasi-Biennial Oscillation In a Changing Climate | 24 |
| **QBO** | Quasi-Biennial Oscillation | 24 |
| **ReLU** | Rectified Linear Unit | 16 |
| **RMSE** | Root Mean Squared Error | 38 |
| **RRTM** | Rapid Radiative Transfer Model | 54 |
| **RRTMG** | RRTM for GCMs | 54 |
| **RRTMGP** | Rapid Radiative Transfer Model for GCMs - Parallel | 14 |
| **RTE+RRTMGP** | Radiative Transfer for Energetics + Rapid Radiative Transfer Model for GCMs - Parallel | 14 |
| **RTE** | Radiative Transfer for Energetics | 14 |
| **SGD** | Stochastic Gradient Descent | 17 |
| **SHAP** | SHapley Additive exPlanations | 27 |
| **SIC** | sea ice concentration | 8 |
| **SST** | sea surface temperature | 8 |
| **SW** | Shortwave | 31 |
| **TOA** | top of the atmosphere | 55 |
| **WRF** | Weather Research and Forcasting model | 18 |

# List of Figures

# List of Tables

# References


Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., ... Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation, 929–947. https://doi.org/10.1145/3620665.3640366

Atkinson, J., Elafrou, A., Kasoar, E., Wallwork, J. G., Meltzer, T., Clifford, S., Orchard, D., & Edsall, C. (2025). FTorch: a library for coupling PyTorch models to Fortran. *J. Open Source Softw.*, *10*(107), 7602.

Barker, H. W., Stephens, G. L., Partain, P. T., Bergman, J. W., Bonnel, B., Campana, K., Clothiaux, E. E., Clough, S., Cusack, S., Delamere, J., Edwards, J., Evans, K. F., Fouquart, Y., Freidenreich, S., Galin, V., Hou, Y., Kato, S., Li, J., Mlawer, E., ... Yang, F. (2003). Assessing 1D Atmospheric Solar Radiative Transfer Models: Interpretation and Handling of Unresolved Clouds. *Journal of Climate*, *16*(16), 2676–2699. https://doi.org/10.1175/1520-0442(2003)016<2676:adasrt>2.0.co;2

Bechtold, P., Köhler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, M. J., Vitart, F., & Balsamo, G. (2008). Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, *134*(634), 1337–1351. https://doi.org/10.1002/qj.289

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., Salzmann, M., & Schemm, S. (2025). Revisiting Machine Learning Approaches for Short- and Longwave Radiation Inference in Weather and Climate Models. https://doi.org/10.22541/essoar.173757583.33108291/v1

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., & Pritchard, M. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bock, L., & Lauer, A. (2024). Cloud properties and their projected changes in CMIP models with low to high climate sensitivity. *Atmospheric Chemistry and Physics*, *24*(3), 1587–1605. https://doi.org/10.5194/acp-24-1587-2024

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., Watanabe, M., & Webb, M. J. (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. https://doi.org/10.1038/ngeo2398

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., Watt-Meyer, O., & Bretherton, C. S. (2020). Machine Learning Climate Model Dynamics: Offline versus Online Performance. https://doi.org/10.48550/ARXIV.2011.03081

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., Perkins, W. A., Clark, S. K., & Harris, L. (2022). Correcting Coarse-Grid Weather and Climate Models by Machine Learning From Global Storm-Resolving Simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2). https://doi.org/10.1029/2021ms002794

Busecke, J. J. M., Balwada, D., Martin, P. E., Nicholas, T. E. G., Johnson, Z. C. P., Nalluri, P., Stern, C. I., & Abernathey, R. P. (2025). The Impact of Sub-Grid Heterogeneity on Air-Sea Turbulent Heat Flux in Coupled Climate Models. *Geophysical Research Letters*, *52*(13). https://doi.org/10.1029/2025gl114951

Cahalan, R. F., Ridgway, W., Wiscombe, W. J., Gollmer, S., & Harshvardhan. (1994). Independent Pixel and Monte Carlo Estimates of Stratocumulus Albedo. *Journal of the Atmospheric Sciences*, *51*(24), 3776–3790. https://doi.org/10.1175/1520-0469(1994)051<3776:ipamce>2.0.co;2

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S., Plattner, G.-K., & Tréguier, A.-M. (2021). Framing, Context, and Methods. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (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). Cambridge University Press. 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

Compton, A. H. (1923). A Quantum Theory of the Scattering of X-rays by Light Elements. *Physical Review*, *21*(5), 483–502. https://doi.org/10.1103/physrev.21.483

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

Cotronei, A., & Slawig, T. (2020). Single-precision arithmetic in ECHAM radiation reduces runtime and energy consumption. *Geoscientific Model Development*, *13*(6), 2783–2804. https://doi.org/10.5194/gmd-13-2783-2020

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

Dinh, T., Gasparini, B., & Bellon, G. (2023). Clouds and Radiatively Induced Circulations. https://doi.org/10.1002/9781119700357.ch11

Doktorowski, S., Kretzschmar, J., Quaas, J., Salzmann, M., & Sourdeval, O. (2024). Subgrid-scale variability of cloud ice in the ICON-AES 1.3.00. *Geoscientific Model Development*, *17*(8), 3099–3110. https://doi.org/10.5194/gmd-17-3099-2024

Doms, G., Förstner, G., Heise, E., Herzog, H.-J., Mironov, D., Raschendorfer, M., Reinhardt, T., Ritter, B., Schrodin, R., Schulz, J.-P., & Vogel, G. (2011). A Description of the Nonhydrostatic Regional COSMO Model. Part II: Physical Parameterization. *Consortium for Small-Scale Modelling*. http://www.cosmo-model.org/

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

Eyring, V., Gillett, N., Achuta Rao, K., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack, P., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., & Sun, Y. (2021). Human Influence on the Climate System. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (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. 423–551). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M., & Reichstein, M. (2024). AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nature Geoscience*, *17*(10), 963–971. https://doi.org/10.1038/s41561-024-01527-w

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., & Zhang, H. (2021a). The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (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. 923–1054). Cambridge University Press. https://doi.org/10.1017/9781009157896.009

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., & Zhang, H. (2021b). 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.

Fu, Q., & Liou, K. N. (1992). On the Correlated k-Distribution Method for Radiative Transfer in Nonhomogeneous Atmospheres. *Journal of Atmospheric Sciences*, *49*(22), 2139–2156. https://doi.org/10.1175/1520-0469(1992)049<2139:OTCDMF>2.0.CO;2

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

Gettelman, A., & Rood, R. B. (2016). *Demystifying Climate Models: A Users Guide to Earth System Models*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48959-8

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., Hohenegger, C., Kornblueh, L., Köhler, M., Manzini, E., Mauritsen, T., Nam, C., Raddatz, T., Rast, S., Reinert, D., Sakradzija, M., Schmidt, H., Schneck, R., Schnur, R., ... Stevens, B. (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

Giorgetta, M. A., Sawyer, W., Lapillonne, X., Adamidis, P., Alexeev, D., Clément, V., Dietlicher, R., Engels, J. F., Esch, M., Franke, H., Frauen, C., Hannah, W. M., Hillman, B. R., Kornblueh, L., Marti, P., Norman, M. R., Pincus, R., Rast, S., Reinert, D., ... Stevens, B. (2022). The ICON-A model for direct QBO simulations on GPUs (version icon-cscs:baf28a514). *Geoscientific Model Development*, *15*(18), 6985–7016. https://doi.org/10.5194/gmd-15-6985-2022

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* [http://www.deeplearningbook.org]. MIT Press.

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.**, Iglesias-Suarez, F., Shamekh, S., Gentine, P., Giorgetta, M. A., Pincus, R., & Eyring, V. (2025a). Interpretable Machine Learning-Based Radiation Emulation for ICON. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(4), e2024JH000501. https://doi.org/10.1029/2024JH000501

**Hafner, K.**, Iglesiaz-Suarez, F., Shamekh, S., Gentine, P., Pincus, R., Giorgetta, M., & Eyring, V. (2025b). Stable Machine Learning based Radiation Emulation for ICON. *Under Review for Journal of Advances in Modeling Earth Systems*. https://doi.org/10.22541/essoar.174708082.27787580/v1

**Hafner, K.**, Shamekh, S., Bertoli, G., Lauer, A., Pincus, R., Savre, J., & Eyring, V. (2025c). Representing Subgrid-Scale Cloud Effects in a Radiation Parameterization using Ma-

chine Learning: MLe-radiation v1.0. *Submitted to Geoscientific Model Development, arXiv preprint*. https://doi.org/10.48550/arXiv.2510.05963

**Hafner, K.** (2025a). Representing Subgrid-Scale Cloud Effects in a Radiation Parameterization using Machine Learning: MLe-radiation v1.0. https://doi.org/10.5281/ZENODO.17280639

**Hafner, K.** (2025b). *Interpretable Machine Learning-based Radiation Emulation for ICON [Software]*. Zenodo. https://doi.org/10.5281/ZENODO.15199158

Henn, B., Jauregui, Y. R., Clark, S. K., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., Pauling, A. G., & Bretherton, C. S. (2024). A Machine Learning Parameterization of Clouds in a Coarse-Resolution Climate Model for Unbiased Radiation. *Journal of Advances in Modeling Earth Systems*, *16*(3). https://doi.org/10.1029/2023ms003949

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

Hines, C. O. (1997). Doppler-spread parameterization of gravity-wave momentum deposition in the middle atmosphere. Part 1: Basic formulation. *Journal of Atmospheric and Solar-Terrestrial Physics*, *59*(4), 371–386. https://doi.org/10.1016/s1364-6826(96)00079-x

Hogan, R. J., & Shonk, J. K. P. (2013). Incorporating the Effects of 3D Radiative Transfer in the Presence of Clouds into Two-Stream Multilayer Radiation Schemes. *Journal of the Atmospheric Sciences*, *70*(2), 708–724. https://doi.org/10.1175/jas-d-12-041.1

Hogan, R. J., & Bozzo, A. (2018). A Flexible and Efficient Radiation Scheme for the ECMWF Model. *Journal of Advances in Modeling Earth Systems*, *10*(8), 1990–2008. https://doi.org/10.1029/2018ms001364

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

Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behravesh, M., Bergemann, M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G., … Stevens, B. (2023). ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development*, *16*(2), 779–811. https://doi.org/10.5194/gmd-16-779-2023

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research: Atmospheres*, *113*(D13). https://doi.org/10.1029/2008jd009944

ICON partnership (DWD and MPI-M and DKRZ and KIT and C2SM). (2024). ICON release 2024.01. https://doi.org/10.35089/WDCC/IconRelease2024.10

IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou, Eds.). Cambridge University Press. https://doi.org/10.1017/9781009157896

Jülich Supercomputing Centre. (2021). JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, *7*(A138). https://doi.org/10.17815/jlsrf-7-183

Jungclaus, J. H., Lorenz, S. J., Schmidt, H., Brovkin, V., Brüggemann, N., Chegini, F., Crüger, T., De-Vrese, P., Gayler, V., Giorgetta, M. A., Gutjahr, O., Haak, H., Hagemann, S., Hanke, M., Ilyina, T., Korn, P., Kröger, J., Linardakis, L., Mehlmann, C., . . . Claussen, M. (2022). The ICON Earth System Model Version 1.0 [e2021MS002813 2021MS002813]. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002813. https://doi.org/10.1029/2021MS002813

Khain, A. P., & Sednev, I. (1996). Simulation of precipitation formation in the Eastern Mediterranean coastal zone using a spectral microphysics cloud ensemble model. *Atmospheric Research*, *43*(1), 77–110. https://doi.org/10.1016/s0169-8095(96)00005-1

Khain, A., Pokrovsky, A., Pinsky, M., Seifert, A., & Phillips, V. (2004). Simulation of Effects of Atmospheric Aerosols on Deep Turbulent Convective Clouds Using a Spectral Microphysics Mixed-Phase Cumulus Cloud Model. Part I: Model Description and Possible Applications. *Journal of the Atmospheric Sciences*, *61*(24), 2963–2982. https://doi.org/10.1175/jas-3350.1

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980

Klöwer, M., Gelbrecht, M., Hotta, D., Willmert, J., Silvestri, S., Wagner, G. L., White, A., Hatfield, S., Kimpson, T., Constantinou, N. C., & Hill, C. (2024). SpeedyWeather.jl: Reinventing atmospheric general circulation models towards interactivity and extensibility. *Journal of Open Source Software*, *9*(98), 6323. https://doi.org/10.21105/joss.06323

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., & Hoyer, S. (2024). Neural general circulation models for weather and climate. *Nature*. https://doi.org/10.1038/s41586-024-07744-y

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). Captum: A unified

and generic model interpretability library for PyTorch. https://arxiv.org/abs/2009.07896

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

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. (2012). Accurate and fast neural network emulations of long and short wave radiation for the NCEP Global Forecast System model. Retrieved 2024, from https://repository.library.noaa.gov/view/noaa/6951

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

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lauer, A., Bock, L., Hassler, B., Schröder, M., & Stengel, M. (2023). Cloud Climatologies from Global Climate Models—A Comparison of CMIP5 and CMIP6 Models with Satellite Data. *Journal of Climate*, *36*(2), 281–311. https://doi.org/10.1175/jcli-d-22-0181.1

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp. In *Neural Networks: Tricks of the Trade* (pp. 9–48). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_3

Lohmann, U., & Roeckner, E. (1996). Design and performance of a new cloud microphysics scheme developed for the ECHAM general circulation model. *Climate Dynamics*, *12*(8), 557–572. https://doi.org/10.1007/bf00207939

Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. https://doi.org/10.48550/ARXIV.1711.05101

Lott, F., & Miller, M. J. (1997). A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, *123*(537), 101–127. https://doi.org/10.1002/qj.49712353704

Lott, F. (1999). Alleviation of Stationary Biases in a GCM through a Mountain Drag Parameterization Scheme and a Simple Representation of Mountain Lift Forces. *Monthly Weather Review*, *127*(5), 788–801. https://doi.org/10.1175/1520-0493(1999)127<0788:aosbia>2.0.co;2

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Manabe, S., & Wetherald, R. T. (1967). Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity. *Journal of the Atmospheric Sciences*, *24*(3), 241–259. https://doi.org/10.1175/1520-0469(1967)024<0241:teotaw>2.0.co;2

Manabe, S., & Bryan, K. (1969). Climate Calculations with a Combined Ocean-Atmosphere Model. *Journal of the Atmospheric Sciences*, *26*(4), 786–789. https://doi.org/10.1175/1520-0469(1969)026<0786:ccwaco>2.0.co;2

Manabe, S., Hasselmann, K., & Parisi, G. (2021). Nobel Prize in Physics 2021 [Accessed: 01.06.2025].

Mauritsen, T., Svensson, G., Zilitinkevich, S. S., Esau, I., Enger, L., & Grisogono, B. (2007). A Total Turbulent Energy Closure Model for Neutrally and Stably Stratified Atmospheric Boundary Layers. *Journal of the Atmospheric Sciences*, *64*(11), 4113–4126. https://doi.org/10.1175/2007jas2294.1

Meador, W. E., & Weaver, W. R. (1980). Two-Stream Approximations to Radiative Transfer in Planetary Atmospheres: A Unified Description of Existing Methods and a New Improvement. *Journal of the Atmospheric Sciences*, *37*(3), 630–643. https://doi.org/10.1175/1520-0469(1980)037<0630:tsatrt>2.0.co;2

Meyer, D., Hogan, R. J., Dueben, P. D., & Mason, S. L. (2022). Machine Learning Emulation of 3D Cloud Radiative Effects [e2021MS002550 2021MS002550]. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002550. https://doi.org/10.1029/2021MS002550

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

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997b). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, *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. https://doi.org/10.21957/ulnklieu

Muhlbauer, A., McCoy, I. L., & Wood, R. (2014). Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmospheric Chemistry and Physics*, *14*(13), 6695–6716. https://doi.org/10.5194/acp-14-6695-2014

National Oceanic and Atmospheric Administration. (2023). The Atmospheric Window [Retrieved: 18.04.2025]. https://www.noaa.gov/jetstream/satellites/absorb

Nordeng, T.-E. (1994). Extended versions of the convective parametrization scheme at ECMWF and their impact on the mean and transient activity of the model in the tropics. https://doi.org/10.21957/E34XWHYSW

Olah, C. (2015). Understanding LSTM networks [Retrieved: 22.04.2025]. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Orr, A., Bechtold, P., Scinocca, J., Ern, M., & Janiskova, M. (2010). Improved Middle Atmosphere Climate and Forecasts in the ECMWF Model through a Nonorographic Gravity Wave Drag Parameterization. *Journal of Climate*, *23*(22), 5905–5926. https://doi.org/10.1175/2010jcli3490.1

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

Phillips, N. A. (1956). The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, *82*(352), 123–164. https://doi.org/10.1002/qj.49708235202

Pincus, R., Barker, H. W., & Morcrette, J.-J. (2003). A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. *Journal of Geophysical Research: Atmospheres*, *108*(D13). https://doi.org/10.1029/2002jd003322

Pincus, R., & Stevens, B. (2009). Monte Carlo Spectral Integration: a Consistent Approximation for Radiative Transfer in Large Eddy Simulations. *Journal of Advances in Modeling Earth Systems*, *1*(2). https://doi.org/10.3894/james.2009.1.1

Pincus, R., & Stevens, B. (2013). Paths to accuracy for radiation parameterizations in atmospheric models. *Journal of Advances in Modeling Earth Systems*, *5*(2), 225–233. https://doi.org/10.1002/jame.20027

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

Pincus, R., makepath LLC, & Sehnem, J. M. (2025). *pyRTE-RRTMGP* (Version 0.1.4). https://doi.org/10.5281/zenodo.16644555

Prill, F., Reinert, D., Rieger, D., & Zängl, G. (2024). ICON Tutorial 2024: Working with the ICON Model. https://doi.org/10.5676/DWD_PUB/NWV/ICON_TUTORIAL2024

Räisänen, P., Barker, H. W., Khairoutdinov, M. F., Li, J., & Randall, D. A. (2004). Stochastic generation of subgrid-scale cloudy columns for large-scale models. *Quarterly Journal of the Royal Meteorological Society*, *130*(601), 2047–2067. https://doi.org/10.1256/qj.03.99

Raschendorfer, M. (2001). *The new turbulence parameterization of LM*. Consortium for Small-Scale Modelling. http://www.cosmo-model.org/

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

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg,

C., Sisk, J., Barrington, L., Bell, A., & Sha, F. (2023). WeatherBench 2: A benchmark for the next generation of data-driven global weather models.

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and Lorenz96 case study (v1.0). https://doi.org/10.5194/gmd-2019-319

Reick, C. H., Gayler, V., Goll, D., Hagemann, S., Heidkamp, M., Nabel, J. E. M. S., Raddatz, T., Roeckner, E., Schnur, R., & Wilkenskjeld, S. (2021). JSBACH 3 - The land component of the MPI Earth System Model: documentation of version 3.2. https://doi.org/10.17617/2.3279802

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

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. https://doi.org/10.48550/ARXIV.1402.1128

Schulz, J.-P., Vogel, G., Becker, C., Kothe, S., Rummel, U., & Ahrens, B. (2016). Evaluation of the ground heat flux simulated by a multi-layer land surface scheme using high-quality observations at grass land and bare soil. *Meteorologische Zeitschrift*, *25*(5), 607–620. https://doi.org/10.1127/metz/2016/0537

Seifert, A., & Beheng, K. D. (2005). A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description. *Meteorology and Atmospheric Physics*, *92*(1–2), 45–66. https://doi.org/10.1007/s00703-005-0112-4

Seifert, A. (2008). *A revised cloud microphysical parameterization for COSMO-LME*. Consortium for Small-Scale Modelling. http://www.cosmo-model.org/

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., & Zhou, B. (2021). Weather and Climate Extreme Events in a Changing Climate. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (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. 1513–1765). Cambridge University Press. https://doi.org/10.1017/9781009157896.013

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

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U.,

Pincus, R., Reichler, T., & Roeckner, E. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., Kodama, C., Kornblueh, L., Lin, S.-J., Neumann, P., Putman, W. M., Röber, N., Shibuya, R., Vanniere, B., Vidale, P. L., . . . Zhou, L. (2019). DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. *Progress in Earth and Planetary Science*, *6*(1). https://doi.org/10.1186/s40645-019-0304-z

Strong, J., & Plass, G. N. (1950). The Effect of Pressure Broadening of Spectral Lines on Atmospheric Temperature. *The Astrophysical Journal*, *112*, 365. https://doi.org/10.1086/145352

Sundqvist, H., Berge, E., & Kristjánsson, J. E. (1989). Condensation and Cloud Parameterization Studies with a Mesoscale Numerical Weather Prediction Model. *Monthly Weather Review*, *117*(8), 1641–1657. https://doi.org/10.1175/1520-0493(1989)117<1641:cacpsw>2.0.co;2

Tiedtke, M. (1989). A Comprehensive Mass Flux Scheme for Cumulus Parameterization in Large-Scale Models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493(1989)117<1779:acmfsf>2.0.co;2

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

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., & Hogan, R. J. (2024). Twelve Times Faster yet Accurate: A New State-Of-The-Art in Radiation Schemes via Performance and Spectral Optimization. *Journal of Advances in Modeling Earth Systems*, *16*(1). https://doi.org/10.1029/2023ms003932

Ukkonen, P., & Chantry, M. (2025). Vertically Recurrent Neural Networks for Sub-Grid Parameterization [e2024MS004833 2024MS004833]. *Journal of Advances in Modeling Earth Systems*, *17*(6), e2024MS004833. https://doi.org/10.1029/2024MS004833

Ukkonen, P. (2022). Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer. *14*. https://doi.org/10.1029/2021MS002875

UNFCCC. (2015). Adoption of the Paris Agreement. https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

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). Academic press.

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., Harris, L., & Bretherton, C. S. (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). https://doi.org/10.1029/2023ms003668

Weiss, P., Herbert, R., & Stier, P. (2025). ICON-HAM-lite 1.0: simulating the Earth system with interactive aerosols at kilometer scales. *Geoscientific Model Development*, *18*(12), 3877–3894. https://doi.org/10.5194/gmd-18-3877-2025

Wood, R. (2012). Stratocumulus Clouds. *Monthly Weather Review*, *140*(8), 2373–2423. https://doi.org/10.1175/mwr-d-11-00121.1

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

Yin, J., & Porporato, A. (2017). Diurnal cloud cycle biases in climate models. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02369-4

Young, T. (1804). I. The Bakerian Lecture. Experiments and calculations relative to physical optics. *Philosophical Transactions of the Royal Society of London*, *94*, 1–16. https://doi.org/10.1098/rstl.1804.0001

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

Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2014). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 563–579. https://doi.org/10.1002/qj.2378

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1). https://doi.org/10.1029/2019gl085782

Zhao, Y., Li, J., Wen, D., Li, Y., Wang, Y., & Huang, J. (2024). Distinct structure, radiative effects, and precipitation characteristics of deep convection systems in the Tibetan Plateau compared to the tropical Indian Ocean. *Atmospheric Chemistry and Physics*, *24*(16), 9435–9457. https://doi.org/10.5194/acp-24-9435-2024

Zhong, X., Ma, Z., Yao, Y., Xu, L., Wu, Y., & Wang, Z. (2023). WRF–ML v1.0: a bridge between WRF v4.3 and machine learning parameterizations and its application to atmospheric

radiative transfer. *Geoscientific Model Development*, *16*(1), 199–209. https://doi.org/10.5194/gmd-16-199-2023

# Acknowledgments

I would like to acknowledge the financial and computational support. I was 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). I was also supported by a fellowship of the German Academic Exchange Service (DAAD) that allowed me to visit Columbia University. Additionally, I 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). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179. I gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Jülich Supercomputing Centre 2021) at the Jülich Supercomputing Centre (JSC).

First, I would like to thank my supervisor Prof. Dr. Veronika Eyring for supporting and guiding me throughout this project. I am grateful for my second supervisor Prof. Dr. Pierre Gentine, who hosted my six-month visit at Columbia University and never failed to suggest more ideas. Special thanks also go to Prof. Dr. Sara Shamekh for endless discussions and hosting a visit at New York University. My sincere thanks go to Prof. Dr. Robert Pincus for support from the radiation side. I am grateful for the guidance provided by Dr. Fernando Iglesiaz-Suarez during the first half of this PhD project. I acknowledge the contribution from my co-authors Dr. Marco Giorgetta, Dr. Guillaume Bertoli, Dr. Axel Lauer and Dr. Julien Savre-Piou with special thanks to Julien for creating the QUBICC simulations. I would like to thank everyone who provided helpful comments while proofreading the thesis including my previously mentioned co-authors, as well as Dr. Gunnar Behrens, Dr. Tina Gier, Dr. Mierk Schwabe and Dr. Katja Weigel. Thank you to all my other colleagues, office mates, and people I met during my visits in New York. I enjoyed many coffees and conversations with you.

Additionally, I would like to thank my family and friends for their support throughout the last years.

Finally, to my partner Peter Fackeldey, thank you for your constant understanding and tireless, invaluable encouragement.