Application of Machine Learning in Predicting Gaseous properties of Earth's Atmosphere

Anna Binoy
School of Physical Sciences
National Institute of Science Education and Research
Odisha
anna.binoy@niser.ac.in

Sumegha M. T.
School of Physical Sciences
National Institute of Science Education and Research
Odisha
sumegha.mt@niser.ac.in

Abstract

The simulation of Earth's atmosphere using gas optics computation is a complex task, often requiring significant computational resources. Machine learning techniques have shown great potential for accelerating these simulations, reducing both the time and resources required. This paper presents a study of the use of machine learning methods to predict gas optics computation of Earth's atmosphere. The study compares the performance of traditional gas optics computation methods to a machine learning-based approach, using a range of atmospheric conditions and scenarios. The results demonstrate the accuracy of simple machine learning methods such as Random Forest and XG Boost in predicting atmospheric optical properties, which, in most models are stored in pre-calculated Look Up Tables.

1 Introduction

Radiative transfer is a fundamental process in atmospheric physics that plays a critical role in shaping the Earth's climate. It refers to the transport of energy by electromagnetic waves through a medium. In the context of Earth's atmosphere the transfer of energy occurs between different layers of the atmosphere, as well as between the atmosphere and the Earth's surface. The impact of radiative transfer on the Earth's environment and climate is significant, as it provides the necessary light for human and plant life, and influences the propagation of electromagnetic radiation through the atmosphere. This process is strongly influenced by atmospheric components such as aerosols, clouds, and some gases. The equation for radiative transfer is derived as

$$\frac{dI(\hat{\mathbf{\Omega}})}{d\tau} = -I(\hat{\mathbf{\Omega}}) + (1-\tilde{\omega})B + \frac{\tilde{\omega}}{4\pi}\int_{4\pi} p\left(\hat{\mathbf{\Omega}}', \hat{\mathbf{\Omega}}\right) I\left(\hat{\mathbf{\Omega}}'\right) d\omega' \tag{1}$$

Here, the optical depth, τ, is defined as:

$$\tau = \int \kappa ds \tag{2}$$

Here, I is the specific intensity, s is the distance along the path of the radiation, $B(T)$ is the Planck source function at temperature T, $\tilde{\omega}$ is the single scattering albedo (SSA), κ is the absorption

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

coefficient and the the scattering phase function $p\left(\hat{\mathbf{\Omega}}', \hat{\mathbf{\Omega}}\right)$ is required to satisfy the normalization condition

$$\frac{1}{4\pi}\int_{4\pi} p\left(\hat{\Omega}', \hat{\mathbf{\Omega}}\right) d\omega' = 1$$

Here, $\hat{\Omega}$ is the direction of interest and $\hat{\Omega}'$ is any arbitrary direction. In the Earth's atmosphere, solar radiation is absorbed and scattered by air molecules. An isolated atom can absorb photons at discrete frequencies that correspond to the energy levels of its electron shell, while molecules can absorb photons at a large number of discrete frequencies due to their additional energy levels associated with additional degrees of freedom.

However, in reality, this absorption takes place over a range of frequencies rather than at discrete values due to two primary effects. First, the movement of gas molecules results in Doppler broadening of absorption frequencies, which broadens the range of frequencies over which absorption occurs. Second, collisions between molecules lead to pressure broadening of absorption lines, which also contributes to the broadening of the absorption spectrum.

The Earth's atmosphere is not completely transparent to solar radiation due to the presence of various atmospheric components such as water vapor, clouds, ozone, and carbon dioxide. Water vapor is a particularly strong absorber of solar radiation due to its bent triatomic structure and permanent dipole. Its concentration is primarily controlled by temperature through the Clausius-Clapeyron relation (3), although deviations from this relationship can occur when air is sub-saturated.

$$\frac{dP}{dT} = \frac{Q}{T(V^I - V^{II})} \tag{3}$$

Here Q is the molar heat of transition, T is the temperature, P is the pressure . V^I is the molar volume of phase I and V^{II} is the same for phase II. Ozone, on the other hand, is primarily involved in photodissociation in the stratosphere, which absorbs UV light emitted by the sun. While ozone also absorbs in the infrared, its role in absorbing UV light makes it particularly important for the Earth's climate.

In addition to the atmospheric components mentioned above, aerosol particles also have a significant impact on atmospheric radiative transfer. These particles can scatter and absorb solar and infrared radiation, altering the air temperature and the rates of photochemical reactions. Aerosols can scatter visible light and contribute to the formation of clouds, while absorbing aerosols can warm the atmosphere.

Atmospheric scientists use Radiative Transfer Models (RTMs) to solve the radiative transfer problem. These models make an assumption that the radiation does not propagate horizontally(Independent Column Approximation), which simplifies the complexity of the Radiative Transfer problem by reducing its dimensionality. In addition, the difficulty of solving the radiative transfer equation for all wavelengths is mitigated by grouping optically spectral regions through a technique known as correlated k-distribution ([6]).
RTE+RRTMGP ([4]) is one such state-of-the-art radiative transfer model. It is used to simulate the transfer of solar and thermal radiation through the Earth's atmosphere. RTE) and RRTM for General circulation model applications — Parallel (RRTMGP) computes optical properties and source functions and Radiative Transfer for Energetics (RTE) computes fluxes from the output of RRTMGP, given a description of boundary conditions. The combined package RTE-RRTMGP computes radiative fluxes across the wavelength bands from vertical profiles of temperature, pressure, and the concentration of various atmospheric gases. The combined package can be used to compute broadband radiative fluxes from input profiles of temperature, pressure and gas concentrations. To speed up the calculations, RRTMGP uses pre-calculated look-up tables (LUTs). These LUTs contain pre-computed radiative transfer solutions for a range of atmospheric conditions and are interpolated to obtain the solution of the current atmospheric state. They are generated by running the RRTMGP model for a set of representative atmospheric profiles, and then storing the results in a data structure that can be quickly accessed during runtime. However, LUTs have limitations, especially in terms of speed and accuracy. Moreover,LUTs require large amount of storage space to store the pre-calculated radiative transfer solutions for different atmospheric conditions. RTE+RRTMGP can simulate a wide range of atmospheric conditions. Hence, they are useful in modelling the climate of the past, present and the future. The model is also able to simulate the effects of aerosols, such as dust and pollution,

Topic	Year	ML models used
Predicting atmospheric optical properties for radiative transfer computations using neural networks	2021	Feedforward Neural Network (FNN)
Accelerating Radiation Computations for Dynamical Models With Targeted Machine Learning and Code Optimization	2020	Feedforward Neural Network (FNN)
Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer	2022	Feedforward Neural Network (FNN) and Recurrent Neural Network (RNN)
RadNet 1.0: exploring deep learning architectures for longwave radiative transfer	2021	Feedforward Neural Network (FNN) and Convoluted Neural Network (CNN)

Table 1: Related works and the machine learning models used .

on radiation, to some accuracy. This has a significance in the study of global warming. However, the calculations involved are computationally expensive, especially when performed over diverse atmospheric profiles. Additionally, due to the fact that LUTs are only accurate within the range of atmospheric conditions for which they have been pre-calculated.
Efforts have been made to reduce the computational cost of RRTMGP model by incorporating neural networks ([2], [1]) to predict the gaseous optical properties. In this experiment, machine learning models such as Random Forest, XG Boost etc., are being employed to predict the gaseous optical properties of the atmosphere.

2 Literature Review

The radiative transfer equation describes how radiation propagates through a medium, such as the Earth's atmosphere, In order to solve this equation, it is necessary to determine the optical properties of the medium, such as absorption and scattering coefficients, at different wavelengths. However, the optical properties of the atmosphere exhibit significant variability due to the presence of various gases, including water vapor, carbon dioxide, and ozone, each with distinct absorption spectra. To accurately model the propagation of radiation through the Earth's atmosphere and its interaction with the surface and atmosphere, it is necessary to consider the variability of the optical properties of the atmosphere. This requires a set of state information, including profiles of temperature, pressure, and gas concentrations, which are used to compute the optical properties at discrete spectral quadrature points. By mapping the state information to these optical properties, the radiative transfer equation can be solved accurately. This approach allows for a comprehensive understanding of how radiation propagates through the atmosphere, making it crucial in fields such as atmospheric science and climate modeling. Menno Veerman's study, described in his paper ([1]), explores the possibility of using deep neural networks to predict key atmospheric optical properties such as Planck source function, single scattering albedo (SSA), and optical depth for both long and short wave bands, by exploring their empirical nature. The study uses a dataset of 100 atmospheric profiles obtained from the Radiative Forcing Model Intercomparison Project (RFMIP) ([7]), and only temperature (T), pressure (p), and the concentrations of H_2O and O_3 are used as inputs. It is assumed that all other gases except H_2O and O_3 are well mixed across all atmospheric layers.

To generate additional data, the set of 100 atmospheric profiles is randomly perturbed. The input profiles are then labelled using RTE-RRTMGP, that is, RRTMGP part of the model is

utilised to output the optical properties of the profile. The RRTMGP model employs a correlated k-distribution to cover the relevant spectral range of radiation for atmospheric problems, which includes 14 shortwave bands (0.2-12 m), 16 longwave bands (3-1000 m), and 16 g-points per band. This is done in order to reduce the computational complexity of calculating the optical properties for each wavelength. In correlated-k distribution, the wavelengths that correspond to an absorption cross section is grouped together. To accurately model radiation propagation through the atmosphere, we need to predict 224 shortwave optical property values and 256 longwave optical depth values. Additionally, RRTMGP calculates the Planck source functions at each layer interface separately using the interface temperatures and wavelength to g-point mapping of the layers above and below. Therefore, we also need to predict the upward and downward emissions at each interface, resulting in 768 values per grid cell.The paper has used MSE of the predicted predicted optical properties and the optical properties obtained using RRTMGP as the loss function. Moreover, validation is performed at every 10 epochs.The author calculated the coefficient of determination (R^2) for each g-point individually and then computed the average over all 224 and 256 g-points. This paper also discuss the challenges associated with using machine learning in climate research, such as the need for large and diverse datasets and the potential for biases.

Peter Ukkonnen's paper ([2]) also focuses on the emulating gas optics computations using Neural Networks due to its empirical nature, but uses more input parameters and different datasets. The Neural Networks were FORTRAN based and hence more faster.
Other related papers referred in this study also tries to achieve machine-specific optimisation using complex Neural Networks.
Here, in this paper, simpler models such as Random Forest and XG Boost were used to emulate gas optics computations.

3 Experiments

3.1 Dataset

The main dataset is downloaded from CMIP6 (Coupled Model Intercomparison Project 6) Forcing Datasets (input4MIPs) run by the UColorado [7]). It includes atmospheric profiles from 18 experiments at 100 different locations, each comprising 60 layers and spanning a vertical distance of 10 km. This dataset provides information for radiative transfer calculations, such as temperature, pressure, and gas concentration profiles, as well as boundary conditions like total solar irradiance, solar zenith angle, and surface temperature. From this dataset, layer pressure, level pressure, layer temperature, level temperature, volume mixing ratios of various gases, surface emissivity, ice water path, liquid water path, cosine of solar zenith angle, spectral direct surface albedo, and spectral diffuse surface albedo are extracted and stored in another netCDF file, where the pressure, temperature, and water vapor concentration profiles are randomly perturbed. This is the input file used for training.

This data are fed into the RTE-RRTMGP model to obtain optical properties corresponding to each g-point of all the grid cells. Output file contains information about g-points, long wave and short wave band limits, SSA, Planck Source function, shortwave optical depth (τ_{SW}) and longwave optical depth (τ_{LW}). A single input file contains 60x100 data points The labelled output file contains 256x60x100 data points for long wavelength and 244x60x100 data points for short wavelength. Multiple input files can be concatenated into a single netCDF file, and the corresponding output files can also be concatenated into another netCDF file. Therefore, to train the models, the input files for 6, 10, 20, 60, and 100 atmospheric profiles were concatenated, as well as their corresponding output files.
Additionally, certain input-output file pairs were separated into upper (36) and lower(24) atmospheric profiles.

The codes required to process the datasets are uploaded in the Github repository.

3.2 Model Performance

In this study, we utilized Layer pressure, layer temperature, water vapor concentration, and ozone concentration as input features to forecast the Planck source function, optical depth short wavelength (τ_{sw}), optical depth long wavelength (τ_{lw}), and Single Scattering Albedo. To achieve this, we employed Random Forest and XG Boost algorithms and generated four distinct models for a single

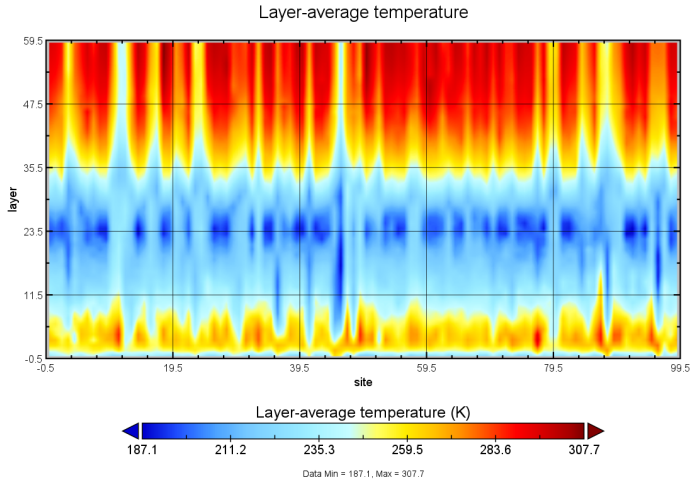

Figure 1: Layer Average Temperature over 100 sites(x-axis)with atmospheric profiles with 60 vertical grid cells(y-axis).

dataset. The accuracy score and MSE value of each model are presented in Table 2.For the data set which contains 60 grid cells per atmospheric profile the hyperparameters used to train the Random Forest Model are $n_{estimator}$=100 and maximum depth of a decision tree max_{depth}=10.

Then we trained the random forest model using the hyperparameter Gridsearch from a range of parameters for $n_estimators$ and maximum depth using the dataset split into lower(first 24 grid cells) and upper(next 36 grid cells) atmosphere as shown in table 3.This is evident in the change in gradient of the layer temperature(the input data we used) as shown in figure 1.

Output	Atmosphere	Dataset	MSE	Accuracy	Maximum Depth of a decision Tree(max. depth)	No. of Decision trees(n-estimators)
Planck Source Function	Lower	60	0.156803	0.687026	20	500
		100	0.084088	0.823788	20	200
	Upper	60	0.044074	0.779488	20	500
optical depth (Short wavelength)	Lower	60	1315943	0.257394	20	500
		100	1767329	0.470325	20	500
	Upper	60	3467961	0.230475	20	100
optical depth (Long wavelength)	Lower	60	29990350	0.596966	20	200
		100	15090078	0.748676	20	500
	Upper	60	6121689	0.187367	20	500
Single Scattering Albedo	Lower	60	0.016479	0.877228	20	500
		100	0.006626	0.951047	20	200
	Upper	60	0.034433	0.752852	20	500

Table 3: Random Forest Algorithm used to predict Planck Source Function,Optical Depth Short Wavelength(τ_{SW}),Optical Depth Long Wavelength(τ_{LW}),Single Scattering Albedo . The MSE and Model Accuracy values were used to evaluate the prediction accuracy.

3.3 Graphs

The correlation plots of the actual value optical property and the value predicted using our machine learning model are plotted in figure 2,3,4,5.The figures on the left is the density scatter plot of the actual vs predicted value of 100 datasets each containing 100 atmospheric profiles with 60 vertical grid cells each.While the figure on the right contains 100 dataset each containing 100 atmospheric profiles with first 24 vertical grid cells each depicting the lower atmosphere.

Output	Dataset (x60x100)	Model	MSE	Model Accuracy
Planck Source Function	6	Random Forest	0.169488	0.550532
		XG Boost	0.668303	0.125143
	10	Random Forest	0.225189	0.385423
	20	Random Forest	0.173947	0.504639
	60	Random Forest	0.206073	0.418546
	100	Random Forest	0.184279	0.459319
optical depth (Short wavelength)	6	Random Forest	547682.3	0.544293
		XG Boost	295694.5	0.754909
	10	Random Forest	1485905	0.12008
	20	Random Forest	6409936	0.373953
	60	Random Forest	2769729	0.118147
	100	Random Forest	4832451	0.165518
Single Scattering Albedo	6	Random Forest	0.040752	0.702703
		XG Boost	0.034746	0.746511
	10	Random Forest	0.050555	0.630096
	20	Random Forest	0.042998	0.679089
	60	Random Forest	0.050555	0.631237
	100	Random Forest	0.048486	0.649238
optical depth (Long wavelength)	6	Random Forest	0.169488	0.550532
		XG Boost	0.668303	0.125143
	10	Random Forest	0.225189	0.385423
	20	Random Forest	0.173947	0.504639
	60	Random Forest	0.206073	0.418546
	100	Random Forest	0.184279	0.459319

Table 2: Two machine learning methods (Random Forest, XG Boost) used to predict Planck Source Function,Optical Depth Short Wavelength(τ_{SW}),Optical Depth Long Wavelength(τ_{LW}),Single Scattering Albedo. The MSE and Model Accuracy values were used to evaluate the prediction accuracy.

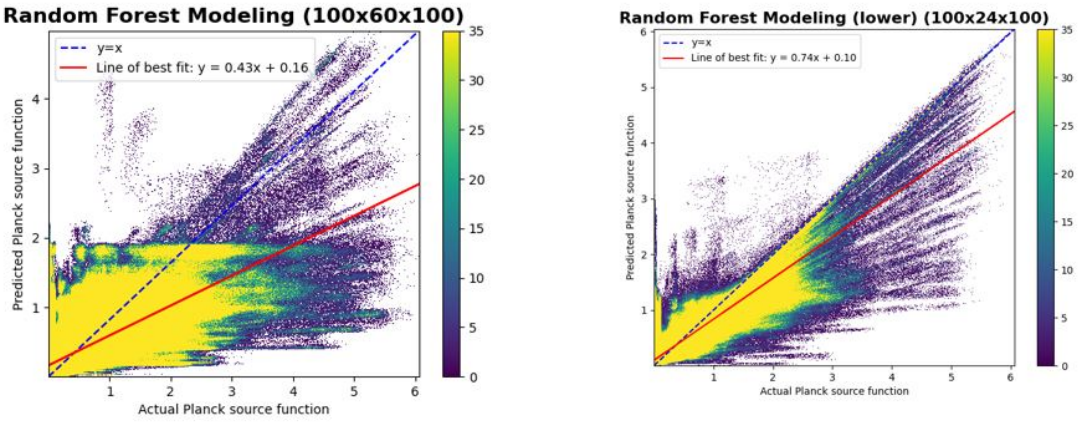

Figure 2: Density Scatter plot of actual value of planck source function obtained from RRTMGP vs value of planck source function predicted using random forest model.

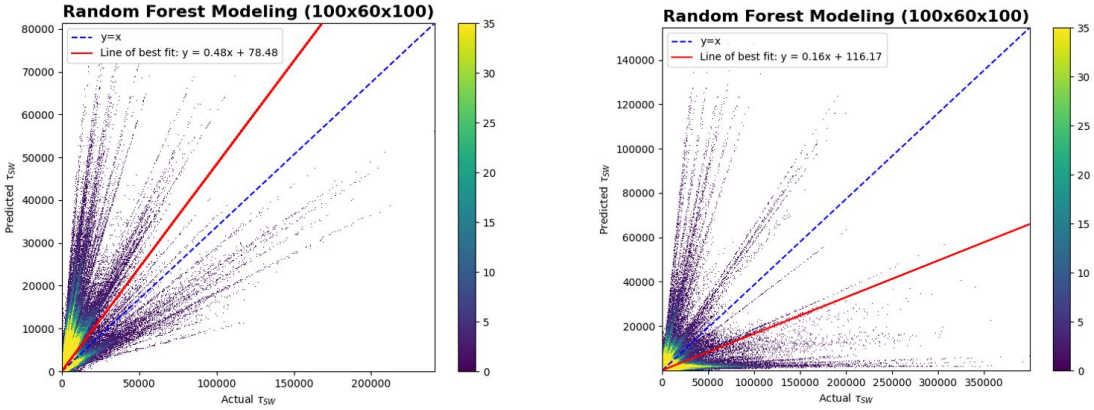

Figure 3: Density Scatter plot of actual value of Optical Depth short wavelength,τ_{sw} obtained from RRTMGP vs value of Optical Depth short wavelength,τ_{sw} predicted using random forest model.

4 Conclusions

From the accuracy score it is clear that the Random Forest is working relatively good for predicting planck source function and single scattering albedo when compared to the predicting of optical depth.Also from fig 2,3,4, and 5 we see that even a simple model like Random forest ,XG Boost etc. can give really good result when trained using huge set of data and taking into account the right physics behind earth atmosphere.All the papers that we have referred and their connected papers uses only Artificial Neural Networks of Various kind.One major obstacle we encountered was the significant difference in size between the input and output data, which resulted in the GPU memory being saturated and preventing us from training the models(even the simpler ones) we had developed with large datasets.

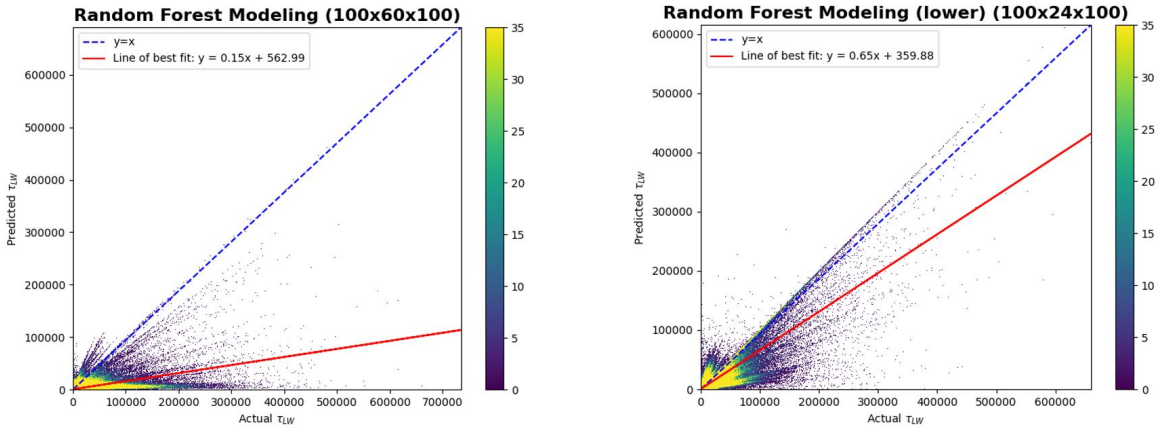

Figure 4: Density Scatter plot of actual value of Optical Depth long wavelength,τ_{lw} obtained from RRTMGP vs value of Optical Depth long wavelength,τ_{lw} predicted using random forest model.

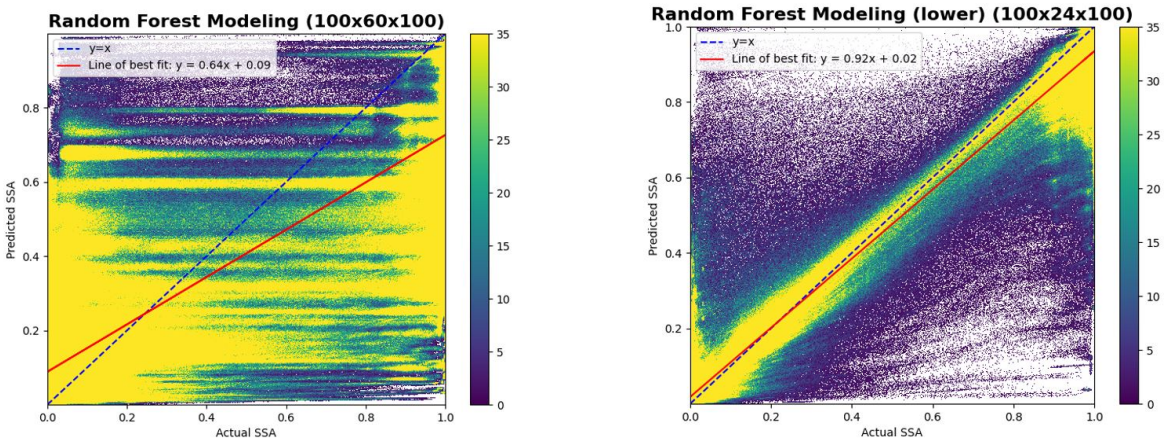

Figure 5: Density Scatter plot of actual value of single scattering albedo obtained from RRTMGP vs value of single scattering albedo predicted using random forest model.

References

[1] Menno A Veerman, Robert Pincus, Robin Stoffer, Caspar M Van Leeuwen, Damian Podareanu, and Chiel C Van Heerwaarden. Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, 379 (2194):20200095, 2021.

[2] Peter Ukkonen, Robert Pincus, Robin J Hogan, Kristian Pagh Nielsen, and Eigil Kaas. Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, 12(12):e2020MS002226, 2020.

[3] Robert Pincus. input4mips.cmip6.rfmip.ucolorado.ucolorado-rfmip-1-2, 2019. URL `https://doi.org/10.22033/ESGF/input4MIPs.10421`.

[4] Robert Pincus, Eli J Mlawer, and Jennifer S Delamere. Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, 11(10):3074–3089, 2019.

[5] Joy Merwin Monteiro and Rodrigo Caballero. The climate modelling toolkit. In *Proceedings of the 15th Python in Science Conference*, pages 69–74, 2016.

[6] Qiang Fu and KN Liou. On the correlated k-distribution method for radiative transfer in nonhomogeneous atmospheres. *Journal of Atmospheric Sciences*, 49(22):2139–2156, 1992.

[7] Robert Pincus, Piers M Forster, and Bjorn Stevens. The radiative forcing model intercomparison project (rfmip): experimental protocol for cmip6. *Geoscientific Model Development*, 9(9): 3447–3460, 2016.