Estimating full longwave and shortwave radiative transfer with neural

Ryan Lagerquist[1], David D Turner[1], Imme Ebert-Uphoff[2], and Jebb Q Stewart[1]

[1]Affiliation not available
[2]Department of Electrical and Computer Engineering, Colorado State University

July 17, 2023

Estimating full longwave and shortwave radiative transfer with neural networks of varying complexity

This article has been conditionally accepted to the AMS *Journal of Atmospheric and Oceanic Technology*.

Ryan Lagerquist[a,b], David D. Turner[b], Imme Ebert-Uphoff[a,c], and Jebb Q. Stewart[b]

[a] *Cooperative Institute for Research in the Atmosphere (CIRA)*

[b] *National Oceanic and Atmospheric Administration (NOAA) Global Systems Laboratory (GSL), Boulder, Colorado*

[c] *Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado*

Corresponding author: Ryan Lagerquist, ralager@colostate.edu

ABSTRACT: Radiative transfer (RT) is a crucial but computationally expensive process in numerical weather/climate prediction. We develop neural networks (NN) to emulate a common RT parameterization called the Rapid Radiative-transfer Model (RRTM), with the goal of creating a faster parameterization for the Global Forecast System (GFS) v16. In previous work we emulated a highly simplified version of the shortwave RRTM only – excluding many predictor variables, driven by Rapid Refresh forecasts interpolated to a consistent height grid, using only 30 sites in the northern hemisphere. In this work we emulate the full shortwave and longwave RRTM – with all predictor variables, driven by GFSv16 forecasts on the native pressure-sigma grid, using data from around the globe. We experiment with NNs of widely varying complexity, including the U-net++ and U-net3+ architectures and deeply supervised training, designed to ensure realistic and accurate structure in gridded predictions. We evaluate the optimal shortwave NN and optimal longwave NN in great detail – as a function of geographic location, cloud regime, and other weather types. Both NNs produce extremely reliable heating rates and fluxes. The shortwave NN has an overall RMSE/MAE/bias of 0.14/0.08/-0.002 K day^{-1} for heating rate and 6.3/4.3/-0.1 W m^{-2} for net flux. Analogous numbers for the longwave NN are 0.22/0.12/-0.0006 K day^{-1} and 1.07/0.76/+0.01 W m^{-2}. Both NNs perform well in nearly all situations, and the shortwave (longwave) NN is 7510 (90) times faster than the RRTM. Both will soon be tested online in the GFSv16.

SIGNIFICANCE STATEMENT: Radiative transfer is an important process for weather and climate. Accurate radiative-transfer models exist, such as the RRTM, but these models are computationally slow. We develop neural networks (NN), a type of machine-learning model that is often computationally fast after training, to mimic the RRTM. We wish to accelerate the RRTM by orders of magnitude without sacrificing much accuracy. We drive both the NNs and RRTM with data from the GFSv16, an operational weather model, using locations around the globe during all seasons. We show that the NNs are highly accurate and much faster than the RRTM, which suggests that the NNs could be used to solve radiative transfer inside the GFSv16.

1. Introduction

Radiative heating is a main driver of the Earth's climate and the only process by which the Earth can exchange energy with the rest of the universe; radiative transfer (RT) is the governing theory. In RT studies the electromagnetic spectrum is often separated into the shortwave part (wavelength $\lesssim 4\ \mu$m), which is mostly emitted by the Sun, and the longwave part ($\gtrsim 4\ \mu$m), which is mostly emitted by the Earth – both its surface and atmosphere.[1] The global distribution of top-of-atmosphere (TOA) incoming shortwave radiation is controlled largely by geographic variations in the solar zenith angle and surface albedo, with low (high) zenith angle and albedo at the low (high) latitudes.[2] This sets up a strong meridional gradient in TOA incoming shortwave radiation, with higher values at lower latitudes. The global distribution of TOA outgoing longwave radiation is somewhat similar, because warmer surfaces (at lower latitudes) emit more longwave radiation than colder surfaces. However, the longwave distribution is more complicated, because longwave radiation interacts more strongly with atmospheric gases. Overall, the low latitudes have a surplus of net radiation (TOA incoming shortwave minus TOA outgoing longwave), while the high latitudes have a deficit. This imbalance maintains the meridional temperature gradient we observe, as well as driving the global atmospheric circulation, including a strong poleward heat flux produced by baroclinic waves. (Wallace and Hobbs 2006)

RT is also crucially important for day-to-day weather prediction, because it causes differential diabatic heating. In numerical weather prediction (NWP), this diabatic heating is a subgrid-scale process and is therefore parameterized by a separate RT model. The most accurate RT models are

[1]The 4-μm threshold is not an exact constant; sometimes other values are used.

[2]Clouds (both liquid and ice; Tang et al. 2020) and aerosols (Myhre et al. 2013) also play a major, though highly uncertain, role in the Earth's shortwave-radiation budget.

line-by-line models (Turner et al. 2004; Mlawer and Turner 2016), but these are far too slow for NWP. A popular compromise is the Rapid Radiative-transfer Model (RRTM; Mlawer et al. 1997), a hybrid physical/statistical model that is nearly as accurate as line-by-line models but millions of times faster. The RRTM, like most RT models, performs 1-dimensional RT, assuming that RT occurs only in the vertical. Faster variants – such as the RRTM for global climate models (RRTMG; Pincus and Stevens 2013), RRTMG Parallel (RRTMGP; Mlawer and Delamere 2019), and RRTMG-K (Baek 2017) – are often used in NWP as well. However, the RRTM and its variants are still computationally expensive, accounting for 20 to 50% of the total computing of the host NWP model (*e.g.*, Cotronei and Slawig 2020). We have elected to emulate the RRTM[3] because, by using more quadrature points, it is more accurate than the RRTMG.

This has motivated a body of work on using neural networks (NN; Part II of Goodfellow et al. 2016), an algorithm from machine learning (ML), to emulate RT models, dating back to Chevallier et al. (1998). ML-based emulation of RT and other subgrid-scale processes almost always uses NNs, so we omit other ML algorithms from this review. The main advantage of NNs is that they can accurately model complex relationships (hence "universal function-approximators"; see, *e.g.*, Sonoda and Murata 2017) and are much faster than the RRTM and its variants at inference time, *i.e.*, when applying a trained NN to predict on new data. The main disadvantage is that they are purely statistical models and, without physical constraints, may not generalize well to conditions outside the range of their training data, such as future climates. Also, adding predictor variables to a NN requires complete retraining. An overall disadvantage of replacing parameterizations such as the RRTM is that the host NWP models are very sensitive to changes in parameterizations (Boukabara et al. 2019; Rasp 2020; Muñoz-Esparza et al. 2022). Thus, even if the RT-emulator has very small errors in offline testing (outside the NWP model), during online testing (inside the NWP model) these errors may accumulate or cause undesired feedbacks to other components of the NWP model, degrading the quality of the overall weather forecast. However, if successfully integrated into an NWP model, a NN-based RT-emulator can decrease computing requirements by orders of magnitude.

The current article expands on work presented in Lagerquist et al. (2021), henceforth L21. Differences between this work and L21 are listed at the end of the introduction. The following

[3]Specifically version 2.7.1 of the shortwave model, covering the 0.2–12.2-μm band, and version 3.3 of the longwave model, covering the 3.07–1000-μm band.

review focuses on recent work in RT emulation, especially work published after L21. We divide recent work into four categories: emulating RT in climate models, emulating RT in weather models, emulating only part of an RT model such as gas optics, and miscellaneous.

In climate-modeling, Pal et al. (2019) developed an RT-emulator for the super-parameterized Energy Exascale Earth System Model (SP-E3SM) and found in online testing that the emulator produces a similar climate to the original RT model. Beucler et al. (2021) used climate-invariant NNs to emulate both RT and other subgrid-scale processes in climate models. They ensured climate-invariance by rescaling three predictor variables for the NN – temperature, humidity, and latent-heat flux – to forms that are not projected to increase with global warming. Without rescaling, applying the trained NN to future climates involved extrapolating (*e.g.*, applying the NN to temperatures higher than any seen in the training data), which degraded performance. Beucler et al. found that rescaling allows their NN to predict subgrid-scale processes well, including RT, in a climate 8 K warmer than the climate used for training. Belochitski and Krasnopolsky (2021) used an emulator developed in 2011 for the Climate Forecast System (CFS) and integrated it into version 16 of the Global Forecast System (GFSv16). They found that the emulator generalized well between the host models without retraining – *i.e.*, the GFSv16 with the emulator produced a similar climate to the GFSv16 with the original RRTMG parameterization. However, this success was achieved only after changing the number of heights and prognostic variables in the GFSv16 to match the CFS.

In weather-modeling, much recent work has been done at the Korean Meteorological Agency (KMA). Roh and Song (2020) became the first to emulate RT at cloud-resolving resolution, developing NNs for a 250-metre version of the Weather Research and Forecasting (WRF) model. However, this work was limited by focusing on a single idealized squall-line simulation. Song and Roh (2021) developed a more general RT-emulator for use in the Korea Local Analysis and Prediction System (KLAPS), an operational version of the WRF used by the KMA. When tested online in KLAPS, the NN produced similar instantaneous temperature and precipitation fields to the original RRTMG-K parameterization, suggesting that the NN may be suitable for operational use. Kim and Song (2022) used automatic hyperparameter-tuning[4] to find the best learning rate and training-batch size for the same KLAPS application, improving the performance of the NN further.

[4]A hyperparameter is a NN parameter that, unlike the weights and biases, cannot be adjusted during training. A hyperparameter must be tuned by trial and error, *i.e.*, training many NNs with different values.

Lastly, researchers at the ECMWF are currently working to integrate NN-based RT-emulators into an operational model, namely the Integrated Forecasting System (Chantry et al. 2022, 2023).

Some groups have used NNs to emulate only the gas-optics step of an RT model. Gas optics maps the physical/chemical state of the atmosphere to a profile of spectral optical depths, and the solver – the second and last step of an RT model – maps the optical depths to heating rates and fluxes (Veerman et al. 2020). Specifically, gas optics converts temperature, pressure, and chemical concentrations into quantities that directly determine how much radiation is emitted, absorbed, and scattered in different directions (Veerman et al. 2020). Gas optics is an empirical algorithm in many RT models, relying on observations stored in large lookup tables, whereas the RT-solver is a physical algorithm, relying on well known equations. Because large lookup tables are computationally slow, gas optics is ripe for acceleration by NNs; because gas optics is already empirical, acceleration by NNs does not remove physical knowledge from the RT model. Ukkonen et al. (2020) emulated the gas-optics scheme in the RRTMGP and found that at most locations on Earth, the emulator introduces an RMSE of < 0.5 W m^{-2} in fluxes and < 0.1 K day^{-1} in heating rates for both the shortwave and longwave. Veerman et al. (2020) also emulated gas optics in the RRTMGP, obtaining similar results. Stegmann et al. (2022) emulated gas absorption in the Community Radiative-transfer Model, which is used in the observation operator for satellite-data assimilation. Lastly, Ukkonen (2022) tested the use of NNs for three different emulation tasks: only the gas-optics scheme, only the reflectance-transmission calculations in the RT-solver, and the full RT model. They found that replacing only the gas-optics scheme leads to the most accurate emulation, followed by replacing the full RT model. However, this study is limited by focusing only on shortwave RT for cloudy profiles. Geiss et al. (2022) emulated the aerosol-optics scheme of an RT model, using NNs with novel architectures, and found that connections between non-adjacent NN layers – which are uncommon in the literature – yielded the best performance.

NNs have additionally been used to simulate 3-dimensional RT (Meyer et al. 2022; Yang et al. 2022) and hyperspectral RT (Le et al. 2020). Also, one study (Liu et al. 2020) has explored the effect of NN-architectural complexity on RT accuracy. They compared fully connected and convolutional NNs[5], finding that convolutional NNs achieve slightly better performance but not

[5]Fully connected (or "dense") NNs treat the predictor variables as independent scalars, while convolutional NNs treat the predictors as images. Thus, convolutional NNs can leverage spatial structure in gridded data, while fully connected NNs cannot. While convolutional NNs are typically applied to 2-D or 3-D images, they can be applied just as easily to 1-D "images" – such as the vertical profiles in this study – and leverage spatial structures therein.

enough to justify the added computational cost. However, they focused only on longwave RT in clear-sky conditions, and their errors were quite large (*e.g.*, heating-rate errors often $\gg$ 1 K day^{-1} near the surface). L21 explored U-net (Ronneberger et al. 2015) and U-net++ models (Zhou et al. 2019), convolutional NNs designed for image-to-image translation. In offline evaluation, they found that U-net++ models outperform fully connected NNs in general and outperform traditional U-nets for profiles with multi-layer cloud, where RT is the most complex. See their Supplemental Section Cd for this architectural comparison.

In this work we use NNs – specifically the U-net++ and U-net3+ architectures – to emulate the full RRTM. "Full" means that we emulate everything: both gas optics and the RT-solver, for both the shortwave and longwave, including all predictor variables. This contrasts with L21, where we emulated a simplified shortwave RRTM without aerosols, trace gases, or information on the particle-size distribution (PSD) of hydrometeors. Our eventual aim is to integrate the NN-based emulators into the GFSv16, a global model with hybrid pressure-sigma coordinates in the vertical. Thus, we train the NNs with GFSv16 data from locations around the globe on the native pressure-sigma grid – in contrast to L21, we trained with data from 30 sites in the northern hemisphere on a standard height grid.

2. Data

This section discusses predictor (input) variables and target (output) variables. The RRTM and the NNs we use to emulate the RRTM have the same target variables and mostly the same predictor variables; the NNs have two extra predictor variables, as discussed in Section 2a. Most predictor variables come from the GFSv16, but some are synthetic, because they are difficult to observe and not available in the GFSv16 output files. Because the NNs are built to emulate the RRTM, target variables produced by the RRTM are considered ground truth – "labels" in ML terminology.

a. GFSv16-based predictors

The GFSv16 is a global, non-hydrostatic, operational model with 0.25° horizontal spacing and 127 vertical levels in hybrid pressure-sigma coordinates, extending to the mesopause at ~80 km above sea level[6]. We have obtained 0000 UTC model runs from the National Environmental Security Computing Center's (NESCC) High-performance Storage System (HPSS). The

[6]See 2021 update here: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs/documentation.php

HPSS archive contains most days from Sep 1 2018 to Dec 23 2020 and forecast lead times of $\{0, 6, 12, 18, 24, 30, 36\}$ hours. We extract 6-, 12-, 18-, 24-, 30-, and 36-hour forecast profiles (columns) from locations around the globe. Specifically, for each model solution (*i.e.*, each combination of initialization time and valid time), we randomly select 4000 grid points from the global grid. We extract all predictor variables used by the RRTM that are in GFSv16 output files, listed in Table 1. We also extract two extra variables – the height thickness and pressure thickness of each layer – for use by the NNs but not the RRTM. For the work in L21, where all profiles have the same physical height grid (*i.e.*, the k^{th} pixel always corresponds to the same height in metres), the thickness variables were not necessary. But for the current work, where all profiles have a different physical height grid due to the hybrid coordinates, we found that the thickness variables improve RT estimation by the NNs. These variables are important because they tell the NNs how much "stuff" is in each layer – *i.e.*, how much air there is to heat and how many other molecules there are to interact with radiation, which cannot be determined from molecular concentrations alone.

b. Synthetic predictors

Predictors not in GFSv16 output files are listed in Table 2. We create synthetic data for these predictors, which fall into three categories: particle sizes, trace gases, and aerosols.

Particle sizes

The two relevant variables are ice effective radius ($r_{\text{eff}}^{\text{ice}}$) and liquid effective radius ($r_{\text{eff}}^{\text{liq}}$), both summaries of the particle-size distribution (PSD; Mitchell et al. 2011). To create a synthetic profile of $r_{\text{eff}}^{\text{ice}}$, we apply the following equation from Mishra et al. (2014, their Figure 6b) independently to each height in the profile:

$$r_{\text{eff}}^{\text{ice}} = 86.73\ \mu\text{m} + \left(1.07\ \frac{\mu\text{m}}{^{\circ}\text{C}}\right)T, \tag{1}$$

where T is the temperature (°C) and each height has a different temperature (Figure 1a). After Equation 1, we apply two types of noise to the profile: bulk noise, which shifts the whole profile to higher or lower values, and structure noise, which changes the structure of the profile (Figure 1b). For bulk noise, we multiply the whole $r_{\text{eff}}^{\text{ice}}$ profile by $1 + \epsilon_b$, where ϵ_b is a random variable drawn from a normal distribution with mean = 0 and standard deviation = 0.5, denoted as $\mathcal{N}(0, 0.5)$. In

Table 1: Description of GFSv16-based predictor variables. "Vector?" asks whether the variable is a profile or a scalar, and "AGL" = above ground level. Downward LWP at height z is LWC integrated from the top of the profile down to z, and upward LWP at height z is LWC integrated from the bottom of the profile up to z. The definitions of downward IWP, upward IWP, downward WVP, and upward WVP are analogous.

Variable	**Units**	**Predictor for shortwave RT?**	**Predictor for longwave RT?**	**Vector?**
Solar zenith angle	°	✓		
Surface albedo	—	✓		
Surface temperature	K		✓	
Surface emissivity	—		✓	
Temperature	K	✓	✓	✓
Pressure	Pa	✓	✓	✓
Specific humidity	kg kg^{-1}	✓	✓	✓
Relative humidity	—	✓	✓	✓
Liquid-water content (LWC)	kg m^{-3}	✓	✓	✓
Ice-water content (LWC)	kg m^{-3}	✓	✓	✓
Downward liquid-water path (LWP)	kg m^{-2}	✓	✓	✓
Downward ice-water path (IWP)	kg m^{-2}	✓	✓	✓
Downward water-vapour path (WVP)	kg m^{-2}	✓	✓	✓
Upward LWP	kg m^{-2}	✓	✓	✓
Upward IWP	kg m^{-2}	✓	✓	✓
Upward WVP	kg m^{-2}	✓	✓	✓
O_3 mixing ratio	kg kg^{-1}	✓	✓	✓
Height	m AGL	✓	✓	✓
Height thickness	m	✓	✓	✓
Pressure thickness	Pa	✓	✓	✓

other words, the standard deviation of bulk noise is 50% of the value generated by Equation 1. For structure noise, we multiply the $r_{\text{eff}}^{\text{ice}}$ value at every height by $1 + \epsilon_s$, where ϵ_s is drawn anew at every height from $\mathcal{N}(0, 0.05)$. After adding noise, we bound $r_{\text{eff}}^{\text{ice}}$ values to the range $[17.18, 65.33]$ μm, which is the same as bounding temperature to $[-65, -20]$ °C, the range of validity for Equation 1. See Figure 1c.

Table 2: Description of synthetic predictor variables.

Variable	Units	Predictor for shortwave RT?	Predictor for longwave RT?	Vector?
Aerosol single-scattering albedo	—	✓		
Aerosol asymmetry parameter	—	✓		
Aerosol extinction coefficient	m^{-1}	✓		✓
Liquid effective radius	m	✓	✓	✓
Ice effective radius	m	✓	✓	✓
N_2O concentration	ppmv	✓	✓	✓
CH_4 concentration	ppmv	✓	✓	✓
CO_2 concentration	ppmv	✓	✓	✓

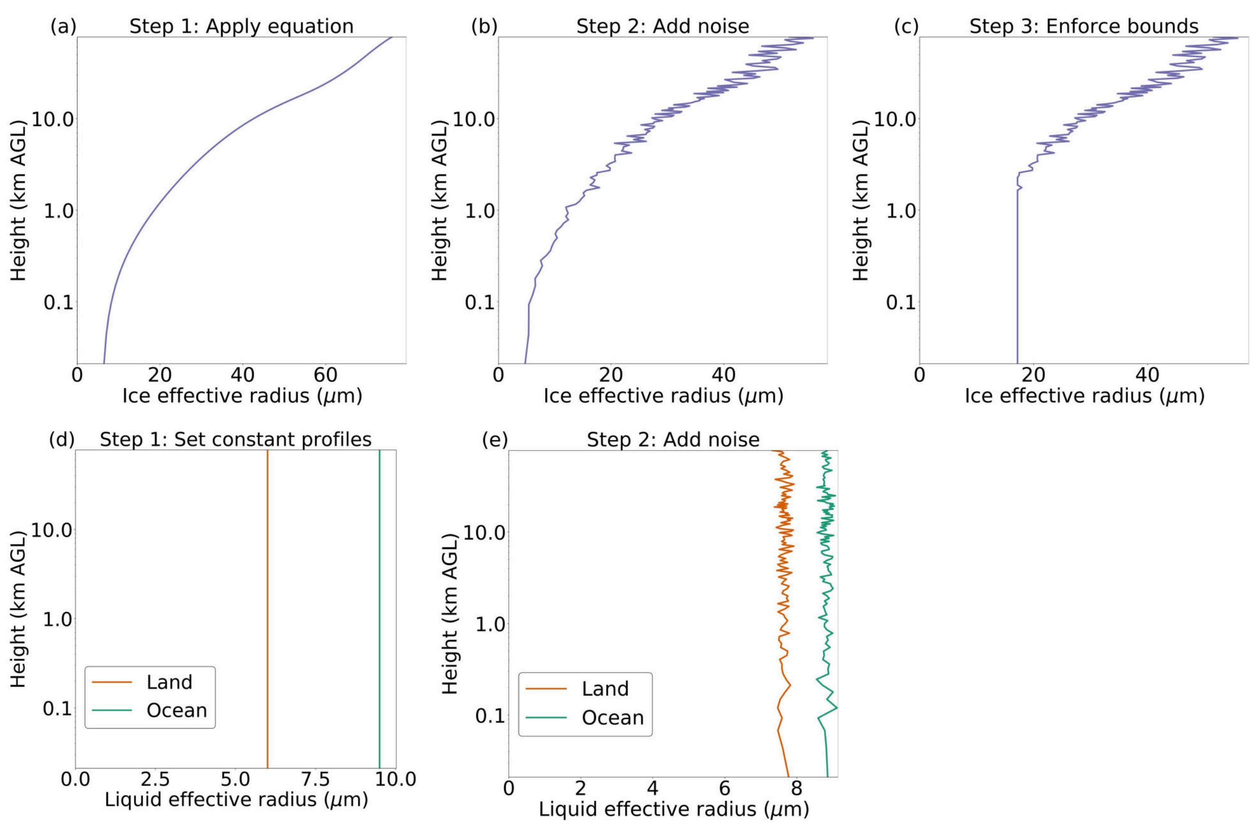

Figure 1: Procedure for creating synthetic profiles of [a-c] ice effective radius and [d-e] liquid effective radius.

Table 3: Definition of standard atmospheres. The categorization is mutually exclusive and collectively exhaustive, *i.e.*, every profile is assigned to exactly one of the five standard atmospheres.

Standard atmosphere	**Months**	**Latitudes**
Mid-latitude summer	May – Oct	$[20, 65]$ °N
Mid-latitude summer	Nov – Apr	$[20, 65]$ °S
Mid-latitude winter	Nov – Apr	$[20, 65]$ °N
Mid-latitude winter	May – Oct	$[20, 65]$ °S
Polar summer	May – Oct	$[65, 90]$ °N
Polar summer	Nov – Apr	$[65, 90]$ °S
Polar winter	Nov – Apr	$[65, 90]$ °N
Polar winter	May – Oct	$[65, 90]$ °S
Tropical	All	$[-20, 20]$ °N

To create a synthetic profile of $r_{\text{eff}}^{\text{liq}}$, we start with the distribution discovered by Miles et al. (2000). They found that $r_{\text{eff}}^{\text{liq}}$ roughly follows the distribution $\mathcal{N}(6\ \mu\text{m}, 1\ \mu\text{m})$ over land and $\mathcal{N}(9.5\ \mu\text{m}, 1.2\ \mu\text{m})$ over ocean. See Figure 1d. However, using this information alone would lead to constant $r_{\text{eff}}^{\text{liq}}$ profiles, which are unrealistic. Thus, we add structure noise to each profile, using the same method as for $r_{\text{eff}}^{\text{ice}}$. See Figure 1e.

Trace gases

For trace gases not in the GFSv16 output files – N_2O, CH_4, and CO_2 – we use canonical profiles provided by Anderson et al. (1986). There is one canonical profile for each gas and each standard atmosphere, the latter defined in Table 3. For example, the five canonical N_2O profiles are shown in Figure 2a. As for $r_{\text{eff}}^{\text{ice}}$, we add both bulk and structure noise to each profile of trace-gas concentrations. We use the same noise distributions as for $r_{\text{eff}}^{\text{ice}}$. See Figure 2b.

Note that the values provided in Anderson et al. (1986) are outdated, corresponding to a past climate. However, by adding noise we sample a wide range of atmospheric conditions, corresponding to both present-day and hypothetical future climates. For example, Supplemental Figure S3 shows that our dataset includes many CO_2 concentrations well above the present-day value of ~412 ppm.

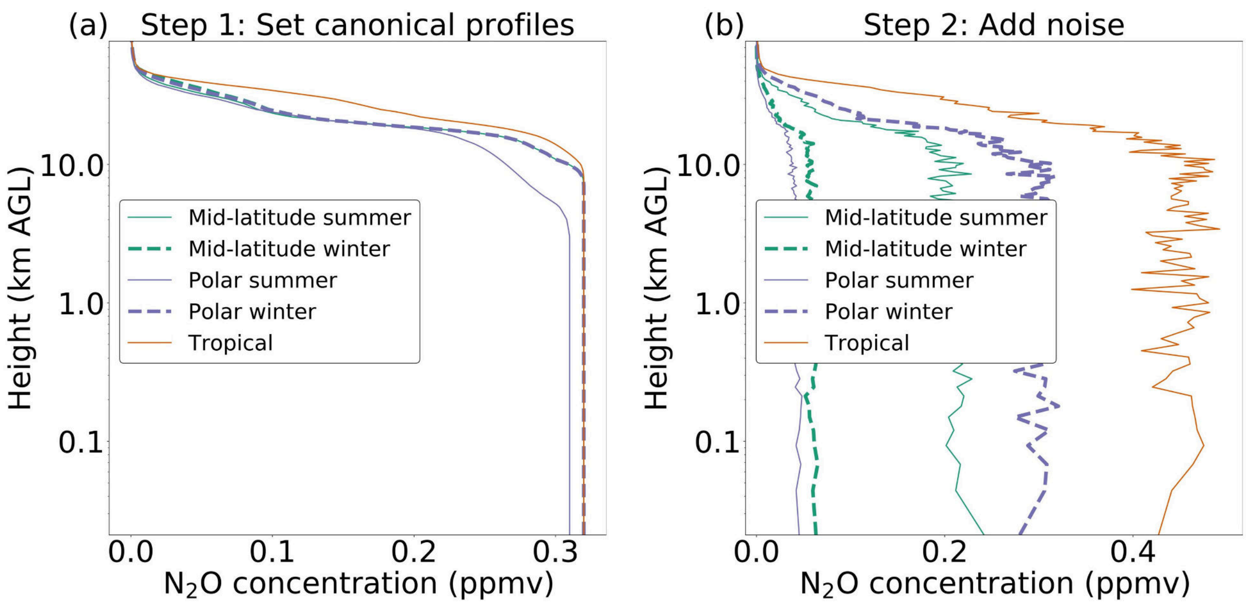

Figure 2: Procedure for creating synthetic profiles of trace-gas concentration – in this example, N_2O concentration.

Aerosols

Due to its complexity, we have relegated our method for creating synthetic aerosol variables – single-scattering albedo (SSA), asymmetry parameter, and extinction coefficient – to Supplemental Section 1.

c. Target variables

We run the shortwave and longwave RRTM separately for each profile. The target variables are those needed by an NWP model from its embedded RT model: a profile of heating rates (HR), surface downwelling flux ($F_{\text{down}}^{\text{sfc}}$), top-of-atmosphere upwelling flux ($F_{\text{up}}^{\text{TOA}}$), and net flux ($F_{\text{net}}$). All four of these variables have both a shortwave and a longwave version. In machine learning the goal is often to improve accuracy, but our goal is to improve efficiency – *i.e.*, to accelerate the RRTM – while emulating it as faithfully as possible. This means that we treat the RRTM as a perfect model, considering its HRs and fluxes to be the correct answers. Although the RRTM is imperfect, its errors are quite small, at less than 0.1 K day^{-1} for HRs and less than 1 W m^{-2} for fluxes (Iacono et al. 2008).

d. Pre-processing

We apply two types of pre-processing to the data: splitting and normalization. As in L21, we use isotonic regression (IR) to bias-correct the NNs, which requires a separate training set. Thus, we split the data into four temporally independent subsets: NN-training, IR-training, validation, and testing (Table 4). Each subset covers locations around the globe during all seasons. For normalization, we use the same methods described in Section 3b of L21, except that we do not normalize any target variables. In L21 we normalized the flux variables, but we have since found that this has a deleterious effect on the quality of NN predictions.

3. Deep-learning methods

This section provides a minimal background on the NN architectures used in L21, followed by a more extensive background on the architectures new to the current work, and finally information on the loss functions used to train NNs.

a. U-net and U-net++ without deep supervision

L21 considered two NN architectures, namely the U-net and U-net++, for shortwave RT. They found that the U-net++ outperforms the U-net in situations with multi-layer cloud (their Supplemental Section Cd), which are the most complex situations for RT and also vitally important for weather/climate prediction. In this article we consider the U-net++ architecture and a new architecture called U-net3+. L21 contains a detailed background on the U-net and U-net++ (their Section 2), and we attempt to reproduce as little of this background as possible – only that which is necessary for understanding the current article.

The U-net (Ronneberger et al. 2015) is a type of NN designed for making predictions on a spatial grid, often called "image-to-image translation" in the ML literature. U-nets are typically applied to images with two or three spatial dimensions, but in our case the "images" are vertical profiles, containing only one spatial dimension. The task is to translate a 127-by-M image of predictors (M, the number of variables, is different for longwave vs. shortwave RT) into a 127-by-1 image of HRs[7].

[7]There is a second learning task, which involves image-to-scalar translation – namely to translate the same 127-by-M image of predictors into 3 flux components.

Table 4: Partitioning of data into temporally independent subsets. “SW” = shortwave; “LW” = longwave; and “sample size” = number of profiles. SW and LW sample sizes are different because the SW radiation scheme (RRTM or NN-based emulator) is not run when the Sun is below the horizon, *i.e.*, when solar zenith angle > 90°. Also, “Number of days” ≠ length of “Time period,” because some days are missing from the archive.

Data subset	**Time period**	**Number of days**	**SW sample size**	**LW sample size**
NN-training	Sep 1 2018 – Dec 21 2019	237	873 086	3 503 226
IR-training	Dec 24-30 2019, Feb 3-9 2020, Mar 15-21 2020, Apr 26 – May 2 2020, Jun 7-13 2020, Jul 18-24 2020, Aug 28 – Sep 3 2020, Oct 10-16 2020, Nov 21-27 2020	63	213 275	939 181
Validation	Jan 2-15 2020, Feb 12-25 2020, Mar 24 – Apr 6 2020, May 5-18 2020, Jun 16-29 2020, Jul 27 – Aug 9 2020, Sep 6-19 2020, Oct 19 – Nov 2 2020, Nov 30 – Dec 13 2020	126	479 806	1 934 460
Testing	Jan 18-31 2020, Feb 28 – Mar 12 2020, Apr 9-22 2020, May 22 – Jun 4 2020, Jul 2-15 2020, Aug 12-25 2020, Sep 22 – Oct 7 2020, Nov 5-18 2020, Dec 16-23 2020	120	474 726	1 929 078

U-nets contain four key components (Figure 3a): convolutional layers, pooling (downsampling) layers, upsampling layers, and skip connections. The role of the convolutional layers is to detect spatial and multivariate features – *i.e.*, features including many pixels and predictor variables – using convolutional filters with weights optimized during training to detect the most useful features for prediction. The role of the pooling and upsampling layers is to change the resolution of the feature maps – a "feature map" being either the original or a transformed version of the predictors – so that convolutional layers at different depths in the network can detect features at different spatial scales. The role of the skip connections is to preserve high-resolution information – *i.e.*, to carry through the network high-resolution information that has not been degraded by downsampling, a lossy operation that cannot be fully reversed by upsampling. The left side of the U-shaped network (Figure 3a) is the encoder side, where the predictors are converted to feature maps with decreasing spatial resolution (fewer height levels) and increasing spectral resolution (more channels). The right side is the decoder side, where feature maps are upsampled and converted to the final prediction – an image of HRs. To make our networks also predict the three flux variables, which are scalars and not images, we attach fully connected layers to the deepest encoder layer, as shown in Figure 3a. These are the layers used in fully connected NNs (Chapter 6 of Goodfellow et al. 2016), which are still a popular choice for scalar data.

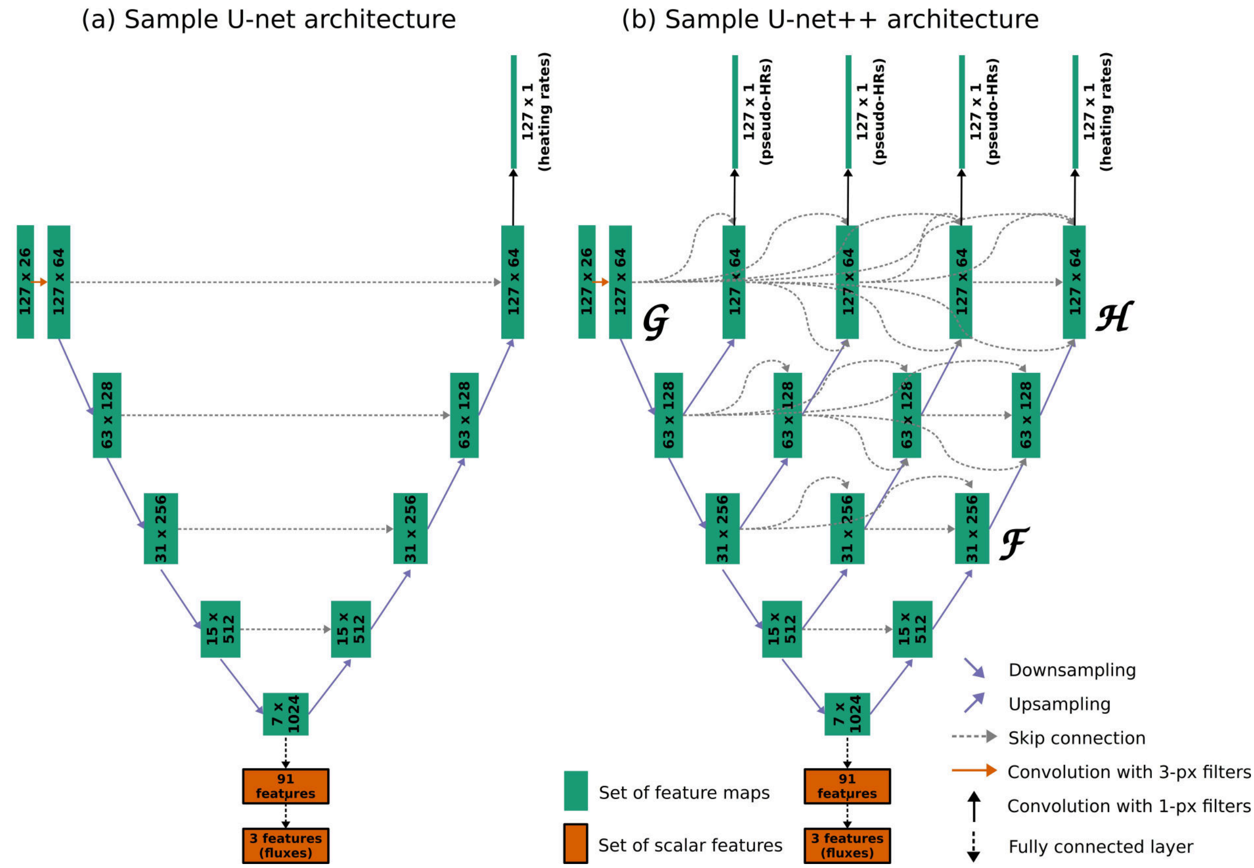

Figure 3: Sample architectures for [a] U-net and [b] U-net++. Labels $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ are referred to in the main text. Actual models used in this study differ in the number of channels and depth (number of encoder/decoder layers, *i.e.*, number of horizontal rows in this figure). For each set of feature maps (green box), the two dimensions are number of heights and channels, respectively. When the U-net++ is trained without deep supervision, all feature maps labeled "pseudo-HRs" go away, along with the arrows pointing to them. In the remaining discussion, let K be the number of convolutional layers per block, a user-chosen hyperparameter. Each orange "convolution" arrow corresponds to K convolutional layers with 3-pixel filters; each "downsampling" arrow corresponds to K convolutional layers with 3-pixel filters, followed by a maximum-pooling layer with a 2-pixel window; each "upsampling" arrow corresponds to an upsampling layer with a 2-pixel window, followed by a convolutional layer with 3-pixel filters; each "skip connection" arrow includes K convolutional layers with 3-pixel filters; each black "convolution" arrow corresponds to one convolutional layer with 1-pixel filters; and lastly, each "fully connected layer" arrow corresponds to one fully connected layer.

The U-net++ (Zhou et al. 2019) contains more skip connections than the U-net, which more effectively preserve small-scale features such as cloud boundaries, leading to better predictions for multi-layer cloud in L21. The U-net3+ (Huang et al. 2020) contains even more skip connections than the U-net++, so we hypothesize that the U-net3+ will perform even better in situations with multi-layer cloud and perhaps overall. Also, the U-net++ and U-net3+ may be trained with deep supervision, which was not used in L21.

b. U-net++ with deep supervision

When a NN is trained without deep supervision, the loss function optimized by the NN compares the ground truth (here, a length-127 profile of HRs) only to the final prediction, *i.e.*, output from the last NN layer. With deep supervision, the ground truth is also compared to intermediate representations, *i.e.*, layer outputs that are ultimately transformed to the final prediction. Zhou et al. (2019) found that deep supervision improves image segmentation for phenomena that occur at different scales, such as lung nodules. We hypothesize that deep supervision will also improve RT estimation, since relevant features for RT estimation also occur at different scales – *e.g.*, cloud depths range from $O(10\text{ m})$ to $O(10\text{ km})$.

Figure 3b shows a sample U-net++ architecture with and without deep supervision. The only difference is that deep supervision requires extra convolutional layers – those producing pseudo-HRs – to transform the intermediate representations from many channels to one channel. With deep supervision, all four outputs (the three pseudo-HR profiles and the actual-HR profile) are produced; without deep supervision, only one output (the actual-HR profile) is produced. For details on the loss function, which compares both psuedo-HRs and actual HRs to the ground truth, see Section 3d. Note that deep supervision is applied only to the spatial outputs (HRs) and not the scalar outputs (fluxes). Deep supervision was invented for spatial data, and there is no clear analogue for scalars.

c. *U-net3+ with and without deep supervision*

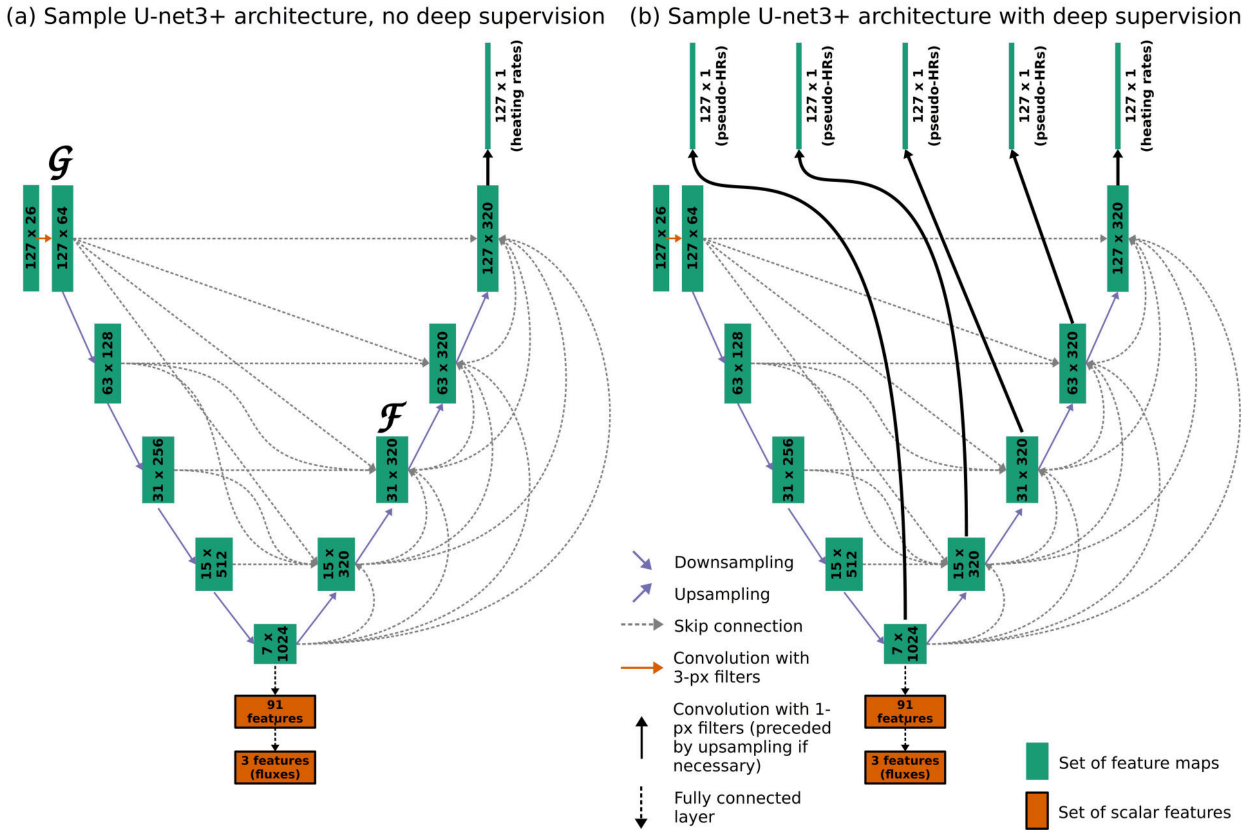

Figure 4: Sample architectures for U-net3+ [a] without and [b] with deep supervision. Labels $\mathcal{F}$ and $\mathcal{G}$ are referred to in the main text. Actual models used in this study differ in the number of channels and depth. Formatting is explained in the caption of Figure 3, except that the solid black arrows are slightly different in this figure. The solid black arrow pointing to actual HRs (top right) corresponds to one convolutional layer with 1-pixel filters, while a solid black arrow pointing to pseudo-HRs corresponds to an upsampling layer followed by a convolutional layer with 1-pixel filters.

The U-net3+ has one property that distinguishes it from the U-net++, namely full-scale skip connections. Full-scale skip connections pass information from all scales to each decoder layer, whereas skip connections in the U-net++ pass information from only two scales to each decoder layer. For example, in the U-net++ shown in Figure 3b, the feature maps labeled $\mathcal{F}$ combine information from the same scale (other feature maps with 31 heights) and the next-largest scale (feature maps with 15 heights). But in the U-net3+ shown in Figure 4a, the feature maps labeled $\mathcal{F}$ combine information from equal and smaller scales (feature maps with $\geq$ 31 heights) on the encoder side, as well as information from larger scales (feature maps with < 31 heights) on the decoder side.

Stated differently, full-scale skip connections more effectively carry high-resolution information through the network. For example, the feature maps labeled $\mathcal{G}$ (in both Figures 3b and 4a) contain information at the smallest scale that has not been degraded by downsampling. In the U-net++ (Figure 3b), skip connections carry this information to only one level on the decoder side, namely the feature maps labeled $\mathcal{H}$. Other levels on the decoder side cannot access the undegraded high-resolution information in $\mathcal{G}$. But in the U-net3+ (Figure 4a), full-scale skip connections carry the information in $\mathcal{G}$ to all levels on the decoder side, allowing this information to be used in decoded feature maps at all resolutions.

Figures 4a and 4b show how to add deep supervision to the U-net3+ architecture. For the U-net3+, deep supervision requires two architecture changes. The first is extra convolutional layers to reduce the number of channels to one (pseudo-HR), as in the U-net++. The second is extra upsampling layers to increase the number of heights to 127.

d. Loss function

In machine learning, the standard loss function for regression tasks – where the model predicts a continuous value instead of a category – is the mean squared error (MSE). However, in L21 we found that using the MSE causes two problems. First, the MSE does not adequately emphasize large HRs, which are rare but important for weather/climate prediction, causing the NN to dramatically underpredict large HRs. Second, the MSE does not ensure that the following conservation law is respected:

$$F_{\text{net}}^{(b)} = F_{\text{down}}^{\text{sfc}\ (b)} - F_{\text{up}}^{\text{TOA}\,(b)}, \tag{2}$$

where the superscript (b) denotes that all three variables must come from the same band, either shortwave or longwave. To remedy the first problem, we used the dual-weighted MSE (DWMSE) for HRs, which emphasizes cases with a large actual or predicted HR, "nudging" the NN to predict these cases correctly. See Section 3c2 of L21. To remedy the second problem, we used the basic MSE for flux variables *but* enforced the law of Equation 2 inside the NN. See Section 3c1 of L21.

Because L21 is concerned with shortwave RT only, the present work requires two updates to the loss function. First, the weight in the DWMSE becomes the maximum of the *absolute* actual and predicted HRs, because although shortwave HR is always ≥ 0, longwave HR may be negative (*i.e.*, longwave cooling). Second, the flux law must be applied to both shortwave and longwave RT. The total loss function becomes the following:

$$\mathcal{L}^{(b)} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{j=1}^{H} \max\left\{|r_{ij}^{(b)}|, |\hat{r}_{ij}^{(b)}|\right\} \left[r_{ij}^{(b)} - \hat{r}_{ij}^{(b)}\right]^2 + \frac{1}{NM} \sum_{i=1}^{N} \sum_{k=1}^{M} \left[F_{ik}^{(b)} - \hat{F}_{ik}^{(b)}\right]^2, \tag{3}$$

where N is the number of examples; $H = 127$ is the number of heights per example; $r_{ij}^{(b)}$ is the actual HR for the j^{th} height in the i^{th} example; $\hat{r}_{ij}^{(b)}$ is the corresponding prediction; $M = 3$ is the number of flux components; $F_{ik}^{(b)}$ is the actual value of the k^{th} flux component in the i^{th} example; and $\hat{F}_{ik}^{(b)}$ is the corresponding prediction. There is one version of Equation 3 for the shortwave, where the superscript (b) is SW, and one version for the longwave.

For NNs without deep supervision, Equation 3 is the whole story. However, for NNs with deep supervision, the loss function includes extra terms for the pseudo-HRs. Specifically, the loss function becomes

$$\mathcal{L}_{\text{deep-sup}}^{(b)} = \mathcal{L}^{(b)} + \frac{1}{PNH} \sum_{p=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{H} \max\left\{|r_{ij}^{(b)}|, |\hat{r}_{pij}^{(b)}|\right\} \left[r_{ij}^{(b)} - \hat{r}_{pij}^{(b)}\right]^2, \tag{4}$$

where P is the number of layers with deep supervision and thus the number of pseudo-HR profiles, and $\hat{r}_{pij}^{(b)}$ is the pseudo-HR produced by the p^{th} layer with deep supervision for the j^{th} height in the i^{th} example.

4. Experiment with neural networks of varying complexity

This section describes a hyperparameter-tuning experiment used to find the optimal level of NN complexity for estimating RT. We tune four hyperparameters: the NN type (U-net++ or U-net3+ with or without deep supervision), NN depth, NN width, and spectral complexity. NN depth is the number of encoder/decoder levels (*e.g.,* all architectures shown in Figures 3-4 have a depth of 4); NN width is the number of convolutional layers per set (K in the caption of Figure 3); and spectral complexity is the number of feature maps produced by the first set of convolutional layers (*e.g.,* all architectures shown in Figures 3-4 have a spectral complexity of 64). Following common practice, we always double the number of feature maps with each downsampling operation. For example,

Table 5: Experimental hyperparameters.

Hyperparameter	Values attempted
NN type	U-net++ without deep supervision, U-net++ with deep supervision, U-net3+ without deep supervision, U-net3+ with deep supervision,
NN depth	3, 4, 5
NN width	1, 2, 3, 4
Spectral complexity	4, 8, 16, 32, 64, 128

Figure 3 shows that with a depth of 4 and spectral complexity of 64, the deepest set of feature maps (*i.e.,* that with the coarsest spatial resolution, designed to capture the largest-scale features) has 1024 feature maps. We chose to experiment with NN type so that we could try new methods (deep supervision and U-net3+) from the ML literature. We chose to experiment with the other three hyperparameters because they strongly control overall NN complexity, *i.e.*, the number of trainable weights. As shown in Supplemental Figures S10 and S18, the number of trainable weights varies from $O(10^5)$ to $O(10^{8.5})$.

Table 5 lists the exact values attempted for each hyperparameter. We perform a grid search (Section 11.4.3 of Goodfellow et al. 2016), training one NN for every combination of values, which leads to $4 \times 3 \times 4 \times 6 = 288$ NNs for each band (shortwave and longwave). Most constant hyperparameters (those not varied during the experiment) are illustrated in Figures 3 and 4. Constants not included in these figures are documented in Supplemental Table S3.

a. Evaluation methods used for model selection

Model evaluation is a multi-faceted problem, and there are many possible ways to choose the best model. Most hyperparameter experiments optimize one evaluation metric, often the loss function used for training. However, we care about several aspects of model performance. In previous work we have noticed that even when overall performance is acceptable, the following regime-based errors are unacceptably high:

- HR errors near the surface, especially in the longwave;

Table 6: Metrics used for model selection. “Column-averaged” = averaged over all 127 heights; “near-surface” = at the lowest grid level, which averages 21 m AGL; and “all-flux RMSE” is the square root of the MSE averaged over all three flux variables. Metrics computed on fog profiles are used only to evaluate longwave models, not shortwave models.

Set of profiles	Metrics used
All	Column-averaged HR DWMSE, column-averaged HR bias, near-surface HR DWMSE, near-surface HR bias, all-flux RMSE, net-flux RMSE, net-flux bias
Profiles with multi-layer liquid-only cloud	Column-averaged HR DWMSE, column-averaged HR bias, near-surface HR DWMSE, near-surface HR bias, all-flux RMSE, net-flux RMSE, net-flux bias
Profiles with fog (longwave only)	Near-surface HR DWMSE, near-surface HR bias, all-flux RMSE, net-flux RMSE, net-flux bias

- flux and HR errors in profiles with multi-layer liquid-only cloud, in both the shortwave and longwave;

- longwave HR errors near the surface in profiles with fog, *i.e.*, cloud reaching the lowest grid level.

Thus, we use the metrics listed in Table 6, computed on validation data only, for model selection. Our choice of the best model is based on a subjective combination of these metrics.

b. Evaluation methods used for best models

As in L21, we evaluate the best models (shortwave and longwave) on the testing dataset as a whole and on meaningful subsets of the testing data. We split the testing data in four ways.

First, we split by cloud regime, because clouds add immense complexity to RT, making the process difficult to emulate, and can result in extreme HRs (large absolute values in both the shortwave and longwave), which are important for weather and climate. For a more detailed explanation of these effects, see Section 5a of L21. We focus on liquid-only cloud, which we have found to have a greater effect on RT than ice-only, mixed-phase, or any-phase cloud. We define a liquid-only cloud layer as a contiguous set of model heights with liquid-water content (LWC) > 0 g m^{-3}, total liquid-water path $\geq$ 25 g m^{-2}, and total ice-water path = 0 g m^{-2}. As in L21, we define three cloud regimes, which are mutually exclusive and collectively exhaustive (MECE): no cloud, single-layer cloud, and multi-layer cloud. For the longwave we add a fourth cloud regime – fog – defined as a cloud reaching the surface (*i.e.*, LWC > 0 g m^{-3} at the lowest model height). Thus, cloud regimes for the longwave are not MECE, as every profile with fog is also a profile with single- or multi-layer cloud. We include fog because it causes large longwave errors near the surface.

Second, we split the testing data by geographic location, specifically on a global latitude-longitude grid with 5° spacing. This spacing highlights large RT errors due to features such as high terrain and persistent stratocumulus cloud. Third, for the shortwave model only, we split the testing data by aerosol optical depth (AOD) and solar zenith angle (SZA). In earlier work we found that shortwave errors increase with higher AOD, which adds complexity to RT, and lower SZA[8], which increases HRs and the frequency of extreme HRs. Fourth, for the longwave model only, we split the testing data by near-surface thermodynamics, specifically temperature lapse rate (Γ_T^{sfc}) and humidity lapse rate (Γ_q^{sfc}). These are defined as

$$\begin{cases} \Gamma_T^{\text{sfc}} & = \frac{T_1 - T_2}{z_2 - z_1}, \\ \Gamma_q^{\text{sfc}} & = \frac{q_1 - q_2}{z_2 - z_1}, \end{cases} \tag{5}$$

where T_1 and T_2 are temperature (K) at the lowest and second-lowest model heights (sigma levels), respectively; q_1 and q_2 are specific humidity (kg kg^{-1}) at the same heights; and z_1 and z_2 are the

[8]Lower SZA means that the Sun is higher above the horizon. Specifically, SZA is 0° when the Sun is directly overhead, and 90° when the Sun is on the horizon.

corresponding physical heights (m AGL). Longwave RT near the surface is highly sensitive to the near-surface temperature and moisture profiles (Schmetz 1989). We also experimented with splitting by surface temperature and humidity, instead of their near-surface lapse rates, but found that lapse rates have a greater impact on longwave-RT errors.

We use several evaluation metrics and plotting tools, most of which are familiar to atmospheric scientists, such as the mean absolute error and bias (mean signed error). We also use the attributes diagram, which is a reliability curve with added reference lines (Hsu and Murphy 1986). However, we have adapted this plot for regression (predicting a continuous value, like flux in W m^{-2}) instead of their typical use, which is binary classification (predicting the probability of an event). For readers interested in the details, see Section 5a of L21. You can interpret the regression- and classification-based version of the attributes diagram in roughly the same way: the curve should be close to the diagonal reference line, indicating perfect reliability, and inside the shaded area, indicating a positive skill score. For the regression-based attributes diagram, this is the MSE skill score. A positive MSE skill score means that the NN model has a better MSE than the climatological model. The climatological model is a simple model that always predicts the climatological mean, estimated as the average in the training data. For example, if the mean net flux in the training data is 100 W m^{-2}, the climatological model will predict a net flux of 100 W m^{-2} for every case.

5. Results and discussion

We start with a brief discussion of the hyperparameter experiment (used to determine the best models), followed by a comparison of computing time between the RRTM and our NN-based emulators, then an in-depth discussion of the best shortwave model and best longwave model.

a. Hyperparameter experiment

Results are discussed briefly here and at length in Supplemental Section 3. For both shortwave and longwave RT, the most important hyperparameter is spectral complexity, while NN depth and width are of secondary importance. The better NNs have large spectral complexity, large depth, and small width. In other words, the better NNs are deep and narrow with many feature maps. For the other hyperparameter – NN type – we hypothesized that the U-net3+ architecture would outperform U-net++ (Section 3a) and that NNs trained with deep supervision would outperform

Table 7: Timing tests for the RRTM and NN-based emulators, based on the testing dataset. All computing times are given in wall-clock time. Because the RRTM is slower for cloudy profiles and faster for cloud-free profiles, the "Time per profile" reported is an average over all atmospheric conditions represented in the dataset. Meanwhile, the NNs have constant computing time for each profile, regardless of atmospheric conditions.

Model **Model**	**Number of profiles**	**Total time (seconds)**	**Time per profile (seconds)**
Shortwave RRTM	472 412	4 207 793	0.11
Shortwave NN	474 726	563	843
Longwave RRTM	1 894 239	369 363	5.13
Longwave NN	1 929 078	4194	460

those with no deep supervision (Section 3b). We are unable to confirm either hypothesis – deep supervision leads to *worse* performance, and architecture has little effect on performance. The best shortwave model – based on our subjective assessment of the metrics listed in Table 6 – is a U-net++ with no deep supervision, depth of 3, width of 1, and spectral complexity of 128, leading to $10^{7.52}$ trainable weights. The best longwave model – again based on Table 6 – is a U-net3+ with no deep supervision, depth of 5, width of 1, and spectral complexity of 64, leading to $10^{7.28}$ trainable weights. Therefore, the best models are on the high end of the overall-complexity range in our experiment, with number of weights ranging from $O(10^5)$ to $O(10^{8.5})$. This is because spectral complexity is the main control on both performance (allowing the models to represent and leverage many features of the input data) and number of weights (see Supplemental Figures S10 and S18).

b. Computing time

The original motivation for NNs was to decrease computing time. To this point, we have compared the wall-clock time of the RRTM and best NNs when run on the same hardware – *i.e.*, one node with 24 CPUs and no GPUs – in stand-alone mode. See Table 7 for details. In summary, the shortwave RRTM (NN) processes 0.11 (843) profiles per second, resulting in a speedup factor of 7510; while the longwave RRTM (NN) processes 5.13 (460) profiles per second, resulting in a speedup factor of 90. Thus, we have accelerated the RRTM by orders of magnitude.

c. Best shortwave model

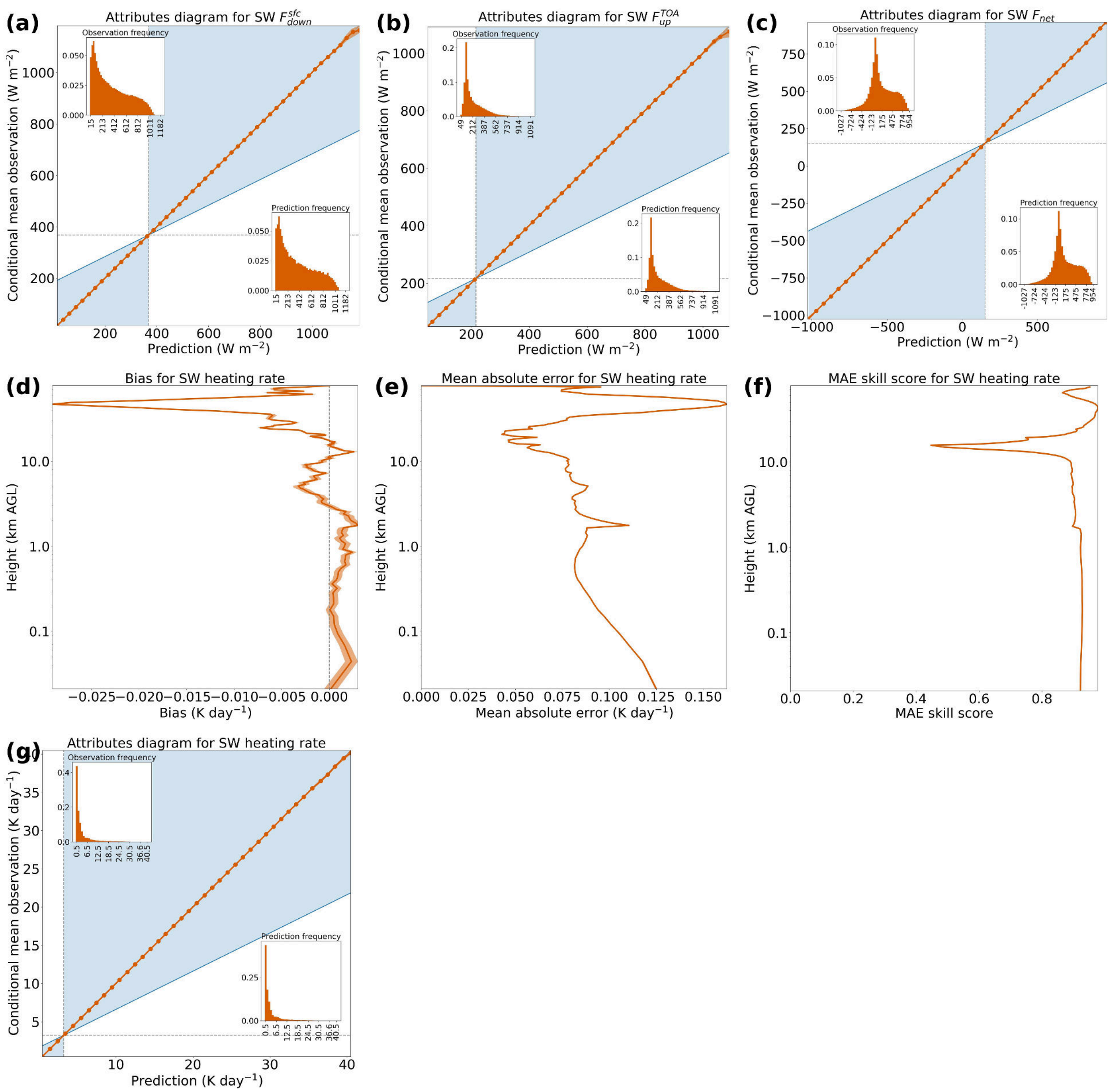

Figure 5: Performance of best shortwave model on testing data. [a-c] Attributes diagram for each flux variable. The orange curve is the reliability curve; the diagonal grey line is the perfect-reliability line; the vertical grey line is the climatology line; the horizontal grey line is the no-resolution line; the blue shading is the positive-skill area, where MSE skill score > 0; and the inset histograms show the distributions of predicted and observed values. [d-f] Profiles of bias, MAE, and MAE skill score for HR. [g] Attributes diagram for HR, including all heights. In all panels, the orange line represents the mean and the lighter shading around it represents the 99% confidence interval, both estimated from a bootstrapping test with 1000 replicates. However, in some panels the 99% confidence interval is narrower than the line representing the mean and is therefore invisible.

Figure 5 shows the overall performance – *i.e.*, averaged over the whole testing set – of the best shortwave model. For all flux variables (Figures 5a-c), the model is almost perfectly reliable (see overlap between reliability curve and diagonal reference line) and almost perfectly reproduces the observed distribution (see similarity between the two histograms). However, the model has slight conditional biases, namely an overprediction of ~10 W m^{-2} for the highest $F_{\mathrm{down}}^{\mathrm{sfc}}$ and $F_{\mathrm{up}}^{\mathrm{TOA}}$ predictions. In other words, when the model predicts an extremely large downwelling or upwelling flux, the prediction is slightly too extreme. However, these two biases offset in the calculation of F_{net} (Equation 2), resulting in near-zero bias for all predicted F_{net} values. The model has an absolute bias < 0.1 K day^{-1} for HR at every height (Figure 5d); this suggests that it could be stably integrated into an NWP system such as the GFS (Iacono et al. 2008), as systematic errors for an RT parameterization are much more important than random errors (Pincus et al. 2003). The model has a substantially larger MAE than bias for HR at every height (Figures 5d-e), which indicates that most of the model's HR error is random instead of systematic. Both bias and MAE are largest in the upper stratosphere, where shortwave RT is dominated by O_3 absorption. The bias and MAE profiles in L21 were similar – even with a dataset that used a constant profile for trace gases such as O_3 – which suggests that O_3 absorption is a fundamentally difficult process to emulate. Since the average HR in the upper stratosphere is large (*e.g.*, 21.6 K day^{-1} at 47 km AGL), the climatological model also has a large MAE here, so the NN's spike in MAE translates to only a small dip in its MAE skill score (Figure 5f). Lastly, the attributes diagram for HR (Figure 5g) tells a similar story to those for the flux variables: the model is almost perfectly reliable and almost perfectly reproduces the observed distribution. However, the model has a slight positive bias ($\ll$ 1 K day^{-1}) for the highest predicted HR values.

Supplemental Figures S22-S23 are analogous to Figure 5 but only for extreme cases – *i.e.*, the 3% of testing profiles with the greatest height-maximum and height-averaged HR, respectively. Although errors are expectedly higher for the extreme cases, HR and flux predictions are still almost perfectly reliable and absolute HR bias is well below 0.1 K day^{-1} at almost every height.

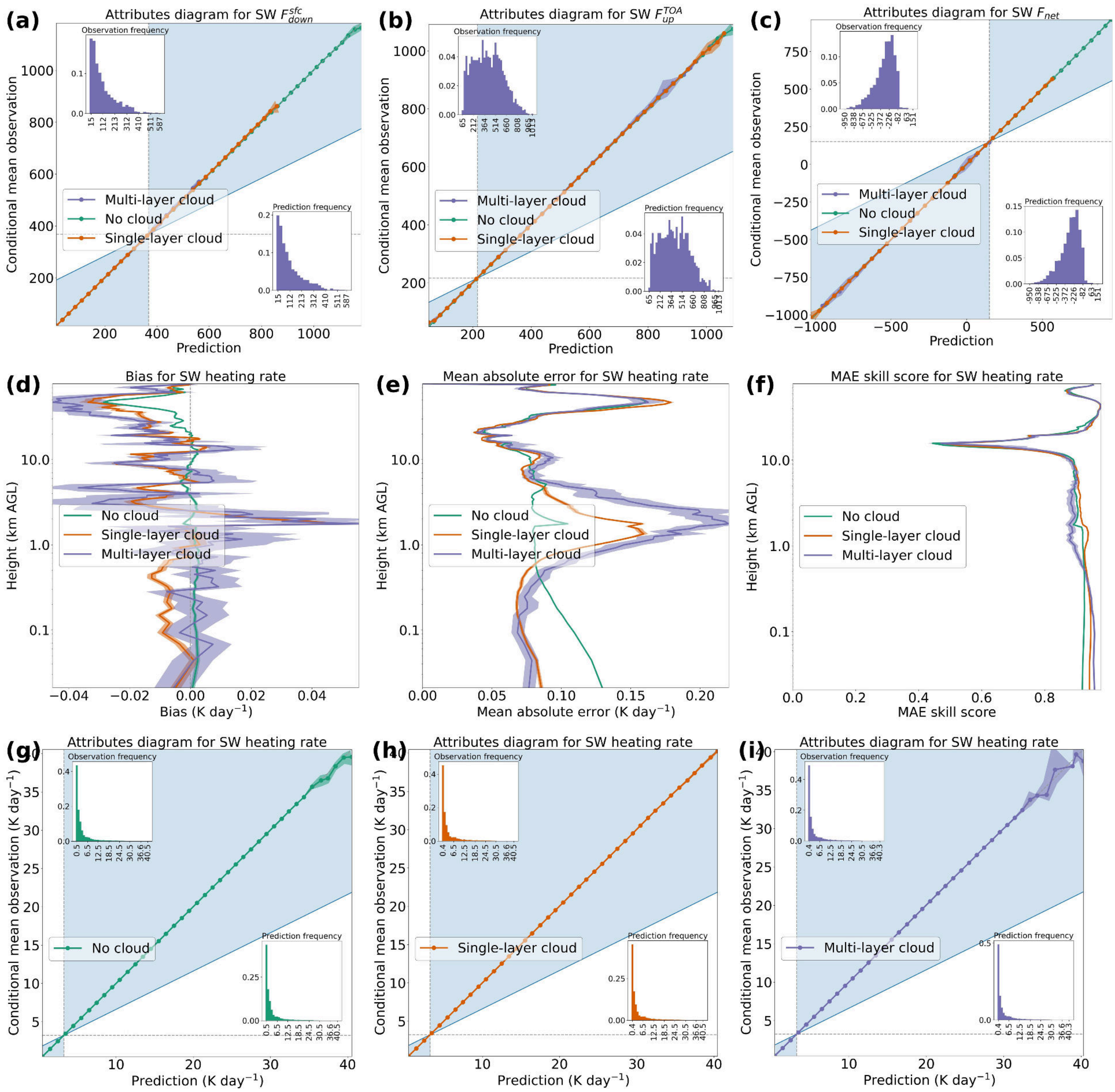

Figure 6: Performance of best shortwave model on testing data, separated by liquid-only cloud regime. [a-c] Attributes diagram (formatting explained in the caption of Figure 5) for each flux variable. The inset histograms are based only on cases with multi-layer cloud. [d-f] Profiles of bias, MAE, and MAE skill score for HR. [g] Attributes diagram for HR, including all heights, only for cases with no cloud (89.67% of the testing data). [h] Same but for single-layer cloud (9.98% of the testing data). [i] Same but for multi-layer cloud (0.35% of the testing data). In all panels, the green/orange/purple line represents the mean and the lighter shading around it represents the 99% confidence interval, both estimated from a bootstrapping test with 1000 replicates.

Figure 6 shows the model's performance as a function of liquid-only cloud regime. Performance for other cloud phases (ice-only, mixed-phase, and any-phase) is shown in Supplemental Figures

S19-S21. The attributes diagram for each flux variable (Figures 6a-c) tells a similar story to its cloud-agnostic analogue (Figures 5a-c): slight conditional bias for extreme predictions of $F_{\text{down}}^{\text{sfc}}$ and $F_{\text{up}}^{\text{TOA}}$ but with no absolute bias exceeding 20 W m^{-2}. The following discussion of error profiles for HR (Figures 6d-f) focuses on the troposphere (below ∼15 km AGL), where shortwave heating is dominated by cloud rather than O_3. In the bottom few 100 m, errors are largest for clear-sky profiles and smallest for cloudy profiles, because in cloudy profiles most of the incoming solar radiation has already been absorbed by clouds above, which leaves little shortwave radiation in the bottom few 100 m, thus making shortwave RT an easier problem here. Meanwhile, in the troposphere above ∼1 km, errors are smallest for clear-sky profiles and largest for cloudy profiles, because this is the region where most clouds and their associated extreme HRs occur. Also, errors for multi-layer cloud are greater than for single-layer cloud, because multi-layer cloud produces non-local effects that are difficult to emulate. For example, consider a profile with two clouds of equal thickness and structure (*i.e.*, equal series of LWC values), one based at 10 km AGL and the other based at 1 km AGL. The upper cloud will absorb most of the incoming solar radiation, leaving little shortwave radiation to be absorbed by the lower cloud; thus, the upper cloud will cause much larger HRs, even though the two clouds are identical except for location. This is a non-local effect, as the two clouds are far (more than a few grid cells) apart. Lastly, the attributes diagrams for HR (Figures 6g-i) tell a similar story to their cloud-agnostic analogue (Figure 5g): an overall positive bias for the highest predicted HR values and near-zero bias for all other values. However, this positive bias is largest for multi-layer cloud (up to ∼2 K day^{-1}) – likely due to a small sample size for the highest predicted HR values, indicated by the wide confidence intervals in Figure 6i.

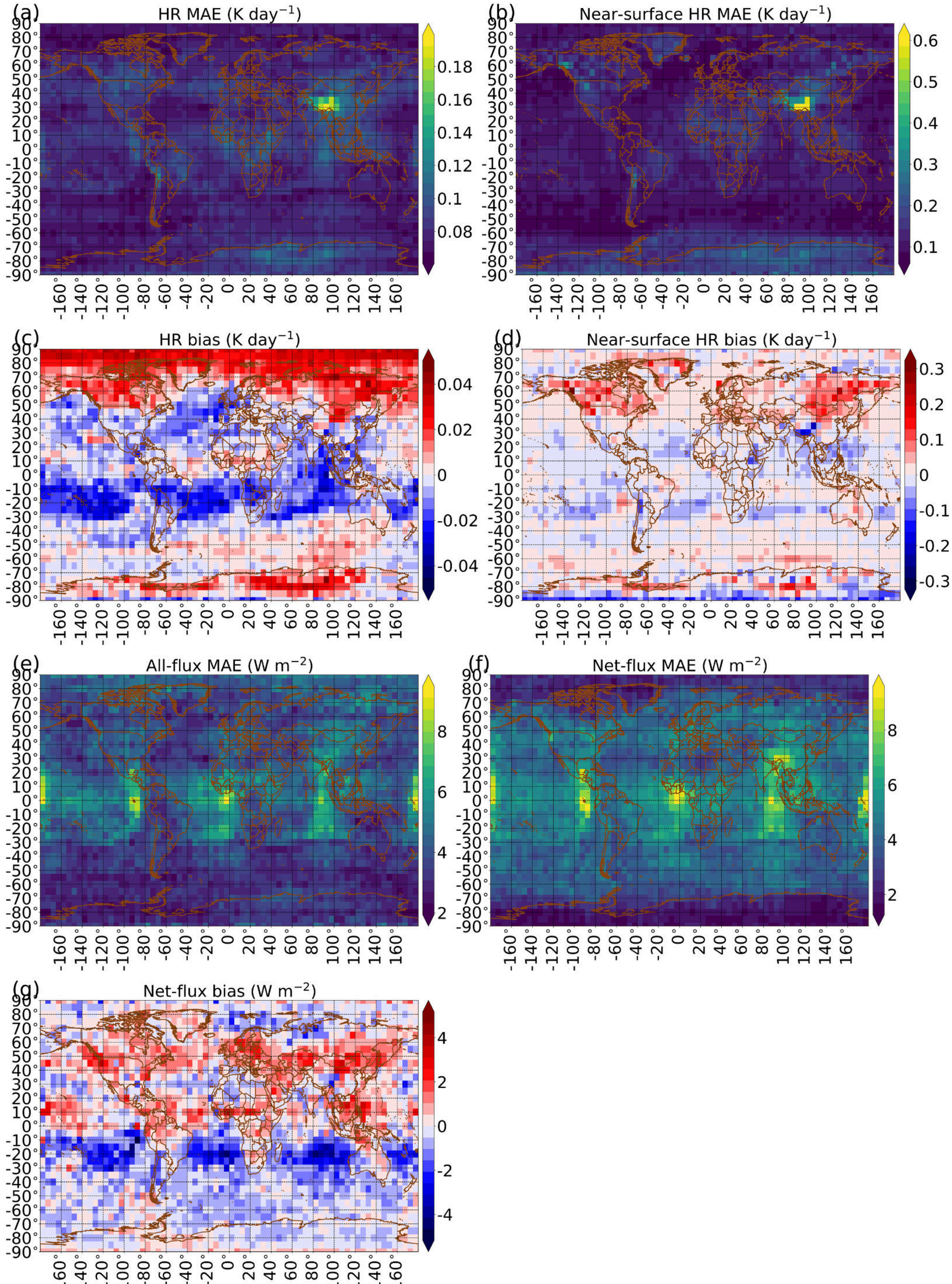

Figure 7: Performance of best shortwave model on testing data, binned by geographic location on a 5°-by-5° grid. [a] Column-averaged MAE for HR. [b] MAE for near-surface HR. [c] Column-averaged bias for HR. [d] Bias for near-surface HR. [e] All-flux MAE, averaged over the three flux variables. [f] MAE for net flux only. [g] Bias for net flux only.

Figure 7 shows the model's performance as a function of location. The column-averaged MAE for HR (Figure 7a) is mostly between[9] 0.07 and 0.11 K day^{-1}; it exceeds 0.11 K day^{-1} at a few locations, notably the Tibetan Plateau and east Antarctica. The MAE for near-surface HR (Figure 7b) is larger – mostly between 0.07 and 0.23 K day^{-1}, exceeding 0.23 K day^{-1} at a few locations, again notably Tibet and east Antarctica. The two locations have very high surface elevation and albedo, the latter due to snow/ice cover. High elevation decreases atmospheric thickness and therefore increases near-surface HR; high albedo decreases near-surface HR; and both extremes are globally rare, causing high model error under these extremes. Many error metrics (panels a-b, d, f) are especially large over the Tibetan Plateau, as it is the largest and highest plateau in the world, thus exacerbating both the thickness and albedo effects. The column-averaged bias for HR (Figure 7c) is mostly between -0.02 and +0.03 K day^{-1}, with absolute bias not exceeding 0.05 K day^{-1} at any location. The bias for near-surface HR (Figure 7d) is larger – mostly between -0.09 and +0.09 K day^{-1}, with absolute value exceeding 0.09 K day^{-1} over high-latitude continents such as Canada, Siberia, and Antarctica. The all-flux MAE (Figure 7e) is mostly between 2.5 and 6.4 W m^{-2}, exceeding 6.4 W m^{-2} mainly in the southern-hemisphere stratocumulus regions. These are regions of semi-persistent stratocumulus cloud in the subtropics off the west coast of a continent – including South America, southern Africa, and Australia (Figure 6 of Neubauer et al. 2014). The net-flux MAE (Figure 7f) follows a similar pattern to the all-flux MAE. Lastly, the net-flux bias (Figure 7g) is mostly between -2.2 and +2.0 W m^{-2}, with mostly negative bias in the southern hemisphere and positive bias in the northern hemisphere.

Supplemental Figure S24 is analogous to Figure 7 but shows relative, instead of raw, errors. For example, "relative net-flux MAE" at grid point P is $\frac{\text{raw net-flux MAE at } P}{\text{mean observed net flux at } P}$. We make two observations from the two figures. First, for column-averaged HR MAE (panel a), the highest relative errors are collocated with the highest raw errors – in Tibet and east Antarctica. This indicates that shortwave HR is *fundamentally* harder to predict at said locations – *i.e.*, these maxima in HR error are not just caused by maxima in HR itself. Second, for all other error metrics (panels b-g), the largest relative errors occur at polar latitudes, where raw errors are small. Polar latitudes receive little solar radiation, leading to small shortwave HRs and fluxes, so a small raw error translates to a large relative error. Supplemental Figure S25 is another variant of Figure 7,

[9]Henceforth, "mostly between" corresponds to the middle 95% of the distribution, *i.e.*, the 2.5th to 97.5th percentiles. However, note that the colour bar in each panel shows 100% of the distribution, ranging from the minimum to the maximum.

but showing errors for individual flux variables instead of averaging to produce all-flux quantities. The main conclusion from this figure is that $F_{\mathrm{down}}^{\mathrm{sfc}}$ errors are worst at the low latitudes, including in the stratocumulus-cloud regions, while $F_{\mathrm{up}}^{\mathrm{TOA}}$ errors are worst at the high latitudes.

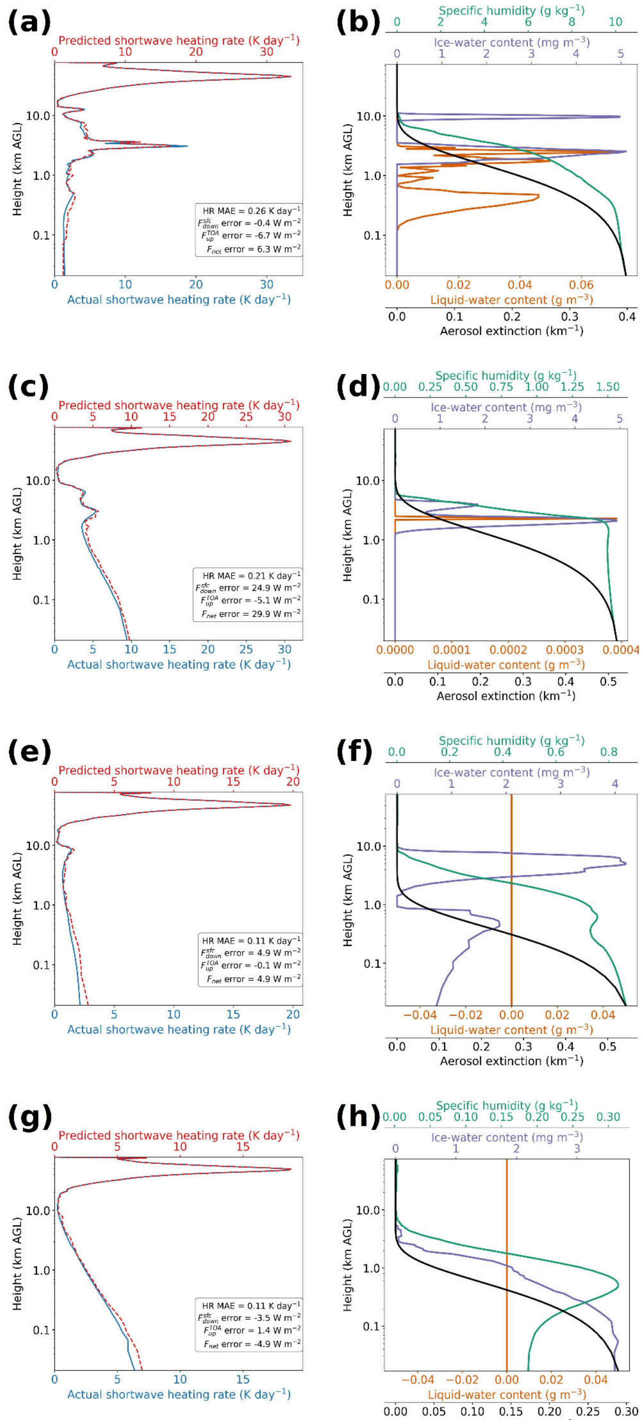

Figure 8: Geography-based case studies for the best shortwave model. [a-b] Case study from Tibet, with AOD of 0.61 and SZA of 16.6°; [c-d] another case study from Tibet, with AOD of 0.72 and SZA of 11.2°; [e-f] case study from east Antarctica, with AOD of 0.23 and SZA of 67.7°; [g-h] another case study from east Antarctica, with AOD of 0.17 and SZA of 70.7°. For each case study, the left panel shows actual and predicted RT solutions, while the right panel shows four of the most important predictor variables for shortwave RT. In each left panel, the legend shows column-averaged MAE for HR (labeled "HR MAE") and errors for the three flux variables (predicted minus actual). AOD is a summary of an important predictor variable (the height-integrated aerosol extinction), while SZA is an important predictor variable itself. These scalars are thus reported in the caption for each panel.

Figure 8 shows case studies from two regions with high model error: Tibet (panels a-d) and east Antarctica (panels e-h). To select these case studies, we first plotted 400 random profiles – 200 from each region – and then manually selected 4 profiles that are representative of the original 400. In the following conclusions, although we reference Figure 8, we have ensured that they represent most of the original 400 profiles as well. First, Tibet experiences a lot of cloud, often complex mixtures of liquid and ice. Second, east Antarctica also experiences a lot of cloud, often ice cloud reaching the surface as fog. Third, although the model matches the shape of the HR profile well, it often misses extreme HRs associated with cloud by > 1 K day^{-1}. Sometimes the model underestimates HR maxima (*e.g.*, ~3 km in panel a, ~6 km in panel c), and sometimes it overestimates (*e.g.*, ~7 km in panel a, ~3 km in panel c, ~8 km in panel e). Fourth, panels e and g are manifestations of the model's positive near-surface HR bias in east Antarctica (Figure 7d).

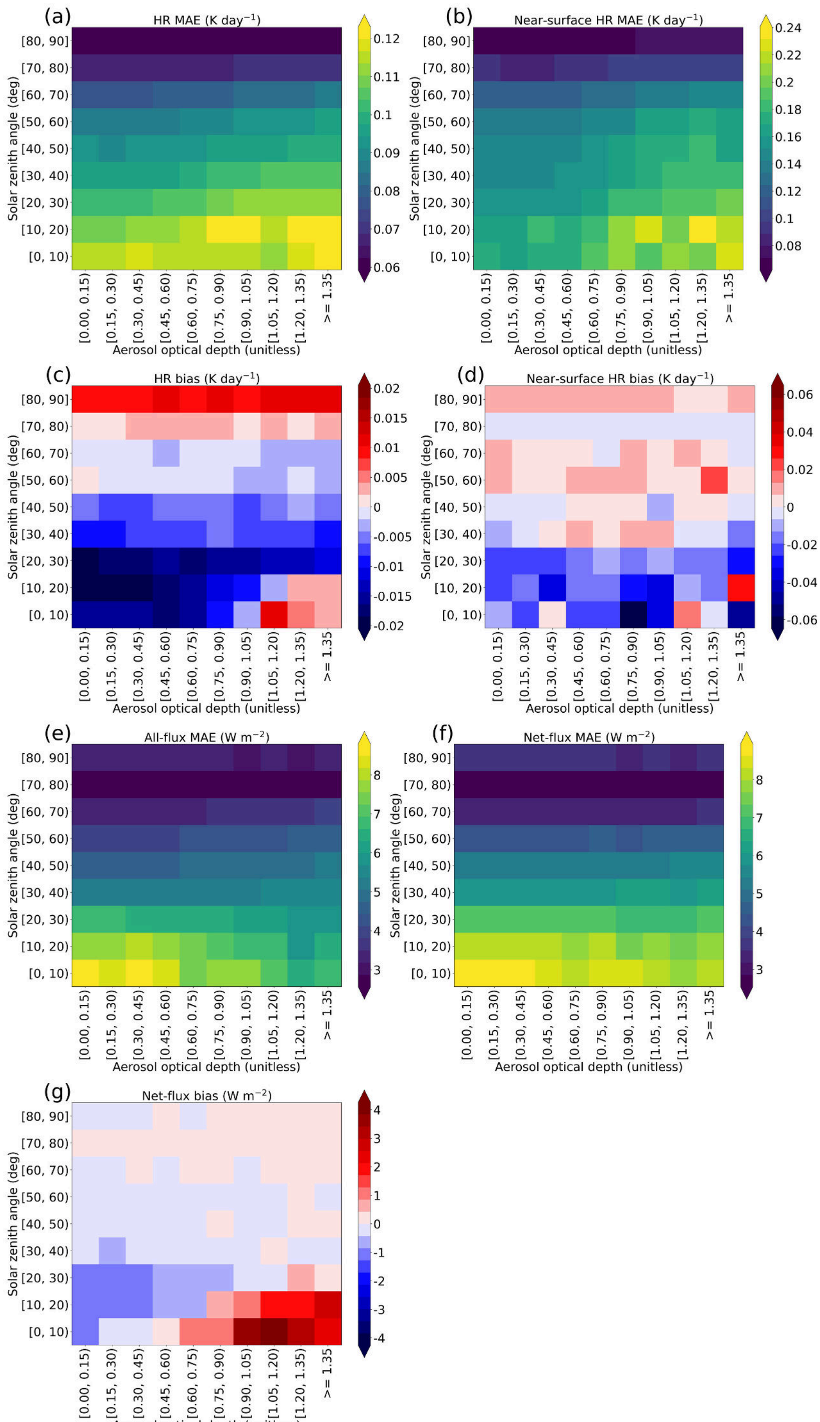

Figure 9: Performance of best shortwave model on testing data, binned by AOD and SZA, with AOD bins of width 0.15 and SZA bins of width 10°. [a] Column-averaged MAE for HR. [b] MAE for near-surface HR. [c] Column-averaged bias for HR. [d] Bias for near-surface HR. [e] All-flux MAE, averaged over the three flux variables. [f] MAE for net flux only. [g] Bias for net flux only.

Figure 9 shows the model's performance as a function of SZA and AOD. Supplemental Figure S26 is analogous but shows relative, instead of raw, errors. We make three observations from the two figures. First, for all error metrics except net-flux bias (panels a-f), raw error decreases strongly with SZA and increases weakly with AOD. In other words, raw errors are worst when there is a lot of incoming solar radiation and a lot of interaction with aerosols. Second, for the same error metrics, relative error increases strongly with SZA (the opposite relationship to raw error) and has no apparent relationship with AOD. Thus, higher solar radiation and aerosol content do not make shortwave RT *fundamentally* harder to predict; raw errors increase because the actual values (HRs and fluxes) increase. Third, for net-flux bias (panel g), when SZA $< 20°$, both raw and relative error increase with decreasing SZA and increasing AOD. In other words, when SZA $< 20°$, higher solar radiation and aerosol content make it fundamentally harder to predict net flux without bias. Supplemental Figure S27 – with errors for individual flux variables rather than all-flux errors – shows that this last relationship is driven primarily by biases in $F_{\mathrm{down}}^{\mathrm{sfc}}$, which are larger than biases in $F_{\mathrm{up}}^{\mathrm{TOA}}$.

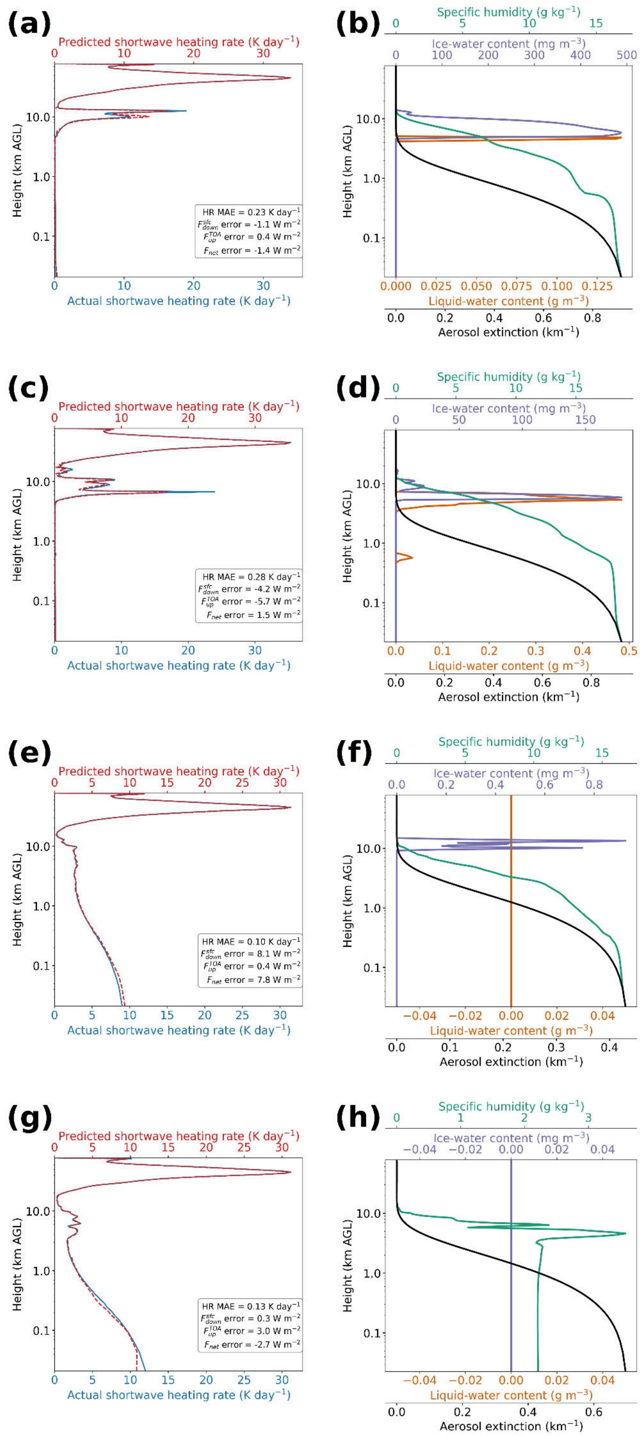

Figure 10: Regime-based case studies for the best shortwave model, specifically from the low-SZA/high-AOD regime, defined as SZA ≤ 20° and AOD ≥ 0.75. Formatting is explained in the caption of Figure 8. For the case in panels [a-b], AOD = 0.85 and SZA = 10.7°; for the case in panels [c-d], AOD = 0.81 and SZA = 7.6°; for the case in panels [e-f], AOD = 0.76 and SZA = 7.9°; for the case in panels [g-h], AOD = 1.44 and SZA = 5.5°.

Figure 10 shows case studies from the low-SZA/high-AOD regime (defined as SZA $\leq 20°$ and AOD ≥ 0.75), where raw errors are highest. The following observations aim to represent 200 random profiles, a superset of the four shown in Figure 10. First, many low-SZA/high-AOD cases feature ice cloud near the tropopause, including the first three in Figure 10. This is a known climatological feature of the tropics (Jensen et al. 2013), where the vast majority of low-SZA/high-AOD cases occur. Second, low-SZA/high-AOD cases without liquid cloud (Figures 10e-h) feature large HRs in the bottom ~1 km of the atmosphere, where the model sometimes overestimates (Figure 10e) but generally underestimates (Figure 10g) – consistent with the bottom grid row in Figure 9d. Third, the model generally overestimates net flux for these cases (by a large amount in Figure 10e). This is due mainly to overestimating $F_{\text{down}}^{\text{sfc}}$ in the low-SZA/high-AOD regime (Supplemental Figure S27).

d. Best longwave model

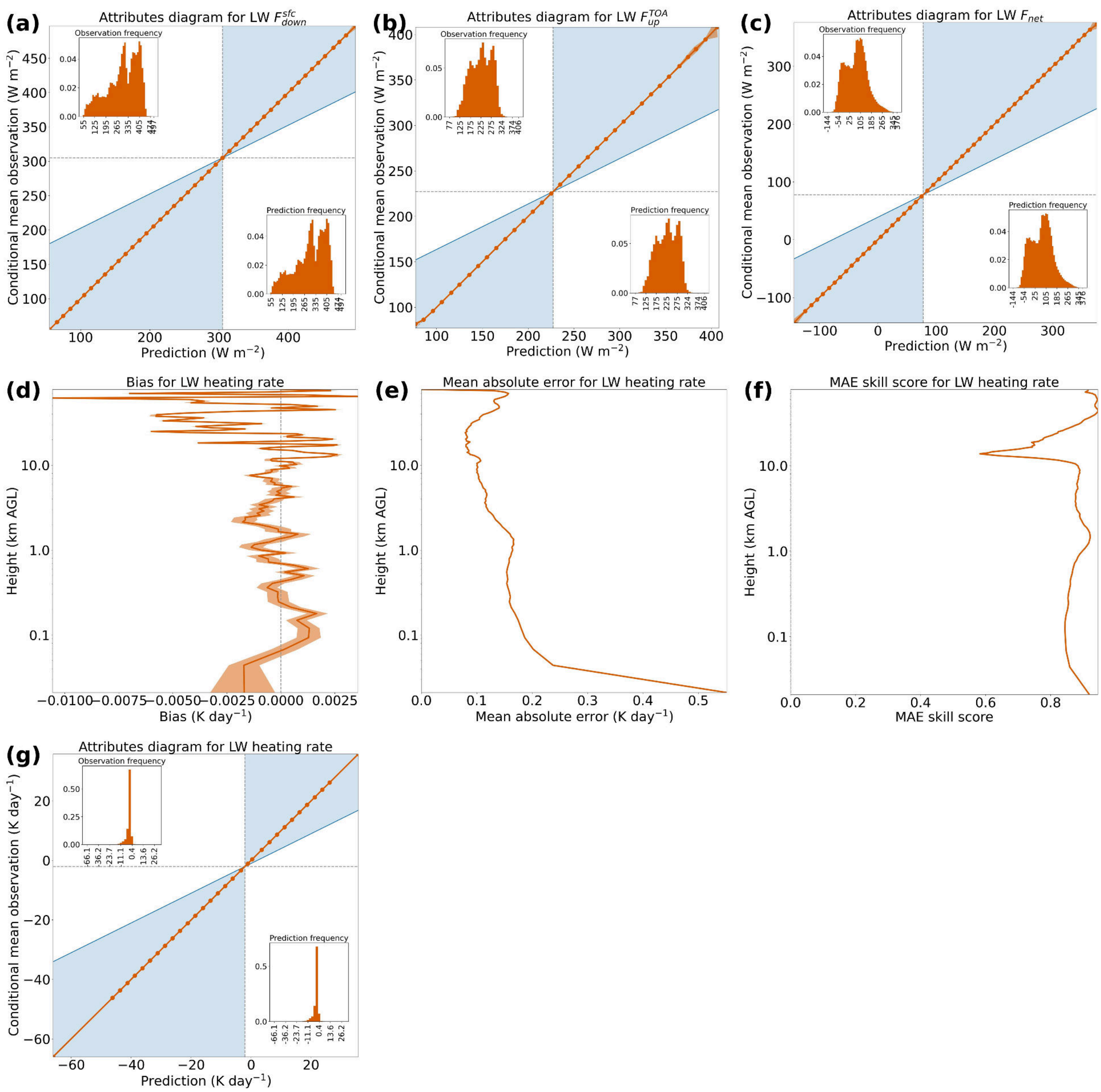

Figure 11: Performance of best longwave model on testing data. Formatting is explained in the caption of Figure 5. [a-c] Attributes diagram for each flux variable. [d-f] Profiles of bias, MAE, and MAE skill score for HR. [g] Attributes diagram for HR, including all heights.

Figure 11 shows the overall performance of the best longwave model. For all flux variables (Figures 11a-c), the model is almost perfectly reliable and almost perfectly reproduces the observed distribution. The model has only one perceptible conditional bias, namely an underprediction of

~10 W m^{-2} for the lowest $F_{\mathrm{up}}^{\mathrm{TOA}}$ predictions. In other words, when the model predicts an extremely low $F_{\mathrm{up}}^{\mathrm{TOA}}$, the prediction is slightly too extreme. The model has an absolute bias $\ll$ 0.1 K day^{-1} for HR at every height (Figure 11d) but much larger MAEs (Figure 11e), reaching 0.55 and 0.24 K day^{-1} at the bottom two grid levels (~21 and ~44 m AGL). As will be shown, longwave RT near the surface is sensitive to fine-scale details of the thermodynamic profile, which the model struggles to capture. Because the climatological model also has its largest HR MAE at the surface, the NN model's local maximum in MAE does not translate to a local minimum in MAE skill score (Figure 11f). Lastly, the attributes diagram for HR (Figure 11g) tells a similar story to those for the flux variables: the model is almost perfectly reliable and almost perfectly reproduces the observed distribution. Supplemental Figures S31-S32 are analogous to Figure 11 but only for extreme cases – *i.e.*, the 3% of testing profiles with the greatest height-maximum and height-averaged absolute HR, respectively. As for the shortwave model, we find that although errors are higher for the extreme cases, HRs and fluxes still have almost perfect reliability and absolute HR bias is well below 0.1 K day^{-1} throughout the profile.

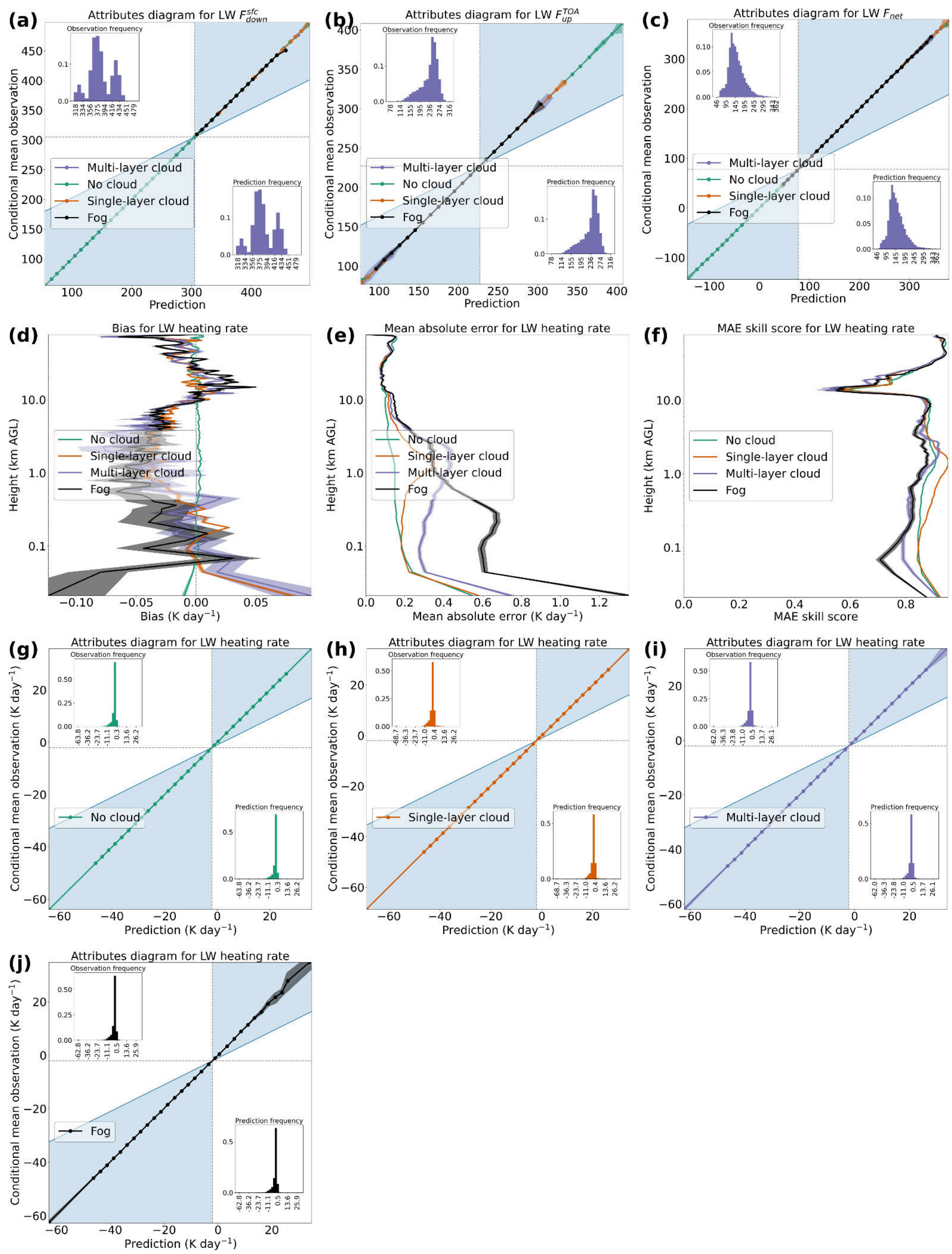

Figure 12: Performance of best longwave model on testing data, separated by liquid-only cloud regime. Formatting is explained in the caption of Figure 6. [a-c] Attributes diagram for each flux variable. [d-f] Profiles of bias, MAE, and MAE skill score for HR. [g] Attributes diagram for HR, including all heights, only for cases with no cloud (90.84% of the testing data). [h] Same but for single-layer cloud (8.74% of the testing data). [i] Same but for multi-layer cloud (0.42% of the testing data). [j] Same but for fog (0.63% of the testing data).

Figure 12 shows the model's performance as a function of liquid-only cloud regime. Performance for other cloud phases (ice-only, mixed-phase, and any-phase) is shown in Supplemental Figures S28-S30. The attributes diagrams for flux variables (Figures 12a-c) tell a similar story to the cloud-agnostic versions (Figures 11a-c): a few slight conditional biases but no absolute bias exceeding 20 W m^{-2}. In the bottom few 100 m of the troposphere, HR errors (Figures 12d-f) are best for clear-sky profiles, followed by single- and multi-layer cloud, and worst for foggy profiles. In other words, the largest HR errors in the bottom few 100 m are caused by clouds, especially clouds that reach the surface. Meanwhile, in the troposphere above ~1 km, HR errors (Figures 12d-f) are best for clear-sky profiles and worst for those with multi-layer cloud. Errors for foggy profiles above ~1 km are intermediate, because many surface-based clouds are not thick enough to reach these heights. Lastly, the attributes diagram for HR (Figures 12g-j) is nearly perfect in all cloud regimes except fog. The model has a considerable negative bias (as large as 1 K day^{-1}) when predicting HR above 20 K day^{-1} in foggy profiles, but as shown by the confidence interval – which overlaps the 1:1 line – this apparent defect could be an artifact of small sample size.

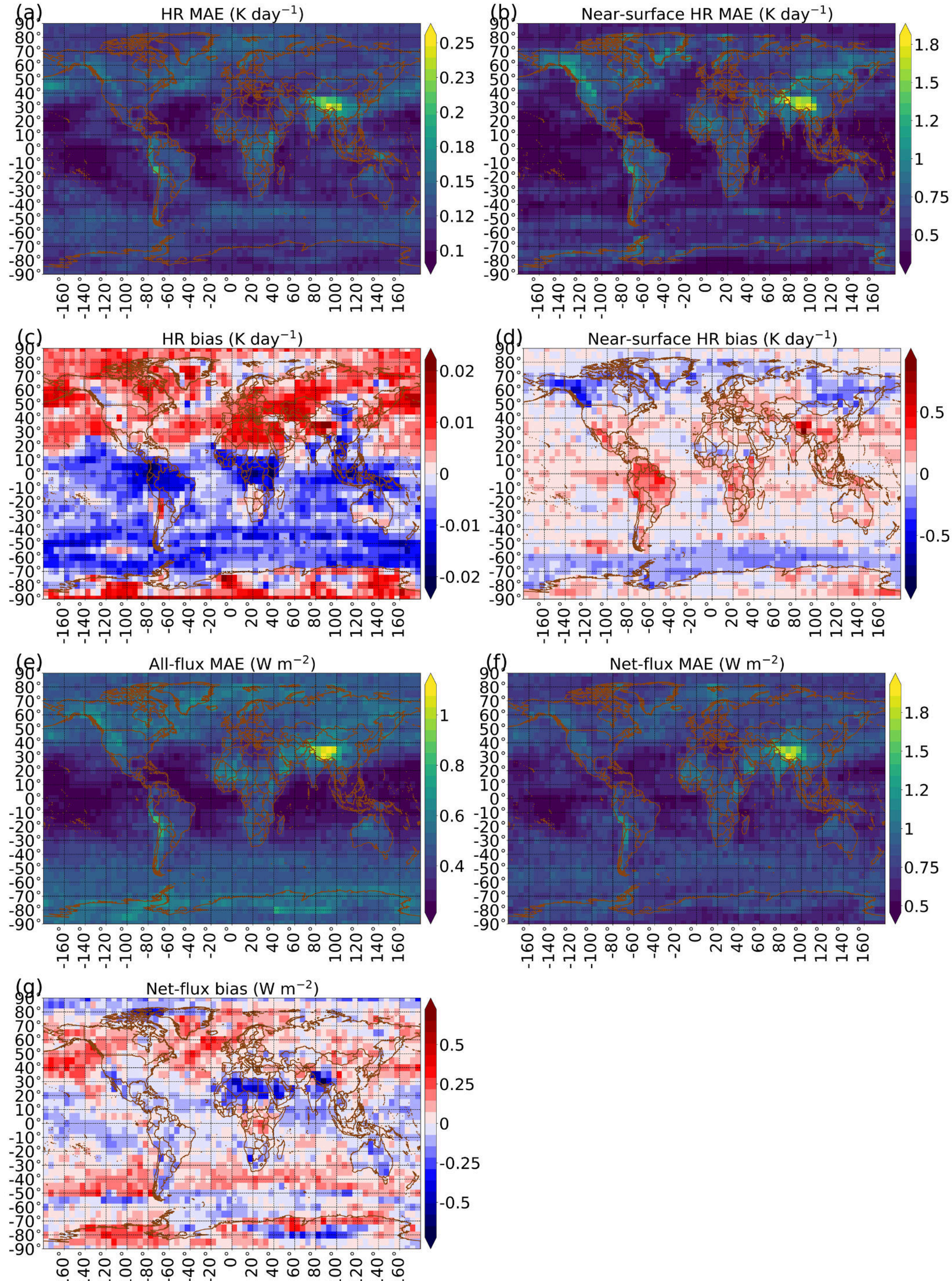

Figure 13: Performance of best longwave model on testing data, binned by geographic location on a 5°-by-5° grid. [a] Column-averaged MAE for HR. [b] MAE for near-surface HR. [c] Column-averaged bias for HR. [d] Bias for near-surface HR. [e] All-flux MAE, averaged over the three flux variables. [f] MAE for net flux only. [g] Bias for net flux only.

Figure 13 shows the model's performance as a function of location. The column-averaged MAE for HR (Figure 13a) is mostly between 0.10 and 0.15 K day^{-1}; it exceeds 0.15 K day^{-1} at a few locations, notably Tibet, southern Peru, and the northwestern Rocky Mountains. The MAE for near-surface HR (Figure 13b) is much larger – mostly between 0.35 and 0.94 K day^{-1}, exceeding 0.94 K day^{-1} at the same locations. The column-averaged bias for HR (Figure 13c) is mostly between -0.01 and +0.01 K day^{-1}, with absolute bias not exceeding 0.02 K day^{-1} at any location. The bias for near-surface HR (Figure 13d) is larger – mostly between -0.24 and +0.22 K day^{-1}, with absolute value exceeding 0.24 K day^{-1} in Tibet, northern South America, and the northwestern Rockies. The all-flux MAE (Figure 13e) is mostly between 0.24 and 0.63 W m^{-2}, exceeding 0.63 W m^{-2} mainly in Tibet. The net-flux MAE (Figure 13f) follows a similar pattern to the all-flux MAE. The net-flux bias (Figure 13g) is mostly between -0.23 and +0.24 W m^{-2}, with absolute bias not exceeding 0.72 K day^{-1} at any location. Maxima in raw error mostly correspond to maxima in relative error (Supplemental Figure S33), which indicates that longwave RT is fundamentally harder to predict in these regions. Lastly, Supplemental Figure S34 shows that, while $F_{\text{down}}^{\text{sfc}}$ and $F_{\text{up}}^{\text{TOA}}$ have similar MAE values over most of the globe, $F_{\text{down}}^{\text{sfc}}$ bias is worse than $F_{\text{up}}^{\text{TOA}}$ bias at most locations. Thus, at most locations, net-flux bias (which equals $F_{\text{down}}^{\text{sfc}}$ bias minus $F_{\text{up}}^{\text{TOA}}$ bias) primarily reflects $F_{\text{down}}^{\text{sfc}}$ bias, with a small contribution from $F_{\text{up}}^{\text{TOA}}$.

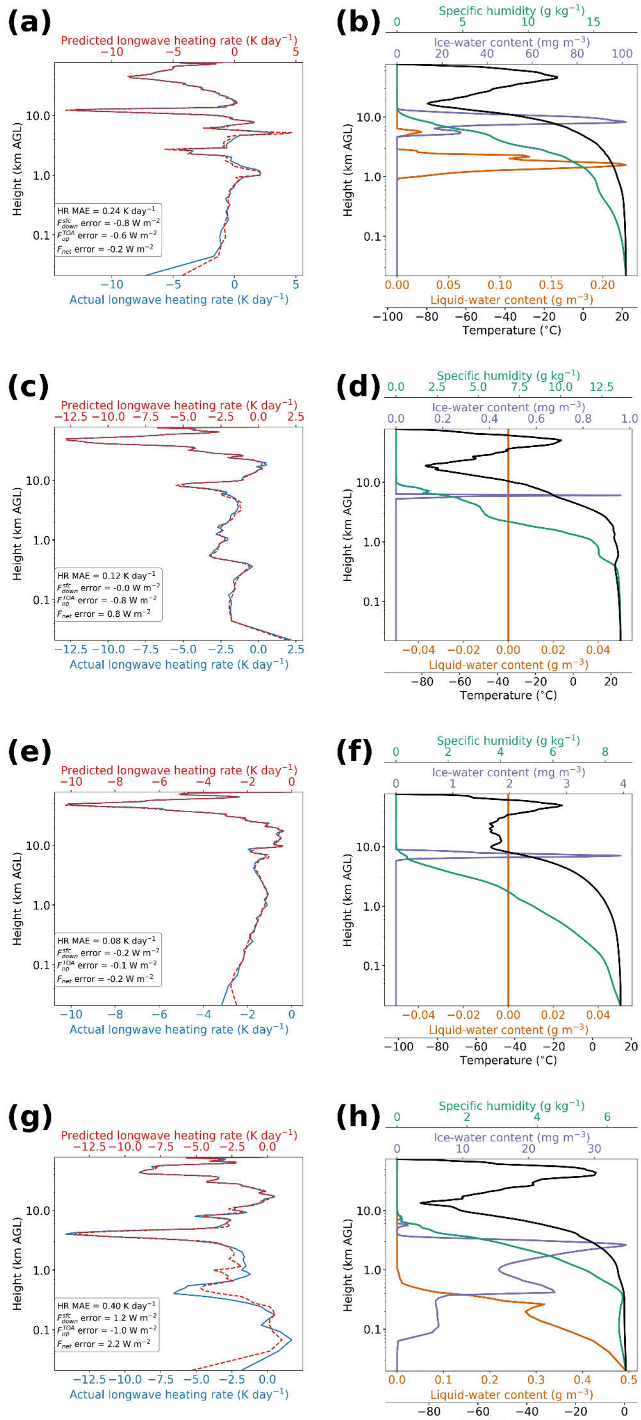

Figure 14: Geography-based case studies for the best longwave model. [a-b] Case study from Tibet, with Γ_T^{sfc} = 4.40 K km^{-1} and Γ_q^{sfc} = 5.9 g kg^{-1} km^{-1}; [c-d] another case study from Tibet, with Γ_T^{sfc} = 11.75 K km^{-1} and Γ_q^{sfc} = 0.7 g kg^{-1} km^{-1}; [e-f] case study from northwestern Rockies, with Γ_T^{sfc} = 4.62 K km^{-1} and Γ_q^{sfc} = 11.9 g kg^{-1} km^{-1}; [g-h] case study from southern Peru, with Γ_T^{sfc} = 10.91 K km^{-1} and Γ_q^{sfc} = 4.3 g kg^{-1} km^{-1}. For each case study, the left panel shows actual and predicted RT solutions, while the right panel shows four of the most important predictor variables for longwave RT. In each left panel, the legend shows column-averaged MAE for HR (labeled "HR MAE") and errors for the three flux variables (predicted minus actual). Γ_T^{sfc} and Γ_q^{sfc} (Equation 5) are summaries of important predictor variables (the thermodynamic profiles). These scalars are thus reported in the caption for each panel.

Figure 14 shows case studies from regions with high model error: Tibet (panels a-d), the northwestern Rockies (panels e-f), and southern Peru (panels g-h). The following observations aim to represent 800 random profiles (200 per region), a superset of the four shown in Figure 14. First, most of the 800 profiles feature liquid and/or ice cloud. Like the shortwave model, the longwave model matches the shape of the HR profile well but often misses extreme HRs associated with cloud by > 1 K day^{-1}. Sometimes the model overestimates longwave cooling above clouds (*e.g.*, ~2.5 and ~10 km in panel a, ~8 km in panel c), and sometimes it underestimates cooling (*e.g.*, ~0.4 and ~4 km in panel g). Second, as for shortwave RT, regions with high longwave error have very high surface elevations, which are globally rare. Third, sometimes longwave HR error near the surface is large even for profiles that appear uncomplicated near the surface (*e.g.*, panels e-f), because near-surface longwave RT is sensitive to fine details of the near-surface thermodynamic profile.

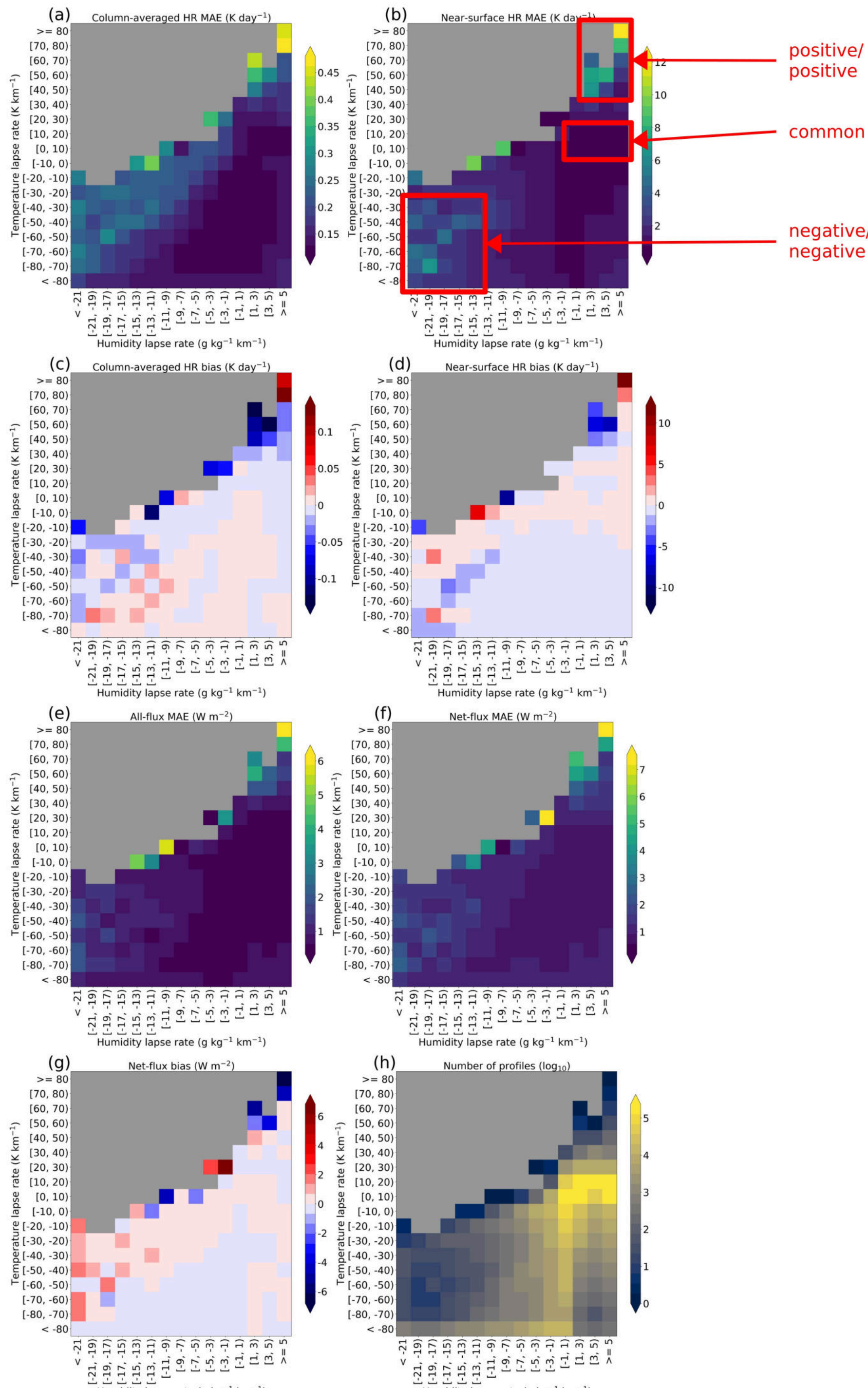

Figure 15: Performance of best longwave model on testing data, binned by near-surface thermodynamic lapse rates, with Γ_T^{sfc} bins of width 10 K km^{-1} and Γ_q^{sfc} bins of width 2 g kg^{-1} km^{-1}. The three labeled regimes (positive/positive, negative/negative, and common) are explained in the main text. [a] Column-averaged MAE for HR. [b] MAE for near-surface HR. [c] Column-averaged bias for HR. [d] Bias for near-surface HR. [e] All-flux MAE, averaged over the three flux variables. [f] MAE for net flux only. [g] Bias for net flux only. [h] Number of testing samples per bin, in logarithmic scale.

Figure 15 shows the model's performance as a function of near-surface thermodynamics, specifically the temperature lapse rate (Γ_T^{sfc} in Equation 5) and humidity lapse rate (Γ_q^{sfc} in Equation 5). First, we note that all error metrics (Figures 15a-g) are worst in two regimes, which we call the positive/positive and negative/negative regimes. The positive/positive regime has large positive Γ_T^{sfc} and Γ_q^{sfc} – *i.e.*, both temperature and humidity decrease strongly with height. The negative/negative regime has large negative lapse rates – *i.e.*, both temperature and humidity exhibit a strong inversion, increasing with height. Second, both the positive/positive and negative/negative regimes are quite rare in our dataset, as shown in Figure 15h. Most profiles have a small positive Γ_T^{sfc} and small positive Γ_q^{sfc}, the "common" regime labeled in Figure 15. Third, while all error metrics are worst in the positive/positive and negative/negative regimes, the most egregious errors are for near-surface HR, where both MAE (Figure 15b) and absolute bias (Figure 15d) can be $\gg 1$ K day^{-1}. Fourth, relative error (Supplemental Figure S35) is also maximized in the positive/positive and negative/negative regimes, which indicates that extreme near-surface thermodynamics make longwave RT fundamentally harder to predict. Lastly, Supplemental Figure S36 that $F_{\text{down}}^{\text{sfc}}$ errors are worse than $F_{\text{up}}^{\text{TOA}}$ errors in both regimes.

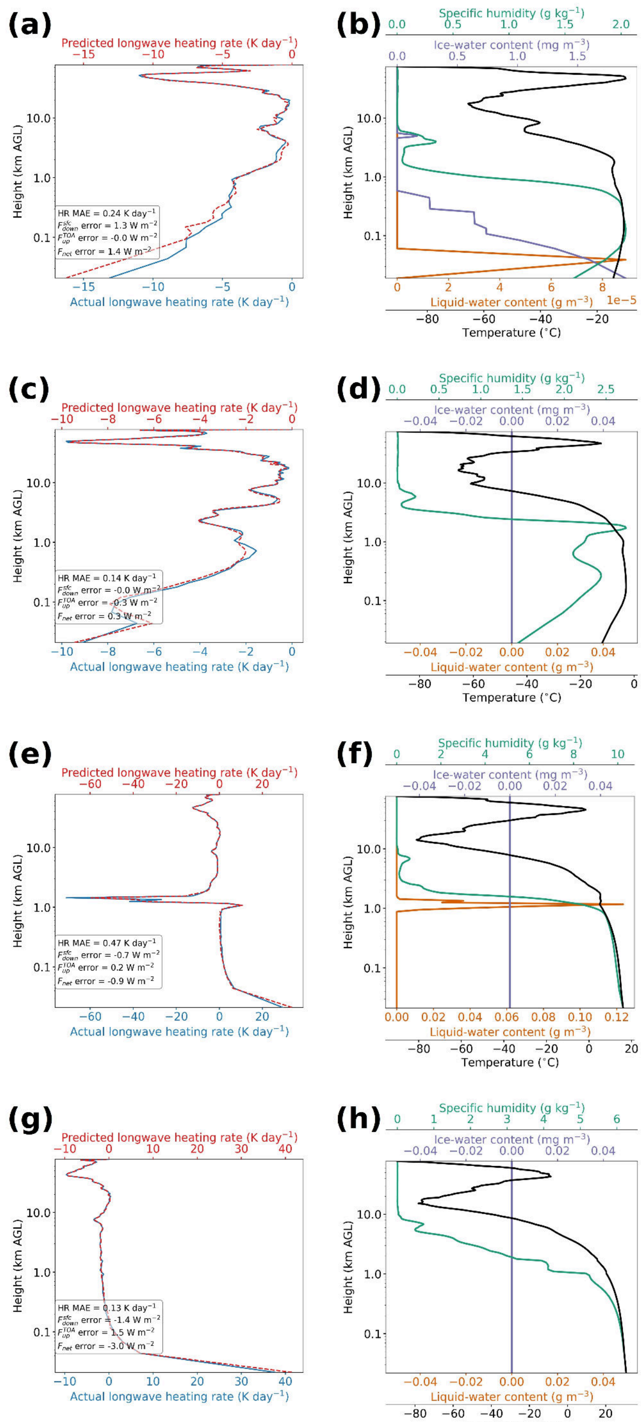

Figure 16: Regime-based case studies for the best longwave model. [a-b] Case study from the negative/negative regime, defined as Γ_T^{sfc} < -30 K km^{-1} and Γ_q^{sfc} < -13 g kg^{-1} km^{-1}. Exact values here are Γ_T^{sfc} = -94.56 K km^{-1} and Γ_q^{sfc} = -14.0 g kg^{-1} km^{-1}. [c-d] Another case study from the negative/negative regime, with Γ_T^{sfc} = -164.83 K km^{-1} and Γ_q^{sfc} = -16.2 g kg^{-1} km^{-1}. [e-f] Case study from the positive/positive regime, defined as Γ_T^{sfc} > 40 K km^{-1} and Γ_q^{sfc} > 1 g kg^{-1} km^{-1}. Exact values here are Γ_T^{sfc} = 44.06 K km^{-1} and Γ_q^{sfc} = 8.1 g kg^{-1} km^{-1}. [g-h] Another case study from the positive/positive regime, with Γ_T^{sfc} = 40.25 K km^{-1} and Γ_q^{sfc} = 2.2 g kg^{-1} km^{-1}. Formatting is explained in the caption of Figure 14.

Figure 16 shows case studies from the negative/negative regime (panels a-d) and positive/positive regime (panels e-h). The following observations aim to represent 400 random profiles (200 per regime), a superset of the four shown in Figure 16. First, we note that most of these profiles feature extreme near-surface heating or cooling. Second, like the geography-based case studies (Figure 14), the model generally performs well for these regime-based case studies, except for near-surface HR and a few extremes associated with cloud (*e.g.*, ~1.5 km in Figure 16e). Third, the model's *fractional* error for near-surface HR is generally quite low; cases like Figure 16a do not occur very often.

6. Summary and future work

We have developed neural networks (NN) to emulate the full RRTM, *i.e.*, the shortwave and longwave RRTM with all predictor variables. Both the RRTM and NN-based emulators are driven by forecast profiles from the GFSv16 on the native vertical grid, which uses hybrid pressure-sigma coordinates. We experimented with novel deep-learning methods designed to produce realistic and accurate spatial structure in gridded predictions: the U-net++ architecture, U-net3+ architecture, and deep-supervision training method. We hypothesized that the best NNs would be those with the U-net3+ architecture and deep supervision. Contrary to our hypotheses, we found that deep supervision leads to worse performance and architecture has little impact. We also experimented with three other hyperparameters – NN width, depth, and spectral complexity – which strongly control the NN's overall complexity, causing the number of trainable weights to vaey from $O(10^5)$ to $O(10^{8.5})$. We found that the best NNs are at the more complex end of the spectrum; the selected shortwave and longwave NNs have $10^{7.52}$ and $10^{7.28}$ trainable weights, respectively. Overall, the better NNs are deep (have encoders and decoders at many spatial resolutions), narrow (have only one convolutional layer per block), and have large spectral complexity (many convolutional filters and thus many feature maps). While NN type (U-net++ or U-net3+) has only a weak effect on performance, the best shortwave NN is a U-net++ model, while the best longwave NN is a U-net3+ model. Our NNs are an example of knowledge-guided machine learning, identified as a major need in ML applications to the geosciences (Gil et al. 2019; Reichstein et al. 2019). Specifically, we enforce energy conservation in the NNs (Equation 2); use a custom loss function to emphasize large heating rates (HR), which are rare but important for weather and climate (Equation 3); and include custom predictors to account for vertically non-local effects (Section 3c3 of L21).

The best shortwave NN model performs extremely well in an aggregate sense, *i.e.*, averaged over all the testing data. Highlights include reliable fluxes, with all conditional biases < 10 W m^{-2} in absolute value; reliable HRs, with all conditional biases $\ll$ 1 K day^{-1} in absolute value; and absolute HR bias < 0.1 K day^{-1} at all heights, suggesting that the NN could be stably integrated into the GFSv16 as a parameterization. The model also performs extremely well in all cloud regimes, at most geographic locations, and in most regimes defined by solar zenith angle (SZA) and aerosol optical depth (AOD). The largest errors occur in Tibet and east Antarctica, which feature high surface elevation/albedo, and in the low-SZA/high-AOD regime, which features a lot of incoming solar radiation and interaction with aerosols. However, even these largest errors are quite small: mean absolute error (MAE) for HR does not exceed 0.6 K day^{-1}, even near the surface; absolute HR bias does not exceed 0.3 K day^{-1}, even near the surface; MAE for flux variables does not exceed 10 W m^{-2}; and net-flux bias does not exceed 5 W m^{-2}. For regimes that make RT *fundamentally* harder to predict – *e.g.*, high elevation/albedo, which increase both raw and relative errors – results could potentially be improved by adding training data from these regimes. Table 8 compares our model to NN-based emulators of shortwave RT from three other studies: Krasnopolsky et al. 2012 (K12), Song and Roh 2021 (SR21), and Kim and Song 2022 (KS22). Although our model appears to perform best, this comparison is not apples-to-apples, due to different vertical resolutions (127 levels here, 64 in K12, 39 in the other two studies), testing cases (time period and spatial domain), and predictor variables. The three comparison studies omit aerosols, all trace gases other than O_3, LWC and IWC (they use cloud fraction instead, with no distinction between liquid and ice), and the particle-size distribution (for which we use liquid and ice effective radii). Lastly, our shortwave NN runs 7510 times faster than the shortwave RRTM.

The best longwave NN model also performs extremely well in an aggregate sense; highlights include near-perfect reliability for both fluxes and HRs and absolute HR bias $\ll$ 0.1 K day^{-1} at every height. The model's main deficiency is a large error in near-surface HR, *e.g.*, an MAE of 0.55 K day^{-1} at the lowest grid level. However, longwave RT near the surface is complicated, and errors here are often quite large. For example, in Veerman et al. (2020), who emulated only the gas-optics part of the RRTMGP, near-surface HR bias is on the order of 1 K day^{-1} (their Figure 2c).

Table 8: Comparison of NN-based emulators for shortwave RT. For our model, we use the testing data only. For the comparison studies, we take results from Table 2 of K12, page 7 of SR21 for HR errors, Table 3 (the "WRF15" column) of SR21 for flux errors, and Figure 1 of KS22 (these values are estimated visually). "Profile RMSE" is defined in Equation A1 of K12; "near-surface" means for the lowest model level; and "N/A" means that the statistic is not reported. Although KS22 reports flux errors, the statistic is all-flux RMSE, computed by averaging over three variables: $F_{\text{down}}^{\text{sfc}}$, $F_{\text{up}}^{\text{TOA}}$, and $F_{\text{up}}^{\text{sfc}}$. We predict a different set of flux variables – F_{net} instead of $F_{\text{up}}^{\text{sfc}}$ – and thus do not compare our flux errors with KS22.

Model	**Ours**	**K12**	**SR21**	**KS22**
Statistic				
Column-averaged HR RMSE (K day^{-1})	0.14	0.26	0.17	~0.2
Column-averaged HR bias (K day^{-1})	-0.002	-0.007	N/A	N/A
HR profile RMSE (K day^{-1})	0.12	0.18	N/A	N/A
Near-surface HR RMSE (K day^{-1})	0.20	0.20	N/A	N/A
Near-surface HR bias (K day^{-1})	+0.0001	-0.03	N/A	N/A
$F_{\text{down}}^{\text{sfc}}$ RMSE (W m^{-2})	5.85	N/A	43.75	N/A
$F_{\text{up}}^{\text{TOA}}$ RMSE (W m^{-2})	3.94	N/A	36.20	N/A

The model performs well in all cloud regimes, at most geographic locations, and in most regimes defined by near-surface thermodynamics. The largest errors occur with liquid-only fog, where the bias and MAE for near-surface HR reach -0.12 and 1.3 K day^{-1} respectively; in Tibet, where near-surface bias and MAE reach almost 1 and 2 K day^{-1} respectively; and under extreme near-surface thermodynamics, where near-surface absolute bias and MAE are $\gg$ 1 K day^{-1}. However, the extreme thermodynamic regimes are quite rare, so this last number is affected by small sample size. Also, even in the aforementioned regimes with large error in near-surface HR, column-averaged bias for HR does not exceed 0.15 K day^{-1} in absolute value; column-averaged MAE for HR does not exceed 0.6 K day^{-1}; MAE for flux variables does not exceed 10 W m^{-2}; and net-flux bias does not exceed 7 W m^{-2}. Table 9 shows that our longwave NN compares very favourably to other studies. Lastly, our longwave NN runs 90 times faster than the longwave RRTM.

Future work will include three items. First, we will develop grid-agnostic NNs that work on profiles with any vertical resolution. This work may benefit from Fourier neural operators (FNO; Lu et al. 2019; Li et al. 2020), which naturally learn physics in a grid-agnostic manner. Second, we will implement the NNs in online mode, *i.e.*, as a parameterization in the GFSv16. To this

Table 9: Comparison of NN-based emulators for longwave RT. For technical notes, see the caption of Table 8.

Model	**Ours**	**K12**	**SR21**	**KS22**
Statistic				
Column-averaged HR RMSE (K day^{-1})	0.22	0.52	0.46	~0.375
Column-averaged HR bias (K day^{-1})	-0.0006	+0.008	N/A	N/A
HR profile RMSE (K day^{-1})	0.20	0.38	N/A	N/A
Near-surface HR RMSE (K day^{-1})	0.83	0.55	N/A	N/A
Near-surface HR bias (K day^{-1})	-0.002	+0.02	N/A	N/A
$F_{\text{down}}^{\text{sfc}}$ RMSE (W m^{-2})	0.64	N/A	5.71	N/A
$F_{\text{up}}^{\text{TOA}}$ RMSE (W m^{-2})	0.81	N/A	7.11	N/A

end we have converted the NNs to a Fortran-friendly format, using the Infero library (ECMWF 2022), and ensured that the NNs yield the same predictions in Fortran as in Python. Note that the NNs alone cannot handle subgrid-scale fractional cloudiness, as cloud fraction is a predictor in neither the RRTM nor the NNs. To handle fractional cloudiness in online mode, we will couple the NNs with the Monte Carlo independent-column approximation (Pincus et al. 2003). Third, we will perform thorough testing of the NNs in online mode. Specifically, we will conduct month-long retrospective simulations in both the summer and winter, using a control model (original parameterization) and experimental model (NN parameterization). We will compare the two models against each other and against observations, using methods as in Turner et al. (2012) and Turner et al. (2020). Given the accuracy and efficiency of modern deep NNs, we expect them to replace many existing parameterizations in weather and climate models. However, operational use should proceed only after thorough NN evaluation and with the caution that NNs may generalize poorly outside the distribution of their training data, *e.g.*, to future climates[10]. Safeguards against this problem should be built into NN parameterizations, such as continued online learning or out-of-distribution detection.

[10]In our case, motivated by the strong influence of clouds on radiation – including their phase and number of layers – we paid particular attention to the NNs' ability to emulate the RRTM for all cloud types.

Acknowledgments. This work was partially supported by the NOAA Global Systems Laboratory, Cooperative Institute for Research in the Atmosphere, and NOAA Award Number NA19OAR4320073. Author Ebert-Uphoff's work was partially supported by NSF AI Institute grant #2019758 and NSF grant #1934668.

Data availability statement. The input data (predictor and target variables for all the time periods: NN-training, IR-training, validation, and testing) and selected models (best shortwave NN, best longwave NN, and IR model used to bias-correct each one) are stored on NOAA's high-performance computing systems and are available from the authors upon request. We used version 2.0.0 of ML4RT (Machine Learning for Radiative Transfer; `https://doi.org/10.5281/zenodo.7378773`) – a Python library managed by author Lagerquist – for all training, evaluation, and analysis.

References

Anderson, G., S. Clough, F. Kneizys, J. Chetwynd, and E. Shettle, 1986: AFGL atmospheric constituent profiles (0-120 km). Tech. rep. URL https://apps.dtic.mil/sti/citations/ADA175173.

Baek, S., 2017: A revised radiation package of G-packed McICA and two-stream approximation: Performance evaluation in a global weather forecasting model. *Journal of Advances in Modeling Earth Systems*, **9 (3)**, 1628–1640, URL https://doi.org/10.1002/2017MS000994.

Belochitski, A., and V. Krasnopolsky, 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, URL https://doi.org/10.5194/gmd-14-7425-2021.

Beucler, T., and Coauthors, 2021: Climate-invariant machine learning. *arXiv e-prints*, **2112 (08440)**, URL https://arxiv.org/abs/2112.08440.

Boukabara, S., V. Krasnopolsky, J. Stewart, E. Maddy, N. Shahroudi, and R. Hoffman, 2019: Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges. *Bulletin of the American Meteorological Society*, **100 (12)**, ES473–ES491, URL https://doi.org/10.1175/BAMS-D-18-0324.1.

Chantry, M., P. Dueben, R. Hogan, and P. Ukkonen, 2022: Progress on emulating the radiation scheme via machine learning. *ECMWF Newsletter*, **173**, URL https://www.ecmwf.int/en/newsletter/173/news/progress-emulating-radiation-scheme-machine-learning.

Chantry, M., P. Ukkonen, R. Hogan, and P. Dueben, 2023: Emulating radiative transfer in a numerical weather prediction model. *European Geophysical Union General Assembly*, Vienna, Austria, URL https://doi.org/10.5194/egusphere-egu23-3256.

Chevallier, F., F. Chéruy, N. Scott, and A. Chédin, 1998: A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology*, **37 (11)**, 1385–1397, URL https://doi.org/10.1175/1520-0450(1998)037%3C1385:ANNAFA%3E2.0.CO;2.

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

ECMWF, 2022: Infero: A lower-level API for machine learning inference in operations. GitHub, URL https://github.com/ecmwf-projects/infero.

Geiss, A., P. Ma, B. Singh, and J. Hardin, 2022: Emulating aerosol optics with randomly generated neural networks. *EGUsphere*, **pre-print**, URL https://doi.org/10.5194/egusphere-2022-559.

Gil, Y., and Coauthors, 2019: Intelligent systems for geosciences: An essential research agenda. *Communications of the Association for Computing Machinery*, **62 (1)**, 76–84, URL https://dl.acm.org/doi/10.1145/3192335.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, URL https://www.deeplearningbook.org.

Hsu, W., and A. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293, URL https://doi.org/10.1016/0169-2070(86)90048-8.

Huang, H., and Coauthors, 2020: UNet 3+: A full-scale connected UNet for medical image segmentation. *arXiv e-prints*, **2004 (08790)**, URL https://arxiv.org/abs/2004.08790.

Iacono, M., J. Delamere, E. Mlawer, M. Shephard, S. Clough, and W. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research: Atmospheres*, **113 (D13)**, URL https://doi.org/10.1029/2008JD009944.

Jensen, E., and Coauthors, 2013: Ice nucleation and dehydration in the tropical tropopause layer. *Proceedings of the National Academy of Sciences*, **110 (6)**, 2041–2046, URL https://doi.org/10.1073/pnas.1217104110.

Kim, P., and H. Song, 2022: Usefulness of automatic hyperparameter optimization in developing radiation emulator in a numerical weather prediction model. *Atmosphere*, **13 (5)**, 721, URL https://doi.org/10.3390/atmos13050721.

Krasnopolsky, V., A. Belochitski, Y. Hou, S. Lord, and F. Yang, 2012: Accurate and fast neural network emulations of long and short wave radiation for the NCEP Global Forecast System model. Tech. rep. URL https://repository.library.noaa.gov/view/noaa/6951.

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

Le, T., C. Liu, B. Yao, V. Natraj, and Y. Yung, 2020: Application of machine learning to hyperspectral radiative transfer simulations. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **246**, 106 928, URL https://doi.org/10.1016/j.jqsrt.2020.106928.

Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, A. Stuart, K. Bhattacharya, and A. Anandkumar, 2020: Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems*, **33**, 6755–6766, URL https://proceedings.neurips.cc/paper/2020/hash/4b21cf96d4cf612f239a6c322b10c8fe-Abstract.html.

Liu, Y., R. Caballero, and J. Monteiro, 2020: RadNet 1.0: Exploring deep learning architectures for longwave radiative transfer. *Geoscientific Model Development*, **13 (9)**, 4399–4412, URL https://doi.org/10.5194/gmd-13-4399-2020.

Lu, L., P. Jin, and G. Karniadakis, 2019: Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv e-prints*, **1910 (03193)**, URL https://arxiv.org/abs/1910.03193.

Meyer, D., R. Hogan, P. Dueben, and S. Mason, 2022: Machine learning emulation of 3D cloud radiative effects. *Journal of Advances in Modeling Earth Systems*, **14 (3)**, URL https://doi.org/10.1029/2021MS002550.

Miles, N., J. Verlinde, and E. E. Clothiaux, 2000: Cloud droplet size distributions in low-level stratiform clouds. *Journal of the Atmospheric Sciences*, **57 (2)**, 295–311, URL https://doi.org/10.1175/1520-0469(2000)057%3C0295:CDSDIL%3E2.0.CO;2.

Mishra, S., D. Mitchell, D. Turner, and R. Lawson, 2014: Parameterization of ice fall speeds in midlatitude cirrus: Results from SPartICus. *Journal of Geophysical Research: Atmospheres*, **119 (7)**, 3857–3876, URL https://doi.org/10.1002/2013JD020602.

Mitchell, D., R. Lawson, and B. Baker, 2011: Understanding effective diameter and its application to terrestrial radiation in ice clouds. *Atmospheric Chemistry and Physics*, **11 (7)**, 3417–3429, URL https://doi.org/10.5194/acp-11-3417-2011.

Mlawer, E., S. Taubman, P. Brown, M. Iacono, and S. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, **102 (D14)**, 16 663–16 682, URL https://doi.org/10.1029/97JD00237.

Mlawer, E., and D. Turner, 2016: Spectral radiation measurements and analysis in the ARM Program. *Meteorological Monographs*, Vol. 57, American Meteorological Society, 14.1–14.17, URL https://doi.org/10.1175/AMSMONOGRAPHS-D-15-0027.1.

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

Muñoz-Esparza, D., C. Becker, J. Sauer, D. Gagne, J. Schreck, and B. Kosović, 2022: On the application of an observations-based machine learning parameterization of surface layer fluxes within an atmospheric large-eddy simulation model. *Journal of Geophysical Research: Atmospheres*, **127 (16)**, URL https://doi.org/10.1029/2021JD036214.

Myhre, G., C. Myhre, B. Samset, and T. Storelvmo, 2013: Aerosols and their relation to global climate and climate sensitivity. *Nature Education Knowledge*, **4 (5)**, 7, URL https://www.nature.com/scitable/knowledge/library/aerosols-and-their-relation-to-global-climate-102215345/.

Neubauer, D., U. Lohmann, C. Hoose, and M. Frontoso, 2014: Impact of the representation of marine stratocumulus clouds on the anthropogenic aerosol effect. *Atmospheric Chemistry and Physics*, **14 (21)**, 11 997–12 022, URL https://doi.org/10.5194/acp-14-11997-2014.

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

Pincus, R., H. Barker, and J. Morcrette, 2003: A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. *Journal of Geophysical Research: Atmospheres*, **108 (D13)**, URL https://doi.org/10.1029/2002JD003322.

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

Rasp, S., 2020: Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: General algorithms and Lorenz 96 case study (v1.0). *Geoscientific Model Development*, **13 (5)**, 2185–2196, URL https://doi.org/10.5194/gmd-13-2185-2020.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, URL https://doi.org/10.1038/s41586-019-0912-1.

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

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, Munich, Germany, Technical University of Munich, URL https://doi.org/10.1007/978-3-319-24574-4_28.

Schmetz, J., 1989: Towards a surface radiation climatology: Retrieval of downward irradiances from satellites. *Atmospheric Research*, **23 (3-4)**, 287–321, URL https://doi.org/10.1016/0169-8095(89)90023-9.

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

Sonoda, S., and N. Murata, 2017: Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, **43 (2)**, 233–268, URL https://doi.org/10.1016/j.acha.2015.12.005.

Stegmann, P., B. Johnson, I. Moradi, B. Karpowicz, and W. McCarty, 2022: A deep learning approach to fast radiative transfer. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **280**, 108 088, URL https://doi.org/10.1016/j.jqsrt.2022.108088.

Tang, T., and Coauthors, 2020: Response of surface shortwave cloud radiative effect to greenhouse gases and aerosols and its impact on summer maximum temperature. *Atmospheric Chemistry and Physics*, **20 (13)**, 8251–8266, URL https://doi.org/10.5194/acp-20-8251-2020.

Turner, D., A. Merrelli, D. Vimont, and E. Mlawer, 2012: Impact of modifying the longwave water vapor continuum absorption model on community Earth system model simulations. *Journal of Geophysical Research: Atmospheres*, **117 (D4)**, URL https://doi.org/10.1029/2011JD016440.

Turner, D., and Coauthors, 2020: A verification approach used in developing the Rapid Refresh and other numerical weather prediction models. *Journal of Operational Meteorology*, **8 (3)**, 39–53, URL http://doi.org/10.15191/nwajom.2020.0803.

Turner, D. D., and Coauthors, 2004: The QME AERI LBLRTM: A closure experiment for downwelling high spectral resolution infrared radiance. *Journal of the Atmospheric Sciences*, **61 (22)**, 2657–2675, URL https://doi.org/10.1175/JAS3300.1.

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

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

Veerman, M., R. Pincus, R. Stoffer, C. V. Leeuwen, D. Podareanu, and C. V. Heerwaarden, 2020: Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, **379 (2194)**, URL https://doi.org/10.1098/rsta.2020.0095.

Wallace, J., and P. Hobbs, 2006: *Atmospheric Science: An Introductory Survey*, Vol. 2. Elsevier.

Yang, C., J. Chiu, J. Gristey, G. Feingold, and W. Gustafson, 2022: Machine learning emulation of 3D shortwave radiative transfer for shallow cumulus cloud fields. *Conference on Atmospheric Radiation, Atmospheric Radiative Transfer, and Light-scattering Theory*, Madison, Wisconsin, American Meteorological Society, URL https://ams.confex.com/ams/CMM2022/meetingapp.cgi/Paper/406293.

Zhou, Z., M. Siddiquee, N. Tajbakhsh, and J. Liang, 2019: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, **39 (6)**, 1856–1867, URL https://doi.org/10.1109/TMI.2019.2959609.

Estimating full longwave and shortwave radiative transfer with neural networks of varying complexity

---

Supplemental material

1. Creating synthetic aerosol variables

We use the following procedure for each profile. Recall that the three aerosol-based predictors are single-scattering albedo (SSA), asymmetry parameter, and extinction coefficient – and that the first two are scalars. All other variables created in this procedure are intermediate.

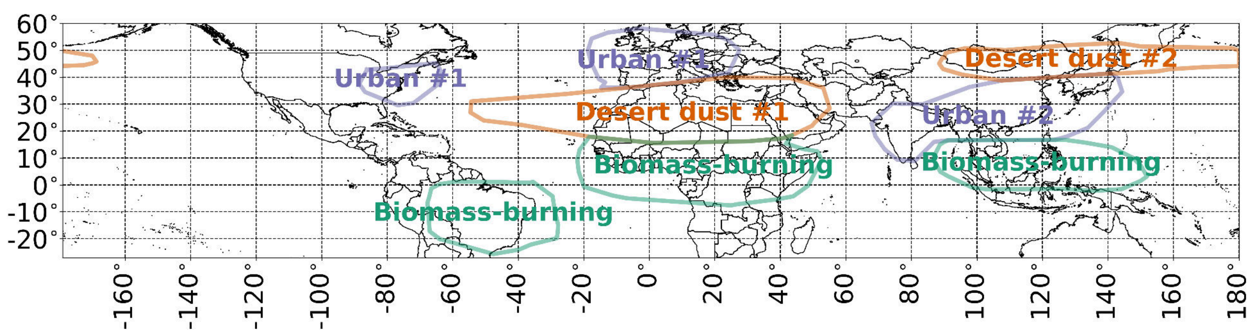

Figure S1: Aerosol regions. Five of the eight regions (urban #1, urban #2, desert dust #1, desert dust #2, and biomass-burning) are outlined in coloured polygons. Outside the coloured polygons, the region defaults to "land" or "ocean" if latitude $\in [-60, 60]$ °N and "polar" otherwise.

1. Determine region. Assign the profile to one of eight regions (Figure S1): polar, land, ocean, urban #1, urban #2, desert dust #1, desert dust #2, and biomass-burning.

2. Determine SSA. Draw the SSA from a normal distribution with region-dependent parameters (Table S1), then bound values to the range $[0, 1]$. Values outside this range are non-physical.

3. Determine asymmetry parameter. Draw the asymmetry parameter from a normal distribution with region-dependent parameters (Table S1), then bound values to the range $[0, 1]$. Values outside this range are non-physical.

4. Determine scale height. Draw the scale height – *i.e.*, the e-folding height for extinction coefficient – from a normal distribution with region-dependent parameters (Table S1), then bound values to the range $[0.1, \infty)$ km.

5. Compute baseline AOD.

   (a) Compute the baseline extinction coefficient at each grid level:

$$\epsilon z = e^{-\frac{z}{H}} \cdot 1 \text{ km}^{-1}, \tag{1}$$

   where z is the grid-point height and H is the scale height computed in step 4, both in km above ground. See Figure S2a.

   (b) Compute the baseline AOD:

$$\text{AOD}_{\text{baseline}} = {}^{z_{\text{top}}}_{z_{\text{bottom}}} \epsilon z dz, \tag{2}$$

   z_{top} and z_{bottom} are the top and bottom heights in the grid (km above ground) and ϵz comes from Equation 1.

6. Determine actual AOD.

   (a) Create narrow AOD distribution, using region-dependent parameters listed in Table S2. See Figure S2b.

   (b) Create wide AOD distribution, using region-dependent parameters listed in Table S2. See Figure S2c.

   (c) Shift wide AOD distribution, giving it the same mean as the narrow distribution. Specifically, subtract $\overline{\text{AOD}_{\text{wide}}} - \overline{\text{AOD}_{\text{narrow}}}$ from every value in the wide AOD distribution, where $\overline{\text{AOD}_{\text{wide}}}$ and $\overline{\text{AOD}_{\text{narrow}}}$ are the means of the two distributions.

   (d) Censor wide AOD distribution, bounding values to the range $[0, 1.5]$. Negative values are non-physical, and values > 1.5 are very rare. See Figure S2d.

7. Compute the actual extinction coefficient at each grid level:

$$\epsilon z = \frac{\text{AOD}_{\text{actual}}}{\text{AOD}_{\text{baseline}}} e^{-\frac{z}{H}} \cdot 1 \text{ km}^{-1}. \tag{3}$$

   Note that, while each level has a different height z, all other variables on the right-hand side are constant throughout the profile. See Figure S2e.

Table S1: Region-dependent distribution parameters for aerosol variables other than AOD. Each cell contains the mean, followed by the standard deviation, of a normal distribution. SSA = single-scattering albedo.

Variable	**SSA (unitless)**	**Asymmetry parameter (unitless)**	**Scale height (m)**
Region			
Polar	0.95, 0.02	0.72, 0.03	500, 100
Land	0.95, 0.02	0.70, 0.03	1500, 300
Ocean	0.96, 0.02	0.75, 0.03	1000, 100
Urban #1	0.94, 0.02	0.70, 0.03	1500, 300
Urban #2	0.91, 0.04	0.70, 0.03	1500, 100
Desert dust #1	0.95, 0.02	0.78, 0.05	1500, 200
Desert dust #2	0.95, 0.02	0.78, 0.03	1500, 200
Biomass-burning	0.91, 0.05	0.72, 0.03	2000, 300

In step 6, the narrow distribution is based on observations of the real atmosphere, while the wide observation is designed to increase the frequency of large AOD values. In previous work we found that NNs trained with AODs from the narrow distribution failed on large AOD values, which were underrepresented in the training data. The distributional parameters in Tables S1 and S2 were selected by co-author Turner, based on numerous presentations and journal papers; our values for SSA, AOD, and asymmetry parameter largely agree with Kinne (2019).

Table S2: Region-dependent distribution parameters for AOD. Each cell contains the shape parameter, followed by the scale parameter, of a gamma distribution. After applying the gamma distribution, all outputs (sampled AOD values) are divided by 10.

Region	**Narrow distribution**	**Wide distribution**
Polar	0.675, 1.333	2.7, 4.0
Land	7.5, 0.4	30.0, 1.2
Ocean	14.7, 0.143	58.8, 0.429
Urban #1	16.875, 0.267	67.5, 0.8
Urban #2	13.333, 0.45	53.333, 1.35
Desert dust #1	13.333, 0.45	53.333, 1.35
Desert dust #2	7.5, 0.6	30.0, 1.8
Biomass-burning	13.333, 0.45	53.333, 1.35

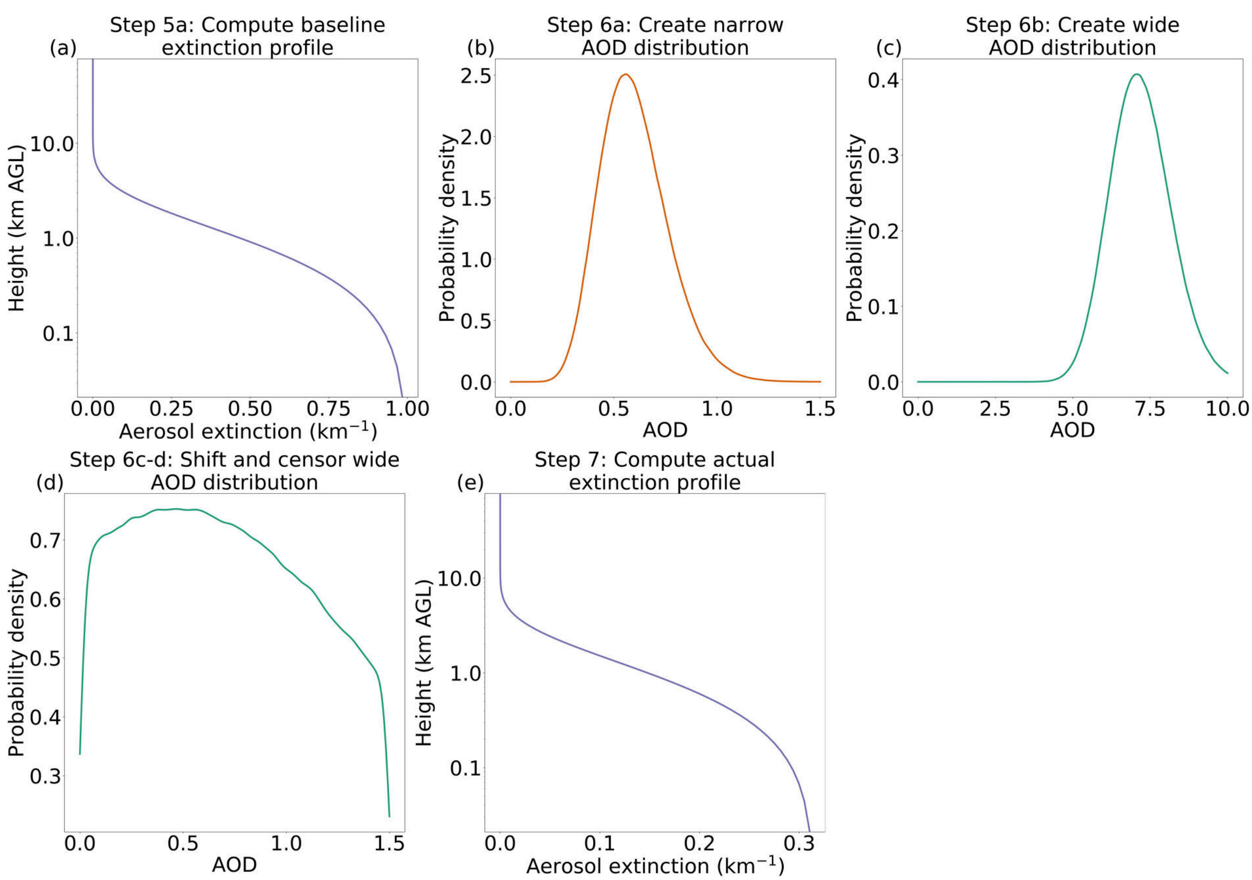

Figure S2: Procedure for creating synthetic profile of aerosol-extinction coefficients. In the case shown, the randomly drawn scale height is 1.318 km; the resulting baseline AOD is 1.297; and the randomly drawn actual AOD, from the distribution in panel d, is 0.409.

2. Distribution of synthetic trace gases

Figure S3 shows the distribution of concentrations for each synthetic trace gas, *i.e.*, those produced from fictitious data as described in Section 2b of the main text. In the canonical profiles taken from Anderson et al. (1986), the maximum values (over all standard atmospheres and heights) are 330 ppmv for CO_2, 1.7 ppmv for CH_4, and 0.32 ppmv for N_2O. The noise included in our procedure yields many values above these maxima, which could be representative of future climates.

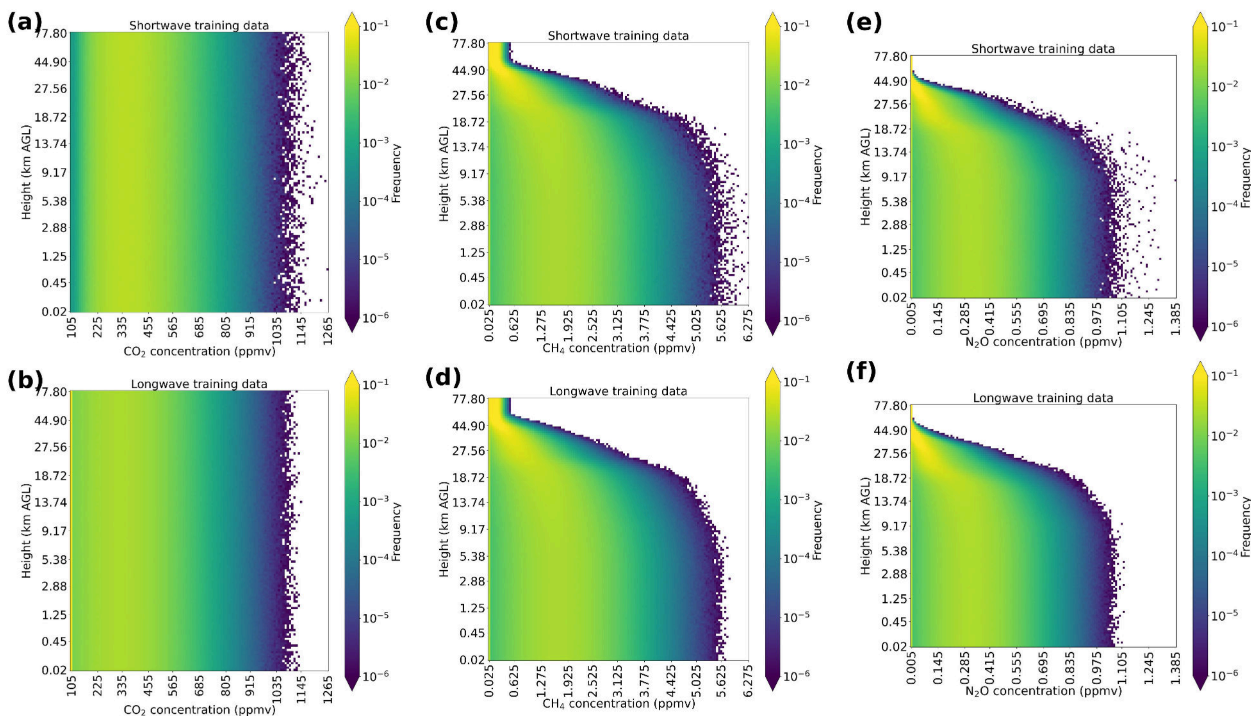

Figure S3: Distribution of trace-gas concentrations in the training data. [a-b] CO_2 concentrations in training data for the shortwave and longwave models; [c-d] same but for CH_4; [e-f] same but for N_2O.

3. Hyperparameter experiment

Table S3 documents constant hyperparameters – *i.e.*, those not varied during the experiment – that are not shown in the architecture schematics (Figures 3-4 in the main text). Subsections a and b discuss results of the experiment.

Table S3: Constant NN hyperparameters, *i.e.*, not varied during the experiment.

Hyperparameter	Value chosen	Justification
Activation function for flux-output layer	Rectified linear unit (ReLU; Nair and Hinton 2010)	ReLU sets negative values to 0 and leaves positive values alone. This is appropriate for the two free flux variables – $F_{\text{down}}^{\text{sfc}}$ and $F_{\text{up}}^{\text{TOA}}$ – which cannot be negative. The other flux variable is F_{net}, which can be negative, but this is computed as $F_{\text{down}}^{\text{sfc}}$ minus $F_{\text{up}}^{\text{TOA}}$ after applying ReLU.
Activation function for HR-output layer	ReLU	ReLU is appropriate for HR, which cannot be negative.
Activation function for internal layers	Leaky ReLU (Maas et al. 2013) with slope of 0.2	The "internal layers" are all non-terminal convolutional and fully connected layers – *i.e.*, all convolutional layers except the HR output and all fully connected layers except the flux output. Leaky ReLU reduces the magnitude of negative values (with the chosen slope, replaces negative values x with $0.2x$) and leaves positive values alone. Strict ReLU solves the problem of vanishing gradients, and leaky ReLU solves the problem of dead neurons that arises from strict ReLU, as discussed in Chapter 4 of Lagerquist (2020).
Batch normalization	Used for internal layers, not output layers	Batch normalization (Ioffe and Szegedy 2015) produces negative values, so it is inappropriate for the output layers. In general, batch normalization alleviates the vanishing-gradient problem (Chapter 4 of Lagerquist 2020).
Number of epochs	1000	In one epoch, each training example is presented to the NN once. Early stopping (below) occurs for all NNs in the experiment, so training never continues for 1000 epochs.
Batch size	724 examples	Each update of the NN's trainable weights is based on 724 profiles. In early experiments (not shown), we found that smaller batches make training susceptible to noise and therefore unstable, while larger batches require too much memory. Both issues are discussed in Li et al. (2014).
Early stopping	Patience of 100 epochs	If the loss on validation data has not reached a new minimum in the last 100 epochs, we stop training and restore NN weights to the epoch with minimum validation loss. In early experiments (not shown) we found that a longer patience merely prolongs training, without helping the NN achieve a lower validation loss.
Optimizer	Adam	Adam (Kingma and Ba 2014) is a sophisticated version of stochastic gradient descent (Section 8.3.1 of Goodfellow et al. 2016). Adam uses a different learning rate for each NN weight and adjusts the learning rates during training, which generally leads to a better model.
Learning rate	Start with 0.001, reduce by 40% upon 10-epoch plateau	A start value of 0.001 is the default in the Keras library. "Reduce by 40% upon 10-epoch plateau" means that, if validation loss has not reached a new minimum in the last 10 epochs and we have not performed a reduction step in the last 10 epochs, we multiply every learning rate by 0.6. The patience (10 epochs) and reduction factor (0.6) are hyperparameters, which we tuned in early experiments (not shown).

a. Results for shortwave RT

Figures S4-S9 show validation error as a function of hyperparameters for a few of the metrics listed in Table 6 of the main text. 12 of the 288 NNs could not be trained, due to memory issues; these NNs are marked by grey squares in Figures S4-S9. NN type has little effect on model performance – note that each figure has one panel per NN type and errors do not vary much across the panels. For unsigned errors (all other than bias; Figures S4-S5 and S8-S9), the most important hyperparameter is spectral complexity, while NN depth and width are of secondary importance. Unsigned errors decrease as spectral complexity increases up to 64, then show little variation as spectral complexity increases beyond 64, which suggests that the optimal value is $\gtrsim$ 64. Also, unsigned errors decreases as NN depth increases and NN width decreases; this suggests that the optimal NN is deep and narrow, with encoders/decoders at many spatial resolutions but only convolutional layer per block.

For HR biases (Figures S6-S7), the most important hyperparameter is again spectral complexity. The relationship between spectral complexity and near-surface HR bias for multi-layer liquid-only cloud (Figure S7) is similar to the above-mentioned relationship between spectral complexity and unsigned errors. Specifically, absolute bias decreases as spectral complexity increases up to 64, suggesting that the optimal value is $\gtrsim$ 64. However, the relationship between spectral complexity and column-averaged HR bias (Figures S6) is quite different, suggesting that the optimal spectral complexity is $\sim$8. In other words, making unbiased predictions of HR in general requires much less spectral complexity than making unbiased predictions of near-surface HR under multi-layer cloud, which is a more difficult problem.

Based on all 14 shortwave error metrics (Table 6 of the main text), we select as "best" the U-net++ trained without deep supervision, with a depth of 3, width of 1, and spectral complexity of 128. The best model achieves the following ranks (1$^{\text{st}}$ being the best and 276$^{\text{th}}$ being the worst) on metrics for all profiles, in the order that they appear in Table 6: 1$^{\text{st}}$, 120$^{\text{th}}$, 9$^{\text{th}}$, 24$^{\text{th}}$, 1$^{\text{st}}$, 1$^{\text{st}}$, and 85$^{\text{th}}$. The model achieves the following ranks on metrics for profiles with multi-layer cloud, in the order that they appear in Table 6: 7$^{\text{th}}$, 93$^{\text{rd}}$, 12$^{\text{th}}$, 53$^{\text{rd}}$, 1$^{\text{st}}$, 1$^{\text{st}}$, 100$^{\text{th}}$. The model contains 33 240 174 ($10^{7.52}$) learned weights, making it one of the more complex models attempted (Figure S10).

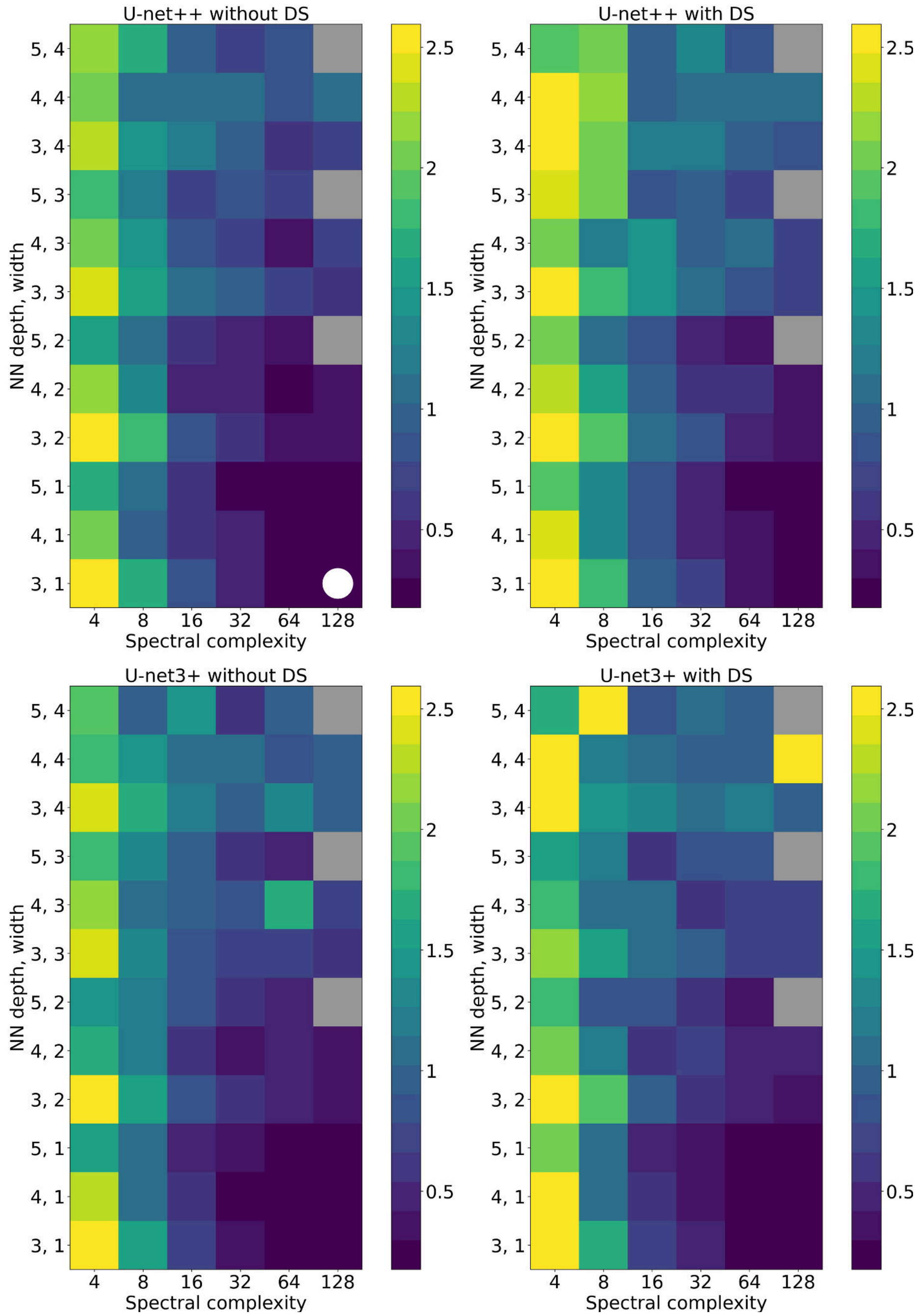

Figure S4: Column-averaged DWMSE for HR on all profiles (K^3 day^{-3}), computed on validation data for each set of hyperparameters. Each panel shows one NN type; within each panel the other three hyperparameters vary. Grey squares correspond to NNs that could not be trained. The white circle marks the selected model, and the white star (hidden behind the white circle) marks the model with the lowest value for this error metric.

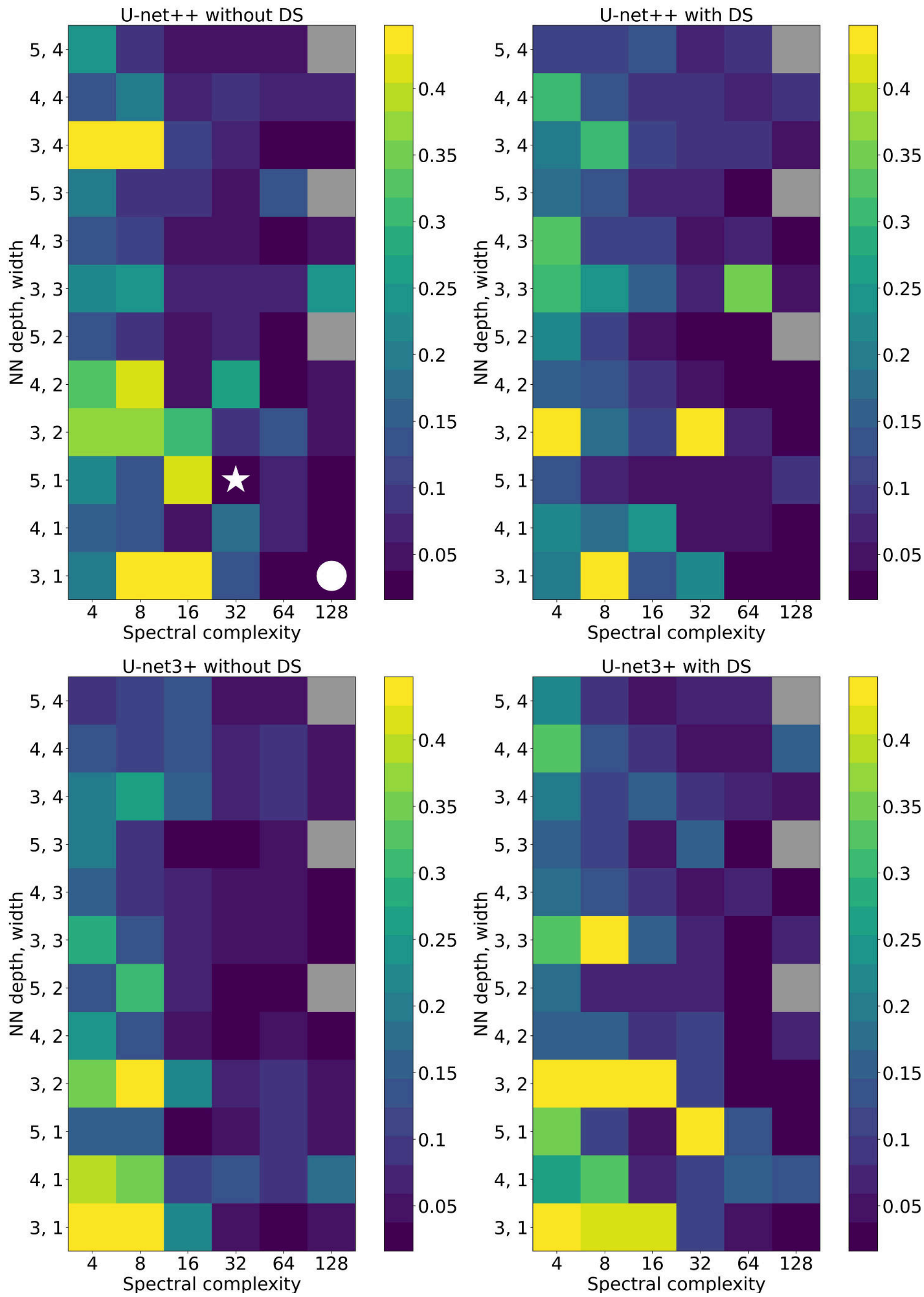

Figure S5: DWMSE for near-surface HR on profiles with multi-layer liquid-only cloud (K^3 day^{-3}), computed on validation data for each set of hyperparameters. The white circle marks the selected model, and the white star marks the model with the lowest value for this error metric. Other formatting is explained in the caption of Figure S4.

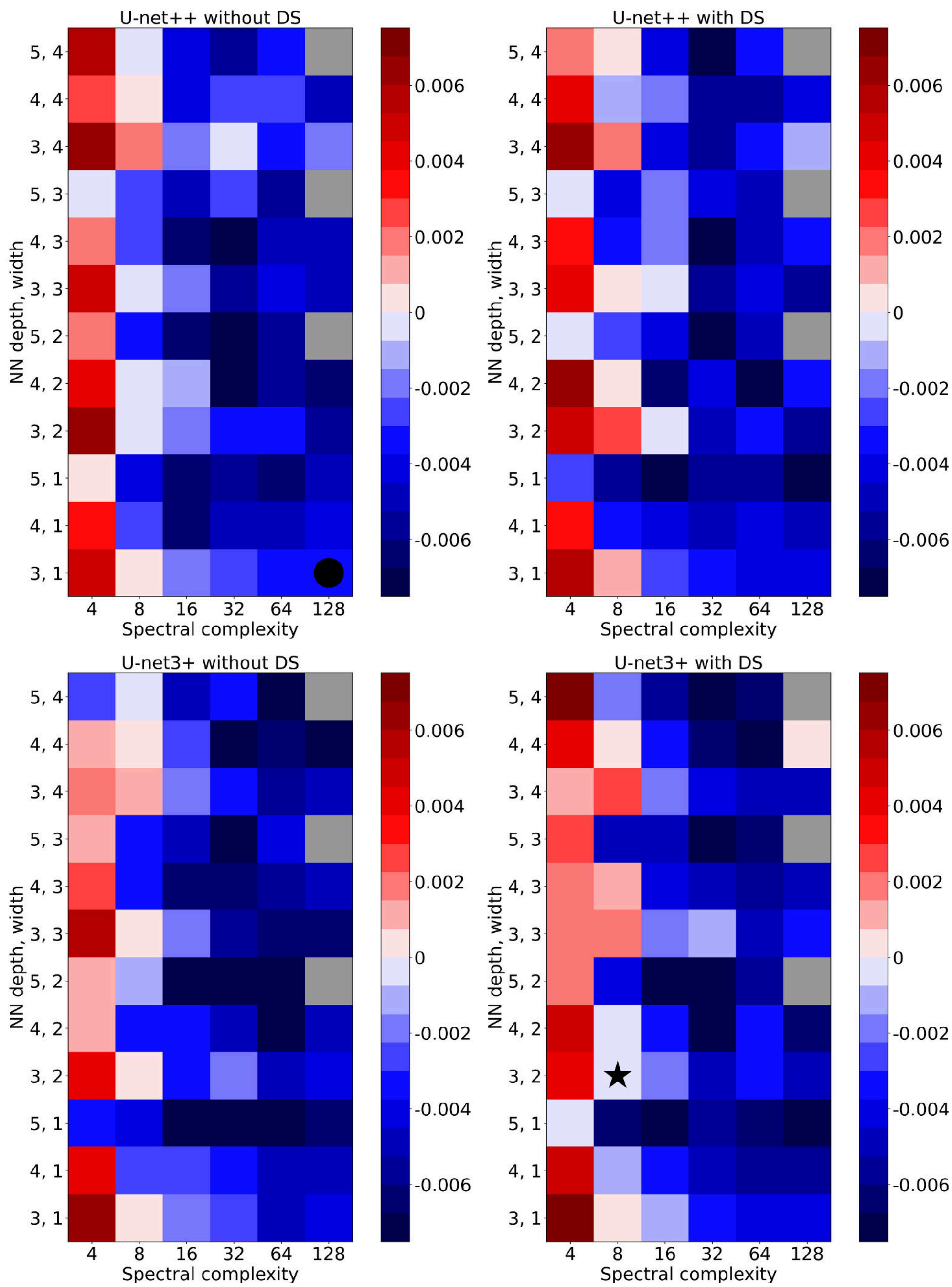

Figure S6: Column-averaged HR bias for all profiles (K day^{-1}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S4.

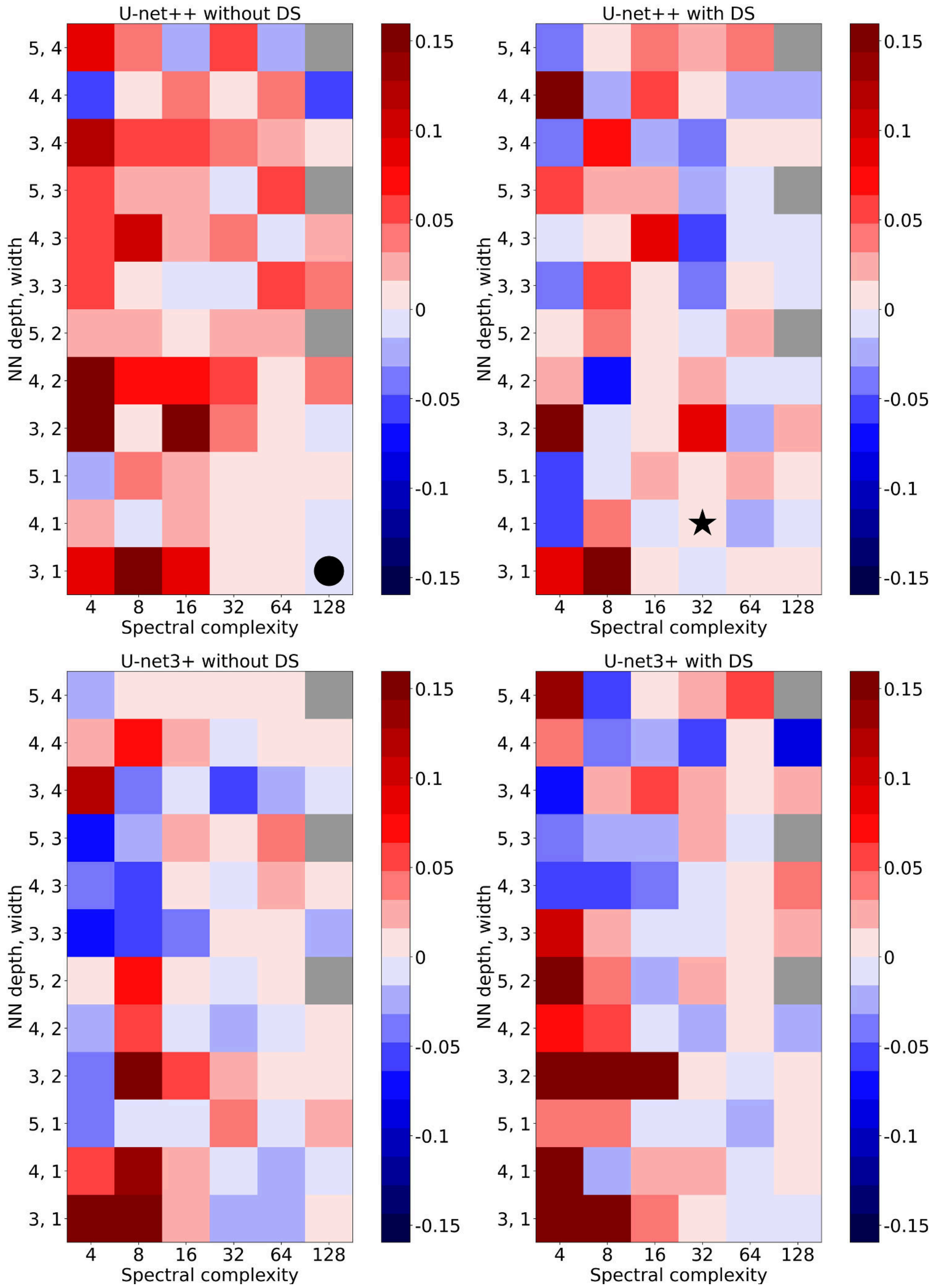

Figure S7: Near-surface HR bias for profiles with multi-layer liquid-only cloud (K day^{-1}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S4.

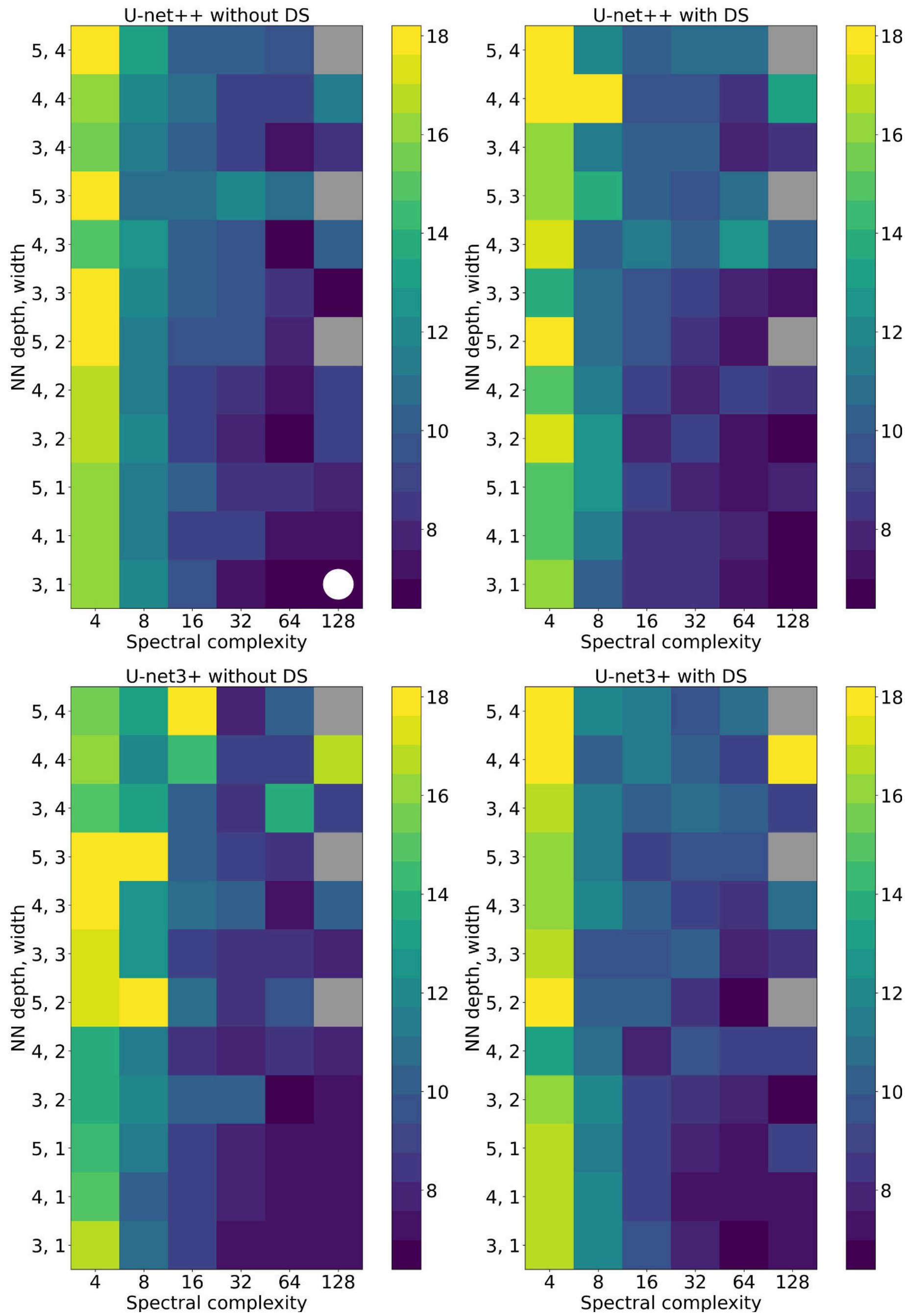

Figure S8: Net-flux RMSE for all profiles (W m^{-2}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S4.

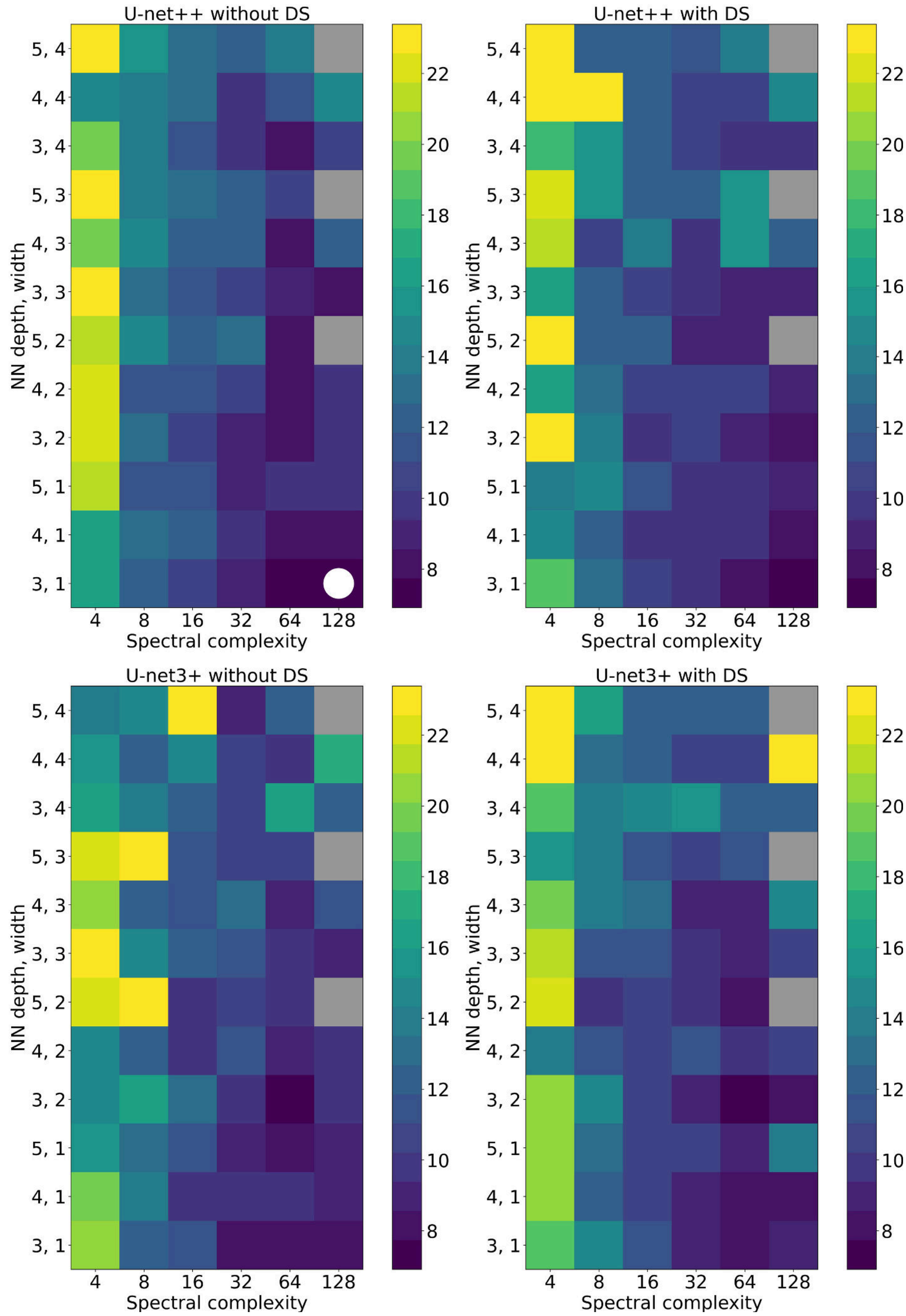

Figure S9: Net-flux RMSE for profiles with multi-layer liquid-only cloud (W m^{-2}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S4.

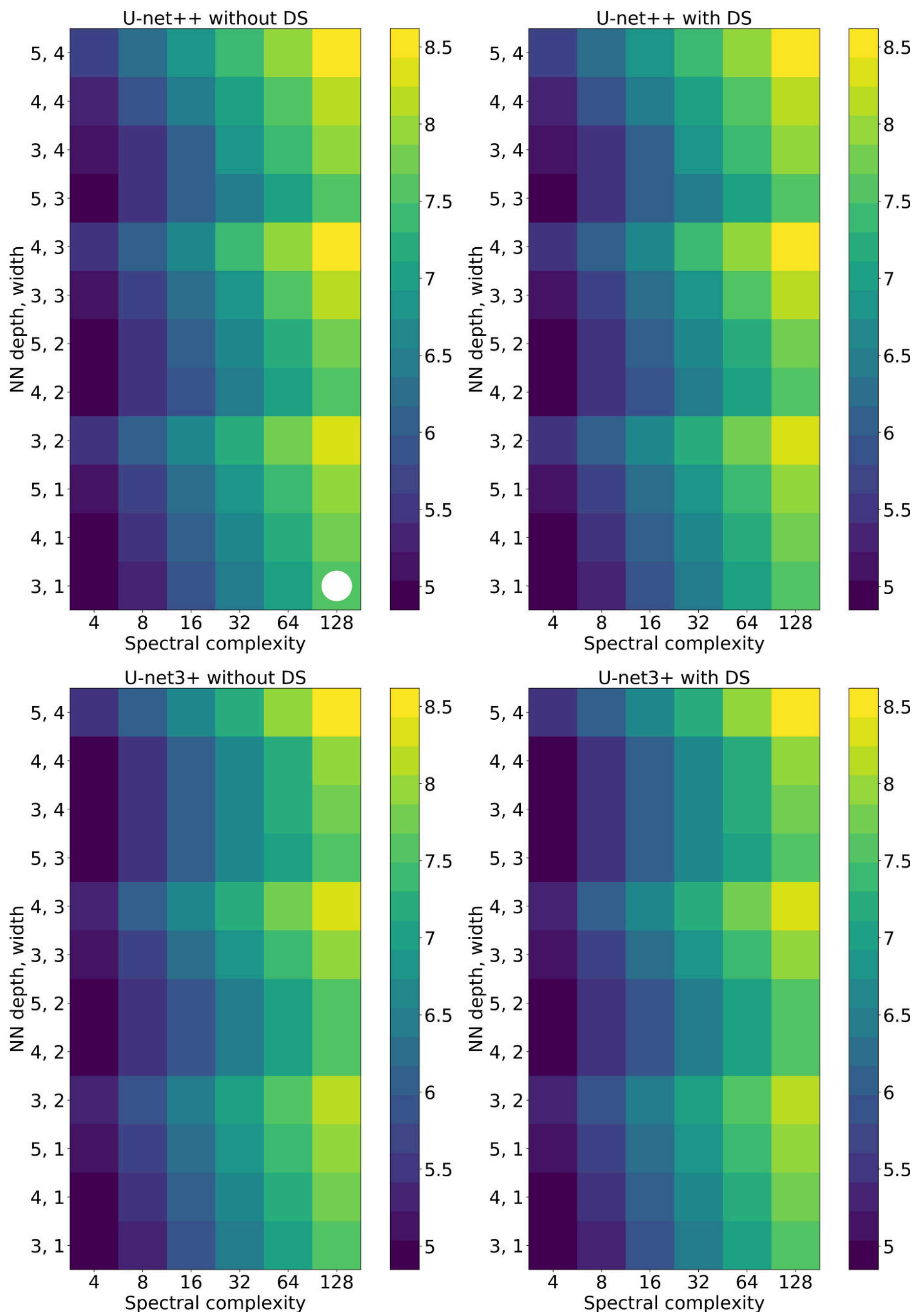

Figure S10: Number of trainable model weights for each set of hyperparameters, in $\log_{10}$ scale. The white circle marks the selected model. Other formatting is explained in the caption of Figure S4.

b. Results for longwave RT

Figures S11-S17 show validation error vs. hyperparameters for a few metrics listed in Table 6 of the main text. As for the shortwave hyperparameter experiment, 12 of 288 NNs could not be trained, due to memory issues – see grey squares in Figures S11-S17. Our broad conclusions for the shortwave experiment (Section 3a) hold for the longwave experiment as well. Specifically, the most important hyperparameter is spectral complexity, with an optimal value of $\gtrsim$64; NN width and depth are of secondary importance, with narrow and deep networks performing best; and NN type appears to be unimportant.

Based on all 19 longwave error metrics, we select as "best" the U-net3+ trained without deep supervision, with a depth of 5, width of 1, and spectral complexity of 64. This model achieves the following ranks (1^{st} being the best and 276^{th} being the worst) on metrics for all profiles, in the order that they appear in Table 6: 1^{st}, 14^{th}, 1^{st}, 16^{th}, 2^{nd}, 2^{nd}, and 83^{rd}. The model achieves the following ranks on metrics for profiles with multi-layer liquid-only cloud, in the order that they appear in Table 6: 1^{st}, 52^{nd}, 1^{st}, 24^{th}, 2^{nd}, 3^{rd}, 86^{th}. Finally, the model achieves the following ranks on metrics for profiles with liquid-only fog, in the order that they appear in Table 6: 1^{st}, 17^{th}, 3^{rd}, 4^{th}, 102^{th}. The model contains 19 189 566 ($10^{7.28}$) learned weights, making it one of the more complex models attempted (Figure S18).

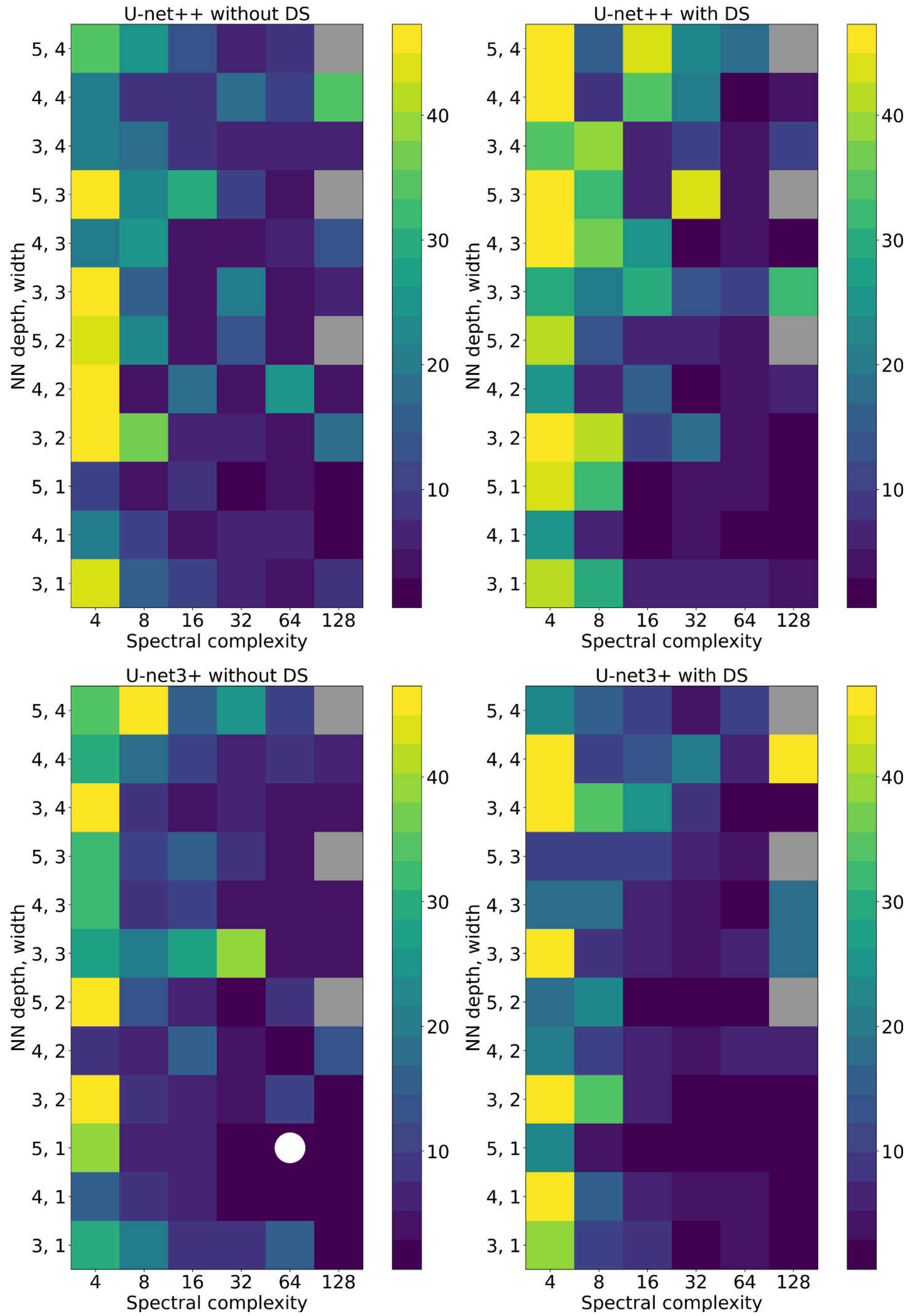

Figure S11: Column-averaged DWMSE for HR on all profiles (K^3 day^{-3}), computed on validation data for each set of hyperparameters. Each panel shows one NN type; within each panel the other three hyperparameters vary. Grey squares correspond to NNs that could not be trained. The white circle marks the selected model, and the white star (hidden behind the white circle) marks the model with the lowest value for this error metric.

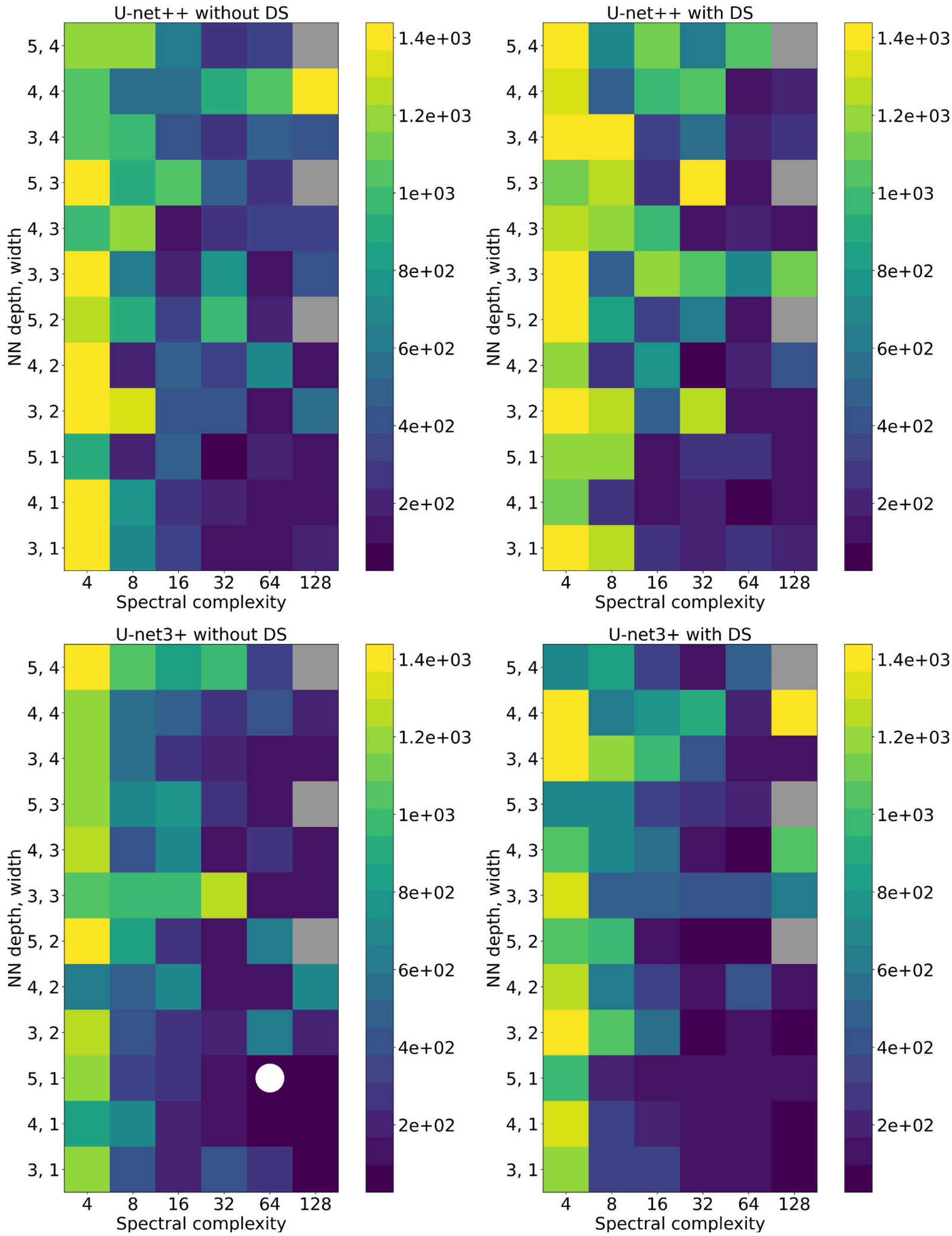

Figure S12: DWMSE for near-surface HR on profiles with multi-layer liquid-only cloud (K^3 day^{-3}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S11.

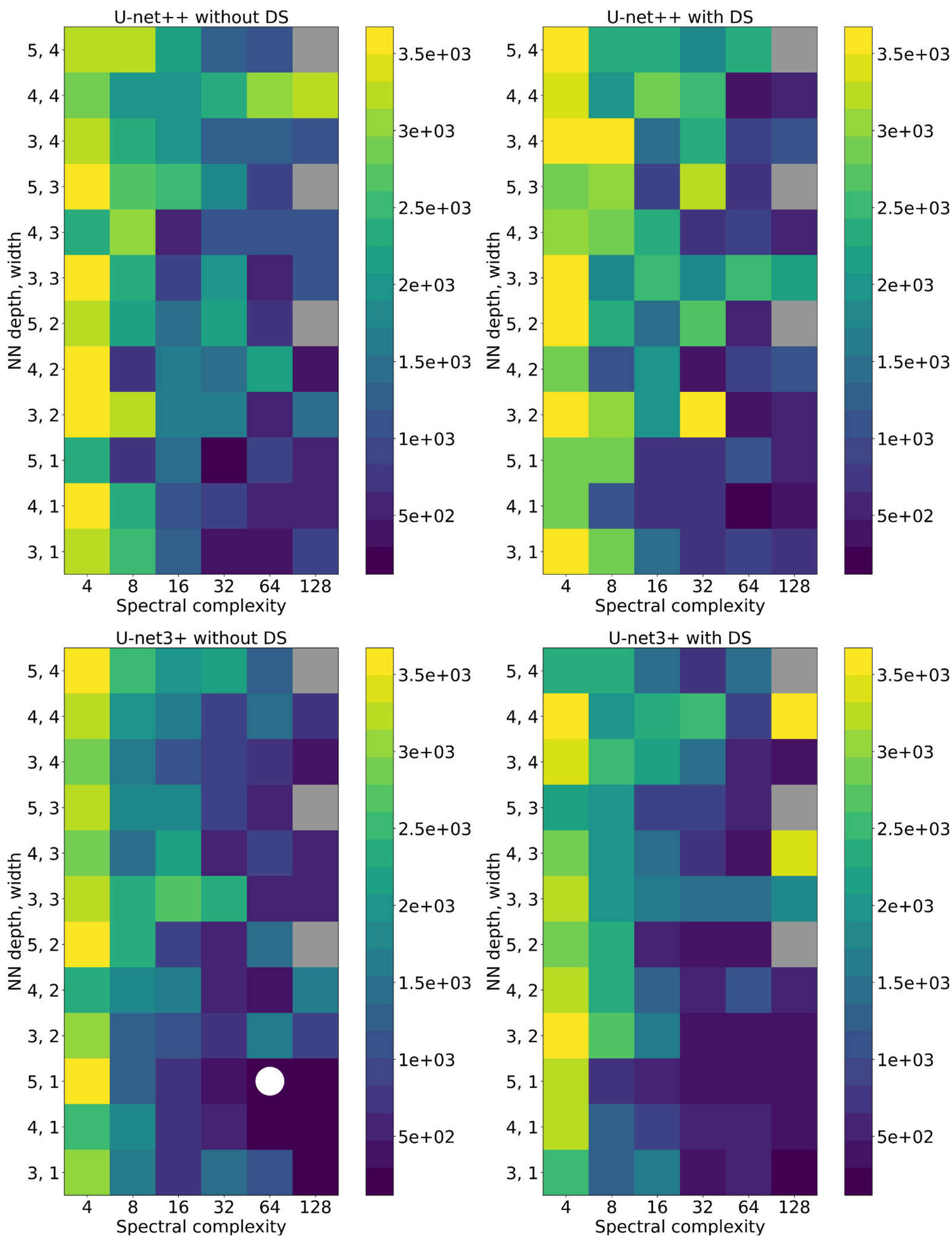

Figure S13: DWMSE for near-surface HR on profiles with liquid-only fog (K^3 day^{-3}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S11.

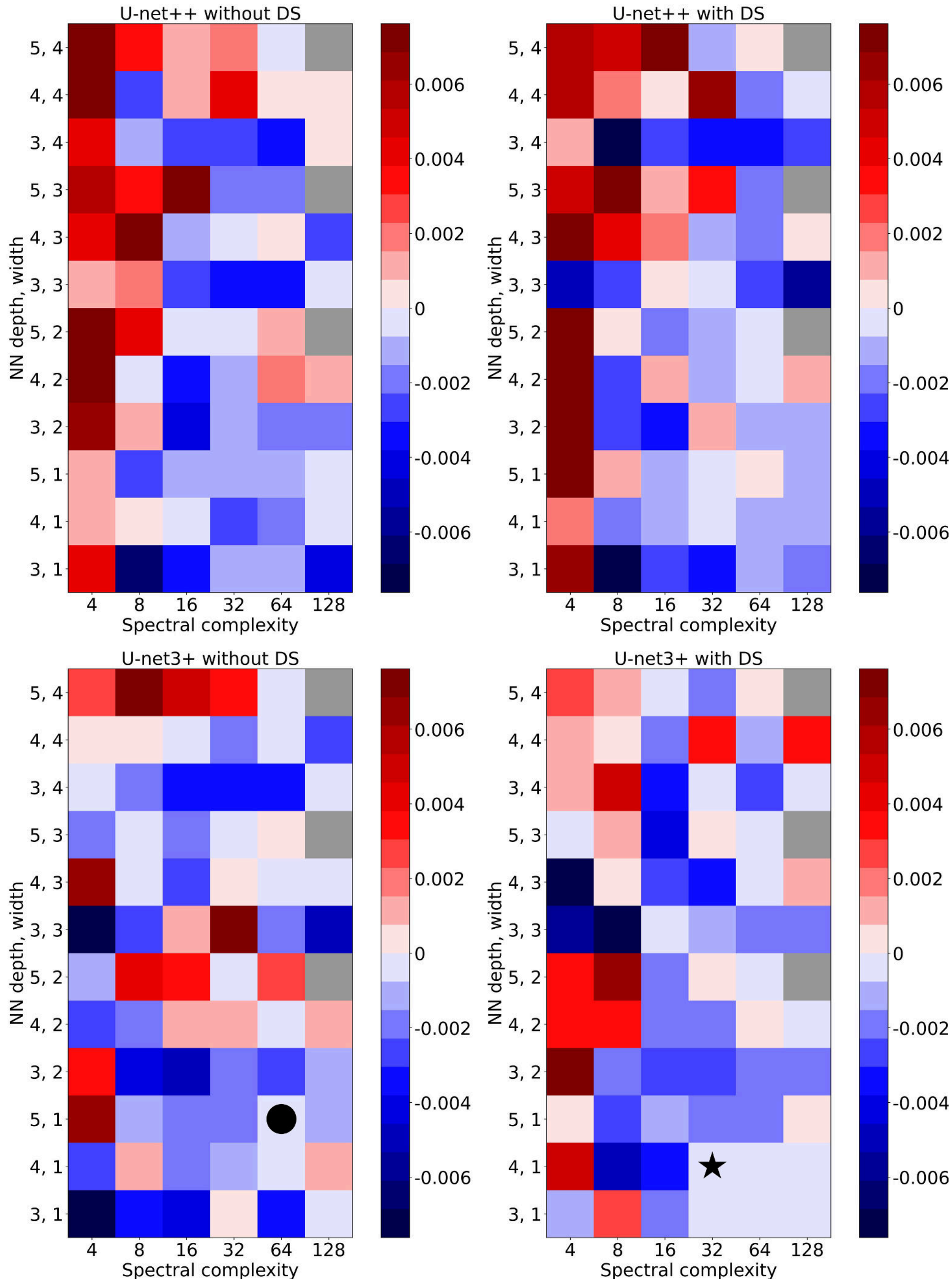

Figure S14: Column-averaged HR bias for all profiles (K day^{-1}), computed on validation data for each set of hyperparameters. The black circle marks the selected model, and the black star marks the model with the lowest value for this error metric. Other formatting is explained in the caption of Figure S11.

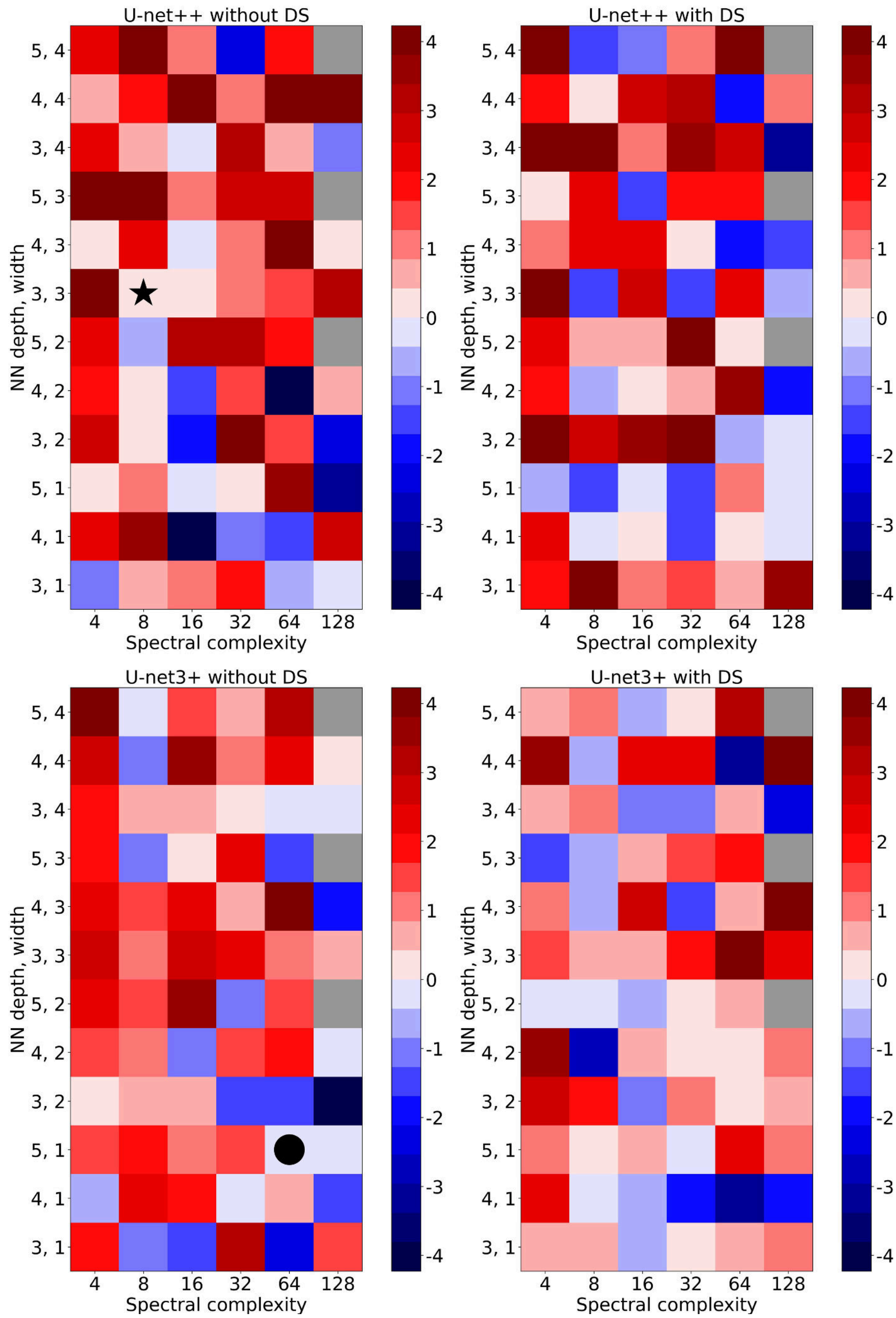

Figure S15: Near-surface HR bias for profiles with liquid-only fog (K day^{-1}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S11.

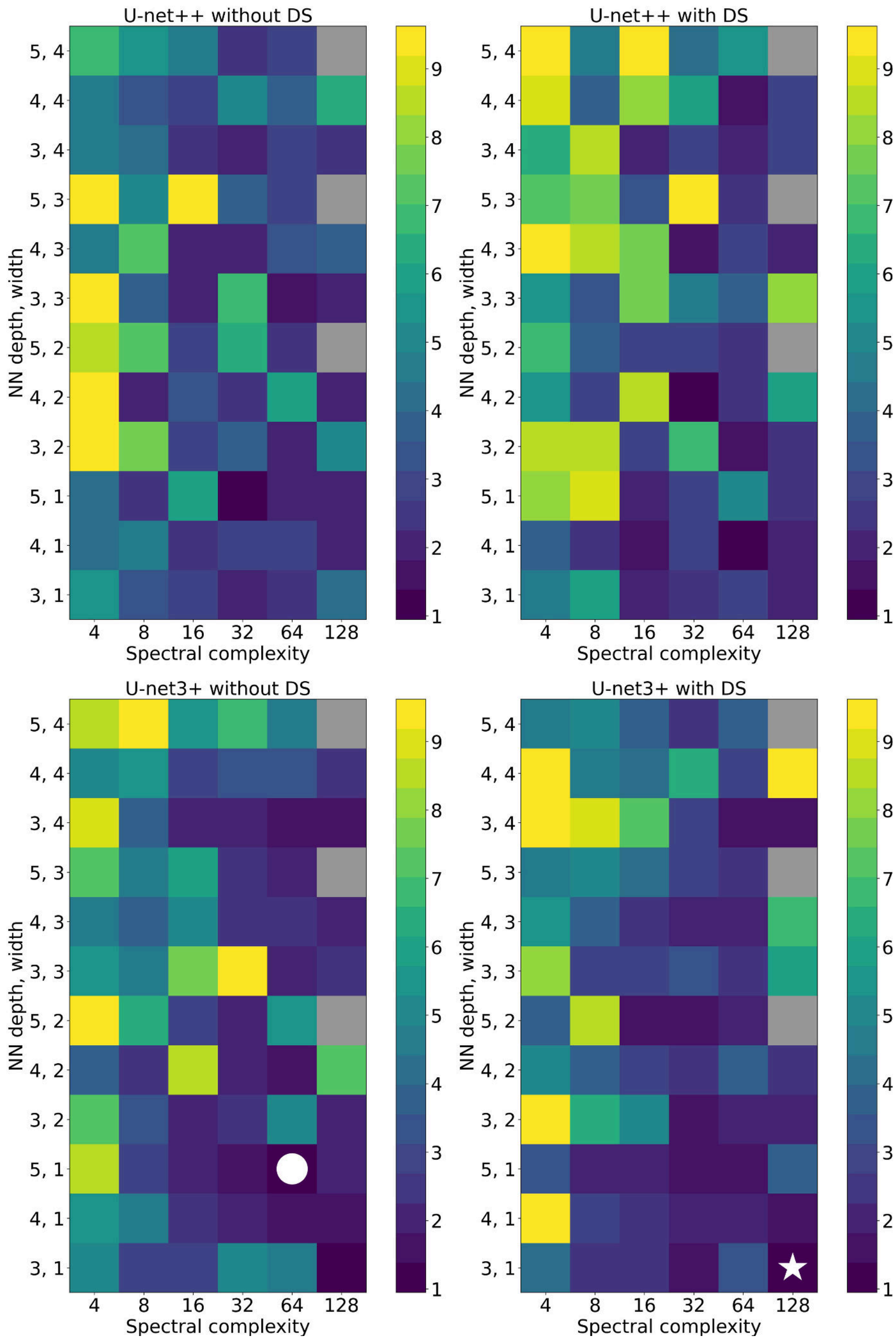

Figure S16: Net-flux RMSE for all profiles (W m^{-2}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S11.

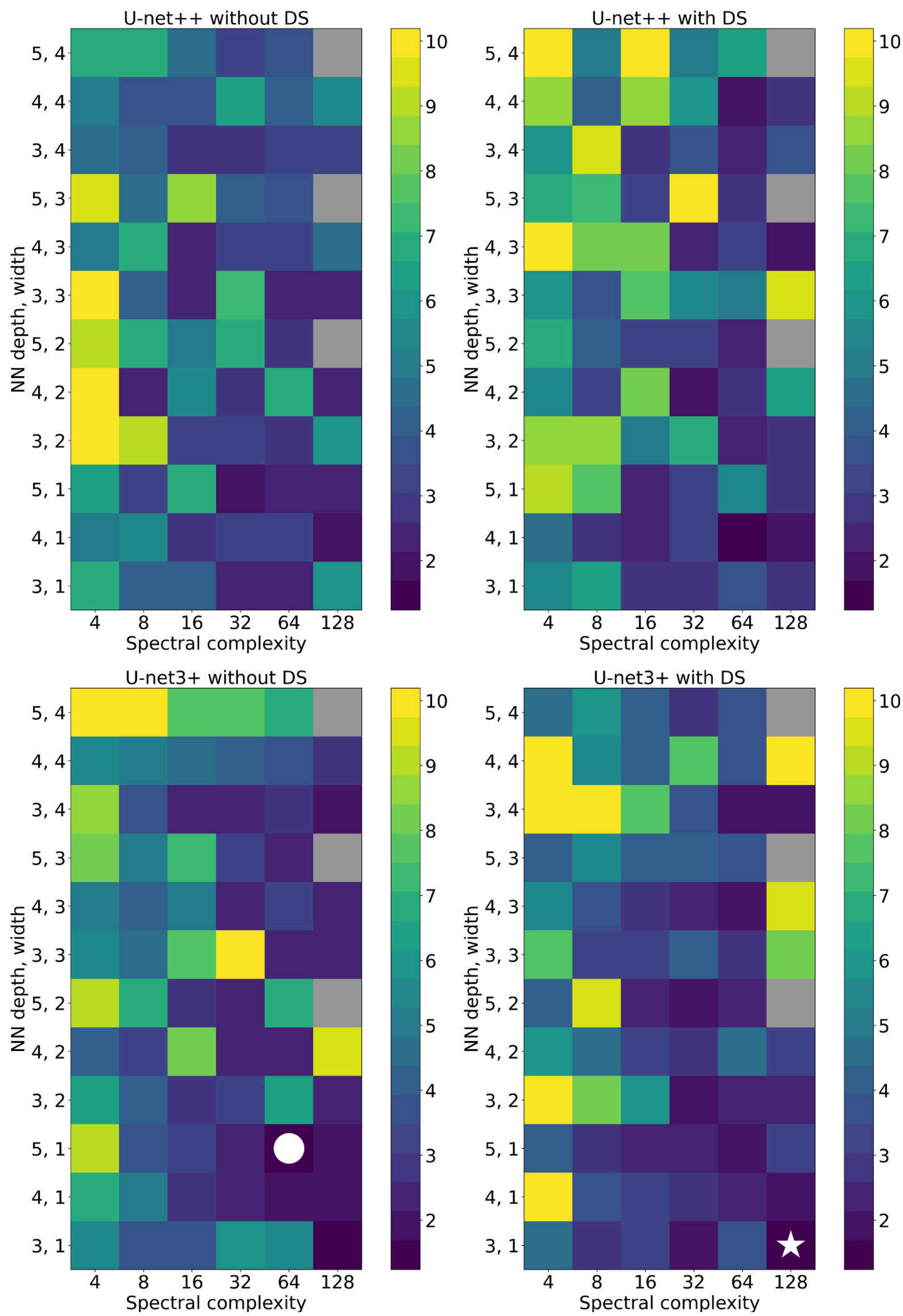

Figure S17: Net-flux RMSE for profiles with liquid-only fog (W m^{-2}), computed on validation data for each set of hyperparameters. Formatting is explained in the caption of Figure S11.

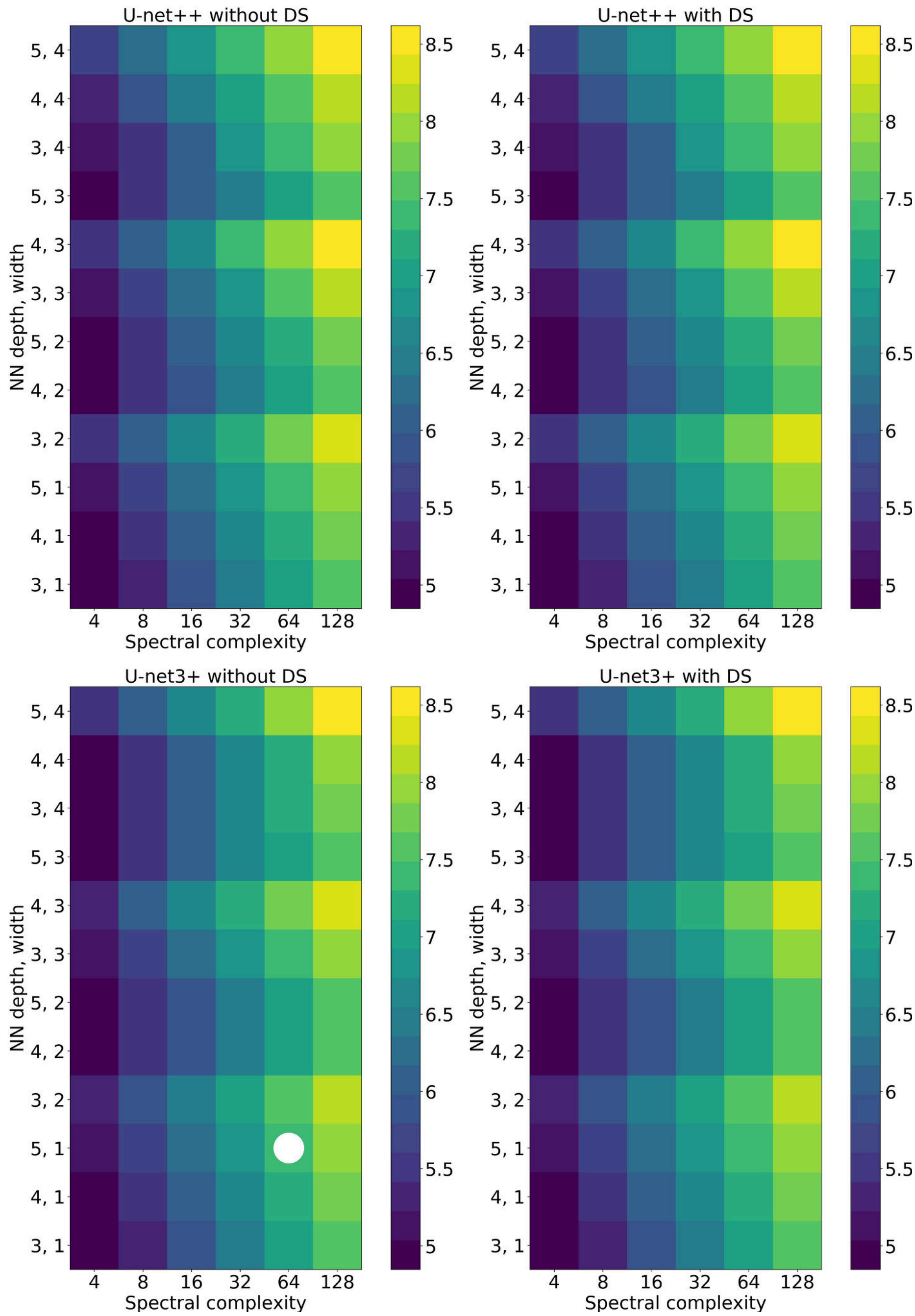

Figure S18: Number of trainable model weights for each set of hyperparameters, in $\log_{10}$ scale. The white circle marks the selected model. Other formatting is explained in the caption of Figure S11.

4. Extended analysis of best models

This section contains figures referenced in the main text, used for extended analysis of the best shortwave and longwave models. For ice-only cloud regimes, a cloud layer is defined as a contiguous set of model heights with ice-water content (IWC) > 0 g m^{-3}, ice-water path (IWP) $\geq$ 25 g m^{-2}, and liquid-water path (LWP) = 0 g m^{-2}. For mixed-phase cloud regimes, a cloud layer is defined as a contiguous set of heights with total water content (LWC + IWC) > 0 g m^{-3}, total water path (LWP + IWP) $\geq$ 25 g m^{-2}, LWP > 0 g m^{-2}, and IWP > 0 g m^{-2}. To put the last two criteria in plain language, the cloud must contain a non-zero amount of both liquid and ice, but the relative contributions do not matter. For any-phase cloud regimes, a cloud layer is defined as a contiguous set of heights with total water content > 0 g m^{-3} and total water path $\geq$ 25 g m^{-2}. The last two criteria have vanished, meaning that the cloud can be liquid-only, ice-only, or mixed-phase – hence "any-phase".

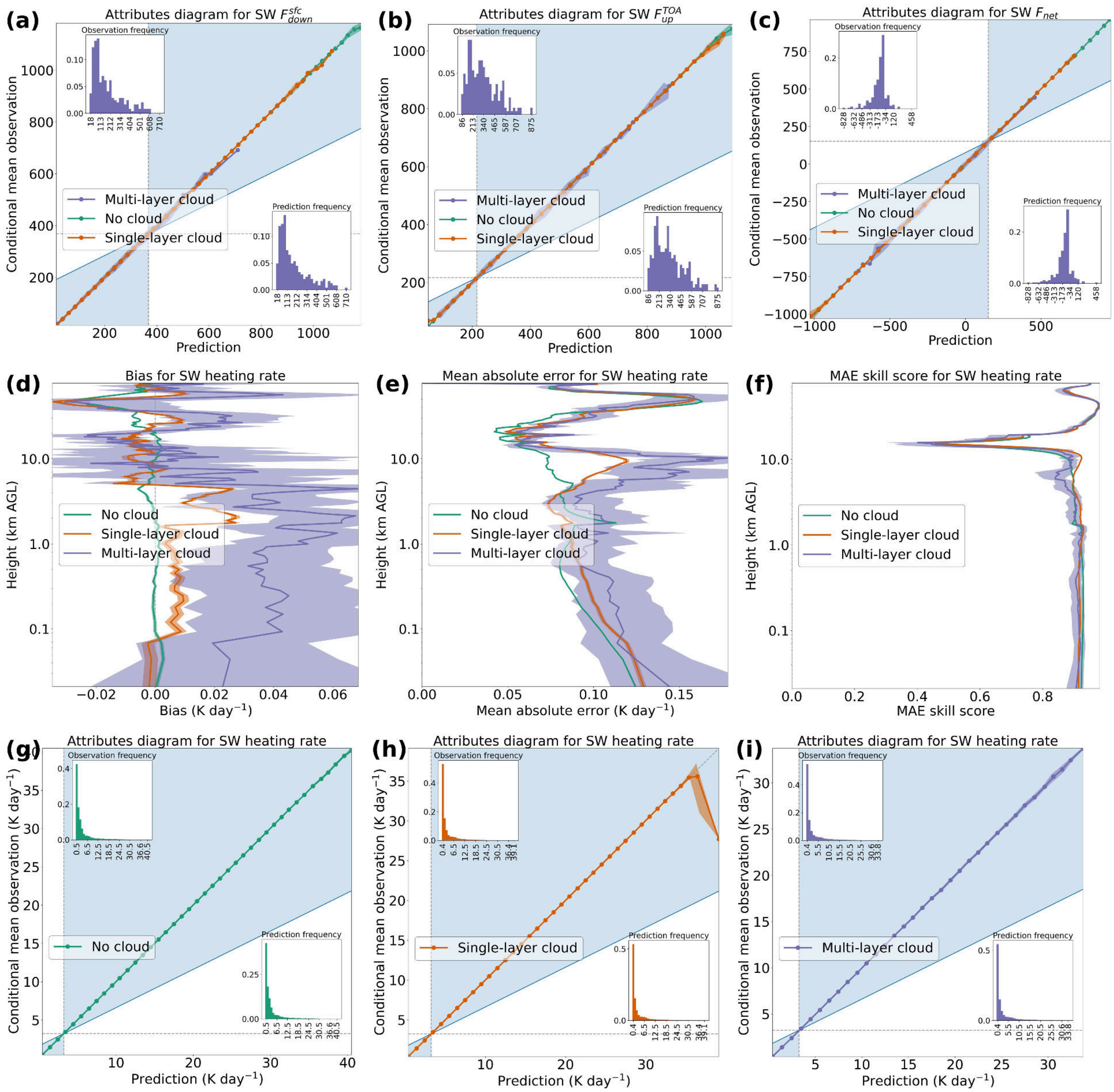

Figure S19: Performance of best shortwave model on testing data, separated by ice-only cloud regime. This is analogous to Figure 6 in the main text but concerns ice-only, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud} account for {89.28%, 10.66%, 0.05%} of the testing data respectively.

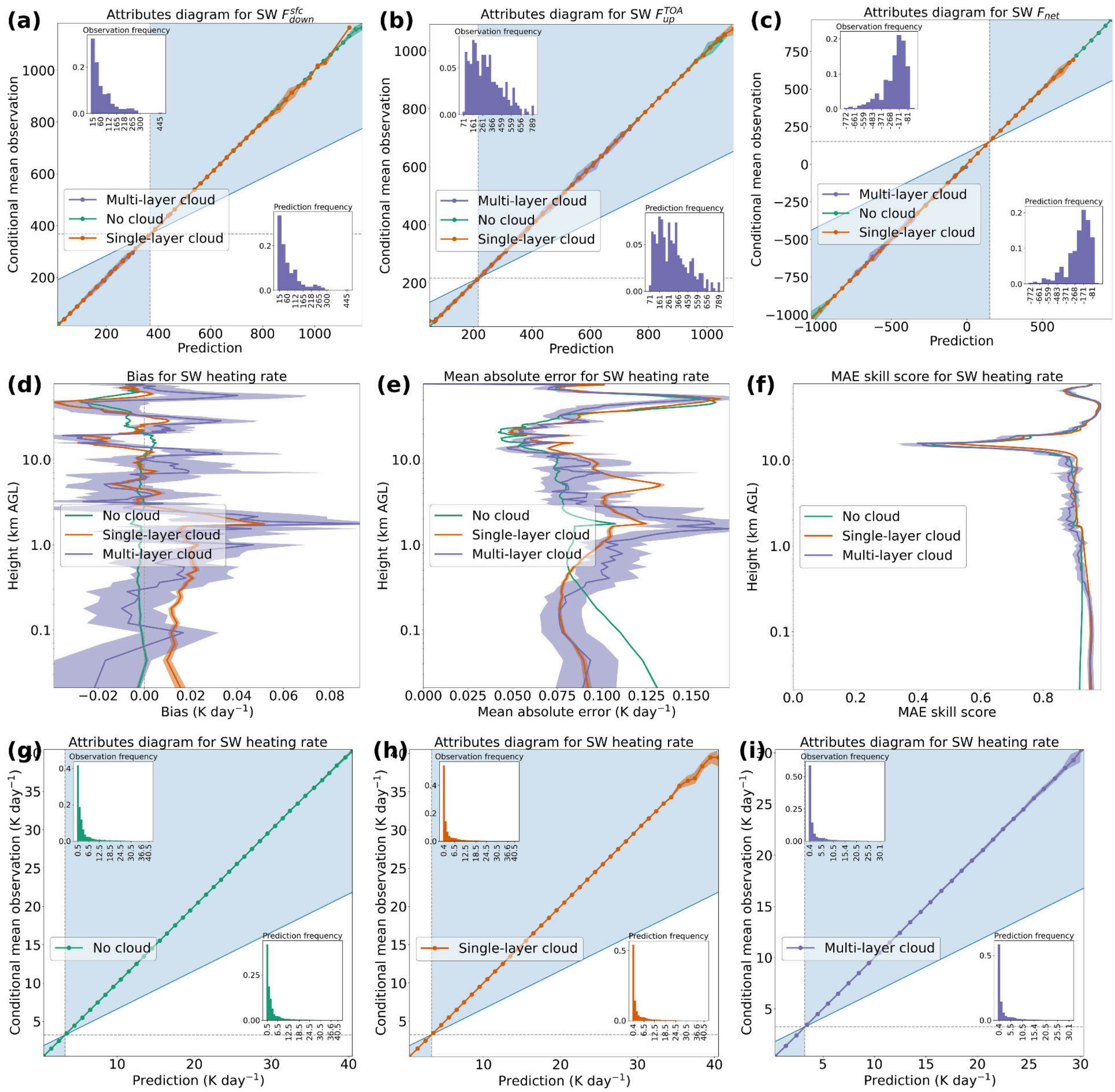

Figure S20: Performance of best shortwave model on testing data, separated by mixed-phase cloud regime. This is analogous to Figure 6 in the main text but concerns mixed-phase, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud} account for {84.50%, 15.44%, 0.07%} of the testing data respectively.

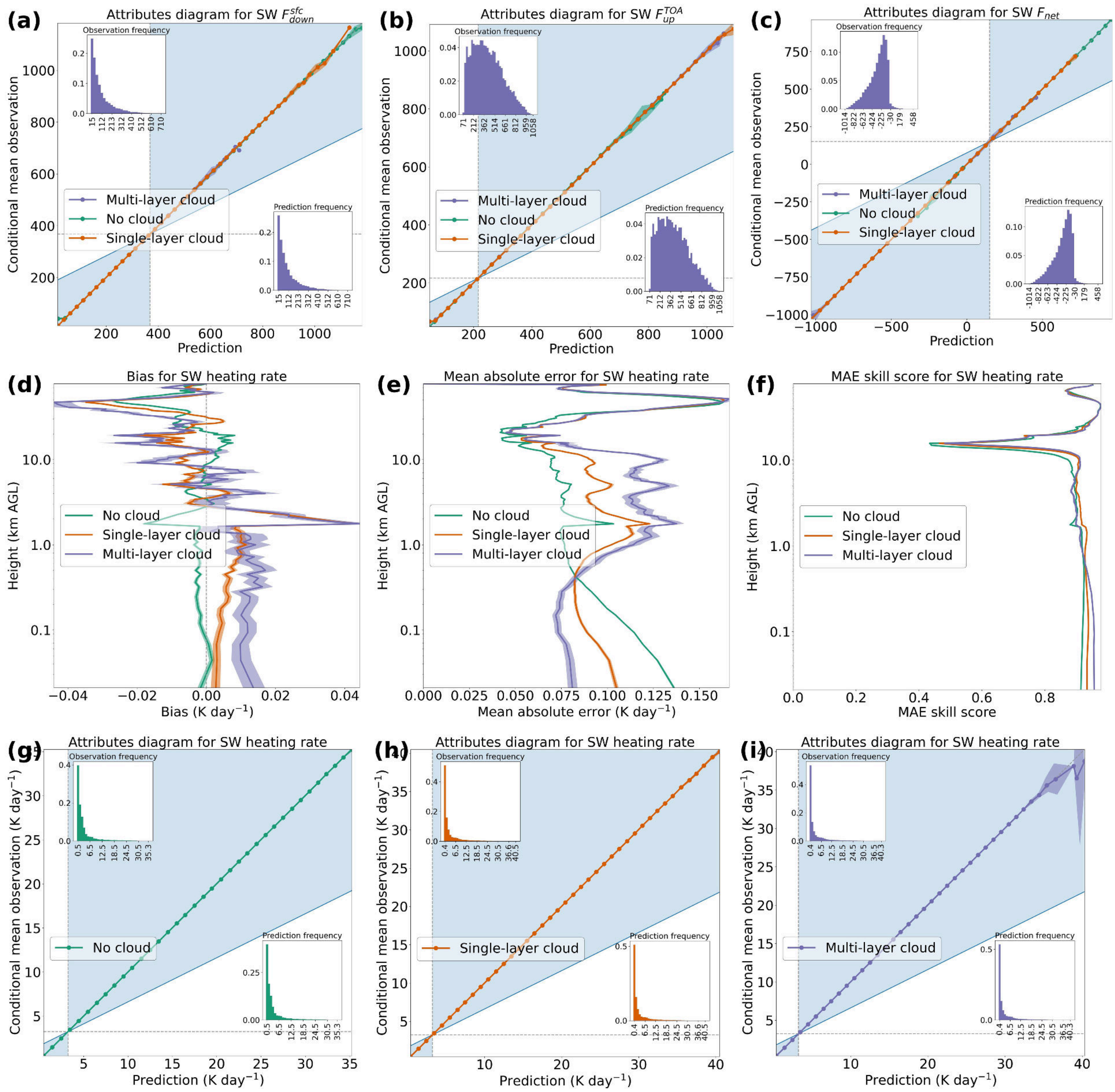

Figure S21: Performance of best shortwave model on testing data, separated by any-phase cloud regime. This is analogous to Figure 6 in the main text but concerns any-phase, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud} account for {66.20%, 30.93%, 2.86%} of the testing data respectively.

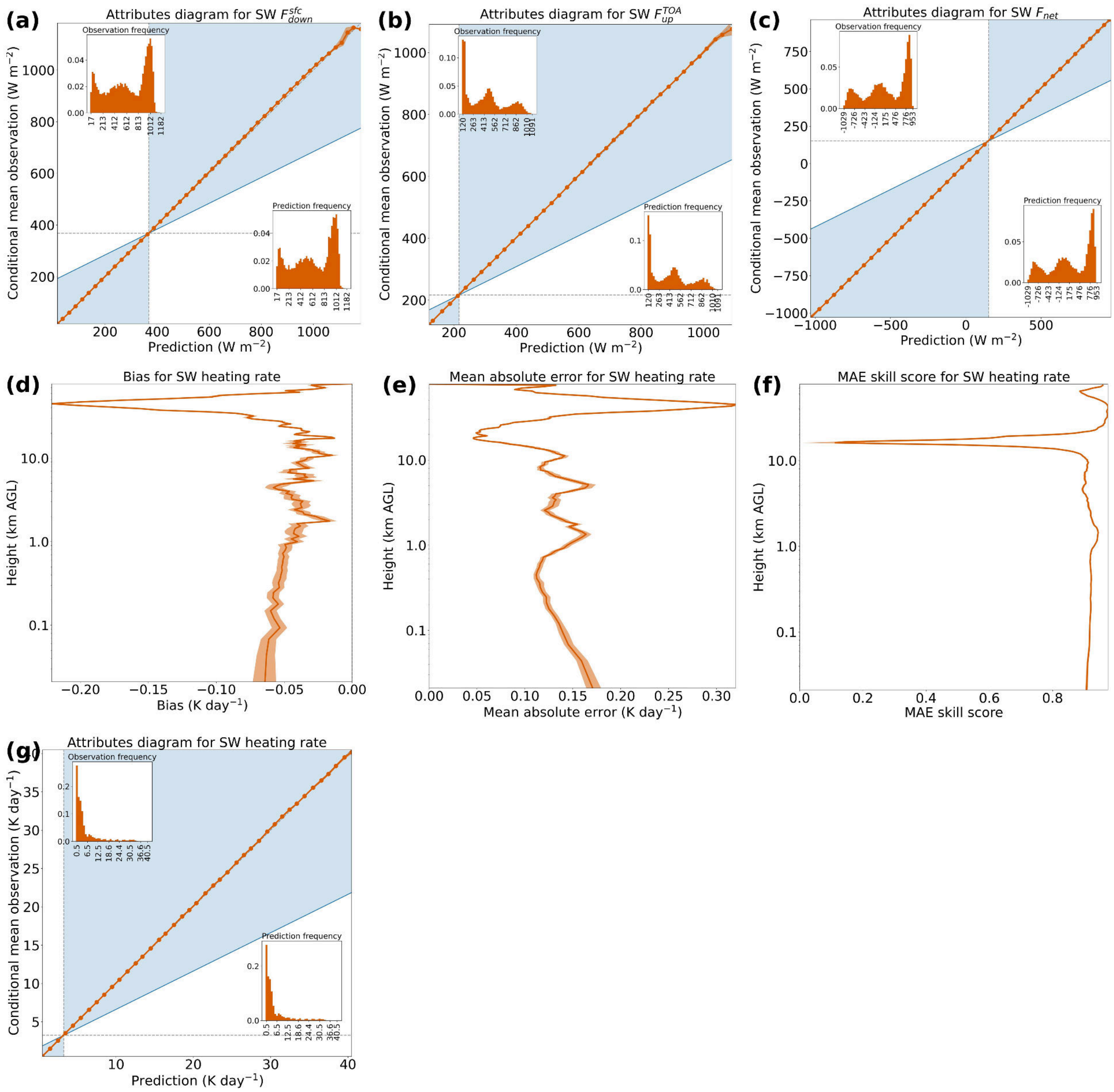

Figure S22: Performance of best shortwave model on single-height extremes, *i.e.*, on the 3% of testing cases with the greatest height-maximum HR. This is analogous to Figure 5 in the main text.

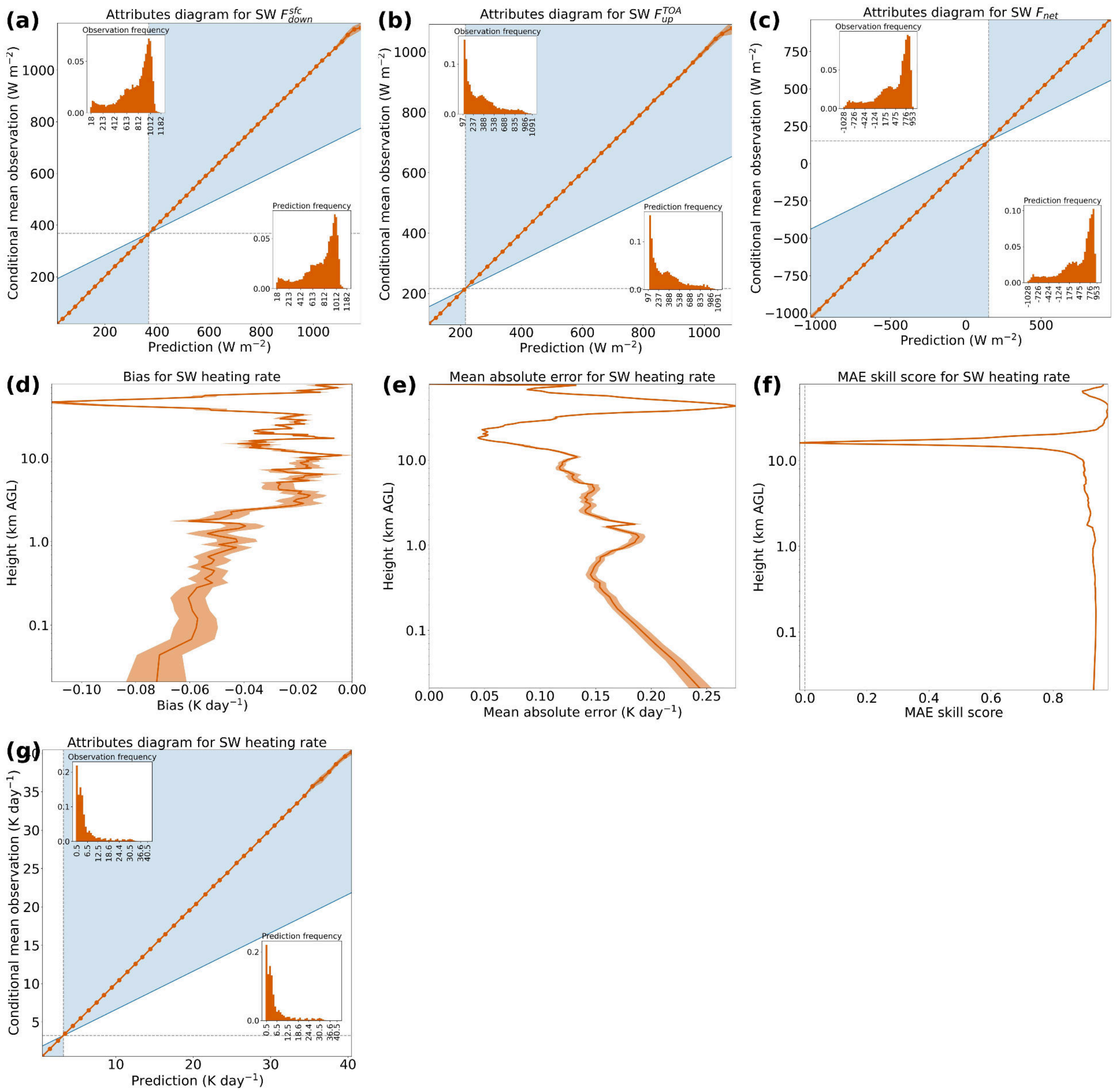

Figure S23: Performance of best shortwave model on full-profile extremes, *i.e.*, on the 3% of testing cases with the greatest height-averaged HR. This is analogous to Figure 5 in the main text.

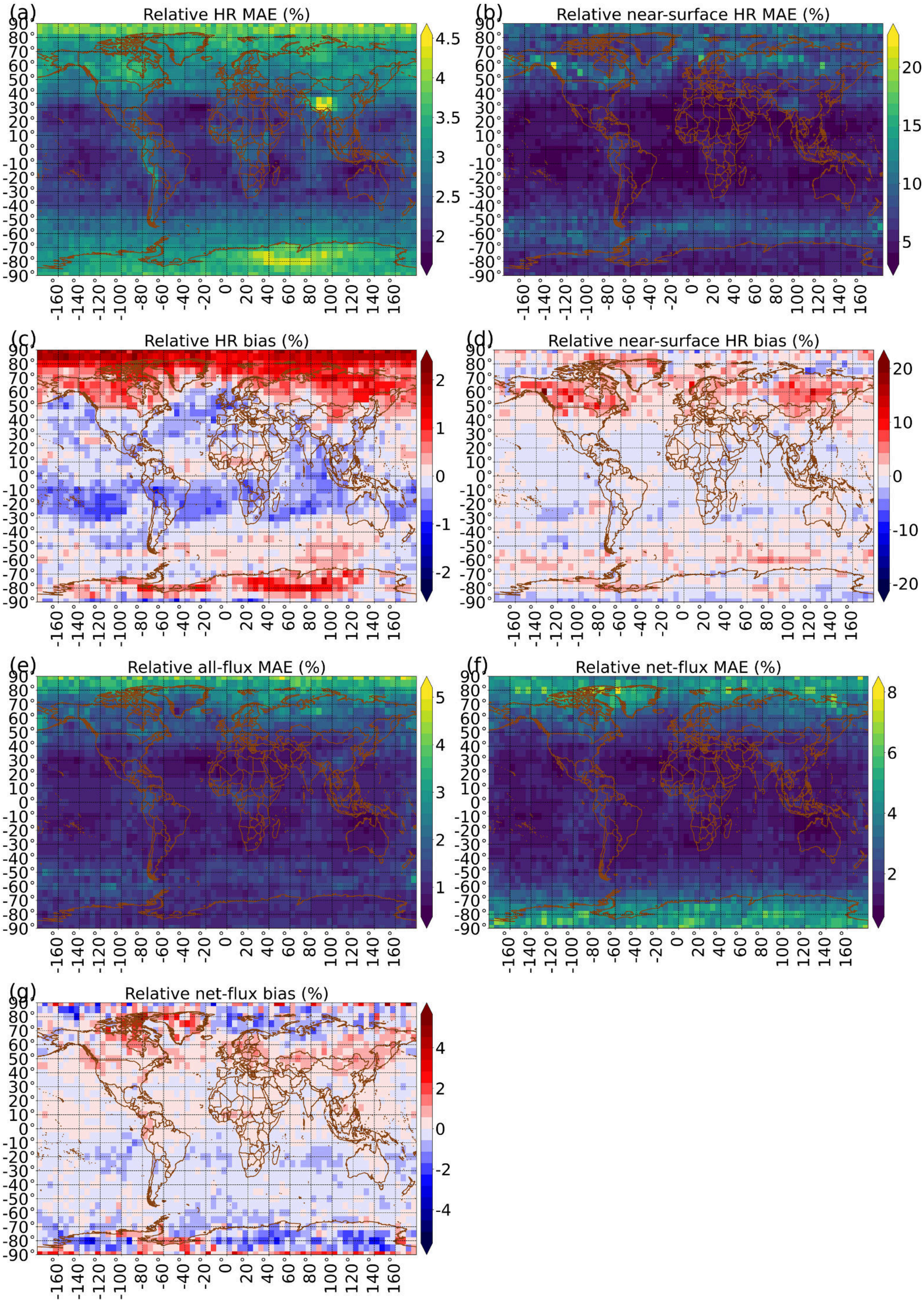

Figure S24: Fractional errors for best shortwave model on testing data, binned by geographic location. This is analogous to Figure 7 in the main text but shows fractional errors instead of raw errors.

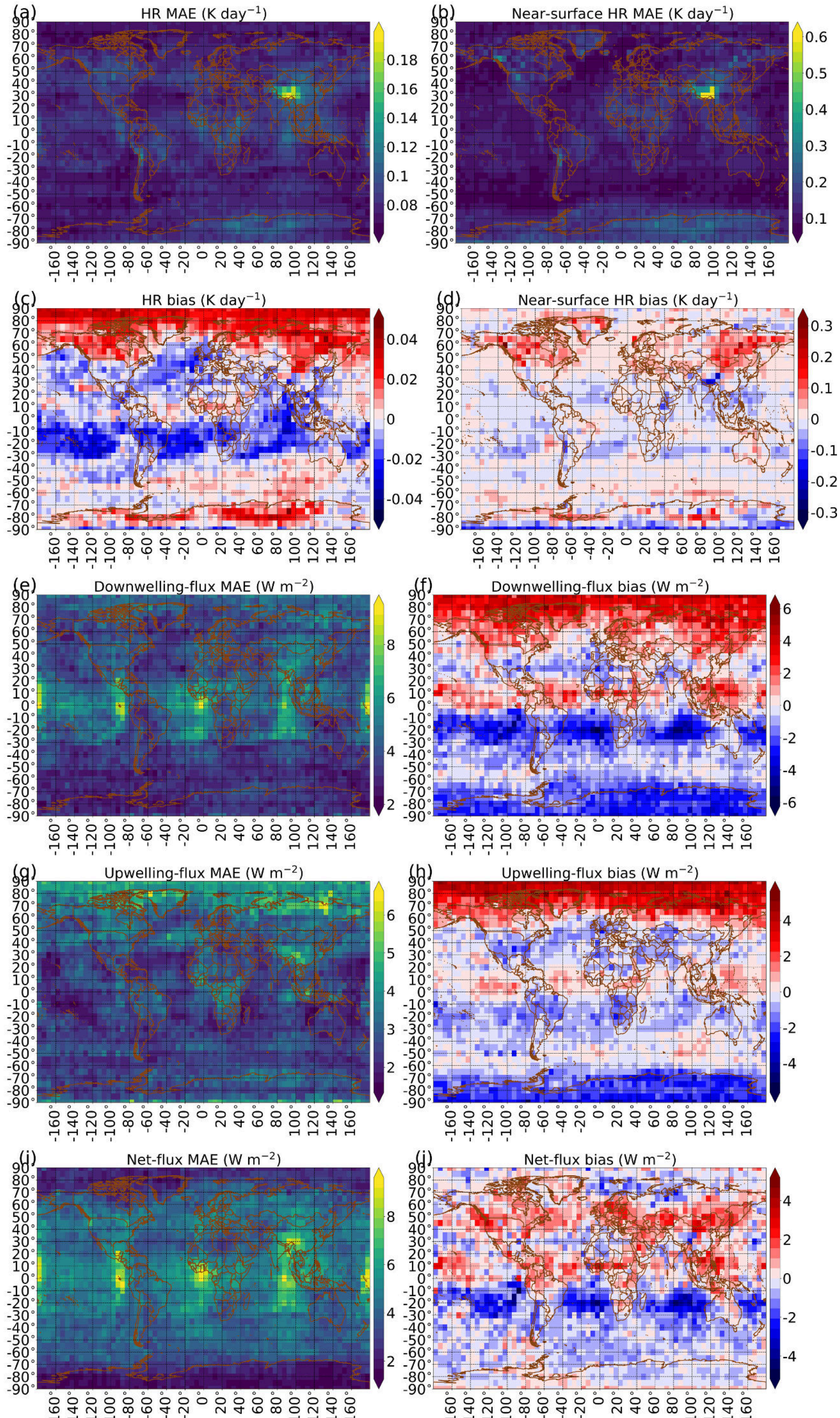

Figure S25: Detailed errors for best shortwave model on testing data, binned by geographic location. This is analogous to Figure 7 in the main text but shows errors for individual flux variables – $F_{\text{down}}^{\text{sfc}}$ in panels e-f, $F_{\text{up}}^{\text{TOA}}$ in panels g-h, and F_{net} in panels i-j – rather than averaging to produce all-flux MAE.

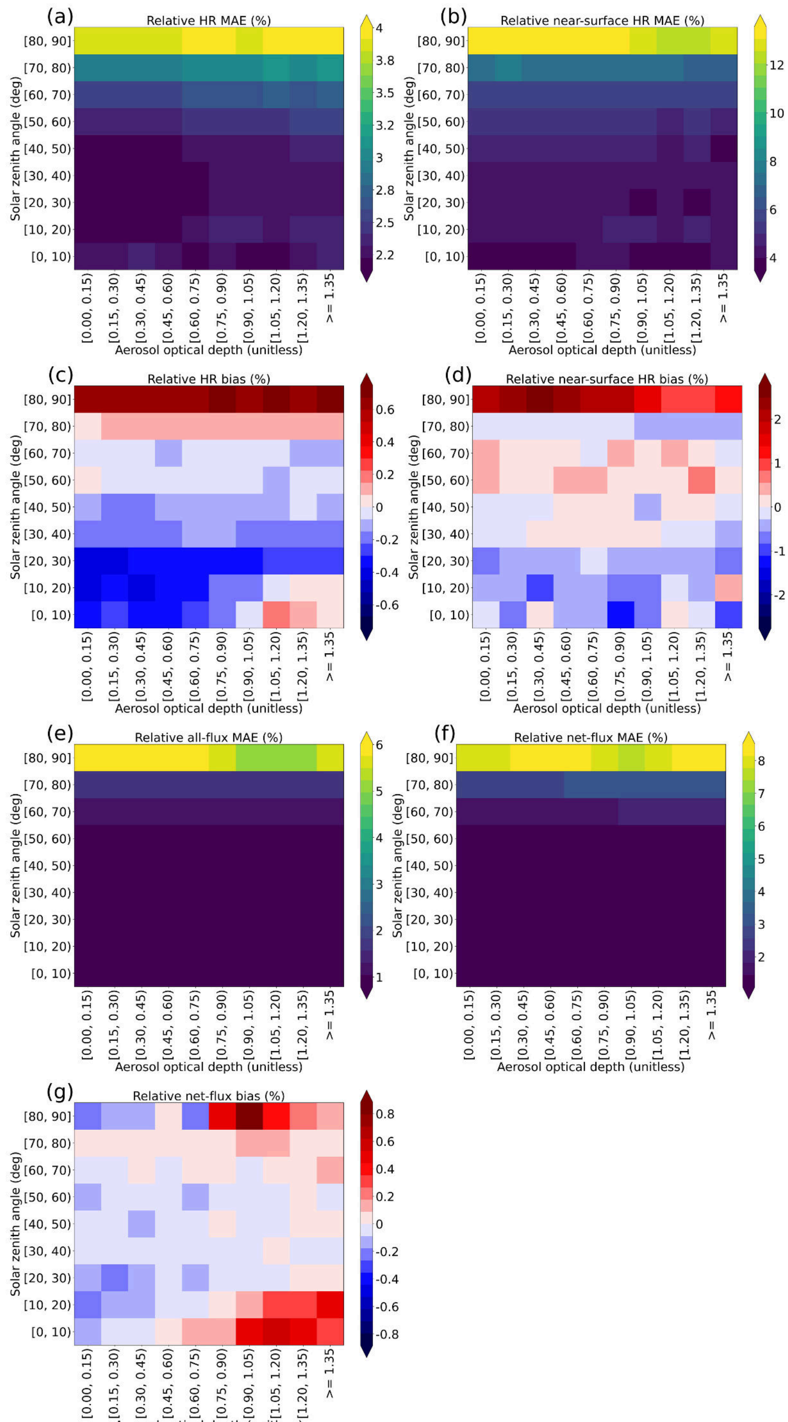

Figure S26: Fractional errors for best shortwave model on testing data, binned by aerosol optical depth (AOD) and solar zenith angle (SZA). This is analogous to Figure 9 in the main text but shows fractional errors instead of raw errors.

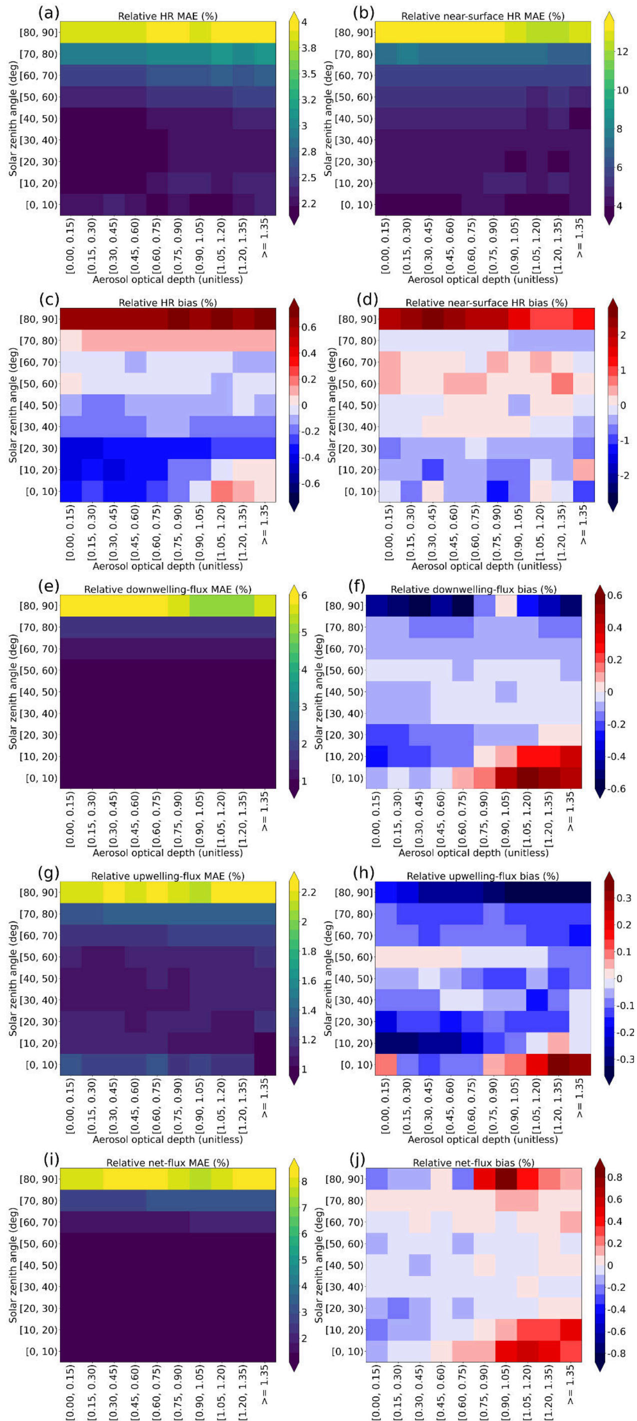

Figure S27: Detailed fractional errors for best shortwave model on testing data, binned by AOD and SZA. This is analogous to Figure S26 but shows errors for individual flux variables – $F_{\mathrm{down}}^{\mathrm{sfc}}$ in panels e-f, $F_{\mathrm{up}}^{\mathrm{TOA}}$ in panels g-h, and F_{net} in panels i-j – rather than averaging to produce all-flux MAE.

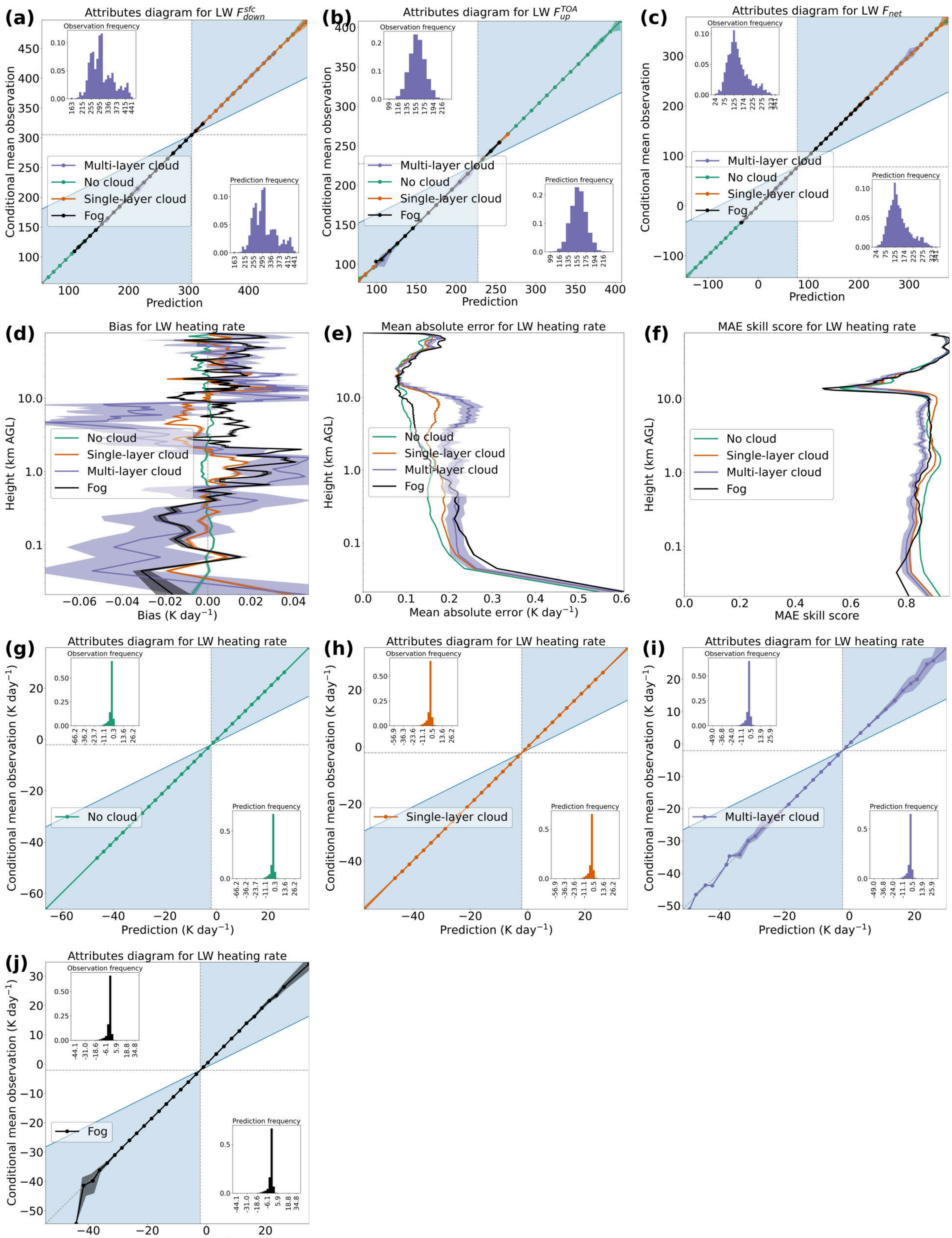

Figure S28: Performance of best longwave model on testing data, separated by ice-only cloud regime. This is analogous to Figure 12 in the main text but concerns ice-only, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud, fog} account for {87.62%, 12.32%, 0.06%, 3.93%} of the testing data respectively.

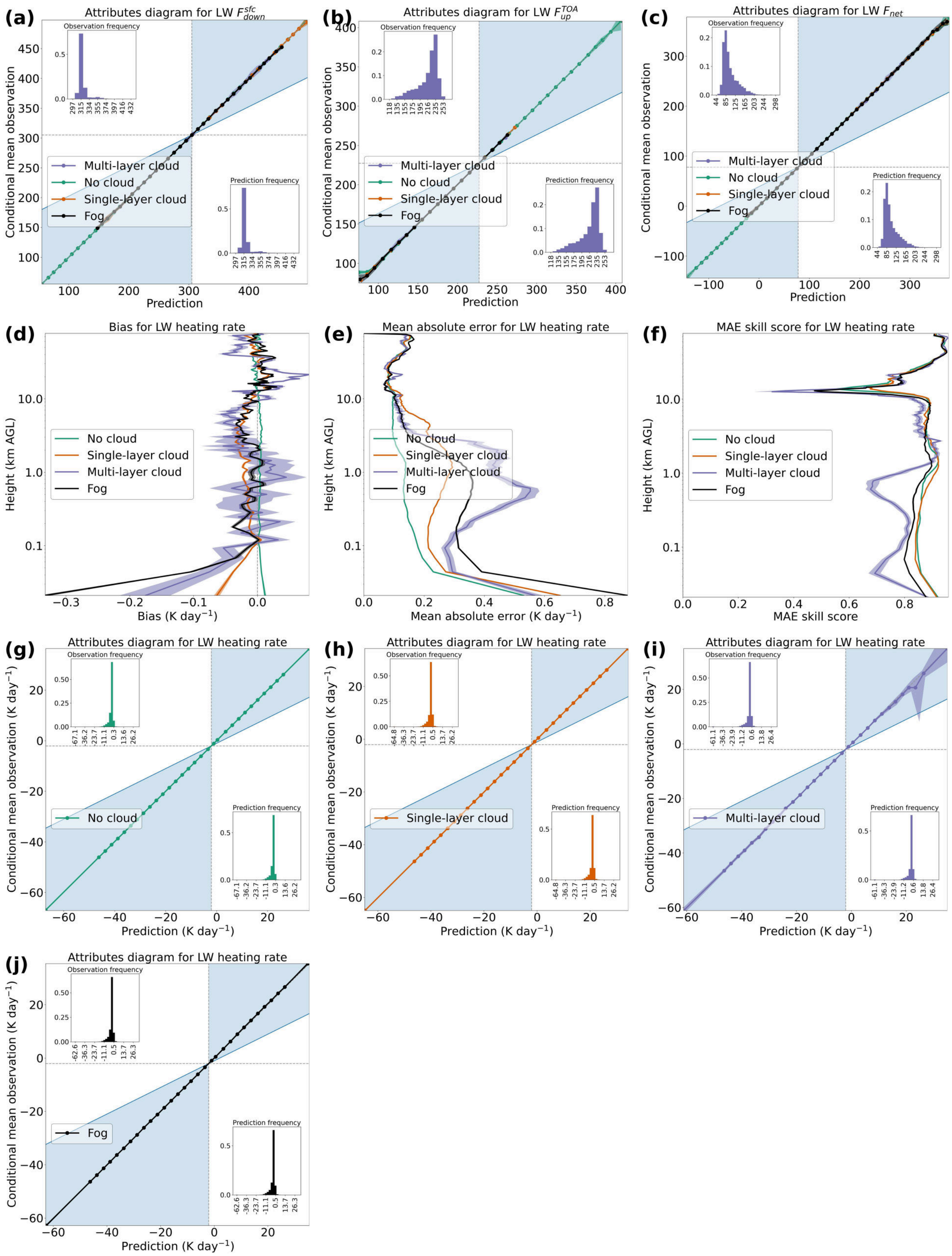

Figure S29: Performance of best longwave model on testing data, separated by mixed-phase cloud regime. This is analogous to Figure 12 in the main text but concerns mixed-phase, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud, fog} account for {82.25%, 17.60%, 0.15%, 4.93%} of the testing data respectively.

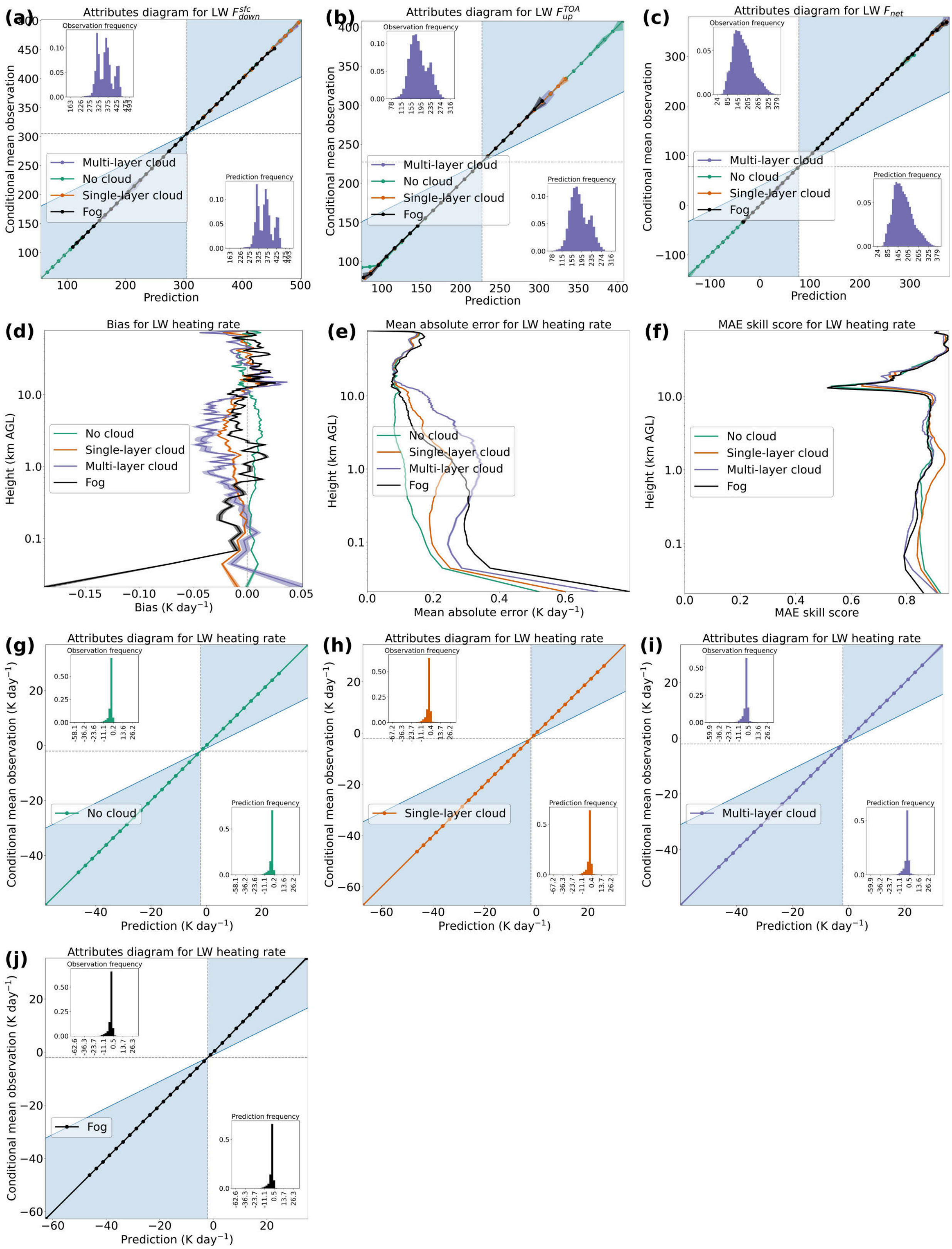

Figure S30: Performance of best longwave model on testing data, separated by any-phase cloud regime. This is analogous to Figure 12 in the main text but concerns any-phase, rather than liquid-only, clouds. Cases with {no cloud, single-layer cloud, multi-layer cloud, fog} account for {63.73%, 32.98%, 3.29%, 9.48%} of the testing data respectively.

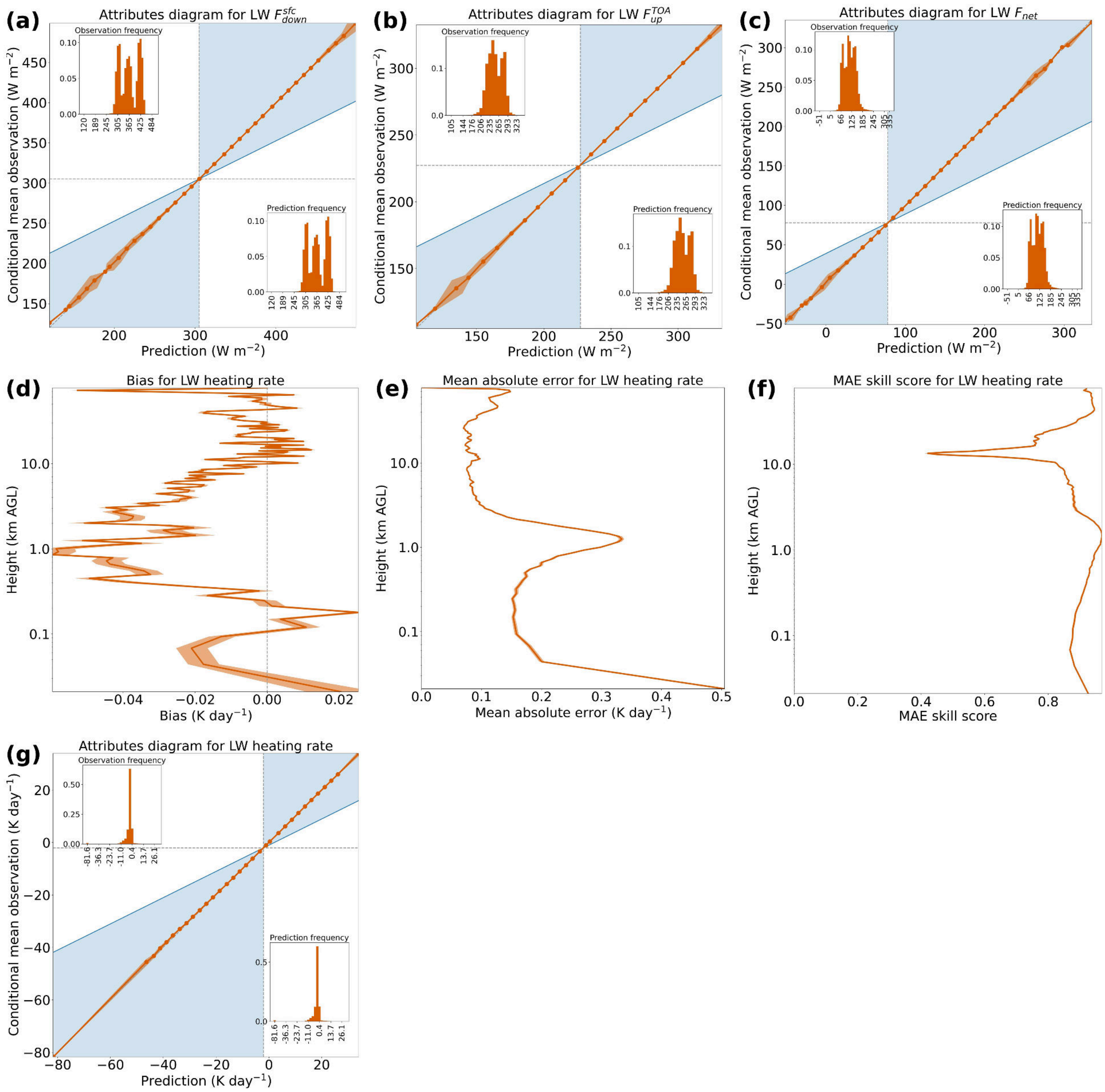

Figure S31: Performance of best longwave model on single-height extremes, *i.e.*, on the 3% of testing cases with the greatest height-maximum absolute HR. This is analogous to Figure 11 in the main text.

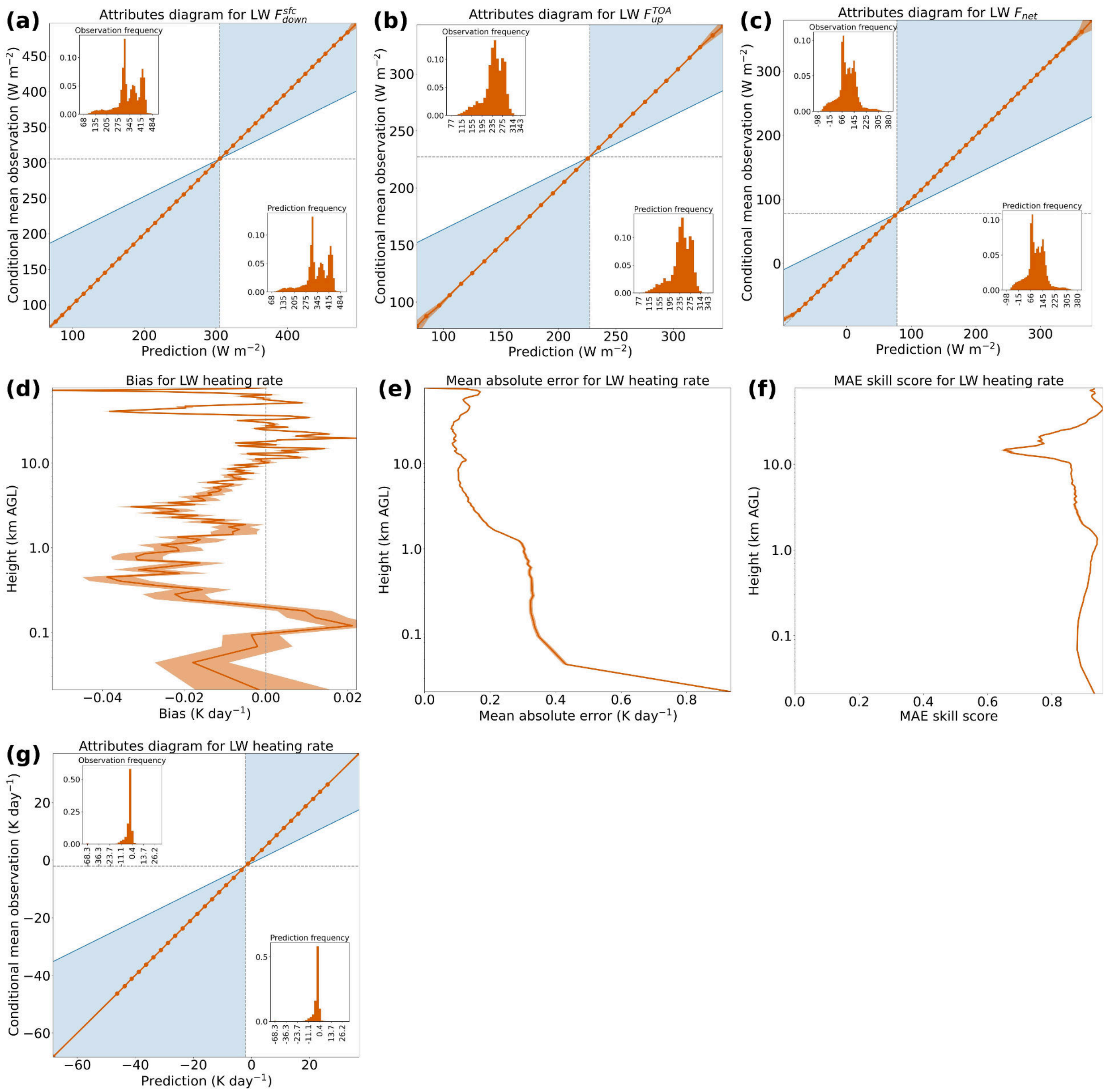

Figure S32: Performance of best longwave model on full-profile extremes, *i.e.*, on the 3% of testing cases with the greatest height-averaged absolute HR. This is analogous to Figure 11 in the main text.

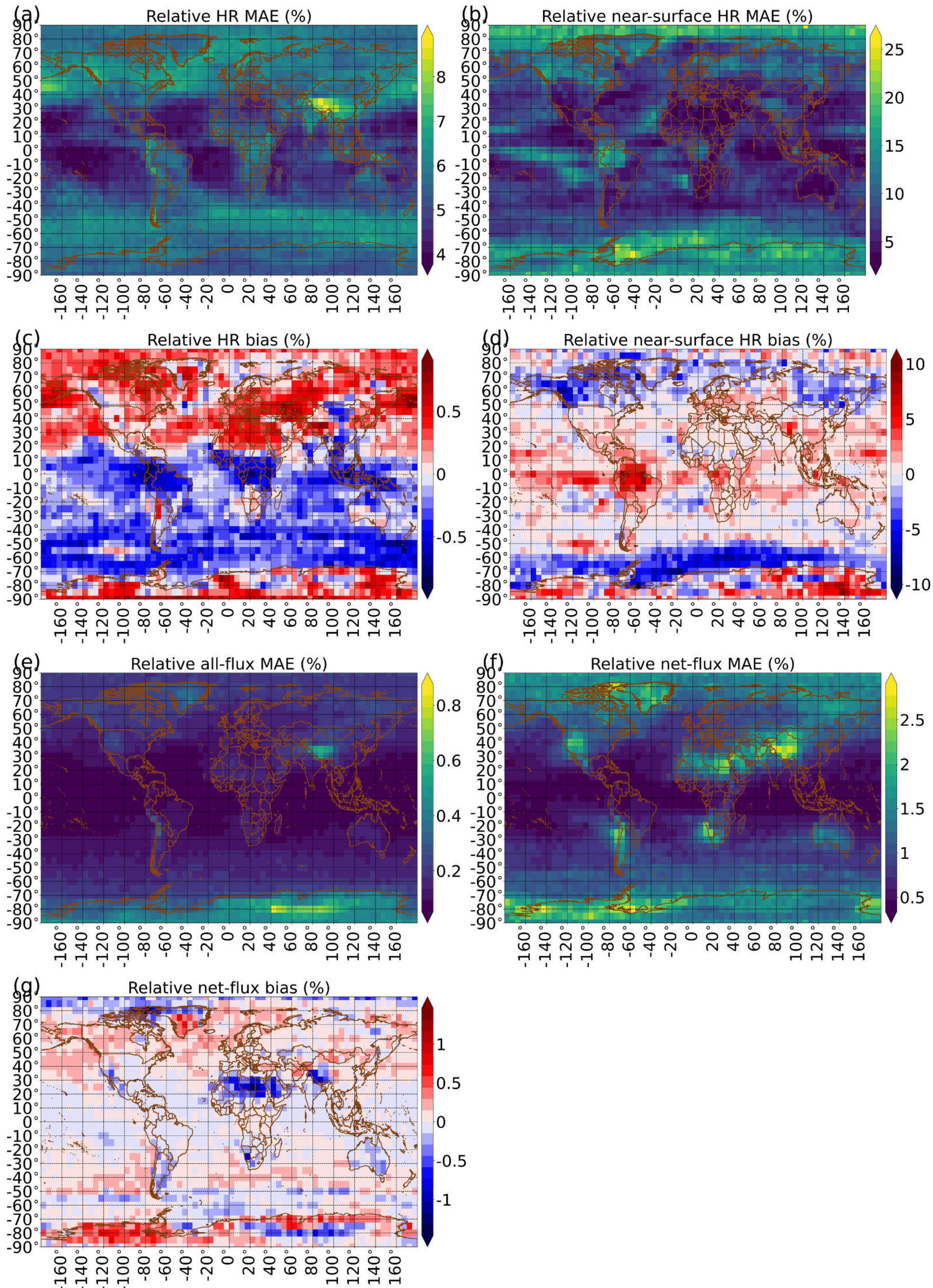

Figure S33: Fractional errors for best longwave model on testing data, binned by geographic location. This is analogous to Figure 13 in the main text but shows fractional errors instead of raw errors.

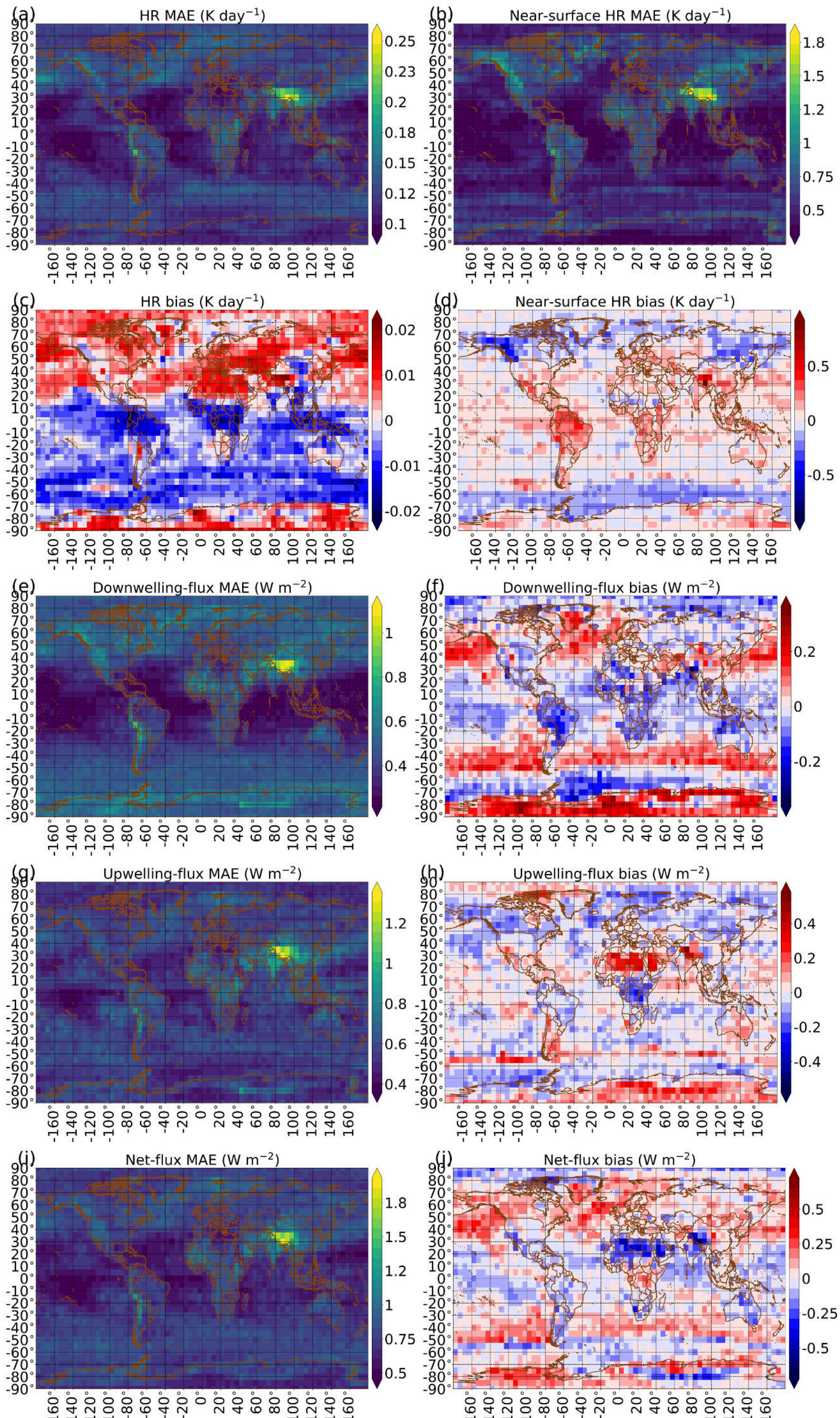

Figure S34: Detailed errors for best longwave model on testing data, binned by geographic location. This is analogous to Figure 13 in the main text but shows errors for individual flux variables – $F_{\mathrm{down}}^{\mathrm{sfc}}$ in panels e-f, $F_{\mathrm{up}}^{\mathrm{TOA}}$ in panels g-h, and F_{net} in panels i-j – rather than averaging to produce all-flux MAE.

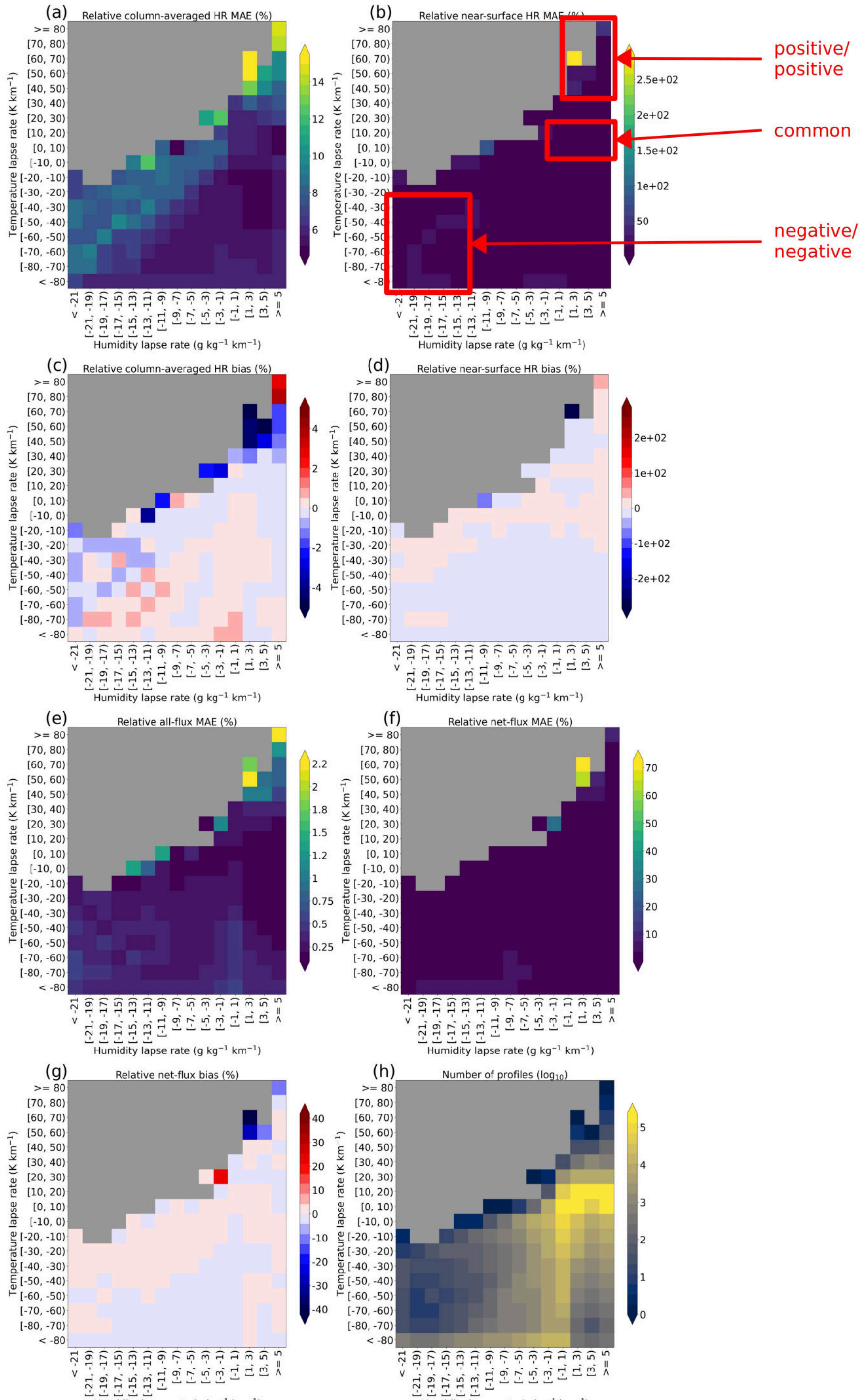

Figure S35: Fractional errors for best longwave model on testing data, binned by near-surface thermodynamic lapse rates. This is analogous to Figure 15 in the main text but shows fractional errors instead of raw errors.

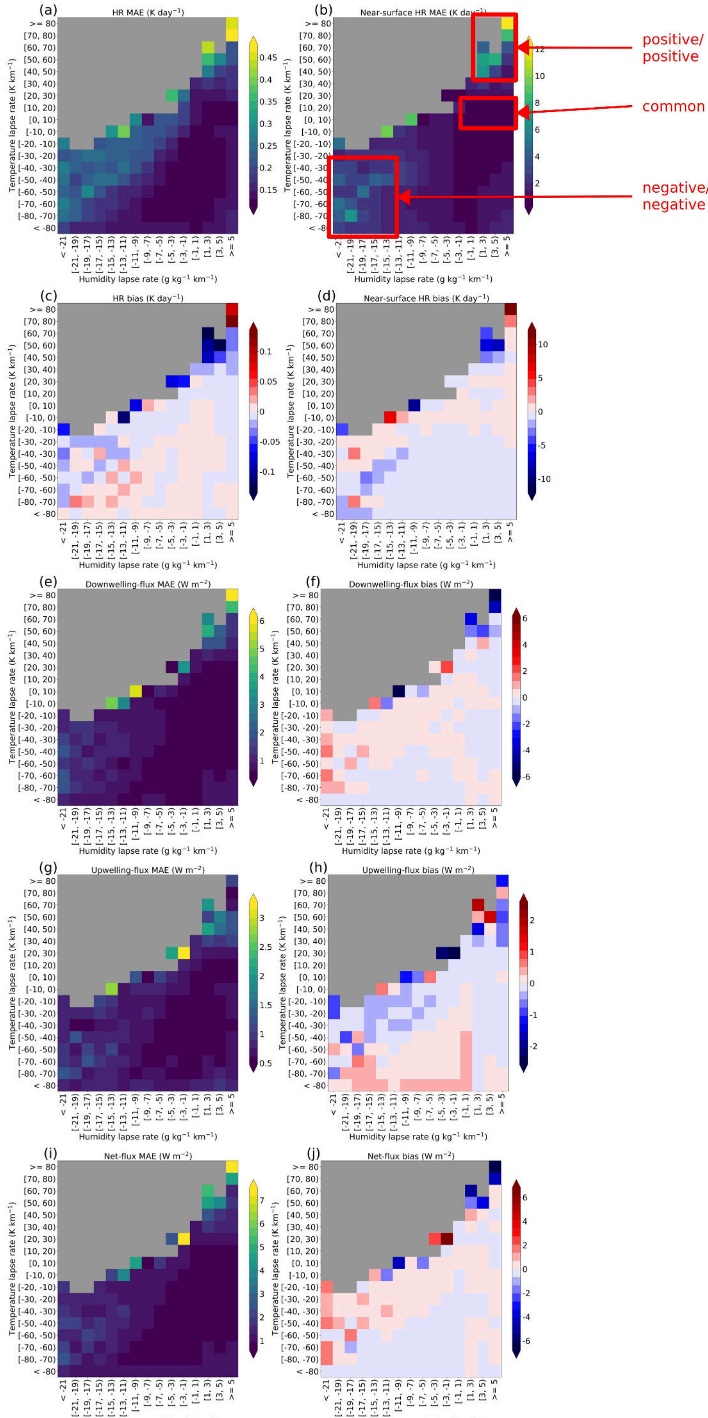

Figure S36: Detailed errors for best longwave model on testing data, binned by near-surface thermodynamic lapse rates. This is analogous to Figure 15 in the main text but shows errors for individual flux variables – $F_{\text{down}}^{\text{sfc}}$ in panels e-f, $F_{\text{up}}^{\text{TOA}}$ in panels g-h, and F_{net} in panels i-j – rather than averaging to produce all-flux MAE.

References

Anderson, G., S. Clough, F. Kneizys, J. Chetwynd, and E. Shettle, 1986: AFGL atmospheric constituent profiles (0-120 km). Tech. rep. URL https://apps.dtic.mil/sti/citations/ADA175173.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, URL https://www.deeplearningbook.org.

Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, Lille, France, International Machine Learning Society, URL https://arxiv.org/abs/1502.03167.

Kingma, D., and J. Ba, 2014: Adam: A method for stochastic optimization. *arXiv e-prints*, **1412 (6980)**, URL https://arxiv.org/abs/1412.6980v8.

Kinne, S., 2019: The MACv2 aerosol climatology. *Tellus B: Chemical and Physical Meteorology*, **71 (1)**, 1–21, URL https://doi.org/10.1080/16000889.2019.1623639.

Lagerquist, R., 2020: Using deep learning to improve prediction and understanding of high-impact weather. URL https://shareok.org/handle/11244/324145, doctoral dissertation, School of Meteorology, University of Oklahoma.

Li, M., T. Zhang, Y. Chen, and A. Smola, 2014: Efficient mini-batch training for stochastic optimization. *International Conference on Knowledge Discovery and Data Mining*, New York, New York, Association for Computing Machinery, URL https://doi.org/10.1145/2623330.2623612.

Maas, A., A. Hannun, and A. Ng, 2013: Rectifier nonlinearities improve neural network acoustic models. *International Conference on Machine Learning*, Atlanta, Georgia, International Machine Learning Society, URL http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Nair, V., and G. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *International Conference on Machine Learning*, Haifa, Israel, International Machine Learning Society.