# ~~WRF-DL~~ WRF-ML v1.0: A Bridge between WRF v4.3 and ~~Deep~~ ~~Machine~~ Learning Parameterizations and its Application to Atmospheric Radiative Transfer

Xiaohui Zhong[1], Zhijian Ma[1], Yichen Yao[1], Lifei Xu[1], Yuan Wu[1], and Zhibin Wang[1]

[1]Damo Academy, Alibaba Group, Hangzhou 311121, China

**Correspondence:** Zhibin Wang(zhibin.waz@alibaba-inc.com)

**Abstract.**

In numerical weather prediction (NWP) models, physical parameterization schemes are the most computationally expensive components, despite being greatly simplified. In the past few years, an increasing number of studies have demonstrated that ~~deep learning (DL~~machine learning (ML) parameterizations of subgrid physics have the potential to accelerate and even outperform conventional physic-based schemes. However, as the ~~DL~~ ML models are commonly implemented using the ~~DL~~ ML libraries written in Python, very few ~~DL-based~~ ML-based parameterizations have been successfully integrated with NWP models due to the difficulty of embedding Python functions into Fortran-based NWP models. To address this issue, we developed a coupler to allow the ~~DL-based~~ ML-based parameterizations to be coupled with a widely used NWP model, i.e., the Weather and Research Forecasting (WRF) model. Similar to the WRF I/O methodologies, the coupler provides the options to run the ~~DL~~ ML model inference with exclusive processors or the same processors for WRF calculations. In addition, to demonstrate the effectiveness of the coupler, the ~~DL-based~~ ML-based radiation emulators are trained and coupled with the WRF model successfully.

## 1 Introduction

Numerical weather prediction (NWP) models have become the most important tools for operational weather forecasting and have many applications in different domains, including energy, traffic, logistics, and planning (Coiffier, 2011; Pu and Kalnay, 2018). The physics-based parameterizations are essential in the NWP models as they approximate some processes that are either too small scale to be explicitly resolved at the model grid resolution or too complex and not fully understood. Although many simplifications are made to parameterizations to reduce the computational cost, the calculations of physical parameterizations still account for a significant portion of the total computational time of NWP models (Wang et al., 2019). Also, the parameterization schemes often contain uncertain parameters estimated from more faithful high-resolution simulations with statistical models (Stensrud, 2013). For example, the parameters in radiative transfer parameterizations can be fitted to the output of the most accurate line-by-line model (Clough et al., 2005), and parameters in cloud turbulence parameterizations can be inferred from large-eddy simulations (LES) (Mason, 1989). However, other parameters can only be learned from observations as the related governing equations are unknown (Schneider et al., 2017).

25    One alternative is to train ~~deep learning (DL~~machine learning (ML)) models to replace the traditional physics-based parameterization schemes. The ~~DL-based~~ ML-based parameterizations have the potential to outperform traditional parameterizations with higher computational efficiency. For example, the radiative transfer parameterization scheme is one of the most computationally expensive components in NWP models, and it has the longest history of developing ~~DL-based~~ ML-based radiation emulators. Chevallier et al. (1998, 2000) developed an NN-based longwave radiation parameterization (NeuroFlux) and has
30    been used operationally in the European Centre for Medium-Range Weather Forecasts (ECMWF) four-dimensional variational data assimilation system. The NeuroFlux is seven times faster than the previous scheme with comparable accuracy (Janisková et al., 2002). Recently, Song and Roh (2021) developed and used the neural network (NN) based radiation emulators in the operational weather forecasting model in the Korea Meteorological Administration. They demonstrated that using NN-based emulators frequently can improve real-time weather forecasting in terms of accuracy and speed compared to the original
35    method, which infrequently uses the original radiation parameterization.

Similarly, ~~DL-based~~ ML-based emulators have been developed for other parameterization schemes in NWP models. Rasp et al. (2018) successfully developed and coupled an NN-based convection parameterization into an aquaplanet general circulation model (GCM). They showed that the NN-based parameterization was able to perform multi-year simulations, of which results were close to that of the original simulations. For planetary boundary layer (PBL) parameterization, Wang et al. (2019)
40    used the inputs and outputs from Yonsei University (YSU) PBL scheme of the Weather Research Forecast (WRF) model to develop the NN-based parameterizations. They showed that the NN-based parameterization successfully simulated the vertical profiles of velocities, temperature, and water vapor within the PBL over the entire diurnal cycle. Urban land surface models (ULSMs) are reasonably fast, but none of them is best at predicting all the main surface fluxes (Grimmond et al., 2010, 2011). One solution is to run an ensemble of ULSMs coupled to the NWP models to improve predictions, which is technically dif-
45    ficult to implement and would require more computational time due to running multiple ULSMs simultaneously. Meyer et al. (2022a) developed an urban NN (UNN) to emulate the ensemble mean of 22 ULSMs. They demonstrate that the WRF model coupled with the UNN is more accurate than WRF with a single well-known USLM. Grundner et al. (2021) demonstrated the potential of an NN-based cloud cover parameterization to accurately reproduce the sub-grid scale cloud fraction to improve the low-resolution climate model predictions.

50    In many cases, ~~DL-based~~ ML-based parameterization schemes are only evaluated in offline settings due to the practical challenge of coupling ~~DL-based~~ ML-based emulators with NWP models. However, the offline performance of ~~DL-based~~ ML-based emulators does not necessarily reflect their online performance when coupled with other components of NWP models. ~~DL-based~~ ML-based parameterizations are often trained using high-level programming languages like Python and easy-to-use ~~DL~~ ML libraries such as Pytorch (Paszke et al., 2017) and Keras (Gulli and Pal, 2017). However, the NWP codes
55    are mainly written in Fortran, making it difficult to integrate directly with ~~DL-based~~ ML-based emulators. Researchers used multiple approaches to circumvent these issues. Krasnopolsky (2014) developed a single-layer NN software for developing NN-based radiation emulators in NWP models (Belochitski and Krasnopolsky, 2021; Krasnopolsky et al., 2010; Song and Roh, 2021). Some researchers wrote a Fortran module that reproduces NN architectures using matrix-matrix multiplication with the weights saved from offline training (Chantry et al., 2021; Hatfield et al., 2021; Rasp et al., 2018), which only works

well for simple NN architectures such as the fully connected (FC) NNs, and becomes increasingly difficult to implement for more complex NN structures. Ott et al. (2020) produced an open-source software library, the Fortran-Keras Bridge (FKB), which enables users to access many features of the Keras API in Fortran and implement NNs in Fortran code. However, the FKB can only support FC or dense layers NNs. Therefore, researchers cannot fully use the most advanced NN structures, such as convolution, self-attention, and variational autoencoder structures, to build powerful ~~DL-based~~ ML-based emulators for online applications. Wang et al. (2022) proposed an NN-GCM coupler to allow Python-based NNs to communicate with the Fortran-based GCM. Although using the coupler can make the ~~DL-based~~ ML-based emulators be deployed efficiently, its requirement for data transfer on a hard disk makes it hard to achieve speedup.

To address the abovementioned problems, we developed a ~~WRF-DL~~ WRF-ML coupler that allows any ~~DL-based~~ ML-based parameterization schemes to be coupled with the WRF model. The WRF model is selected as it is a popular open-source NWP model used by many researchers and operational organizations. Also, to demonstrate the applicability of the coupler, we train and couple the ~~DL-based~~ ML-based radiation emulators with the WRF model. The ~~DL-based~~ ML-based radiation emulators were studied most among all the physical parameterization schemes and coupled with the ECMWF Integrated Forecasting System (IFS), the National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS), and the WRF model in the previous studies (Song and Roh, 2021; Meyer et al., 2022b). Also, the rapid radiative transfer model for general circulation models (RRTMG; Iacono et al., 2008) is selected for radiation as it is widely used in weather and climate models, such as ~~ECMWF IFS,~~ NCEP Global Forecast System (GFS), and the China Meteorological Administration (CMA) Global/Regional Assimilation and Prediction System (GRAPES). To train the ~~DL-based~~ ML-based radiation emulators, the WRF model is run to generate a dataset for training and evaluation. We also illustrate the performance of the ~~DL-based~~ ML-based radiation emulators in the online setting. Lastly, we compare the accuracy and computational costs of the WRF simulations coupled with ~~DL-based~~ ML-based radiation emulators with the WRF simulations using original RRTMG schemes.

The remainder of the paper is organized as follows. Section 2 describes the original and ~~DL-based~~ ML-based RRTMG module. Section 3 introduces the WRF I/O quilt processors and the ~~WRF-DL~~ WRF-ML coupler. Section 4 briefly presents the WRF model setup and ~~DL-based~~ ML-based radiation emulators. Section 5 presents the analysis and results in online settings. Conclusions and discussions are in Section 6.

## 2 Description of Original and ~~DL-based~~ ML-based RRTMG model

### 2.1 Description of Original RRTMG Module

The RRTMG is developed by the Atmospheric and Environmental (AER) and comprises shortwave (SW) and longwave (LW) schemes that calculate SW and LW radiation fluxes and heating rates, respectively. Figure 1 presents the pseudo-code for the original radiation module in the WRF model. At the beginning of the radiation driver module, the input and output variables are initialized. Then, the corresponding radiation calculations are started according to the parameters (i.e., ra_lw_physics and ra_sw_physics) specified in the configuration file (i.e., namelist.input). When the RRTMG schemes are selected, the subroutine

```
subroutine radiation_driver
    defining variables
    initializing data
    call data preprocessing related subroutines, e.g.: call
    radconst, cldfra, ozone or aerosol related interpolation

    lwrad_select:
    select case (config_flag%ra_lw_physics)
        case (RRTMG_LWSCHEME)
            call rrtmg_lwinit
            call rrtmg_lwrad
        case (other lw schemes)
    end select lwrad_select
    swrad_select:
    select case (config_flag%ra_sw_physics)
        case (RRTMG_SWSCHEME)
            call rrtmg_swrad
        case (other sw schemes)
    end select swrad_select
end subroutine radiation_driver
```

**Figure 1.** Pseudo-code for the original RRTMG module

rrtmg_lwinit is called to perform calculations needed to initialize the longwave calculations. Next, the subroutines rrtmg_lwrad and rrtmg_swrad are called to perform the RRTMG SW and LW radiation calculations.

### 2.1.1   Description of ~~DL-based~~ ML-based RRTMG Module

95   The original radiation computations process an atmospheric column at a time and loop over all the horizontal grid points. The three-dimensional variables associated with the grid cell (indexed as $(i, j, k)$ in the WRF model output files) are indexed as ~~$(i, j, k)$~~ $(i, k, j)$ in memory by WRF to make it more cache friendly and increase the cache hit rate. The index $i$, $j$, and $k$ represent the west-east, south-north, and vertical directions, respectively. With the ~~DL-based~~ ML-based RRTMG model, inference can be implemented on a batch of data, which requires the input data to be packed into batches. The input features

100   of ~~DL-based~~ ML-based radiation emulators have dimensions of $W \times H$, where $W$ and $H$ represent the number of features and vertical levels, respectively. Therefore, the original three-dimensional variables are ~~converted~~ reshaped to be indexed as $(i, j, k)$ to match the dimensions requirement of ~~DL-based~~ ML-based input features (done by preprocess_layout in Figure 2). The ~~DL~~ ML model inference is completed within the function infer_run. Unlike physics-based radiation schemes, ~~DL-based~~ ML-based schemes do not need to calculate intermediate variables such as optical depth. Instead, the ~~DL-based~~ ML-based

```
subroutine radiation_driver
        defining variables
        initializing data
        call preprocess_layout
        call infer_run
        call postprocess
end subroutine radiation_driver
```

**Figure 2.** Pseudo-code for the ~~DL-based~~ ML-based RRTMG module.

105 model accomplishes the mapping from the input features to outputs. After inference is finished, ~~DL~~ ML model outputs will be post-processed to match the dimensions of the original WRF data array to continue model simulations.

## 3    ~~WRF-DL~~ WRF-ML coupler

This subsection briefly describes the WRF I/O quilt server techniques as a similar method is adopted in the ~~WRF-DL~~ WRF-ML coupler to have exclusive servers for ~~DL~~ ML model inference. Then the methodologies of the ~~WRF-DL~~ WRF-ML coupler are 110 described in detail.

### 3.1    Description of WRF I/O quilt severs

The WRF model is designed to run efficiently on parallel distributed-memory computing systems. To do this, WRF applies domain decomposition to split the entire domain into nprocs (equal to $n_x \times n_y$) number of subdomains so that the total amount of work is divided into nprocs compute tasks, where $n_x$ and $n_y$ are the number of processors along the west-east and south-north 115 directions. Furthermore, the WRF model contains an I/O layer to gather all the distributed data arrays onto the master Message Passing Interface (MPI) rank 0 while other MPI ranks block until the master completes the write. However, as the number of domain sizes increases or resolution increases, leading to an increasing amount of information to write, the increasing time cost of writing output files becomes a bottleneck for the overall performance. Therefore, the WRF model also provides an option to use asynchronous I/O (quilt) servers that deal exclusively with I/O so that the compute processors can continue without 120 waiting for writing output files. More specifically, WRF allows users to specify the number of groups of I/O servers to allocate (nio_groups) and the number of I/O ranks per group (nio_tasks_per_group). Similar to the I/O quilt servers, the ~~WRF-DL~~ WRF-ML coupler also provides the option to use several processors for ~~DL~~ ML model inference exclusively, of which more details are described in the following subsection. While GPUs are typically more powerful than CPUs, both CPUs and GPUs are supported for inference. CPUs are cheaper than GPUs and more widely available, and using CPUs for inference can also 125 avoid extra costs due to GPU-CPU data transfer.
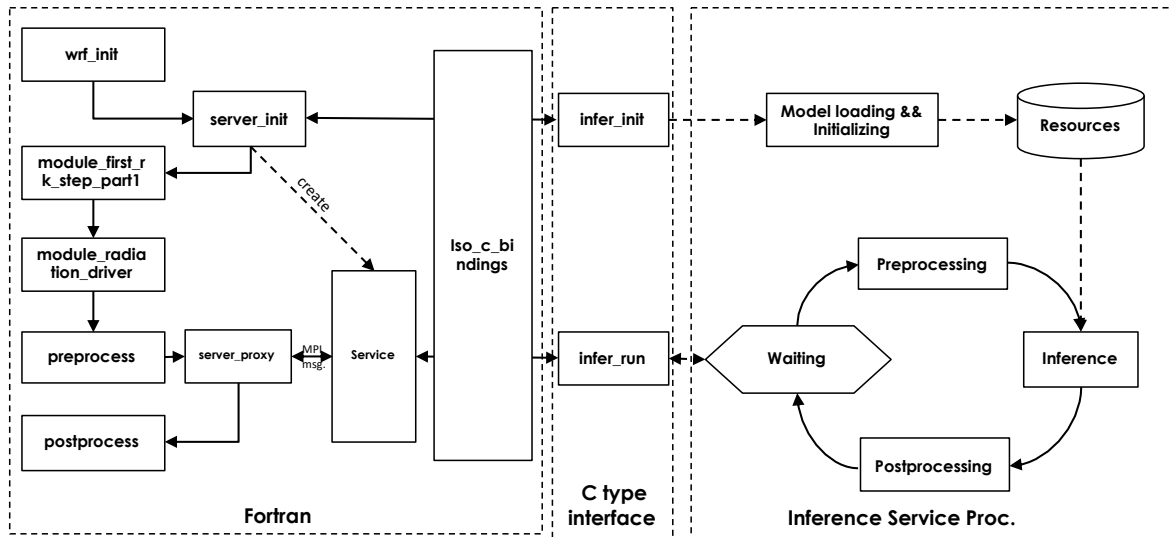
**Figure 3.** Framework of ~~WRF-DL~~ WRF-ML coupler

### 3.1.1  Description of the ~~WRF-DL~~ **WRF-ML** coupler

Figure 3 shows that when the WRF model launches the initialization process, the subroutine infer_init is called to initialize the ~~DL~~ ML model inference services. Similar to the asynchronous I/O servers, several processes are assigned to exclusively work on ~~DL~~ ML model inference using either GPUs or CPUs. The number of inference processors and the number of inference ranks

130  per group are specified in the configuration file by setting the newly added variables ninfer_tasks_per_group and ninfer_groups in the namelist.input file. The difference between the inference processors and the asynchronous I/O servers is that the inference processors apply the synchronous method, as WRF calculations run sequentially, and the subsequent calculations of the WRF model rely on the outputs from ~~DL-based~~ ML-based parameterizations. Usually, the number of inference processors is much smaller than that of WRF compute processors, so a single inference processor must process data from multiple processors. The

135  inference processor listens and receives data from corresponding WRF compute processors through the MPI transmission while blocking the compute processes (see the right part in Figure 4). After inference is finished, the inference process sends outputs of ~~DL-based~~ ML-based schemes back to the compute processors to continue the standard WRF calculations. In addition, the ninfer_tasks_per_group and ninfer_groups can be set to -1 to execute the ~~DL~~ ML model inference using the same processors as for WRF compute processes. Then, memory copy is not needed between WRF compute processors and inference processors.

140  The functions for implementing the ~~DL-based~~ ML-based model inference are all written in Python, commonly used by the machine learning community. To allow the Fortran-based WRF code to call those Python functions, it is necessary first to link the Fortran executables to the system Python library. iso_c_binding is an intrinsic module defined in Fortran 2003 for interoperability with C. As shown in the left part of Figure 3, the iso_c_binding is imported within the Fortran code. Although no C code is needed, the C Foreign Function Interface (CFFI) for Python is used to generate c-style binding interfaces within
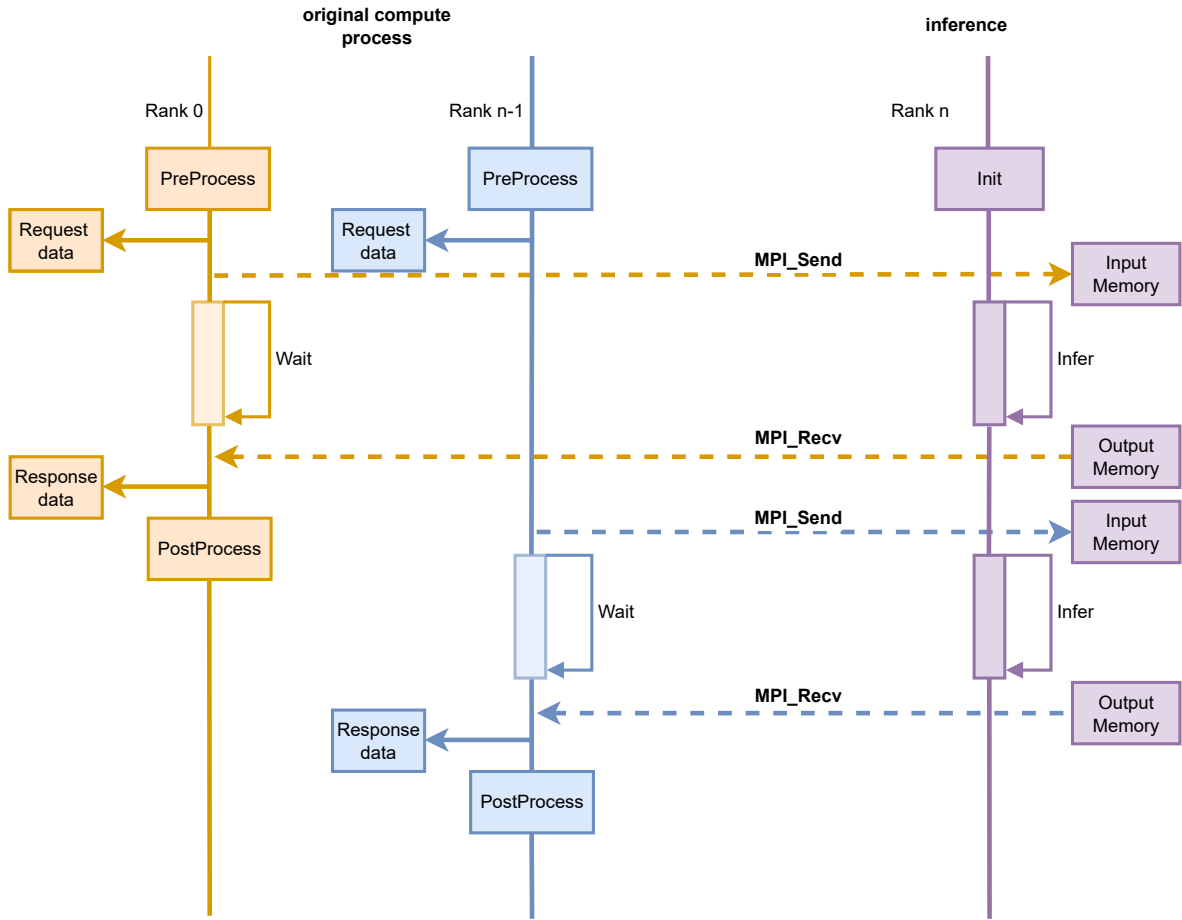
**Figure 4.** Synchronous sequence implementation between WRF and ~~DL-based~~ ML-based parameterization

145   the Python ~~scripts~~ script (see the middle part of Figure 3). The interfaces for C-function are also defined in the Python scripts, which are used to generate a dynamic library that Fortran modules can link.

    This ~~WRF-DL coupler is superior because it makes the DL models~~ WRF-ML coupler allows the ML models to be easily coupled with the WRF model with minimal effort. Also, researchers are able to make full use of the state-of-art ~~DL~~ ML model structures. This methodology is made open-source in the hope that more and more machine learning researchers will participate

150   in improving the traditional NWP models.
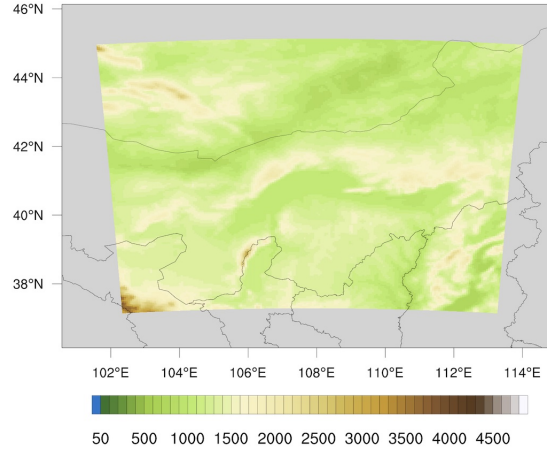
**Figure 5.** Digital evaluation data of the single WRF domain with horizontal resolution at 5 km.

## 4 Experiment Setup

The ~~DL-based~~ ML-based radiation emulator is coupled with the WRF model to demonstrate the coupler's practicality. This section firstly describes the WRF model setup, ~~DL-based~~ ML-based radiation emulators' network structures, and offline evaluation metrics.

155 ### 4.1 WRF Model Setup

In this work, the WRF model version 4.3 is compiled using the GNU Fortran (gfortran version 6.5.1) compiler with the "dmpar" option. To generate a dataset for model training, the WRF model is run using the domain configuration shown in Figure 5. The WRF model is set up within the GFS model grid with a single domain at a horizontal grid spacing of 5 km and has $190 \times 170$ grid points in the west-east and north-south directions (i.e., 32300 grid cells). The number of vertical levels is 57, and the

160 model top is 10 hPa. In addition, the WRF model is configured using physics schemes, including New Thompson et al. scheme (Thompson et al., 2008), Bougeault-Lacarrère (BouLac) planetary boundary layer (PBL) scheme (Bougeault and Lacarrère, 1989), RUC land surface model (Smirnova et al., 1997; Smirnova et al., 2000), and RRTMG for both SW and LW radiation (Iacono et al., 2008). The WRF simulations were run for three days per month, and the initialization days were randomly selected. The first two days of every three days' simulations are used for training, and the last day is used for testing. The

165 model generates radiation inputs and outputs every 30 model minutes.

### 4.2 ~~DL-based~~ ML-based radiation emulators and offline evaluations

Many researchers (Chevallier et al., 2000; Krasnopolsky et al., 2010; Song and Roh, 2021) have previously used the FC networks to replace the radiation schemes within the operational NWP models. Additionally, ~~Yao et al. (2022)~~ Ukkonen (2022)

**Table 1.** Description of different ~~DL~~ ML models used for emulating radiative transfer calculations, including description of model structures and total number of parameters.

| Model Name | Network Structure | ~~Layer Number~~ Number of Layers (Bidirectional Layers for Bi-LSTM models) | ~~Node Number~~ Number of Nodes | ~~Parameter Nu~~ |
|---|---|---|---|---|
| Model A | FC | 5 | 30 | |
| Model B | FC | 10 | 200 | |
| Model C | Bi-LSTM | 1 | 16 | |
| Model D | Bi-LSTM | 3 | 32 | |
| Model E | Bi-LSTM | 5 | 96 | |

showed that bidirectional recurrent NNs (RNNs) are more accurate than the feed-forward NNs (FNNs) for emulating atmospheric
radiative transfer. Yao et al. (2022) also demonstrated that the bidirectional long short-term memory (Bi-LSTM) achieves the best accuracy in ~~emulating radiation~~ radiative transfer parameterization. Therefore, this study has trained and tested both FC networks and Bi-LSTM models. As an increasing number of trainable parameters of a ~~DL~~ ML model increases both accuracy and computational cost, finding a balance is crucial for applying ~~DL~~ ML models in operational NWP models. Table 1 illustrates the structures of five ~~DL~~ ML models, including two FC networks (models A and B) and three Bi-LSTM models (models C to E). Models A and D, as well as models B and E, have approximately the same number of parameters, respectively. More details about the model training and comparisons can be referenced in Yao et al. (2022). All the ~~DL-based~~ ML-based emulators have been converted to ONNX and run using the ONNX Runtime library version 1.7.

Figure 6 shows the inference time of different ~~DL-based~~ ML-based radiation emulators on one Intel Xeon Platinum 8269CY CPU and one NVIDIA Tesla T4 GPU, respectively. Model E has the longest inference time and is much longer than Model B, even though they have a similar number of parameters. This is because the FC networks have much lower time complexity than LSTM models. Similarly, although the number of parameters of model A is about 10.3 and 1.03 times that of models C and D (see Table 1), Model A is still faster than models C and D.

The performance of different ~~DL-based~~ ML-based radiation emulators is evaluated on the testing data in the offline setting. Table 2 summarizes the RMSE of fluxes and heating rates which are the final outputs of the radiation schemes. Since SW fluxes at the surface and top-of-atmosphere (TOA) radiation is particularly important to climate and weather prediction, their RMSE is also shown in Table 2. It is shown that model E is the most accurate, with RMSE of SW and LW fluxes of about 6.157 and 1.274 $W/m^2$, RMSE of SW and LW heating rates of about $5.091 \times 10^{-2}$ and $9.520 \times 10^{-2}$ $K/day$, and SW fluxes at the surface and TOA of about 6.871 and 5.576 $W/m^2$. Model D has comparable accuracy to the best-performing model E while having only 1/14 of the parameters and a much shorter inference time. Furthermore, considering that model C only has the smallest number of parameters, model C has relatively good accuracy while the RMSE of SW heating rate is five times higher than model E. On the other hand, models A and B have at least four times higher RMSE than model E in terms of heating rates, with RMSE of SW fluxes at the surface and top-of-atmosphere at least 20 times higher than model E. In summary, it is
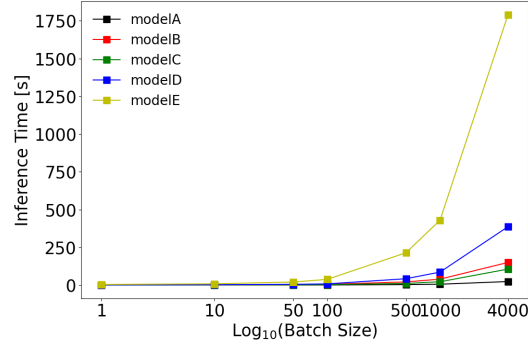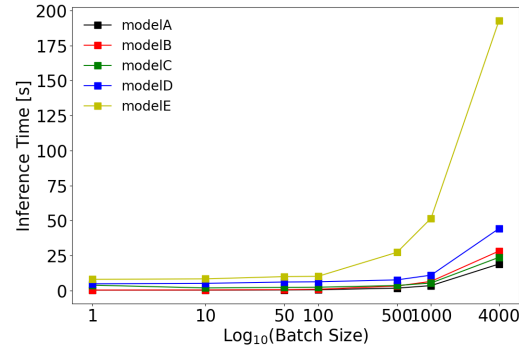
**Figure 6.** Inference time of different ~~DL-based~~ ML-based emulators run on one CPU core (6a) and GPU (6b) with different batch size. The number of vertical levels is 57.

**Table 2.** RMSE of different ~~DL-based~~ ML-based radiation emulators for test data in offline evaluation

|  | SW Flux | LW Flux | SW Heating Rate | LW Heating Rate | Surface SW Flux | TOA SW Flux |
|---|---|---|---|---|---|---|
|  | $W \cdot m^{-2}$ | $W \cdot m^{-2}$ | $K \cdot d^{-1}$ | $K \cdot d^{-1}$ | $W \cdot m^{-2}$ | $W \cdot m^{-2}$ |
| Model A | 2.470e+02 | 4.146e+01 | 1.172e+00 | 2.035e+00 | 2.092e+02 | 2.735e+02 |
| Model B | 1.952e+02 | 2.964e+01 | 9.455e-01 | 2.172e+00 | 1.733e+02 | 2.163e+02 |
| Model C | 7.658e+00 | 2.032e+00 | 2.772e-01 | 3.373e-01 | 8.534e+00 | 6.628e+00 |
| Model D | 5.355e+00 | 1.305e+00 | 9.943e-02 | 1.614e-01 | 5.854e+00 | 4.804e+00 |
| Model E | 6.157e+00 | 1.274e+00 | 5.091e-02 | 9.520e-02 | 6.871e+00 | 5.576e+00 |

demonstrated that the Bi-LSTM model is more accurate than FC networks for radiation emulation, and model D achieves the best balance between accuracy and computing efficiency.

Since it is difficult to know the performance of the ~~DL-based~~ ML-based emulators in the online settings based on offline evaluation, the following Section presents the online evaluations of different ~~DL-based~~ ML-based emulators coupled with the WRF model. In addition, the WRF model is also run with the original RRTMG schemes as a reference to evaluate the performance of the ~~DL-based~~ ML-based emulators.

## 5 Online Evaluation Results

The ~~DL-based~~ ML-based radiation emulators are coupled with the WRF model using the ~~WRF-DL~~ WRF-ML coupler. The WRF simulations coupled with ~~Dl-based emulators~~ ML-based emulators were run using three configurations: zero exclusive processes for inference, four exclusive inference processes on CPUs, and four exclusive inference processes on ~~CPUs~~GPUs. Also, the total number of WRF processes is kept to 24 to ensure fair comparisons. The calculation time of radiation driver using different ~~DL-based~~ ML-based radiation emulators is compared. The WRF model is initialized at 12 UTC on a randomly selected day (i.e., November 9th) in the year 2021 and runs for three days, which is not part of the previously used dataset to ensure the ~~DL-based~~ ML-based emulators are evaluated on unseen data. Finally, the temperature at 2 meters ($T_{2m}$) and wind speed at 10 meters ($WS_{10m}$) output from the WRF model are compared.

### 5.1 Computational efficiency evaluation of radiation driver

The calculation time of the WRF radiation driver using original RRTMG schemes and ~~DL-based~~ ML-based radiation emulators are profiled using the timing function rsl_~~interal~~internal_microclock provided by the WRF model. Table 3 shows that it takes about 1441 milliseconds to run a radiation driver using the original RRTMG schemes. In terms of ~~DL-based~~ ML-based radiation emulators, using GPUs for inference is significantly faster than using CPUs for inference, especially for the Bi-LSTM model, as it is less computationally efficient than the FC networks given a similar number of parameters. When GPUs are used for inference, all ~~DL-based~~ ML-based emulators except model E are at least three times faster than the original RRTMG schemes. However, whether this speed-up is due to applying ML-based emulators or the upgrade in hardware (using CPUs instead of GPUs) is unclear, which is beyond the scope of this paper. On the other hand, when CPUs are used for inference, having four processes exclusively for inference is slower than having inference executed on the WRF compute processors, which is probably due to the more significant increase in time cost due to data copy between CPUs than the decrease in time cost from having CPUs exclusively for inference.

### 5.2 ~~Accuracy evaluation~~Prognostic validation

Figures 7 and 8 present the difference in 24, 48, and 72 hours forecast between the WRF simulations coupled with ~~DL-based~~ ML-based radiation emulators and the WRF simulation with original RRTMG schemes. On the spatial difference, the red and blue patterns indicate significantly positive and negative biases of ~~DL-based~~ ML-based simulations, respectively. On the other hand, green patterns indicate no or small difference from the original WRF simulations. A domain averaged mean absolute difference (MAD) is also calculated to measure the overall performance of ~~DL-based~~ ML-based emulators in online settings.

**Table 3.** Calculation time (in units of milliseconds) of radiation driver when using the original RRTMG schemes and the ~~DL-based~~ ML-based radiation emulators. The batch size is 32300, as the domain has $190 \times 170$ grid points.

| Total Processes | 24 | | |
|---|---|---|---|
| original RRTMG | 1441.24 ms | | |
| ~~DL~~ ML Inference Processes | 0 | 4 ~~CPU~~ CPUs | 4 ~~GPU~~ GPUs |
| Model A | 91.73 | 237.57 | 176.87 |
| Model B | 365.31 | 808.43 | 331.34 |
| Model C | 666.37 | 921.56 | 258.66 |
| Model D | 2620.92 | 3383.42 | 418.28 |
| Model E | 7166.46 | 10330.05 | 1690.76 |

Notably, the difference does not increase with the simulation time, as the difference at 72 forecast hours is similar to that of 24 forecast hours. The WRF simulations coupled with FC networks (models A and B) have the worst performance, with a much larger difference from the original WRF simulation than WRF simulations coupled with Bi-LSTM models (models D and E) over the entire domain. Model B is more accurate than model A in offline settings but performs much worse than model A in the online evaluation. The MAD of $T_{2m}$ and $WS_{10m}$ is greater than ~~2.3 and 1.0~~ 5.6 and 1.3, respectively, for model B, while the MAD of $T_{2m}$ and $WS_{10m}$ is slightly less than ~~3~~ 2.3 and 1.0, respectively, for ~~model~~ models A. Model C performs best among all the three FC networks, with MAD slightly worse than MAD of models D and E. Models D and E have a similar spatial distribution of difference, although model E has about ten times the number of parameters as model D and slightly higher accuracy in offline validation. ~~From the comparisons of offline and online performance of Dl-based parameterization, it can be concluded that the DL-based parameterization's offline skills can not reflect their online skills. Therefore, it is necessary to have DL-based parameterization coupled with the NWP models and evaluated in online settings.~~

Overall, model E demonstrates the best performance, with the domain averaged MAD smaller than 0.10 for both temperature at 2 meters and wind speed at 10 meters. Both model D and model E show a comparable forecast with the original WRF simulation over the entire domain. Overall, model D is more suitable to replace the original RRTMG schemes as it is more computationally efficient than model E. Using model D with GPU as inference processors is about 7 times faster.

In summary, the ~~WRF-DL~~ WRF-ML coupler enables all the ~~DL-based~~ ML-based radiation emulators coupled with the WRF models ~~more~~ efficiently. Moreover, the WRF simulations coupled with ~~DL~~ ML models are run successfully using CPU or GPU for inference. Furthermore, the ~~WRF-DL~~ WRF-ML coupler can be a valuable tool for researchers to evaluate their ~~DL-based~~ ML-based parameterization in online settings.
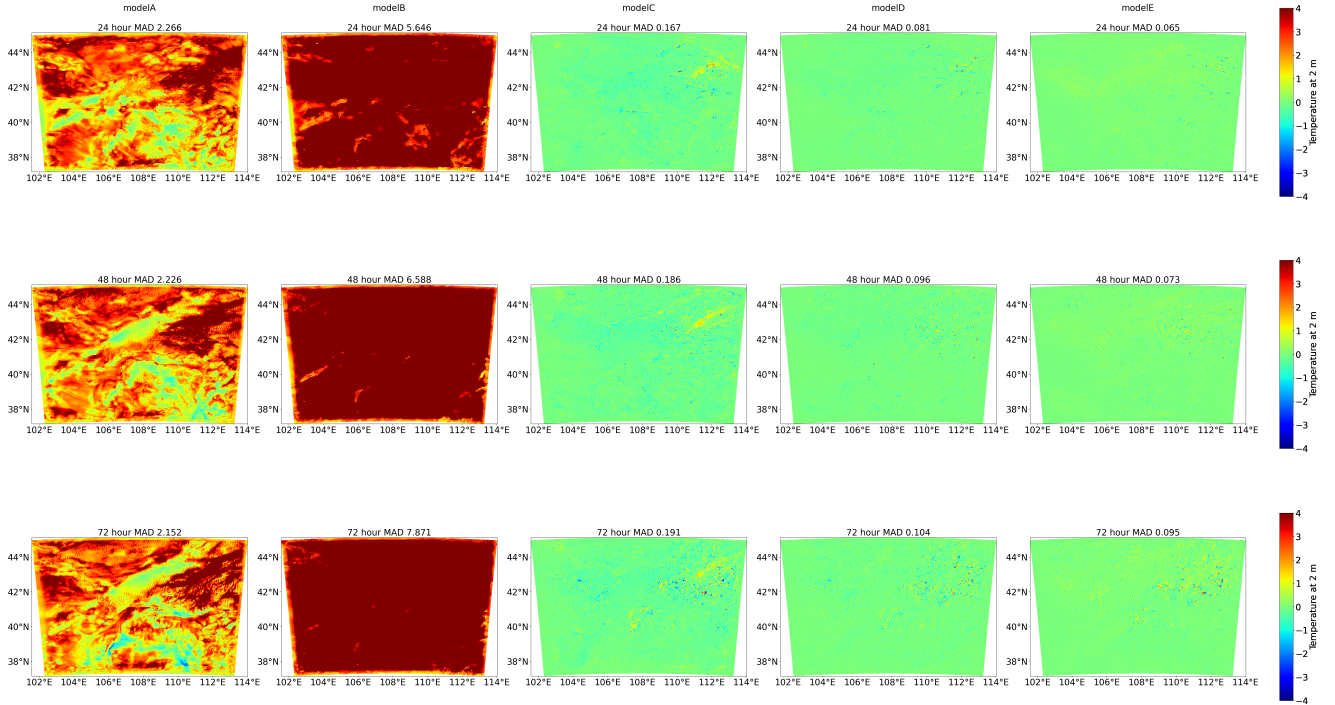
**Figure 7.** Differences in temperature at 2 meters at forecast horizons of 24, 48, and 72 hours between WRF simulations coupled with ~~DL~~ ML models A, B, C, D, and E and WRF simulation with the original RRTMG schemes for radiation emulation. The title of each figure indicates the mean absolute difference (MAD) of the WRF domain.
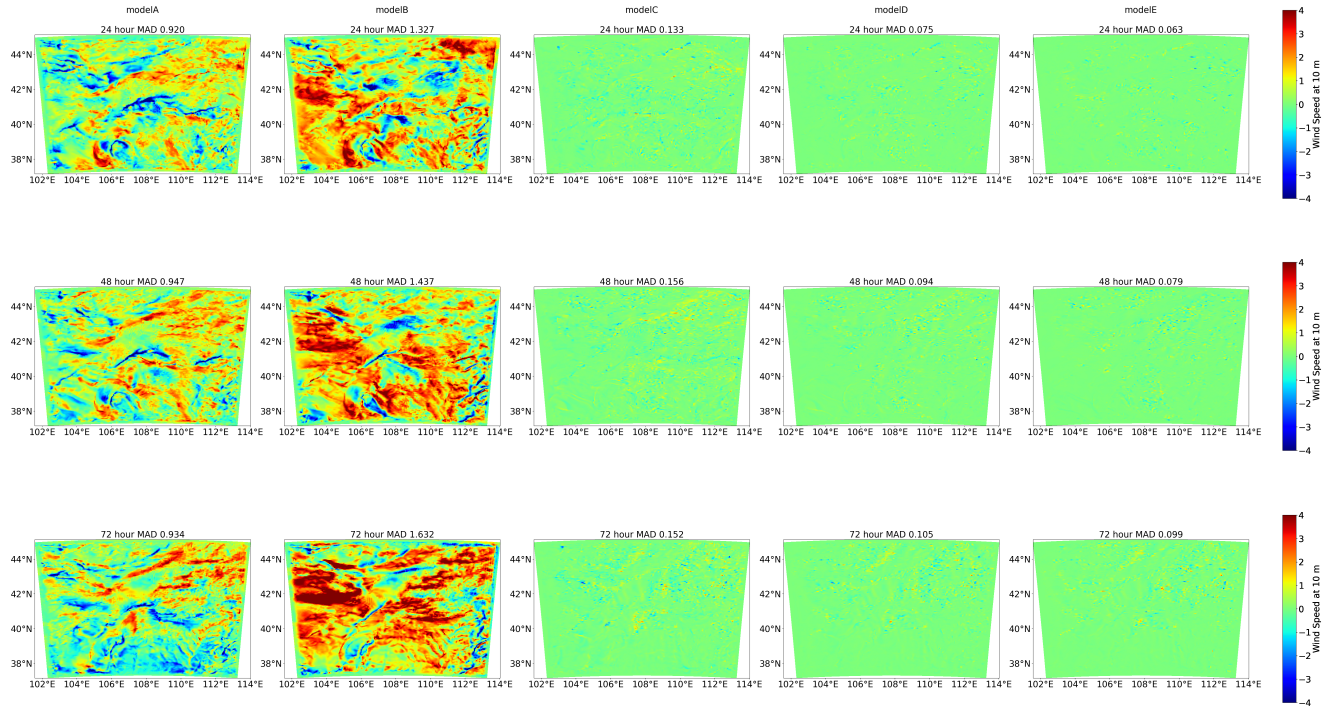
**Figure 8.** Differences in wind speed at 10 meters at forecast horizons of 24, 48, and 72 hours between WRF simulations coupled with ~~DL~~ ML models A, B, C, D, and E and WRF simulation with the original RRTMG schemes for radiation emulation.

## 6 Summary and Conclusions

As demonstrated in many previous studies, emulating the subgrid physics using ~~DL~~ ML models has the potential to be faster~~and~~, or by training on observations or detailed high-resolution models, even more accurate than the conventional ~~physics-based~~ parameterization schemes. However, the fact that ~~DL~~ ML models are commonly implemented using Python while the NWP models are coded in Fortran makes it challenging to implement the ~~DL-based~~ ML-based parameterization in NWP models for operational applications. Previously, researchers usually hard coded the operations of NNs into Fortran, which is time-consuming and prone to error as even minor changes to the ~~DL~~ ML model require rewriting the Fortran code. Some researchers used ~~the~~ simple Fortran-based NN libraries~~to access features of the Python-based deep learning frameworks in Fortran~~, which could convert existing models trained in Python to models usable in the Fortran framework. However, these methods are only feasible for applying simple, dense, layer-based NNs. In this paper, we propose a ~~WRF-DL~~ WRF-ML coupler that provides researchers with the tool to implement ~~DL~~ ML models into NWP models with minimal effort. The advantage of this coupler is that more complex and advanced ~~art DL~~ ML model structures are supported~~as the implementation of DL-based parameterization is achieved in Python~~, as it can directly communicate with Python-based modules. The coupler is made open source in the hope that more and more researchers can integrate the ~~DL-based~~ ML-based parametrization to improve the traditional NWP models for more accurate weather or climate forecast. In addition, the methodologies also allow the integration of any Python scripts into the WRF model. An example of coupling the primitive Python subroutines into the WRF code is also provided in the code repository.

Furthermore, the capability of the ~~WRF-DL~~ WRF-ML coupler was demonstrated by coupling the ~~DL-based~~ ML-based radiation parameterization into the WRF model. It was also illustrated that ~~DL-based~~ ML-based parameterizations' computational efficiency and accuracy should be balanced to achieve the best overall performance for operational applications.

*Code and data availability.* The source code and data used in this are available at https://doi.org/10.5281/zenodo.7056166 (Zhong et al., 2022)

*Author contributions.* Y.Y. trained the machine learning models and calculated the statistics of models' offline performance. X.Z., Y.W., and Z.M. modified the WRF source code and developed the WRF-ML coupler. X.Z. conducted the WRF model simulations coupled with the original RRTMG schemes for ML model training and offline evaluation. X.Z. and Z.M. conducted the WRF simulations coupled with ML-based radiation emulators. X.Z. wrote, reviewed, and edited the original draft; Z.W. supervised and supported this research and gave valuable 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.

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

# References

275    Belochitski, A. and Krasnopolsky, V.: Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art general circulation model, Geoscientific Model Development, 14, 7425–7437, 2021.

Bougeault, P. and Lacarrère, P.: Parameterization of orographic induced turbulence in a mesobeta scale model, Monthly Weather Review, 117, 1872–1890, https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2, 1989.

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., and Palmer, T.: Machine Learning Emulation of Gravity Wave
280    Drag in Numerical Weather Forecasting, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 477, https://doi.org/https://doi.org/10.1029/2021MS002477, e2021MS002477 2021MS002477, 2021.

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

Chevallier, F., Morcrette, J.-J., Chéruy, F., and Scott, N.: Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF
285    atmospheric model, Quarterly Journal of the Royal Meteorological Society, 126, 761–776, 2000.

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

Coiffier, J.: Fundamentals of Numerical Weather Prediction, Cambridge University Press, https://doi.org/10.1017/CBO9780511734458, 2011.

290    Grimmond, C. S. B., Blackett, M., Best, M. J., Barlow, J., Baik, J.-J., Belcher, S. E., Bohnenstengel, S. I., Calmet, I., Chen, F., Dandou, A., Fortuniak, K., Gouvea, M. L., Hamdi, R., Hendry, M., Kawai, T., Kawamoto, Y., Kondo, H., Krayenhoff, E. S., Lee, S.-H., Loridan, T., Martilli, A., Masson, V., Miao, S., Oleson, K., Pigeon, G., Porson, A., Ryu, Y.-H., Salamanca, F., Shashua-Bar, L., Steeneveld, G.-J., Tombrou, M., Voogt, J., Young, D., and Zhang, N.: The international urban energy balance models comparison project: First results from phase 1, Journal of Applied Meteorology and Climatology, 49, https://doi.org/10.1175/2010JAMC2354.1, 2010.

295    Grimmond, C. S. B., Blackett, M., Best, M. J., Baik, J.-J., Belcher, S. E., Bohnenstengel, S. I., Calmet, I., Chen, F., Coutts, A., Dandou, A., Fortuniak, K., Gouvea, M. L., Hamdi, R., Hendry, M., , Kanda, M., Kawai, T., Kawamoto, Y., Kondo, H., Krayenhoff, E. S., Lee, S.-H., Loridan, T., Martilli, A., Masson, V., Miao, S., Oleson, K., Ooka, R., , Pigeon, G., Porson, A., Ryu, Y.-H., Salamanca, F., Steeneveld, G.-J., Tombrou, M., Voogt, J., Young, D., and Zhang, N.: Initial results from Phase 2 of the international urban energy balance model comparison, International Journal of Climatology, 31, https://doi.org/10.1002/joc.2227, 2011.

300    Grundner, A., Beucler, T., Iglesias-suarez, F., Gentine, P., Giorgetta, M. A., and Eyring, V.: Deep Learning Based Cloud Cover Parameterization for ICON, arXiv preprint arXiv, https://doi.org/10.48550/arXiv.2112.11317, 2021.

Gulli, A. and Pal, S.: Deep learning with Keras, Packt Publishing Ltd, 2017.

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., and Palmer, T.: Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 521, 2021.

305    Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D.: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models, Journal of Geophysical Research: Atmospheres, 113, 2008.

Janisková, M., Mahfouf, J. F., Morcrette, J. J., and Chevallier, F.: Linearized radiation and cloud schemes in the ECMWF model: Development and evaluation, Quarterly Journal of the Royal Meteorological Society, 128, 1505–1527, 2002.

Krasnopolsky, V., Fox-Rabinovitz, M., Hou, Y., Lord, S., and Belochitski, A.: 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, 1822–1842, 2010.

Krasnopolsky, V. M.: NN-TSV, NCEP neural network training and validation system; brief description of NN background and training software, https://doi.org/https://doi.org/10.7289/v5qr4v2z, 2014.

Mason, P. J.: Large-Eddy Simulation of the Convective Atmospheric Boundary Layer, Journal of the Atmospheric Sciences, 46, 1492–1516, https://doi.org/10.1175/1520-0469(1989)046, 1989.

Meyer, D., Grimmond, S., Dueben, P., Hogan, R., and Reeuwijk, M. v.: Machine Learning Emulation of Urban Land Surface Processes, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2021MS002744, 2022a.

Meyer, D., Hogan, R. J., Dueben, P. D., and Mason, S. L.: Machine Learning Emulation of 3D Cloud Radiative Effects, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2021ms002550, 2022b.

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., and Baldi, P.: A Fortran-Keras Deep Learning Bridge for Scientific Computing, arXiv preprint arXiv:2004.10652, 2020.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A.: Automatic Differentiation in PyTorch, in: NIPS 2017 Workshop on Autodiff, https://openreview.net/forum?id=BJJsrmfCZ, 2017.

Pu, Z. and Kalnay, E.: Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation, https://doi.org/10.1007/978-3-642-40457-3_11-1, 2018.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the IEEE conference on computer vision and pattern recognition, https://doi.org/10.1073/PNAS.1810286115, 2018.

Schneider, T., Lan, S., Stuart, A., and Teixeira, J.: Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations, Geophysical Research Letters, 44, 12,396–12,417, https://doi.org/10.1002/2017GL076101, 2017.

Smirnova, T. G., Brown, J. M., and Benjamin, S. G.: Performance of Different Soil Model Configurations in Simulating Ground Surface Temperature and Surface Fluxes, Monthly Weather Review, 125, 1870–1884, https://doi.org/10.1175/1520-0493(1997)125<1870:PODSMC>2.0.CO;2, 1997.

Smirnova, T. G., Brown, J. M., Benjamin, S. G., and Kim, D.: Parameterization of cold-season processes in the MAPS land-surface scheme, Journal of Geophysical Research, 105, 4077–4086, https://doi.org/10.1029/1999JD901047, 2000.

Song, H.-j. and Roh, S.: Improved Weather Forecasting Using Neural Network Emulation for Radiation Parameterization, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2021MS002609, 2021.

Stensrud, D. J.: Parameterization schemes: keys to understanding numerical weather prediction models, Cambridge University Press, https://doi.org/10.1017/CBO9780511812590, 2013.

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization, Monthly Weather Review, 136, 5095 – 5115, https://doi.org/10.1175/2008MWR2387.1, 2008.

Ukkonen, P.: Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 875, 2022.

Wang, J., Balaprakash, P., and Kotamarthi, R.: Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model, Geoscientific Model Development, 12, 4261–4274, https://doi.org/10.5194/gmd-12-4261-2019, 2019.

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

350 Yao, Y., Zhong, X., Zheng, Y., and Wang, Z.: A Physics-Incorporated Deep Learning Framework for Parameterization of Atmospheric Radiative Transfer, Journal of Advances in Modeling Earth Systems, https://doi.org/10.1002/essoar.10512741.1, 2022.

Zhong, X., Ma, Z., Yao, Y., Xue, L., Wu, Y., and Wang, Z.: Code and Data for "A Bridge between WRF and Deep Learning Parameterization and Application on Deep Learning Parameterization of Atmospheric Radiative Transfer", Zenodo[code and data], https://doi.org/10.5281/zenodo.7056166, 2022.