A Physics-Incorporated Deep Learning Framework for Parameterization of Atmospheric Radiative Transfer

Yichen Yao[1⋆], Xiaohui Zhong[1⋆], Yongjun Zheng[2], and Zhibin Wang[1]

[1]Damo Academy, Alibaba Group, Hangzhou 311121, China
[2]Nanjing University of Information Science and Technology, Nanjing 210044, China

[⋆] The two authors contributed equally to this paper.

Corresponding author: Yongjun Zheng, zhengyongjun@gmail.com

Abstract

The atmospheric radiative transfer calculations are among the most time-consuming components of the numerical weather prediction (NWP) models. Deep learning (DL) models have recently been increasingly applied to accelerate radiative transfer modeling. Besides, a physical relationship exists between the output variables, including fluxes and heating rate profiles. Integration of such physical laws in DL models is crucial for the consistency and credibility of the DL-based parameterizations. Therefore, we propose a physics-incorporated framework for the radiative transfer DL model, in which the physical relationship between fluxes and heating rates is encoded as a layer of the network so that the energy conservation can be satisfied. It is also found that the prediction accuracy was improved with the physic-incorporated layer. In addition, we trained and compared various types of deep learning model architectures, including fully connected (FC) neural networks (NNs), convolutional-based NNs (CNNs), bidirectional recurrent-based NNs (RNNs), transformer-based NNs, and neural operator networks, respectively. The offline evaluation demonstrates that bidirectional RNNs, transformer-based NNs, and neural operator networks significantly outperform the FC NNs and CNNs due to their capability of global perception. A global perspective of an entire atmospheric column is essential and suitable for radiative transfer modeling as the changes in atmospheric components of one layer/level have both local and global impacts on radiation along the entire vertical column. Furthermore, the bidirectional RNNs achieve the best performance as they can extract information from both upward and downward directions, similar to the radiative transfer processes in the atmosphere.

Plain Language Summary

Numerical weather prediction (NWP) models require a lot of computational resources and time to run. Calculating the atmospheric radiative transfer processes is one of the most computationally expensive parts of the NWP model. One alternative is to model the radiative transfer using deep learning (DL) models, but the deep learning models do not involve physical laws and may have physically inconsistent outputs. This paper proposes a DL model framework to ensure the thermal equilibrium between fluxes and heating rates, which are outputs of radiative transfer models. Also, the accuracy of DL-based radiative transfer prediction is improved when using the framework. Various deep learning models have been trained and compared. The results demonstrate that model structures with global receptive fields work best for emulating radiative transfer calculations.

keyworks

parameterization, atmposhereic radiative transfer, deep learning, physics-incorporated

1 Introduction

Solar (shortwave, SW) and thermal radiation (longwave, LW) are the fundamental drivers of the atmospheric and oceanic circulation by creating the equator-versus-pole energy imbalance. The atmospheric radiative transfer processes are well understood and accurately represented by the line-by-line model LBLRTM (S. Clough et al., 2005; S. A. Clough et al., 1992). However, the LBLRTM requires unaffordable computational costs; thus, it is inappropriate for weather and climate modeling. Therefore, various parameterization methods are proposed to efficiently approximate radiative transfer calculations for application in numerical models (Stephens, 1984).

Despite being simplified, the radiative transfer parameterization is still more computationally expensive than other dynamical or physical processes. Therefore, the radiative transfer parameterization is usually performed less frequently in time and on a coarser spatial grid. For example, in the European Centre for Medium-Range Weather

Forecasts (ECMWF), the radiation scheme is run 8 times less frequently in time and 10.24 times coarser in spatial resolution than the high-resolution deterministic forecast (HRES), which would degrade the precision compared to frequent calls in time and space (Hogan & Bozzo, 2018). While for the ECMWF ensemble forecast with 12 minutes time step, the radiation scheme is only called every 3 hours on a spatial grid 6.25 times coarser than the rest of the model.

To further accelerate the radiation calculations in weather and climate models and make it feasible for more frequent calls of the radiation schemes, many researchers have investigated alternative approaches such as neural networks (NNs). Chevallier et al. (1998) and Chevallier et al. (2000) used shallow NNs with one hidden layer (NeuroFlux) to simulate the LW radiative budget from the top of the atmosphere to the surface in a model with 31 vertical levels. The NeuroFlux achieved comparable accuracy to the accuracy of the ECMWF operational scheme and was also 22 times faster. However, NeuroFlux failed to maintain both accuracy and acceleration when applied to models with 60 vertical layers and above (Morcrette et al., 2008). Pal et al. (2019) developed two dense, fully connected, feed-forward deep NN (DNN) to emulate SW and LW radiative calculations. They replaced the original radiation parameterization in the Super-Parameterized Energy Exascale Earth System Model (SP-E3SM) with these DNN-based emulators and were able to run simulations stably for up to a year. The DNN-based models achieved approximately 90-95% accuracy and sped up by 8-10 times compared to the original parameterizations. Their results demonstrated the applicability of machine learning in modeling radiative transfer calculations in NWP models. Roh and Song (2020) found that the frequent use of the NN-based radiation model improves the forecast accuracy compared to the infrequent use of the original radiation scheme with similar calculation costs. Moreover, Belochitski and Krasnopolsky (2021) showed that the shallow NN-based radiation emulators developed ten years ago for the general circulation model (GCM) are robust despite the structural change in the host model. Moreover, this model can generate realistic and stable radiation results when applied to numerical simulations for up to 7 months. Liu et al. (2020) compared feed-forward NNs with convolutional NNs for radiative transfer computations. Their results showed that the feed-forward NNs demonstrated a better balance between accuracy and computational performance.

In addition, the DL-based parameterization should be not only accurate but also credible by integrating the physical laws into the DL framework (Reichstein et al., 2019). Regarding the physical constraints, there exists a physical relationship between fluxes and heating rates. The previous studies (Krasnopolsky et al., 2010; Lagerquist et al., 2021; Liu et al., 2020; Roh & Song, 2020) trained NN-based emulators to output profiles of heating rates and fluxes at the surface and top-of-atmosphere directly, which causes issues with energy conservation. Cachay et al. (2021) and Ukkonen (2022) chose to predict the radiative fluxes and compute heating rates from fluxes, which ensures physical consistency (Yuval et al., 2021). However, Ukkonen (2022) found that the heating rates are highly sensitive to the continuity in the fluxes profile, and minor errors in fluxes lead to relatively large errors in heating rates. Based on the above research, the satisfaction of physical constraints has become a critical issue in NN-based radiative transfer emulation.

In this paper, we use deep learning models to emulate radiative transfer calculations. We run the Model for Prediction Across Scales - Atmosphere (MPAS-A) (Skamarock et al., 2012) that covers the entire globe and all months to generate the dataset for training and validation. The rapid radiative transfer model for general circulation models (RRTMG) is selected for radiative transfer calculations as the RRTMG model is widely used by many global and regional models. We also propose a physically incorporated training scheme, where the energy conservation is encoded in the network as hard constraints. Based on this framework, we apply and compare different network structures and analyze the advantages and disadvantages of each network structure in detail. Section 2 describes the

dataset used for training and evaluation. The overall physics-incorporated solution and various network structures are described in Section 3. The results related to each type of model and detailed error analysis are demonstrated in Section 4. Finally, section 5 contains the conclusions and discussions.

2 Data

2.1 Data generation

The dataset was generated by running the MPAS-Atmosphere version 7.1 with initial conditions provided by the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). MPAS employs an unstructured centroidal Voronoi mesh, which allows for both quasi-uniform and variable horizontal resolution. In this study, we used a global quasi-uniform horizontal mesh of approximate 60 km grid spacing containing 163842 grid cells and 57 vertical levels with a model top at 30 km.

The experiments used physics packages consisting of the "mesoscale reference" suite in MPAS-A. These packages include the new Tiedtke for cumulus convection (Zhang & Wang, 2017), RRTMG for SW and LW radiation (Iacono et al., 2008), Xu-Randall for subgrid cloud fraction (Xu & Randall, 1996), WRF Single-Moment 6-Class (WSM6) for microphysics (Hong & Lim, 2006), and Yonsei University (YSU) for planetary boundary layer mixing (Hong et al., 2006). The simulations were run for three days per month, and the initialization days were randomly selected (i.e., 20200108, 20200213, 20200302, 20200420, 20200528, 20200615, 20200719, 20200811, 20200927, 20201012, 20201124, and 20201204). The first two days of each three consecutive days are used for training, and the last day is used for testing. The model generates radiation inputs and outputs every 1 hour.

2.2 Input and output data

Table 1 lists all the input and output variables, where the input contains 29 original variables and the output contains 6 variables. Among the input variables, 11 are surface variables, and others are three-dimensional variables (either layer or level). To preprocess the data for the DL models, we pad the surface and layers variables to match the dimensions of the levels variables. Then, the z-score normalization technique is applied to normalize all the input and output variables to ensure they have the same mean and variance. For three-dimensional variables, the mean and standard deviation (std) were determined from values of all the vertical levels or layers.

Table 1. Definition of all the input and output variables, and whether they are of surface, layers, or level type, and units. There are 57 full model levels, and 56 layers.

Type	Variable name	Definition	Location	Unit
Input	aldif	Surface albedo (near-infrared spectral regions) for diffuse radiation	Surface	1
	aldir	Surface albedo (near-infrared spectral regions) for direct radiation	Surface	1
	asdif	Surface albedo (UV/visible spectral regions) for diffuse radiation	Surface	1
	asdir	Surface albedo (UV/visible spectral regions) for direct radiation	Surface	1
	cosz	Cosine solar zenith angle for current time step	Surface	1
	landfrac	Land mask (1 for land, 0 for water)	Surface	1
	sicefrac	Ice fraction	Surface	1
	snow	Snow water equivalent	Surface	kg/m2
	solc	Solar constant	Surface	W/m2
	tsfc	Surface temperature	Surface	K
	emiss	Surface emissivity for 16 LW spectral bands	Surface	1
	ccl4vmr	CCL4 volume mixing ratio	layer	mol/mol
	cfc11vmr	CFC11 volume mixing ratio	layer	mol/mol
	cfc12vmr	CFC12 volume mixing ratio	layer	mol/mol
	cfc22vmr	CFC22 volume mixing ratio	layer	mol/mol
	ch4vmr	Methane volume mixing ratio	layer	mol/mol
	cldfrac	Cloud fraction	layer	1
	co2vmr	CO2 volume mixing ratio	layer	mol/mol
	n2ovmr	N2O volume mixing ratio	layer	mol/mol
	o2vmr	O2 volume mixing ratio	layer	mol/mol
	o3vmr	O3 volume mixing ratio	layer	mol/mol
	play	Layer pressure	layer	hPa
	tlay	Layer temperature	layer	K
	qc	Cloud water mixing ratio	layer	kg/kg
	qg	Graupel mixing ratio	layer	kg/kg
	qi	Cloud ice mixing ratio	layer	kg/kg
	qr	Rain water mixing ratio	layer	kg/kg
	qs	Snow mixing ratio	layer	kg/kg
	qv	Water vapor mixing ratio	layer	kg/kg
Output	swuflx	Layer SW upward fluxes	level	W/m2
	swdflx	Layer SW downward fluxes	level	W/m2
	lwuflx	Layer LW upward fluxes	level	W/m2
	lwdflx	Layer LW downward fluxes	level	W/m2
	swhr	SW hearting rate	layer	K/day
	lwhr	LW heating rate	layer	K/day

3 Method

This section describes the physics-incorporated model framework, different DL model structures, and the evaluation methods.

3.1 Physics-Incorporated Framework

The primary functionality of the physics-based radiation parameterization schemes within the NWP model is to provide the heating rates due to both SW and LW radiation. Given the input variables listed in Table 1, the radiation parameterization first calculates the optical properties of the atmosphere due to various gases and the presence of clouds. Then, the radiation schemes use the optical properties and boundary conditions such as incoming solar flux, solar zenith angle, and surface albedo to calculate and output the vertical profiles of fluxes and heating rates. The flux measures the energy being radiated per unit area, with the unit of watts per meter squared (W/m^2). The heating rate describes the temperature change per unit of time, and it has the units of Kelvin per day (K/d). These two variables are not independent, and there is such a physical relationship:

$$HR_l = \frac{g}{c_p} \frac{(F_{l+1}^{up} - F_{l+1}^{down}) - (F_l^{up} - F_l^{down})}{{p_{l+1}}^{lev} - p_l^{lev}} \tag{1}$$

where g is the gravitational constant, c_p is the specific heat at constant pressure, F_l^{up}, F_l^{down}, and p_l^{lev} are the upward flux, downward flux, and pressure of level $l \in [1, \cdots, nlev]$. The output variables are involved in the subsequent calculations of the NWP models. It is critical to ensure that the relationship described by Equation (1). In addition, the change in atmospheric variables of one layer/level has both local and global impacts on radiation along the entire vertical column. For example, the presence of clouds or liquid water at any layer affects the distribution of fluxes across all the vertical levels by producing local heating rates peaks. Therefore, the related variables can be integrated vertically to allow for the nonlocal effects. Based on the above considerations, the framework is designed as shown in Figure 1, which includes three layers: the differential/integration layer, the radiative transfer layer, and the physics-incorporated layer.

The differential/integral layer is a data preprocessing module to preprocess input variables so that some prior knowledge can be fully utilized. As the cloud fraction (cldfrac in Table 1) and liquid water (qc) can affect fluxes far away from where they are present, these variables are integrated upward and downward along the vertical direction. The vertically accumulated cloud fraction and liquid water allow the models to learn vertically nonlocal effects. Meanwhile, calculating the heating rates requires the pressure difference between the two adjacent layers. Given the same values of fluxes, the smaller values of pressure difference result in larger values in heating rates. Therefore, the air pressure difference is obtained in advance by the differential module. The preprocessed features produced by the differential/integral layer are concatenated with the original features before being input into the models.

The radiative transfer layer contains the DL model to be trained to learn the mapping similar to the physics-based radiative transfer model. The learnable parameters only exist in this layer, as shown in the orange block in Figure 1. Although the model output is fluxes only, a custom loss function is designed as a weighted sum of the flux loss $\mathcal{L}_{flux}$ and heating rate loss $\mathcal{L}_{hr}$, as shown in Equation (2). λ is the weight of heating rate loss. The flux loss is defined as an average of the four groups of dimensionless values calculated as the mean square deviations divided by variance, as shown in Equation (3). Similarly, the heating rate loss averages two groups of dimensionless values, as shown in Equation (4). In the forward propagation stage, the fluxes are first output by this layer, and then heating rates are derived by the physics-incorporated layer (third layer). Fi-

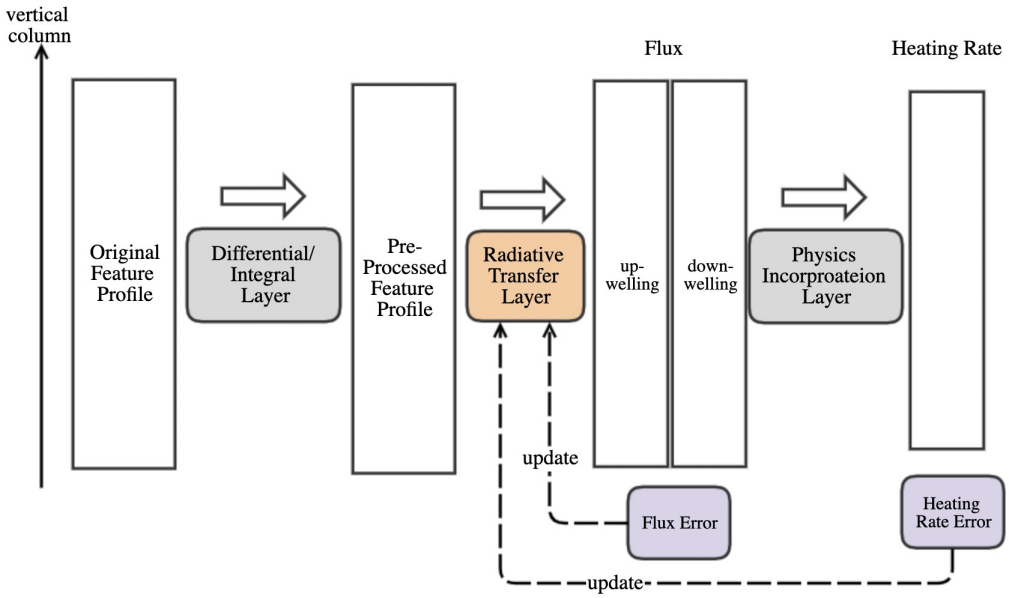

Figure 1. Physics-incorporated framework for emulating atmospheric radiative transfer

nally, the flux and heating rate loss are combined, and then the parameters of this layer will be updated accordingly. Many DL model structures can be implemented in this layer, and the details of some selected models are described in the following subsection.

The last layer is the physics-incorporated layer, which computes heating rates from fluxes based on Equation (1). The equation is treated as an independent layer and encoded into the framework to ensure physical consistency and conservation of energy. The gradient of heating rate loss can be derived using the gradient of flux loss and Equation (1), so there are no learnable parameters within this layer.

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{flux} + \lambda\mathcal{L}_{hr} \tag{2}$$

$$\mathcal{L}_{flux} = \frac{1}{4}[\frac{MSE_{F_{sw-up}}}{\sigma^2_{F_{sw-up}}} + \frac{MSE_{F_{sw-dn}}}{\sigma^2_{F_{sw-dn}}} + \frac{MSE_{F_{lw-up}}}{\sigma^2_{F_{lw-up}}} + \frac{MSE_{F_{lw-dn}}}{\sigma^2_{F_{lw-dn}}}] \tag{3}$$

$$\mathcal{L}_{hr} = \frac{1}{2}[\frac{MSE_{HR_{sw}}}{\sigma^2_{HR_{sw}}} + \frac{MSE_{HR_{lw}}}{\sigma^2_{HR_{lw}}}] \tag{4}$$

3.2 DL Models within the Radiative Transfer Layer

In this section, the detailed DL model structures in the radiative transfer layer are described. Various DL model structures are compared, including fully connected (FC) NNs, convolutional-based NNs (CNNs), recurrent-based NNs (RNNs), transformer-based NNs, and neural operator networks, respectively. For each group of model structures, the total number of parameters is controlled to be around 1 million. This way, the influence of the number of parameters can be ruled out, and the influence of the network structures on the radiative transfer modeling can be examined more clearly. The schematic diagram of the ResNet, Bi-LSTM, FNO, and Transformer model is shown in Figure 2.

- FC NNs: FC NNs are the most classical network structures in the study of DL-based emulators of radiative transfer parameterization. In FC networks, all the input variables are flattened and passed through a series of fully connected layers to obtain the outputs. In this work, the number of hidden layers used is 10, and each layer contains 200 hidden units. After each fully connected layer, batch normalization is performed, and the activation function of ReLU is used. The total number of parameters for this FC model is 0.84 million.

- CNNs: CNNs were firstly designed for image processing and have also become popular tools in the atmospheric science (Bolton & Zanna, 2019; Wimmers et al., 2019; Liu et al., 2020; Lagerquist et al., 2021). The CNNs use convolution kernels to process a small input region at a time, so they are good at extracting local features. However, the convolution kernel also limits CNNs' ability of global perception due to its fixed sizes. Furthermore, although the receptive field increases with the more convolutional layers, adding more layers substantially increases the computational costs of CNN models. There are different types of CNN models, and the two classical CNN models have been implemented in this work: ResNet and U-Net. The ResNet was first proposed by He et al. (2016), and it is built on the concept of shortcut connections between layers to minimize the problem of vanishing gradients. In this work, the input feature dimension is first increased to 128 through a 1D convolution with a kernel size of 7. Then ten residual blocks are applied, each containing three layers of Convolution-BatchNormalization-ReLU operations. Within each residual block, the kernel size is three, and the number of output channels is 128. The total number of parameters is 0.77 million. The U-Net model was first proposed by Ronneberger et al. (2015). It first consists of several convolutional layers and downsampling processes while the number of channels increases. The downsampling module has a layer of Conv-BN-ReLU operation with stride 2. The U-Net structure used here contains four downsampling modules, and the numbers of output channels are 24, 48, 96, and 192, respectively. Next, multiple upsampling steps are performed to recover the original resolution while the number of channels reduces. The upsampling module goes through a single-layer deconvolution module with stride one and then through two layers of Conv-BN-ReLU operations. In addition, the outputs of each downsampling module (except for the last one) are used for the corresponding upsampling module through skip connections. Lastly, the network adopts a 1×1 convolutional layer to map the channel dimension to the output dimension. The total number of parameters of the U-Net model is 1.52 million.

- Recurrent Type: Recurrent NNs (RNNs) are widely used for sequential data such as text data in natural language processing (NLP) tasks and time series. Here, the sequence is represented by the vertical profiles simulated by NWP models, and the input vectors are the variables describing atmospheric conditions at a vertical level. However, standard RNNs are insufficient for modeling the radiative transfer processes. On the one hand, the RNNs are ineffective for modeling long sequences, while the number of vertical levels has been increasing to improve the model forecast. On the other hand, radiative fluxes transfer in upward and downward directions, so the fluxes at a certain level were affected by the atmospheric conditions above and below. Therefore, the unidirectional RNNs are not appropriate. The Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014) are designed to learn long-term dependencies. They use gates to learn which data in a sequence is important to keep or remove. In addition, the bidirectional LSTM and GRU are implemented to extract information from both directions of the sequence (i.e., vertical profiles of the atmosphere). The Bi-LSTM model applied in this paper contains five layers, each with 96 hidden layer units, and the number of model parameters is 1.12 million. For the Bi-GRU, a 5-layer structure is used, with each layer having 128 hidden layer units, and the number of network parameters is 0.77 million.

- Transformer Type: Transformers (Vaswani et al., 2017), a NN architecture built on the self-attention mechanism, were initially designed for NLP tasks and have become a general architecture for almost every ML task. Unlike CNN models that start by learning local features and slowly get a larger receptive field, transformers have a global perspective at each layer due to self-attention. Because of the nature of radiative transfer, changes in atmospheric conditions at any layers affect the entire profile of radiative fluxes. For example, when clouds occur, the fluxes at all levels are changed accordingly. Therefore, transformers are appropriate for emulating radiative transfer as they can extract information from the whole atmospheric column. The transformer model used in this work contains seven self-attention blocks, each having one self-attention layer and two fully connected layers. All the query, key, and value vectors in the model have a dimension of 128. Finally, the dimension of embedding is changed to be the same as the output dimension through a 1×1 convolutional layer. The total number of trainable parameters in this transformer network is 0.71 million.

- Neural Operator Type: The traditional radiative transfer parameterization approximates the full equations of radiative transfer by discretizing the atmosphere in the vertical direction. However, vertical discretization also results in a trade-off between speed and accuracy: low resolution is fast but less accurate, while high resolution is accurate but slower. Unlike traditional grid-dependent methods, the Fourier Neural Operators (FNO) can parameterize the radiative transfer modeling in function space instead of the discretized space. The output of the FNO is the complete wave field solution, similar to the wavelike pattern of fluxes. The FNO (Li et al., 2020) model we implement in this study includes four Fourier modules, each performing convolutions in the frequency domain through the Fourier transform and reverting to the spatial domain through the inverse Fourier transform. The FNO allows a single-layer operator to capture global information of the entire atmospheric column. The total number of trainable parameters in the FNO model is 1.22 million.

All settings of the hyperparameters used for different NNs are the same. Each model is trained with 500 epochs using a batch size of 4096. Adam optimizer is used with the initial learning rate 1e-3. Also, the plateau scheduler is applied to decrease the learning rate by a factor of 0.5 when the loss does not decrease for five consecutive epochs.

3.3 Evaluation methods

All the DL-based radiation emulators are evaluated by comparing against the outputs of the original RRTMG schemes, including upward and downward SW and LW fluxes and heating rates. The overall model performance metrics include root mean squared error (RMSE), and mean bias error (MBE). For each vertical level, the mean absolute error (MAE), MBE, and standard deviation of biases per level or layer were calculated using the following equations:

$$MAE_l = \frac{1}{N}\sum_{i=1}^{N}|Y_{DL}(i,l) - Y(i,l)| \tag{5}$$

$$MBE_l = \frac{1}{N}\sum_{i=1}^{N}Y_{DL}(i,l) - Y(i,l) \tag{6}$$

$$STD_l = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((Y_{DL}(i,l) - Y(i,l)) - MBE_l)^2} \tag{7}$$

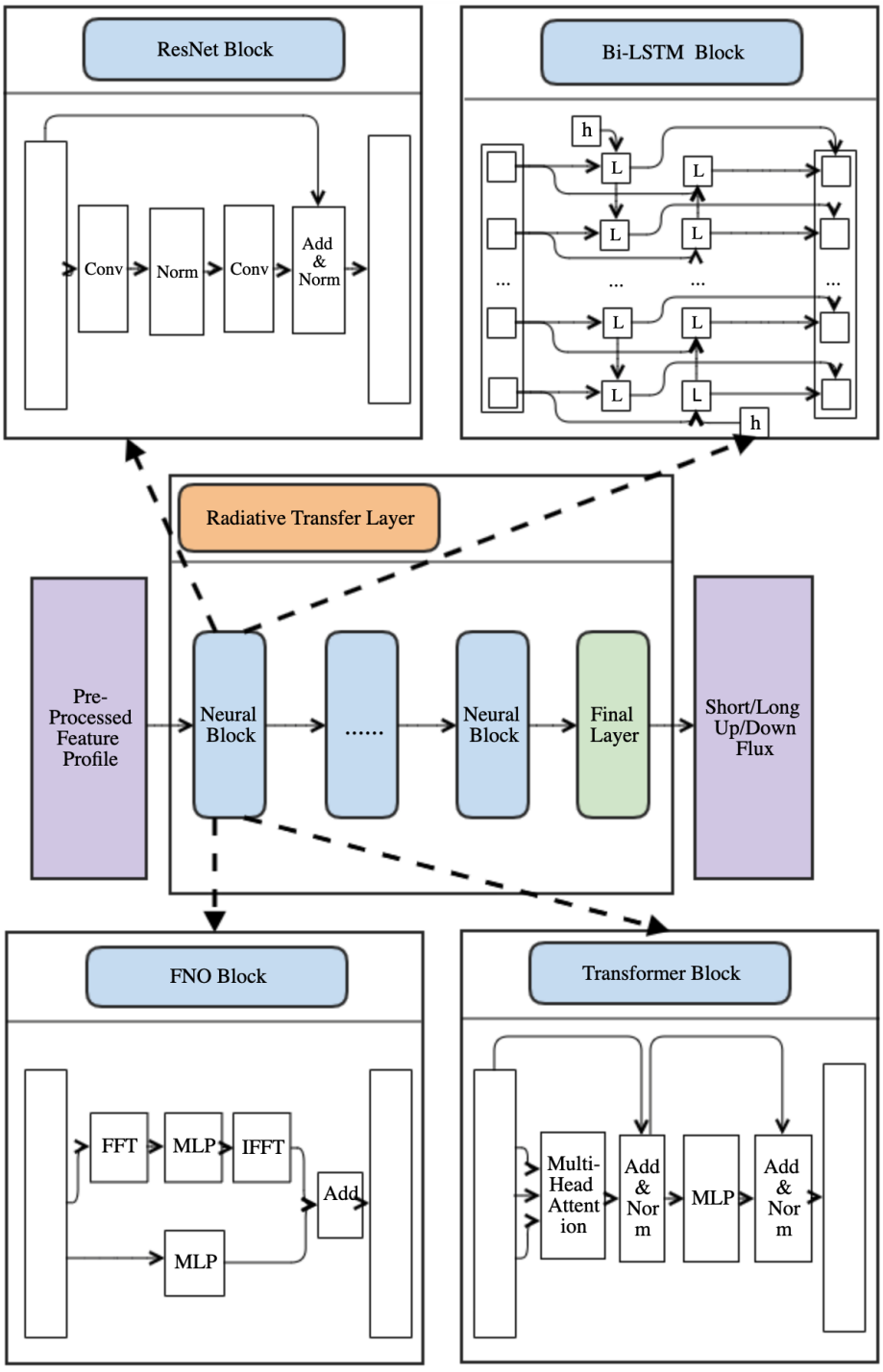

Figure 2. Schematic diagram of the structures of DL models used in the radiative transfer layer, including ResNet, Bi-LSTM, FNO, and Transformer.

where $Y(i,l)$ and $Y_{DL}(i,l)$ are output from the RRTMG radiation schemes and DL-based radiation emulators, respectively, i is the horizontal grid point of a vertical profile, N is the number of the horizontal grid points, l is the vertical level or layer index.

4 Results

4.1 Statistical results

The offline evaluation was done using 12 days of data that was not used for training. Table 2 summarizes the error statistics of different DL-based emulators for fluxes and heating rates averaged over all the testing data. The FC, ResNet, and U-Net models predict far less accurate fluxes and heating rates, with RMSE of SW and LW fluxes higher than 10.9 and 2.4 W/m^2 and RMSE of SW and LW heating rates higher than 0.09 and 0.21 K/day, respectively. The RMSE of LW fluxes is always smaller than that of SW fluxes, as SW fluxes have a greater magnitude than LW fluxes and are more difficult to predict. However, the RMSE of LW heating rates is always higher than the SW heating rates of each corresponding DL-based emulator, as LW heating rates are more sensitive to clouds and more difficult to predict (see Figure 3). Overall, FC and CNN networks perform worse than the RNN, transformer, and FNO models in radiative transfer emulations, which the structural characteristics of these models can explain. For FC networks, the flattening operation erases the vertical distribution of all the features, leading to the loss of important information. Also, FC and CNN networks only have the local receptive fields in the vertical direction for each operation performed.

The Bi-GRU, Bi-LSTM, transformer, and FNO significantly improve forecast accuracy, with RMSE of SW and LW fluxes smaller than 3.8 and 1.3 W/m^2, respectively. In addition, the RMSE of SW and LW heating rates is reduced to less than 0.042 and 0.15 K/day. The change in atmospheric variables of one layer/level has both local and global impacts on radiation along the entire vertical column. For example, the presence of clouds or liquid water at any layer significantly reflects or absorbs radiation which affects the subsequent fluxes across the whole atmospheric column and produces local heating rate peaks. Therefore, having global perception ability is critical for DL-based radiative transfer emulation. The significant improvement in the accuracy of fluxes and heating rates for Bi-GRU, Bi-LSTM, transformer, and FNO models is due to their ability to obtain a global perspective of an entire atmospheric column in single-layer operations. However, the FNO model performs worse than Bi-GRU, Bi-LSTM, and transformer models because it assumes that the input variables have a uniform distribution while the atmospheric layers are not distributed uniformly. The Bi-GRU and Bi-LSTM model have the best performance and outperform the transformer model, with RMSE in SW and LW fluxes around 2.3 and 1.2 W/m^2, and RMSE in SW and LW heating rates about 3.20×10^{-2} and 1.39×10^{-1} K/day, respectively. The Bi-GRU and Bi-LSTM are most accurate because they mimic atmospheric radiative transfer's bidirectional behavior in the vertical direction.

In addition, the biases of the net fluxes at the top-of-atmosphere (TOA) directly determine the energy budget of the global atmosphere. Therefore, if the MBE of net fluxes at the TOA tends to be 0, it represents a more consistent energy budget with the physics-based radiation schemes. Table 2 shows that the Bi-LSTM model has the highest accuracy in terms of net fluxes at TOA, with a value of 4.91×10^{-2} W/m^2, which is one order of magnitude smaller than other models.

For a clearer view of the vertical distribution of errors, Figure 3 presents the vertical profiles of statistics in fluxes and heating rates for FC, U-Net, Bi-LSTM, and transformer. The solid and dotted lines in the figure represent the MAE and MBE of fluxes or heating rates at each vertical level/layer, and the shaded area shows the mean std of biases. The FC and U-Net models have much higher variance, as shown by the vertical profiles of mean std of biases. Also, FC and U-Net models have much higher MAE than the Bi-LSTM and

Table 2. RMSE of SW flux, LW flux, SW heating rate, LW heating rate, and MBE of TOA net flux for DL-based emulators including FC, ResNet, U-Net, Bi-GRU, Bi-LSTM, Transformer, and FNO averaged over all the testing data.

Model	SW Flux $W \cdot m^{-2}$ RMSE	LW Flux $W \cdot m^{-2}$ RMSE	TOA Net Flux $W \cdot m^{-2}$ MBE	SW Heating Rate $K \cdot d^{-1}$ RMSE	LW Heating Rate $K \cdot d^{-1}$ RMSE
FC	14.63	5.28	-3.78	18.85e-2	3.94e-1
ResNet	38.97	8.72	-2.32e-1	22.89e-2	4.14e-1
Unet	10.92	2.46	-7.62	9.58e-2	2.17e-1
Bi-GRU	2.334	1.216	3.97e-1	3.29e-2	1.41e-1
Bi-LSTM	**2.315**	**1.205**	**4.91e-2**	**3.20e-2**	**1.39e-1**
Transformer	2.753	1.286	-5.61	4.06e-2	1.46e-1
FNO	3.755	1.289	-6.77	4.20e-2	1.47e-1

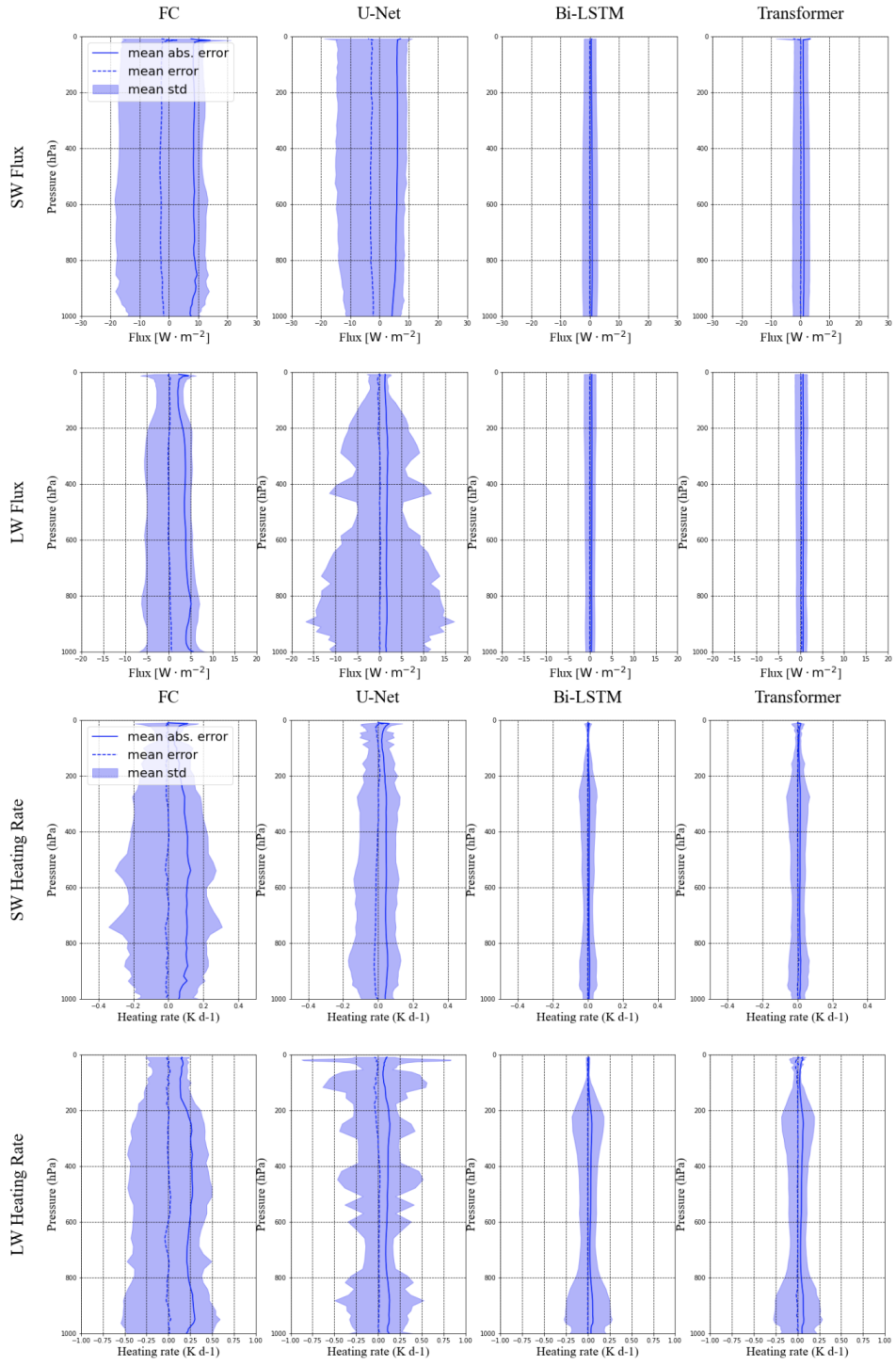

Figure 3. Vertical profiles of the statistics in SW fluxes (first row), LW fluxes (second row), SW heating rates (third row), and LW heating rates (fourth row) for the test data using different NN-based emulators: FC (first column), U-Net (second column), Bi-LSTM (third column), and Transformer (fourth column). The solid and dotted lines show the MAE and MBE profile, respectively, and the shaded area indicates the mean std relative to the bias.

Table 3. Performance of Bi-LSTM on radiative transfer problems with different error types

Loss Type	SW Flux $W \cdot m^{-2}$	LW FLux $W \cdot m^{-2}$	SW Heating Rate $K \cdot d^{-1}$	LW Heating Rate $K \cdot d^{-1}$
only fluxes	2.404	1.222	1.958e-1	1.810e-1
only heating rates	\	\	1.166e-1	1.419e-1
with physics-incorporated layer	2.315	1.205	0.320e-1	1.390e-1

transformer models at all levels. The error distributions of Bi-LSTM and transformer are very similar, and the Bi-LSTM has slightly smaller values in error and std. Both models show a uniform vertical error distribution and std in fluxes. For heating rates, they have relatively higher values in std of biases among the pressure layers between 800-1000 hPa and 200-400 hPa. Those two vertical regions are where liquid and ice clouds occur most frequently and are thus more difficult to predict.

4.2 Benefits of introducing the physics-incorporated layer

In this subsection, we discuss the benefits of introducing the physics-incorporated layer. The physics-incorporated layer ensures the satisfaction of the thermal equilibrium between fluxes and heating rates, as shown in Equation (1), by encoding it as part of network layers. We designed three groups of experiments: only supervising fluxes, only supervising heating rates, and a joint loss with the physics-incorporated layer imposed. The corresponding weights (λ in Equation (2)) are set to 0, 1, and 0.091. The RMSE of these experiments are summarized in Table 3.

When only supervising the fluxes, the heating rates are derived using Equation (1). As the vertical profiles of fluxes are smooth, the model is relatively easy to fit fluxes well. As a result, the RMSE of only supervising fluxes is slightly worse than that using the physics-incorporated layer. However, the RMSE of SW and LW heating rates are 6 times and 1.5 times greater than using the physics-incorporated layer. On the other hand, when the model is trained only to supervise the heating rates, fluxes cannot be derived accordingly. In this case, the model predicted heating rates are still less accurate than the model trained with the physics-incorporated layer, and the RMSE of SW and LW heating rates are 1.5 and 1.25 times greater. The physics-incorporated layer demonstrates its superiority by ensuring a physically consistent relationship between fluxes and heating rates and showing a more accurate prediction of heating rates and fluxes. Overall, the Bi-LSTM model trained using the physics-incorporate layer achieves the most accurate forecast.

4.3 Performance under different cloud conditions

As clouds play an important role in weather and climate prediction, this section analyzes the performance of the DL-based radiation emulators under three typical cloud conditions: profiles with no liquid cloud, single-layer liquid cloud, and multi-layer liquid cloud. The liquid clouds strongly absorb and scatter radiation, and they cause discontinuity in radiative fluxes and heating rates. Therefore, it is more difficult for radiation emulators to perform well under cloudy conditions than in liquid cloud-free conditions. This work defines the liquid cloud layer as a contiguous set of vertical layers with cloud water mixing ratios (qc in Table 1) larger than 0. In all the testing data, no liquid cloud, single-layer, and multi-layer liquid cloud account for 61.3%, 29.6%, and 9.1% (Table 4). Here, ice cloud layers forming at high altitudes are not considered as their impact on fluxes and heating rates are much weaker than liquid cloud layers. Table 4 presents the RMSE of different DL-based emulators to predict heating rates under three liquid cloud conditions. The Bi-LSTM and Bi-GRU predicted heating rates are the most accurate under all three cloud conditions. The

Table 4. RMSE for heating rates under *no liquid cloud*, *single-layer liquid cloud* and *multi-layer liquid cloud* conditions.

Model	no liquid cloud (61.3%) $K \cdot d^{-1}$		single-layer liquid cloud (29.6%) $K \cdot d^{-1}$		multilayer liquid cloud (9.1%) $K \cdot d^{-1}$	
	SW HR	LW HR	SW HR	LW HR	SW HR	LW HR
FC	0.1005	0.2938	0.1612	0.3826	0.2127	0.4835
ResNet	0.1263	0.2284	0.2023	0.4456	0.2165	0.4289
Unet	0.0510	0.1159	0.0837	0.2060	0.1013	0.2674
Bi-GRU	0.0157	0.0554	0.0303	**0.1370**	0.0359	**0.1566**
Bi-LSTM	**0.0152**	**0.0546**	**0.0297**	0.1379	**0.0346**	0.1567
Transformer	0.0201	0.0680	0.0367	0.1437	0.0440	0.1644
FNO	0.0211	0.0683	0.0378	0.1453	0.0463	0.1684

RMSE under the condition of multi-layer liquid clouds is higher than that of single-layer and cloud-free conditions for all models.

To better understand why the error statistics vary significantly under different cloud conditions, we randomly select three vertical profiles for demonstration, as shown in Figure 4. Figures 4 also illustrates the vertical profiles of SW and LW fluxes and heating rates predicted by the original RRTMG scheme, FC, U-Net, Bi-LSTM, and transformer models for the three selected cases, respectively. Under cloud-free conditions, fluxes and heating rates change smoothly from the TOA to the surface. While under single-layer and multi-layer liquid cloud conditions, a large gradient of fluxes and heating rates are shown where liquid clouds are presented. As a result, Figures 4 show that these DL models can accurately predict the vertical profiles of heating rates, while their prediction for the vertical profiles of SW fluxes is worse than that of LW fluxes. Among the DL-based radiation emulators, the Bi-LSTM and transformer models are superior in capturing the discontinuities in vertical caused by liquid water, consistent with Table 2.

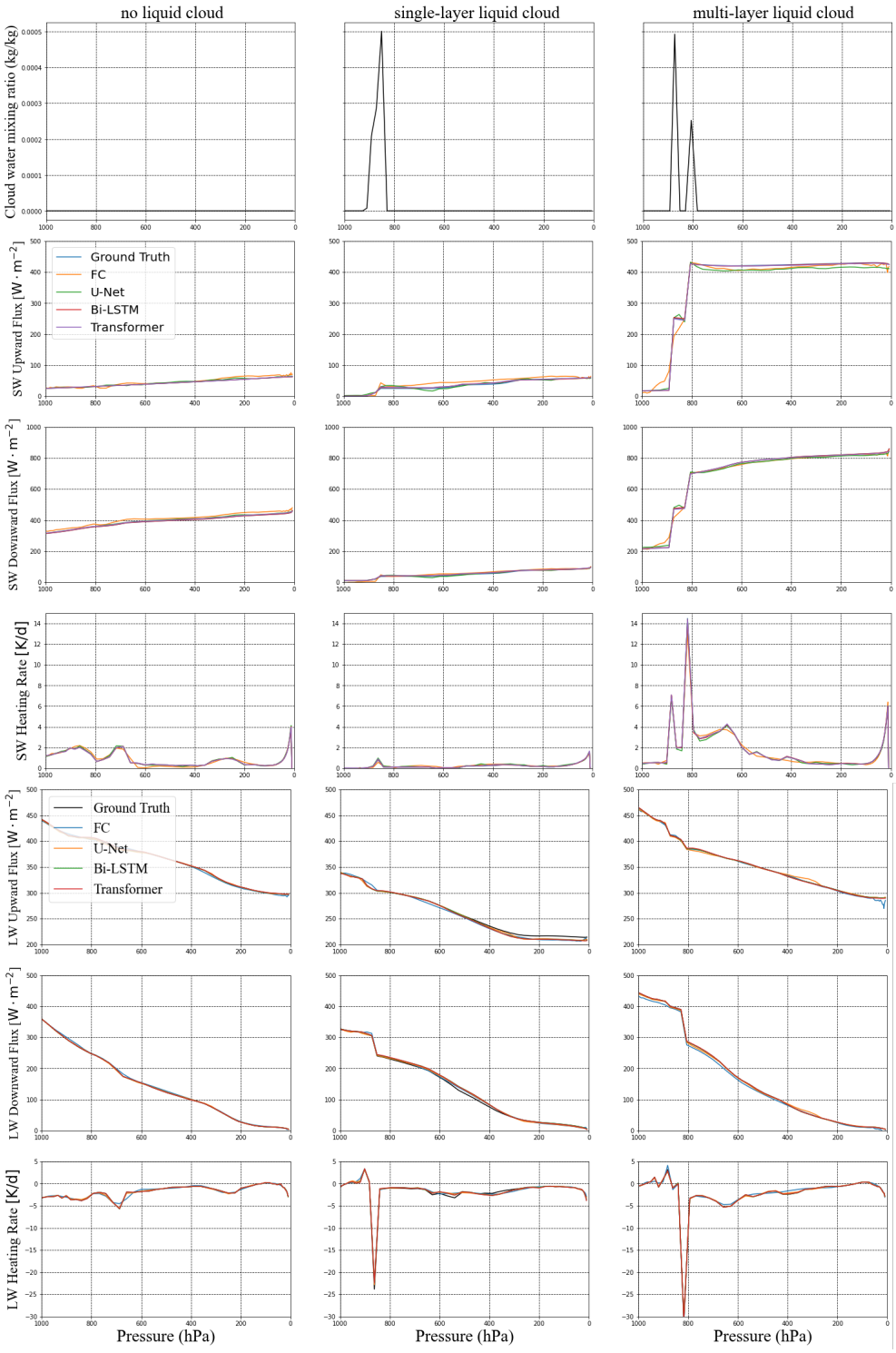

Figure 4. Vertical profiles of the liquid water distribution of the 3 typical case (top row):no liquid cloud (left column), single-layer liquid cloud (middle column), and multi-layer liquid cloud (right column). Vertical profiles of SW upward fluxes (top row), SW downward fluxes (middle row), SW heating rates (bottom row), LW upward fluxes (top row), LW downward fluxes (middle row), and LW heating rates predicted by the original RRTMG scheme, FC, U-Net, Bi-LSTM, and transformer models for the three selected cases.

5 Conclusions

In this paper, we propose a physics-incorporated framework for emulating atmospheric radiative transfer processes. The physical relationship between fluxes and heating rates is considered in our framework and encoded as a layer of the network. Based on this framework, we designed and compared various DL model structures, such as FC NNs, CNNs, bidirectional RNNs (Bi-LSTM and Bi-GRU), transformer-based NNs, and FNO. We found that models with the ability of global perception perform better than FC and CNNs and are thus more suitable for radiative transfer emulation. Among the models with a global perspective of an entire atmospheric column, the Bi-LSTM and Bi-GRU have the best accuracies, outperforming the transformer and FNO, as they benefit from extracting information from two directions. It is also demonstrated that the physics-incorporated layer makes the prediction of the Bi-LSTM model more accurate. Furthermore, evaluations are performed under different liquid cloud conditions due to the importance of clouds to weather and climate prediction. The results suggest the Bi-LSTM performs well at all vertical levels, although there are slightly larger errors and variances where clouds are present.

Future work will investigate the online implementation of the DL-based emulators in an NWP model such as the MPAS model with different vertical levels. Besides, due to the nonlinearity of the radiative transfer models, there is no corresponding tangent-linear and adjoint model of radiative transfer scheme for the MPAS model. Hatfield et al. (2021) demonstrated the feasibility of constructing the tangent-linear and adjoint models from the NN-based gravity wave drag scheme. They showed that the NN-derived tangent-linear and adjoint models successfully passed the standard test and were applied in four-dimensional variational data assimilation. Likewise, our future work includes developing the adjoint model of radiation schemes using NN-based radiation emulators to improve the four-dimensional variational data assimilation system.

Author contributions: Y.Y. trained the deep learning models and calculate the statistics of model performance. Y.Z. conducted the MPAS-A model simulations to provide dataset for training and evaluation, and offered valuable suggestions on the model training and paper revision. X.Z. and Y.Y wrote, reviewed and edited the original draft; Z.W. supervised and supported this research, and gave important opinions. All of the authors have contributed to and agreed to the published version of the manuscript.

Competing interests: The authors declare no conflict of interest.

Acknowledgments

This work was supported in part by the Zhejiang Science and Technology Program under Grant 2021C01017.

Open Research

Data Availability Statement

The data used for training and testing all the deep learning models in this work are available at https://doi.org/10.5281/zenodo.7213941 (Yao et al., 2022).

Code Availability Statement

The source code used for training all the deep learning models in this work are available at https://doi.org/10.5281/zenodo.7213941 (Yao et al., 2022). The source code for the MPAS-A model used in this work is available from https://mpas-dev.github.io.

References

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.

Bolton, T., & Zanna, L. (2019). Applications of Deep Learning to Ocean Data Inference and Subgrid Parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376-399. doi: 10.1029/2018MS001472

Cachay, S. R., Ramesh, V., Cole, J. N., Barker, H., & Rolnick, D. (2021). Climart: A benchmark dataset for emulating atmospheric radiative transfer in weather and climate models. *arXiv preprint arXiv:2111.14671*.

Chevallier, F., Chéruy, F., Scott, N., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of applied meteorology*, *37*(11), 1385–1397.

Chevallier, F., Morcrette, J.-J., Chéruy, F., & Scott, N. (2000). Use of a neural-network-based long-wave radiative-transfer scheme in the ecmwf atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 761–776.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Clough, S., Shephard, M., Mlawer, E., Delamere, J., Iacono, M., Cady-Pereira, K., ... Brown, P. (2005). Atmospheric radiative transfer modeling: A summary of the aer codes. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *91*(2), 233–244.

Clough, S. A., Iacono, M. J., & Moncet, J.-L. (1992). Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor. *Journal of Geophysical Research: Atmospheres*, *97*(D14), 15761–15785.

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., & Palmer, T. (2021). Building tangent-linear and adjoint models for data assimilation with neural networks. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002521.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

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.

Hong, S.-Y., & Lim, J.-O. J. (2006). The wrf single-moment 6-class microphysics scheme (wsm6). *Asia-Pacific Journal of Atmospheric Sciences*, *42*(2), 129–151.

Hong, S.-Y., Noh, Y., & Dudhia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly weather review*, *134*(9), 2318–2341.

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).

Krasnopolsky, V., Fox-Rabinovitz, M., Hou, Y., Lord, S., & Belochitski, A. (2010). Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: Climate simulations and seasonal predictions. *Monthly Weather Review*, *138*(5), 1822–1842.

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.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Liu, Y., Caballero, R., & Monteiro, J. M. (2020). Radnet 1.0: Exploring deep learning architectures for longwave radiative transfer. *Geoscientific Model Development*, *13*(9),

4399–4412.

Morcrette, J.-J., Mozdzynski, G., & Leutbecher, M. (2008). A reduced radiation grid for the ecmwf integrated forecasting system. *Monthly weather review*, *136*(12), 4760–4772.

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.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195-204. Retrieved from https://ideas.repec.org/a/nat/nature/v566y2019i7743d10.1038_s41586-019-0912-1.html doi: 10.1038/s41586-019-0912-1

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

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Skamarock, W. C., Klemp, J. B., Duda, M. G., Fowler, L. D., Park, S.-H., & Ringler, T. D. (2012). A multiscale nonhydrostatic atmospheric model using centroidal voronoi tesselations and c-grid staggering. *Monthly Weather Review*, *140*(9), 3090 - 3105. Retrieved from https://journals.ametsoc.org/view/journals/mwre/140/9/mwr-d-11-00215.1.xml doi: 10.1175/MWR-D-11-00215.1

Stephens, G. L. (1984). The parameterization of radiation for numerical weather prediction and climate models. *Monthly weather review*, *112*(4), 826–867.

Ukkonen, P. (2022). Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002875.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wimmers, A. J., Velden, C. S., & Cossuth, J. (2019). Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, *147*(6), 2261-2282. doi: 10.1175/MWR-D-18-0391.1

Xu, K.-M., & Randall, D. A. (1996). A semiempirical cloudiness parameterization for use in climate models. *Journal of the atmospheric sciences*, *53*(21), 3084–3102.

Yao, Y., Zhong, X., Zheng, Y., & Wang, Z. (2022). A physics-incorporated deep learning framework for parameterization of atmospheric radiative transfer (Version 1.0) [Dataset] [Software]. Zenodo. https://doi.org/10.5281/zenodo.7213941.

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

Zhang, C., & Wang, Y. (2017). Projected future changes of tropical cyclone activity over the western north and south pacific in a 20-km-mesh regional climate model. *Journal of Climate*, *30*(15), 5923–5941.

Figure1.

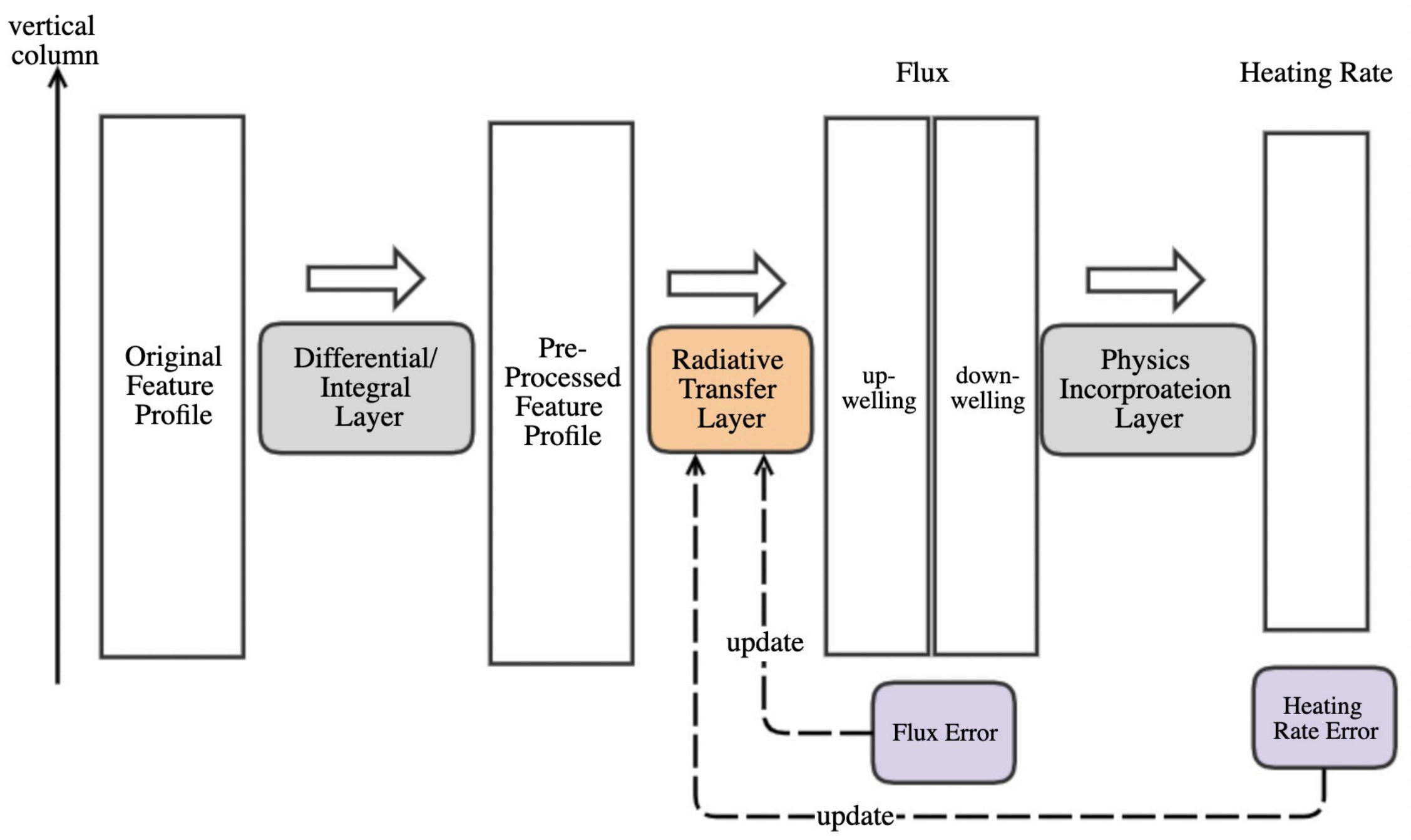
vertical column
Original Feature Profile
Differential/ Integral Layer
Pre-Processed Feature Profile
Radiative Transfer Layer
Flux
up-welling
down-welling
Physics Incorproateion Layer
Heating Rate
update
Flux Error
Heating Rate Error
update

Figure2.

ResNet Block

Conv

Norm

Conv

Add & Norm

Bi-LSTM Block

h

L

...

h

Radiative Transfer Layer

Pre-Processed Feature Profile

Neural Block

.....

Neural Block

Final Layer

Short/Long Up/Down Flux

FNO Block

FFT

MLP

IFFT

Add

MLP

Transformer Block

Multi-Head Attention

Add & Norm

MLP

Add & Norm

Figure3.

FC

U-Net

Bi-LSTM

Transformer

SW Flux

LW Flux

SW Heating Rate

LW Heating Rate

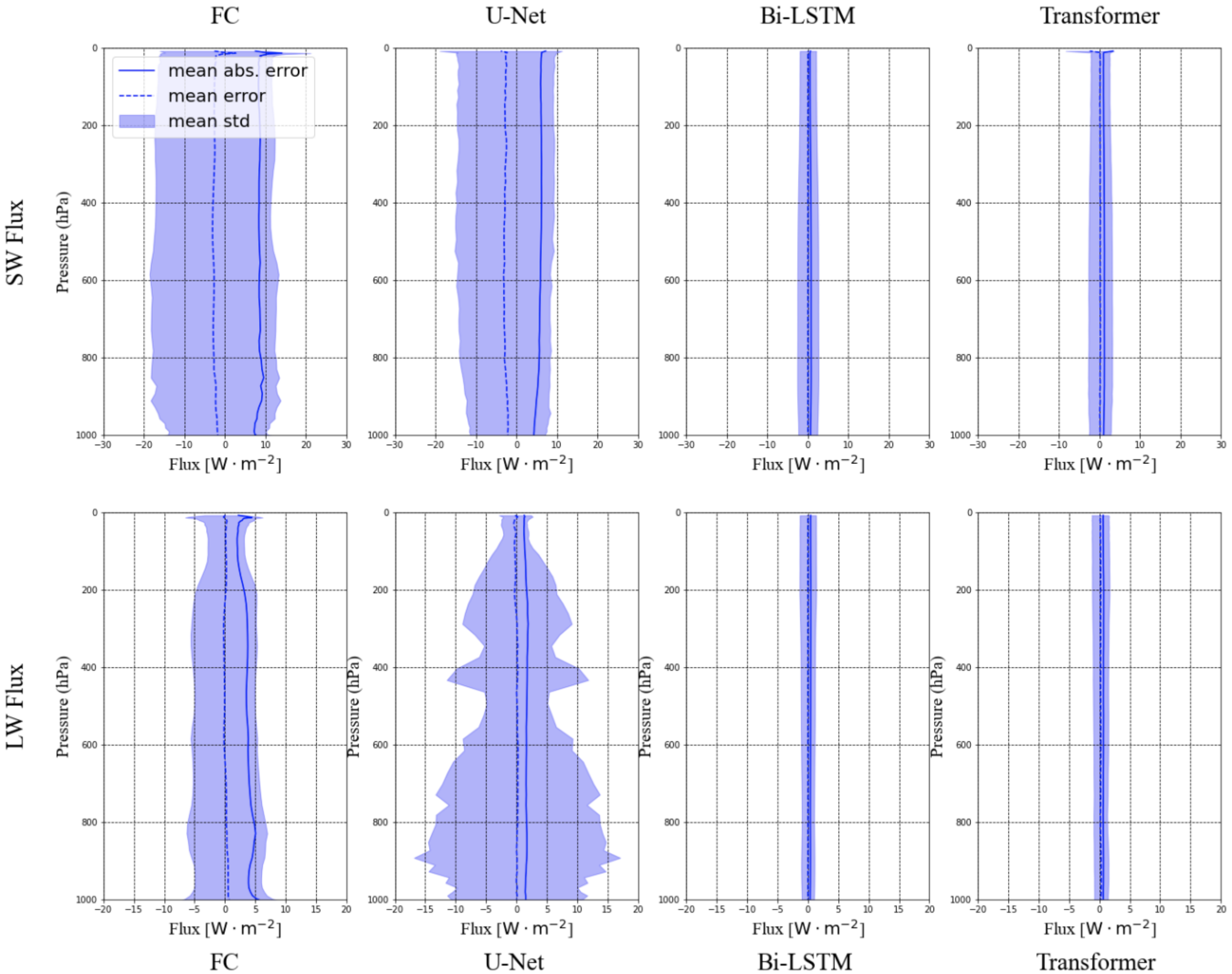

Figure4.

no liquid cloud

single-layer liquid cloud

multi-layer liquid cloud

Cloud water mixing ratio (kg/kg)

SW Upward Flux [W · m^{-2}]

SW Downward Flux [W · m^{-2}]

SW Heating Rate [K/d]

LW Upward Flux [W · m^{-2}]

LW Downward Flux [W · m^{-2}]

LW Heating Rate [K/d]

Ground Truth

FC

U-Net

Bi-LSTM

Transformer

Pressure (hPa)