Using deep learning to emulate and accelerate a radiative-transfer model

Ryan Lagerquist*

Cooperative Institute for Research in the Atmosphere (CIRA), National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory (ESRL) / Global Systems Laboratory (GSL), Boulder, Colorado.

David Turner

NOAA GSL, Boulder, Colorado, USA

Imme Ebert-Uphoff

CIRA / Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado, USA

Jebb Stewart

NOAA ESRL/GSL, Boulder, Colorado, USA

Venita Hagerty

CIRA, NOAA GSL / Assimilation and Verification Innovation Division, Boulder, Colorado, USA

Corresponding author: Ryan Lagerquist, ralager@colostate.edu

Early Online Release: This preliminary version has been accepted for publication in *Journal of Atmospheric and Oceanic Technology*, may be fully cited, and has been assigned DOI 10.1175/JTECH-D-21-0007.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

This paper describes the development of U-net++ models, a type of neural network that performs deep learning, to emulate the shortwave Rapid Radiative-transfer Model (RRTM). The goal is to emulate the RRTM accurately in a small fraction of the computing time, creating a U-net++ that could be used as a parameterization in numerical weather prediction (NWP). Target variables are surface downwelling flux, top-of-atmosphere upwelling flux (F_{up}^{TOA}), net flux, and a profile of radiative-heating rates. We have devised several ways to make the U-net++ models knowledge-guided, recently identified as a key priority in machine learning (ML) applications to the geosciences. We conduct two experiments to find the best U-net++ configurations. In Experiment 1, we train on non-tropical sites and test on tropical sites, to assess extreme spatial generalization. In Experiment 2, we train on sites from all regions and test on different sites from all regions, with the goal of creating the best possible model for use in NWP. The selected model from Experiment 1 shows impressive skill on the tropical testing sites, except four notable deficiencies: large bias and error for heating rate in the upper stratosphere, unreliable F_{up}^{TOA} for profiles with single-layer liquid cloud, large heating-rate bias in the mid-troposphere for profiles with multi-layer liquid cloud, and negative bias at low zenith angles for all flux components and tropospheric heating rates. The selected model from Experiment 2 corrects all but the first deficiency, and both models run $\sim 10^4$ times faster than the RRTM. Our code is available publicly.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

1. Introduction

Radiation is a key component of the global energy budget. In the shortwave part of the spectrum (mostly solar radiation, with wavelengths $\lesssim 4$ μm), incoming radiation is much greater in the tropics than at the poles. This imbalance, which is due to Earth-Sun geometry, sets up a meridional gradient in absorbed shortwave radiation that drives the global circulation (Sections 4.6 and 10.1.1 of Wallace and Hobbs 2006). Surface albedo has a secondary effect on absorbed shortwave radiation: at high latitudes the surface is often covered by snow and ice, which increases albedo and causes less shortwave radiation to be absorbed. This enhances the meridional gradient in absorbed shortwave radiation. In the longwave part of the spectrum (mostly terrestrial radiation, with wavelengths $\gtrsim 4$ μm), there is also an albedo effect: areas with high albedo, typically at high latitude, are colder and emit less longwave radiation. In terms of net radiation (absorbed shortwave minus emitted longwave), the two albedo effects approximately cancel out. Thus, in a globally and annually averaged sense, the meridional distribution of net radiation is similar to that of absorbed shortwave radiation. (Stone 1978)

When radiation propagates through the atmosphere, heating (cooling) occurs in areas of radiative-flux convergence (divergence). Most radiative-transfer models (RTM) assume horizontal independence, *i.e.*, that radiation is transferred only in the vertical dimension. In this case, radiative transfer is governed by the following equation:

$$\frac{dT}{dt} = \frac{g}{c_p}\frac{\Delta F_{net}}{\Delta p}, \tag{1}$$

where g is the gravitational constant ($\sim$9.81 m s^{-2}); c_p is the specific heat of dry air (1004 J kg^{-1} K^{-1}); Δp is the thickness of a layer in pressure coordinates (Pa); $\Delta F_{net} = \Delta F_{down} - \Delta F_{up}$ is the net flux into the layer (W m^{-2}); and $\frac{dT}{dt}$ is the resulting heating rate (K s^{-1}). Radiative transfer is extremely important in numerical weather prediction (NWP) and climate models. However,

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

because radiative transfer is a subgrid-scale process, it must be parameterized – *i.e.*, estimated outside the dynamical core by a separate RTM, rather than explicitly resolved.

Radiative transfer is inherently complex, due to the spectral (wavelength-dependent) nature of gaseous absorption, as well as changes in the refractive index and shape of particles acting to scatter and absorb radiation. The most accurate RTMs are line-by-line models, which explicitly simulate gaseous absorption in each spectral band (Turner et al. 2004; Mlawer and Turner 2016). However, the radiative properties of clouds and aerosols are much smoother in spectral space than those of gaseous molecules. Thus, simpler scattering models can be used for clouds and aerosols (*e.g.*, Stamnes et al. 1988). Nonetheless, both line-by-line and scattering models are extremely computationally expensive, so cannot be used as parameterizations in NWP. There is an inherent trade-off between computational cost and accuracy, and the goal is typically to reduce computational cost by orders of magnitude without a large reduction in accuracy.

Perhaps the most common approach is correlated-k models, like the Rapid Radiative-transfer Model (RRTM; Mlawer et al. 1997), which emulates line-by-line models but is many orders of magnitude faster. When implemented as a parameterization, an RTM must provide three variables to the parent NWP model for both the shortwave and longwave spectra: a vertical profile of radiative-heating rates, surface downwelling flux (F_{down}^{sfc}), and top-of-atmosphere upwelling flux (F_{up}^{TOA}). For the RRTM, F_{down}^{sfc} and F_{up}^{TOA} are accurate within $\sim$1 W m^{-2}, while heating rates are accurate within $\sim$0.1 K day^{-1} (Iacono et al. 2008). The longwave RRTM has been used in NWP since the early 2000s (Iacono et al. 2000), and the shortwave RRTM since the mid-2000s (Iacono et al. 2005). Although the RRTM is much faster than line-by-line models, it is still too slow for operational NWP. The RRTMG (RRTM for global climate models; Pincus and Stevens 2013) makes additional simplifications and is approximately twice as fast as the RRTM, but it is still too slow to call at every atmospheric time step in NWP. Thus, while other parameterizations

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

(microphysics, boundary layer, etc.) are called at every time step, the RRTMG is called less often, which makes the NWP model less accurate. Also, even when called less often, the RRTMG still accounts for ∼50% of the computation of the overall NWP model (Krasnopolsky 2020).

Due to these issues, some groups have used neural networks (Part II of Goodfellow et al. 2016), a type of machine learning (ML), to emulate RTMs (Krasnopolsky 2020 and references therein). Neural networks are also popular for emulating other atmospheric processes, especially subgrid-scale convection in NWP models (Gentine et al. 2018; Brenowitz and Bretherton 2018; Brenowitz et al. 2020; Krasnopolsky 2020; Beucler et al. 2021). Because neural networks can theoretically approximate a function of arbitrary complexity, they are often called "universal function-approximators". Although neural networks are often slow to train, at inference time (when applying a trained neural network to new data), they are much faster than process-based RTMs, even the RRTMG. Neural networks often contain many layers with many weights in each layer, allowing them to represent important features at various levels of abstraction, which they ultimately transform into predictions. However, each weight is one degree of freedom and neural networks often contain millions of weights, which makes them prone to overfitting. Also, ML is typically poor at extrapolating to conditions outside those seen in the training data. This diminishes the trustworthiness of ML, which is a key requirement for transitioning ML to operational products such as NWP (Gil et al. 2019).

We have developed neural networks to emulate shortwave radiative transfer, with three main characteristics that make our work unique. First, we use U-net++ models (Zhou et al. 2019), as opposed to the fully connected networks (sometimes called "dense" or "feed-forward"; see Chapter 6 of Goodfellow et al. 2016) used in previous work. U-net++ models are a type of deep learning, which can exploit spatial patterns in gridded data to make better predictions. Second, we have built physical constraints and vertical non-locality into the U-net++ models, allowing them to handle

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

non-adjacent cloud layers and better extrapolate to different conditions (*e.g.*, from non-tropical to tropical sites). Third, we train U-net++ models to emulate the RRTM, instead of the less accurate RRTMG used in previous work (Krasnopolsky et al. 2010; Krasnopolsky 2020). Although line-by-line models are the most accurate, they are only slightly more accurate than the RRTM (Iacono et al. 2008) and many orders of magnitude slower, so emulating line-by-line models would vastly increase the time required to create training data for the U-net++ models.

The rest of this paper is organized as follows. Section 2 describes the inner workings of a U-net++; Section 3 describes the input data and methods used to train the U-net++ models; Section 4 describes experiments to find the best U-net++ configuration (hyperparameters); Sections 5 and 6 evaluate and interpret the selected U-net++ models; and Section 7 concludes.

2. Background on U-net++

This section focuses mainly on traditional U-nets, extending the discussion to U-net++ at the end. We use the Keras library for Python (Chollet et al. 2020) to implement all U-net++ models, and our code is freely available on the internet (see data-availability statement).

U-nets are a specialized type of convolutional neural network (CNN; Fukushima 1980; Fukushima and Miyake 1982). CNNs are a deep-learning method (Section 1.1.4 of Chollet 2018) designed to exploit spatial patterns in gridded data, which they achieve via convolution and pooling, spatial operations defined later in this section. CNNs have become popular tools in atmospheric science (Wang et al. 2016; Racah et al. 2017; Kurth et al. 2018; Bolton and Zanna 2019; Gagne et al. 2019; McGovern et al. 2019; Wimmers et al. 2019; Lagerquist et al. 2019; Ebert-Uphoff and Hilburn 2020; Lagerquist et al. 2020a,b). U-nets (Ronneberger et al. 2015) retain all the advantages

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

of CNNs but are designed for pixelwise prediction[1] – *i.e.*, to make a prediction at every grid point. CNNs are typically used for full-image prediction – *i.e.*, to make one prediction based on the full grid. There are several U-net applications to atmospheric science in the refereed literature (Chen et al. 2020; Kumler-Bonfanti et al. 2020; Sadeghi et al. 2020; Sha et al. 2020a,b), and we are aware of several other atmospheric scientists currently adopting U-nets (Stewart et al. 2020; Berthomier and Pradel 2021; Felt et al. 2021; Hayatbini et al. 2021).

As shown in Figure 1, a U-net contains four types of specialized components: convolutional layers, pooling (downsampling) layers, upsampling layers, and skip connections. The left side of the U-shape is the downsampling side, where spatial resolution decreases with depth, and the right side is the upsampling side, where resolution increases with depth. The convolutional layers detect spatial features, and the other components allow convolutional layers to detect features at various spatial resolutions, which is important due to the multi-scale nature of atmospheric phenomena. Inputs to the first convolutional layer (top-left green box in Figure 1) consist of raw predictors (here, physical variables like temperature and pressure), while inputs to all other layers consist of feature maps, which are transformed versions of the raw predictors. As the spatial resolution decreases, the number of feature maps ("channels") typically increases, to offset the loss of spatial information. Convolution is both a spatial and multivariate transformation, so the feature maps encode spatial patterns that include all predictor variables. Most CNN applications involve data with two spatial dimensions (2-D), for which the inner workings of a convolutional layer are illustrated in Supplemental Figure S1 of Lagerquist et al. (2020b). For 1-D data like those used in the current work, see our Supplemental Figure S1 (an animation). In general, a convolutional layer is followed by an activation function and possibly batch normalization (Supplemental Table S2).

[1]U-nets are not the only type of CNN designed for pixelwise prediction. Other examples, in the encoder-decoder family along with U-nets, include convolutional auto-encoders (Chen et al. 2017) and fully convolutional networks (Long et al. 2015).

Each pooling layer downsamples the feature maps to a lower resolution (larger grid spacing), using either a maximum or mean filter. On the downsampling side of the U-net (left side in Figure 1), feature maps at deeper layers contain higher-level abstractions, because they contain information from a wider variety of spatial scales and have passed through more convolutions. For 2-D data, the inner workings of a pooling layer are illustrated in Supplemental Figure S2 of Lagerquist et al. (2020b). For 1-D data, see our Supplemental Figure S2 (an animation).

Each upsampling layer upsamples the feature maps to a higher resolution, using an interpolation method such as nearest-neighbour or linear. In this work we use nearest-neighbour. However, the choice of interpolation method is unimportant: upsampling always consists of interpolation followed by convolution, because interpolation cannot adequately reconstruct high-resolution information from low-resolution information. On the upsampling side of the U-net (right side of Figure 1), while spatial resolution increases the number of channels decreases, terminating in the number of output channels. In this work there is one output channel (radiative-heating rate, as discussed in Section 3). For 1-D data, the inner workings of an upsampling layer are shown in Supplemental Figure S3 (an animation).

Skip connections preserve high-resolution information from the downsampling side of the U-net and carry it to the upsampling side, as shown in Figure 1. Without skip connections, the U-net would simply perform downsampling followed by upsampling, which is a lossy operation. In other words, upsampling cannot fully recover the high-resolution information lost during downsampling. On the upsampling side of the U-net, at each spatial resolution r (each row in Figure 1), some feature maps are provided by the upsampling layer at the next-coarsest resolution (the row below in Figure 1), while some are provided by a skip connection. The advantage of feature maps from the upsampling layer is that they contain higher-level abstractions, because they include information from more spatial scales and more convolutions. The advantage of feature maps from the skip

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

connection is that they are truly at resolution r, not merely upsampled to r. In other words, for the skip connection the nominal and effective resolutions are both r, whereas for the upsampling layer the effective resolution is coarser than r. Feature maps from the skip connection and upsampling layer are both passed through a convolutional layer, which combines information from both ("the best of both worlds").

Fully connected layers (sometimes called "dense"; see Chapter 6 of Goodfellow et al. 2016) are designed for full-image prediction, so they are not typically included in a U-net. However, we include fully connected layers in our U-nets, because the task is a combination of pixelwise prediction (a vertical profile of radiative-heating rates) and full-image prediction (scalar fluxes). See Section 3 for more on the output variables. Since fully connected layers are spatially agnostic, feature maps are flattened into a vector before they are passed to the fully connected layers (in Figure 1 this is a vector of length $4 \times 1024 = 4096$). Each feature in one fully connected layer is a weighted sum of those in the previous layer. Like convolutional layers, each fully connected layer is followed by an activation function and possibly batch normalization.

Figure 1 shows a U-net with the traditional architecture (Ronneberger et al. 2015), but we have adopted the U-net++ architecture (Zhou et al. 2019), shown in Figure 2. The U-net++ architecture contains more skip connections, allowing features from more than two scales to be combined at each level. For example, the set of feature maps labeled "D" in Figure 2 is produced by combining A, B, and the upsampled version of C. Although these feature maps all have a nominal resolution of 18h (18 heights in the profile, or $\sim \frac{1}{4}$ the resolution of the predictors), their effective resolutions, due to upsampling, are respectively 18h, 9h, and 4h. This ability to combine information from many scales at once can allow the U-net++ to make better predictions than the U-net (Zhou et al. 2019).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Before training, all weights (in the convolutional, fully connected, and batch-normalization layers) are initialized to random values; during training, they are adjusted to minimize the loss function. Our particular loss function is discussed in Section 3c2.

3. Data and Methods

a. Data Description

Like the RRTM, our U-net++ models assume horizontal independence and thus treat each vertical column separately. To create inputs (predictors) for the RRTM and U-net++ models, we use data from the Rapid Refresh model (RAP; Benjamin et al. 2016). The RAP is a non-hydrostatic, mesoscale, operational NWP model, run every hour with 13-km horizontal grid spacing and 51 vertical levels. We have obtained RAP data from an internal NOAA archive in height coordinates, running from 10 to 50 000 metres above ground level (m AGL), with 20-m vertical spacing near the surface and 4000-m vertical spacing near the top. We extract 0-hour analyses of 14 variables (Table 1 and Figure 3) from 30 sites throughout the northern hemisphere (Figure 4), at every hour in the years 2017-2020. We are currently emulating a simplified version of the RRTM, which assumes a climatological profile of trace gases (O_3, CO_2, CH_4, etc.) and does not consider aerosols or precipitation (see future work in Section 7), which is why the predictors do not include this information. Other than trace gases, aerosols, and precipitation, the main controls on radiative transfer are the solar zenith angle, albedo, profiles of atmospheric state variables (temperature and pressure), and profiles of the three water species. This explains our choice of predictors (Table 1).

To create desired outputs ("targets" or "labels" in the ML literature), we run the RRTM separately for each example, where one "example" is one profile at one time. The output variables are those

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

required by an NWP model from a shortwave RTM, namely the heating-rate profile and the two flux components: F_{down}^{sfc} and F_{up}^{TOA}. See Figure 3d.

b. Pre-processing

Before training U-net++ models, we pre-process the data in two ways. First, we split the data into training, validation, and testing sets. We split the data differently for the two experiments (Section 4), as shown in Table 2. For each experiment, the datasets are mutually independent – *i.e.*, any pair of datasets contains different years and/or different sites. Also, there is a one-week gap between each pair of consecutive datasets, to eliminate temporal autocorrelation. Second, we normalize predictor and target variables, using the methods listed in Table 3. The procedure is described below for each scalar predictor[2] x; only step 1 is applied to the target variables. Note that only the U-net++-training data (Table 2) are used for scaling, *i.e.*, to compute percentiles in step 1. This ensures that no information from the isotonic-regression-training, validation, or testing set is used to train the U-net++. If it were, the four datasets would no longer be independent.

1. Uniformization. Transform x to a uniform distribution over $[0, 1]$, by converting each value to its percentile over all x-values in the U-net++-training set. Let the transformed variable be x'.

2. z-score normalization. Transform x' to a standard Gaussian distribution (with mean of 0.0 and variance of 1.0), using the inverse of the cumulative density function (CDF).

The purpose of normalizing predictors is to ensure that they have equal variance, which prevents the U-net++ models from unduly focusing on predictors with higher variance due to physical units. For example, in our dataset, specific humidity has a variance of 2.4 $\times$ 10^{-5} kg^2 kg^{-2}, while temperature has a variance of 672.1 K^2. z-score normalization is common practice for neural networks (Section 3.6.2 of Chollet 2018; Shanker et al. 1996), but the standard approach

[2]A scalar predictor may be zenith angle, albedo, or one vector predictor at one height.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

is to divide each variable by its standard deviation in the raw data. We use a different approach (uniformization followed by the inverse CDF) because the standard approach assumes that the raw data follow a Gaussian distribution, which is untrue for our predictors.

The purpose of normalizing target variables is similar: to ensure that they have equal ranges, so that one target variable cannot dominate the loss function. For example, in our dataset, F_{up}^{TOA} ranges from 0-993.3 W m^{-2} with a median of 118.7 W m^{-2}, while F_{down}^{sfc} ranges from 0-1198.9 W m^{-2} with a median of 322.1 W m^{-2}. Without normalization, errors for F_{down}^{sfc} would generally be larger, causing F_{down}^{sfc} to have a greater influence on the loss function. Unlike the predictors, we apply only uniformization, not z-score normalization, to the target variables. Normalizing to a distribution without negative values allows us to use the rectified linear unit, which prohibits negative values, as the activation function for the output layers (Supplemental Table S2).

Note that we normalize only two target variables: F_{up}^{TOA} and F_{down}^{sfc}. We do not normalize heating rate, for reasons discussed in Section 3c2.

c. Knowledge-guided Machine Learning

We have devised three ways to make the U-net++ models knowledge-guided – *i.e.*, to include physical relationships in the training – which is a key priority in ML applications to the geosciences (Reichstein et al. 2019; Gil et al. 2019).

1) Physically consistent and skillful net flux

The U-net++ models predict three flux components, but they predict only F_{down}^{sfc} and F_{up}^{TOA} independently, with the net flux (F_{net}) predictions constrained by the following law:

$$F_{net} = F_{down}^{sfc} - F_{up}^{TOA}. \tag{2}$$

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

F_{net} is included in an output neuron (in the last fully connected layer, at the bottom of Figure 2) and is therefore included in the loss function. Equation 2 could have been easily satisfied by post-processing (*i.e.*, computing F_{net} outside the U-net++ models), but this would leave F_{net} out of the loss function. As discussed in Section 5b, the post-processing approach allowed the models to make poor predictions of F_{net}, while our approach forces predictions of all flux components to be both physically consistent and skillful.

2) Custom loss function to emphasize large heating rates

We use a custom loss function:

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

where N is the number of examples; $H = 73$ is the number of heights per example; r_{ij} is the actual heating rate for the j^{th} height in the i^{th} example; $\hat{r}_{ij}$ is the corresponding prediction; $M = 3$ is the number of flux components; F_{ik} is the actual value of the k^{th} flux component in the i^{th} example; and $\hat{F}_{ik}$ is the corresponding prediction. α is a coefficient that will be discussed later.

The first term in Equation 3 is the dual-weighted mean squared error (MSE) for heating rates, and the second term is the MSE for fluxes. Using the dual-weighted MSE for heating rates, rather than the standard MSE, weights points with a large predicted or actual heating rate more heavily. In early experiments (not shown), we found that this is necessary to skillfully predict large heating rates. Large heating rates are important in many atmospheric regimes, including stratocumulus clouds and the upper stratosphere. Shortwave radiation is absorbed by liquid water at the top of a stratocumulus cloud, leading to diabatic heating and a turbulent circulation that maintains the cloud; this is why stratocumulus clouds tend to be long-lived (Morrison et al. 2012; Wood 2012). In the upper stratosphere, shortwave radiation is absorbed by ozone, leading to extreme diabatic heating (Iacono et al. 2008); this is why the temperature profile of the stratosphere increases with

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

height. However, large heating rates in the troposphere are rare (Figure 5d), making them difficult to predict unless they are emphasized with a custom loss function such as dual-weighted MSE. The flux components follow less skewed distributions (Figure 5a-c), so no custom loss function is needed to make the U-net++ models skillfully predict extreme fluxes.

The U-net++ models predict heating rates in raw physical units (K day^{-1}), and values in our dataset range from 0-42 K day^{-1}, so the weight ranges from approximately 0-42. Meanwhile, the U-net++ models predict flux components in normalized units, ranging from 0-1. In early experiments (not shown), we tried balancing the two terms by setting $\alpha \geq 1$ in Equation 3. However, we found that regardless of α, training is effectively partitioned into two phases. During early training, heating-rate predictions improve rapidly while flux predictions improve slowly; during late training, heating-rate predictions improve slowly while flux predictions improve rapidly. In other words, the U-net++ models learn to predict heating rates well, then learn to predict fluxes well. Thus, for models shown in the paper, we use α = 1.

3) Custom predictors to account for non-local effects

Our choice of predictors allows the U-net++ models to consider vertically non-local effects, which occur when the heating rate at height z is affected by predictors far away from z. Specifically, we include height-integrated paths of the three water species: downward and upward LWP, IWP, and WVP (Table 1). The raw RAP data include only concentrations of the three water species: LWC, IWC, and humidity. Height-integrated paths are crucial in many scenarios – *e.g.*, to predict the heating-rate profile in a column with multi-layer liquid cloud, like that shown in Figure 3. The top cloud layer attenuates a lot of downwelling solar radiation, leading to large heating rates in the top cloud layer (around 5.5 km AGL in Figure 3; the cloud layer itself is shown in panel b, and the resulting radiative heating is shown in panel d). However, lower cloud layers do not produce

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

large heating rates, because at lower heights most downwelling solar radiation has already been attenuated by the top cloud layer (*e.g.*, Turner et al. 2018). This is exemplified in Figure 3 for the lower cloud layer, stretching from 0-2.4 km AGL. When trained with only concentrations and not paths, the U-net++ models cannot represent these relationships, which are typically vertically non-local because the cloud layers are far apart (more than a few grid cells from each other).

d. Isotonic Regression for Bias Correction

We bias-correct predictions from each U-net++ with isotonic regression (Barlow and Brunk 1972), which is an ML method commonly used to bias-correct other ML methods. The ML method being corrected is called the "base model". For each target variable y, isotonic regression creates a mapping of the following form:

$$y_i \rightarrow y_i', \tag{4}$$

where y_i is the i^{th} cutoff point for base-model predictions and y_i' is the bias-corrected value. For y-values that fall between two cutoff points, isotonic regression uses linear regression. For example, if a base-model prediction falls halfway between y_j and y_k, the bias-corrected prediction will fall halfway between y_j' and y_k'. During training, the mapping is adjusted to minimize mean squared error (MSE), subject to the isotonic constraint: if $y_k > y_j$, then $y_k' > y_j'$. In other words, isotonic regression cannot change the rank order of predictions.

Because isotonic regression is a univariate method (with one input variable and one output variable), we apply isotonic regression separately to heating rate at each height, F_{down}^{sfc}, and F_{up}^{TOA}. We do not apply isotonic regression to F_{net}, so F_{net} predictions from isotonic regression are computed outside the model, via Equation 2. Thus, unlike for the U-net++ models, F_{net} is not included in the loss function for isotonic regression (which is MSE). However, we have found

that F_{net} predictions are still better with isotonic regression than without. In other words, bias-correcting F_{down}^{sfc} and F_{up}^{TOA} bias-corrects F_{net} as a side effect.

We use separate training data (sites and times) for U-net++ and isotonic regression, as shown in Table 2. If we used the same training data, isotonic regression would learn to bias-correct the U-net++ models only for data that they have already "seen," for which the U-net++ predictions are unrepresentatively good.

4. Hyperparameter Experiments

A hyperparameter is a property of an ML model that, unlike the weights (sometimes called "parameters"), cannot be adjusted by training. We conduct two experiments to find the best U-net++ hyperparameters for emulating the shortwave RRTM. In Experiment 1, we train ML models (U-net++ and isotonic regression) with data from non-tropical sites in 2018-2020, then test with data from tropical sites in 2017 (Table 2). This tests the ability of the ML models to generalize in both space and time. It is crucial that we test the ability to generalize in space, because although 30 sites are used for model development[3] (Figure 4), an ML-based parameterization would be applied to every site (horizontal grid location) in the NWP model. Also, extreme differences between the training and application data might be seen in other scenarios, such as climate change (if an ML model remains in production for long enough, it may be applied to a different climate than in the training data) and rare events (the application data may contain a weather pattern not found in the training data). In Experiment 2, we train ML models with data from "Assorted1" sites in 2018-2020, then test with data from "Assorted2" sites in 2017 (Table 2). The difference here is

[3]We have obtained RAP data from only 30 sites, because (a) the native RAP-output files are large and stored on a tape archive, which makes processing computationally slow; (b) the 30 sites chosen are important for other NOAA projects, so the data will be reused; (c) extracting millions of examples from 30 sites yields a large sample size at each site, as opposed to extracting millions of examples from thousands of sites. This allows us to robustly test the models' generalization ability to each site in the testing data.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

that both the Assorted1 and Assorted2 sites include all three regions: Arctic, mid-latitude, and tropical. Thus, although to some extent the testing data for Experiment 2 test the models' ability to generalize in space (to different sites), this test is less stringent than in Experiment 1 (to a completely different region). The goal of Experiment 2 is to create the best possible ML model for use as a parameterization in NWP. We hypothesize that a model trained with data from all three regions will perform better than one trained with only non-tropical data.

In both experiments we perform a grid search (Section 11.4.3 of Goodfellow et al. 2016) to optimize hyperparameters. A grid search involves four steps: (1) define the experimental hyperparameters and values to be attempted for each, (2) train a model with every possible combination of values, (3) evaluate all models on the validation data, (4) select the model that performs best on validation data and evaluate it on testing data. We choose three experimental hyperparameters and attempt the values listed in Table 4: the number of fully connected layers, dropout rate for fully connected layers, and L_2 weight for convolutional layers. The number of fully connected layers (dashed black arrows in Figure 2) controls the complexity of features used to predict flux components, with more layers allowing for higher complexity. Although higher complexity would ideally improve predictions, the number of weights increases dramatically with the number of fully connected layers, which can lead to overfitting. Meanwhile, dropout (Hinton et al. 2012) and L_2 are both regularization methods; regularization encourages a simpler model, which reduces overfitting. The amount of regularization increases with both the dropout rate and L_2 weight (see Section 4b of Lagerquist et al. 2020b for details).

U-net++ models have many hyperparameters, and it is impossible to experiment with them all, due to combinatorial explosion. For example, at a conservative estimate of 20 hyperparameters, if we attempted 5 values for each, we would need to train $5^{20} = 9.5 \times 10^{13}$ U-net++ models. Training one U-net++ takes approximately 192 core-hours on graphics-processing units (GPU) and 480 core-

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

hours on central processing units (CPU), so training more than a few hundred to a few thousand U-net++ models is infeasible. Some important fixed (non-experimental) hyperparameters are listed in Supplemental Tables S1-S2, along with the value chosen for each and a justification. This leaves the three experimental hyperparameters listed in Table 4.

5. Model Evaluation

a. Evaluation Methods

For both Experiments 1 and 2, we evaluate the selected model overall (on the whole testing set) and in three regime-based settings. First, we evaluate the model by cloud regime: on profiles with no liquid cloud, single-layer liquid cloud, and multi-layer liquid cloud. For this purpose, a cloud layer is defined as a contiguous set of heights with LWC > 0 g m^{-3} and total LWP ≥ 25 g m^{-2}. Clouds add immense complexity to radiative transfer, because they both absorb and scatter radiation, creating a discontinuity in the profile of extinction optical depth. Thus, a model that performs well in cloud-free situations, is not guaranteed to perform well in cloudy situations. Also, radiative heating is a key process in the maintenance of stratocumulus clouds, which makes it key for climate prediction. Second, we evaluate the model by solar zenith angle. The zenith angle determines the amount of incoming top-of-atmosphere solar radiation, as well as its incidence angle, which determines the amount of atmosphere through which radiation must pass en route to the surface. A model that performs well for intermediate zenith angles, may not perform well when the Sun is directly overhead (zenith angle of 0°) or on the horizon (90°). Third, we evaluate the model by site. Different sites around the globe have different properties not accounted for in the partitioning by cloud regime and zenith angle, such as temperature, albedo, and cloud type (*e.g.*, stratocumulus clouds are very common in the Arctic).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

We make abundant use of the reliability curve and attributes diagram. Although both graphics were initially developed for classification (*i.e.*, to evaluate probabilistic predictions of an event), we have adapted them for regression (*i.e.*, to evaluate real-valued predictions). For classification, the reliability curve plots predicted probability vs. conditional event frequency and answers the question: "For a given probability, what is the expected event frequency?" For regression, the reliability curve plots the predicted value vs. conditional mean observed value and answers the question: "For a given prediction, what is the expected observation?" For both classification and regression, a perfect reliability curve follows the $x = y$ line (diagonal grey line in Figure 6a). Meanwhile, the attributes diagram (Hsu and Murphy 1986) is a reliability curve with extra reference lines: the no-resolution line (horizontal grey in Figure 6a), climatology line (vertical grey in Figure 6a), and positive-skill area (blue shading in Figure 6a). For classification, the no-resolution and climatology lines both correspond to the event frequency in the dataset; for regression, these lines correspond to the mean observation (in Figure 6a, mean F_{down}^{sfc}) in the dataset. For a model with no resolution, the reliability curve follows the no-resolution line – *i.e.*, the conditional mean observation is the same for every prediction. For a climatological model (one that always predicts the mean value), the reliability curve consists of one point, at the intersection of the no-resolution and climatology lines. Where the reliability curve passes through the positive-skill area, the model has a lower MSE (for classification, MSE is called the Brier score) than a climatological model. Lastly, the inset histograms show the distribution for both the predictions and observations. In a perfect attributes diagram, the reliability curve is perfect (follows the $x = y$ line) and the two histograms are identical.

Both the reliability curve and attributes diagram are useful for diagnosing conditional bias. For example, if a model has positive bias for low predictions and negative bias for high predictions, these biases may offset, making overall bias (on the whole testing set) negligible. Thus, using the

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

reliability curve and attributes diagram fits our motif of conducting regime-based evaluation, since averaging over the whole testing set may obscure issues that occur in certain regimes. For the scalar target variables (flux components), we plot one attributes diagram for each (*e.g.*, Figures 6a-c). For the vector target variable (heating rate), we plot one reliability curve for each height (*e.g.*, Figures 6g-i), omitting the reference lines in the attributes diagram. The reference lines would be different for each of the 73 heights, and it is not feasible to show 73 sets of reference lines.

b. Experiment 1

Results of the hyperparameter experiment, used to select the preferred model, are relegated to the Supplemental Material. The main conclusion to note here is that the U-net++ performs best when the dropout rate and L_2 weight are small (less regularization), which suggests that overfitting is not a serious problem for emulating the shortwave RRTM. This is surprising, as our experience with ML for atmospheric science indicates that overfitting is a serious problem and aggressive regularization is needed (*e.g.*, Lagerquist et al. 2019, 2020b). We suspect that overfitting is less problematic for our task because it is a perfect-model experiment, where the ML model is trained to emulate another model (the shortwave RRTM), rather than to fit real-world observations, which have more noise and uncertainty. Ultimately, we select the model with 3 fully connected layers, a dropout rate of 0.1, and L_2 weight of $10^{-6.5}$. Results shown in the rest of this section, for the selected model only, are based on testing data rather than validation data.

Figure 6 shows the model's performance on the whole testing set (tropical sites in 2017). The mean absolute error (MAE) skill score is defined as $\frac{\text{MAE}_{\text{climo}} - \text{MAE}_{\text{actual}}}{\text{MAE}_{\text{climo}}}$, where $\text{MAE}_{\text{climo}}$ is the MAE that would result from always predicting the climatological mean, estimated here as the mean over the U-net++-training data. The definition of MSE skill score is analogous. Both skill scores range from $(-\infty, 1]$; the optimal value is 1; and values > 0 signal an improvement over

climatology. Figures 6a-c show the attributes diagram for each flux component; the reliability curves are nearly perfect, and as shown by the inset histograms, the predictions and observations are similarly distributed. Figure 6d shows the bias profile for heating rates; nearly all heights have an absolute bias < 0.1 K day^{-1}, which is considered a threshold for stable integration into NWP (Iacono et al. 2008). Figure 6e shows the MAE profile for heating rates, which has a similar shape but with slightly larger values, because MAE includes both systematic error (bias) and random error. Both absolute bias and MAE are largest in the upper stratosphere, specifically at 46 km. This is the height with the largest climatological mean (32.2 K day^{-1} in the U-net++-training data), due to absorption of solar radiation by ozone. Thus, both the actual and climatological models have a large MAE at 46 km, leading to only a small dip in the MAE skill score (Figure 6f). Figure 6g shows the reliability curve for heating rate at each height; all curves nearly follow the line of perfect reliability.

Figure 7 shows the model's performance by cloud regime. In the attributes diagram for F_{down}^{sfc} (Figure 7a), reliability is nearly perfect for all three cloud regimes, except a general underprediction up to $\sim$20 W m^{-2} for no-cloud examples. For F_{up}^{TOA} (Figure 7b), reliability is good for all three cloud regimes, except a general underprediction up to $\sim$20 W m^{-2} for single-layer cloud and $\sim$50 W m^{-2} for multi-layer cloud, as well as a large underprediction for single-layer cloud in the two lowest bins. In other words, the lowest predicted F_{up}^{TOA} values for single-layer cloud tend to be far too low. For F_{net} (Figure 7c), reliability is nearly perfect for all three cloud regimes, except a general overprediction up to $\sim$20 W m^{-2} for multi-layer cloud. In the bias profile for heating rate (Figure 7d), examples with no cloud and single-layer cloud have an absolute bias < 0.1 K day^{-1} except in the upper stratosphere, as for the whole testing set (Figure 6d). However, for examples with multi-layer cloud, absolute bias slightly exceeds 0.1 K day^{-1} at a few heights in the mid-troposphere. Also, in the profiles of MAE and MAE skill score (Figures 7e-f), the worst values in the troposphere are for

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

multi-layer cloud in the middle to upper troposphere. This is because (a) multi-layer clouds lead to the most complex heating-rate profiles, due to the non-local effects discussed in Section 3c3; (b) examples with multi-layer cloud are rare (0.86% of U-net++-training examples), and rare events are inherently hard to predict. For all three cloud regimes, the reliability curves for heating rate (Figures 7g-i) are near the perfect line. However, the reliability curves are jagged for multi-layer cloud, due to small sample size.

Supplemental Figure S20 is analogous to Figure 7, except for a U-net++ that does not include F_{net} in the loss function (*i.e.*, one that uses the post-processing approach discussed in Section 3c1). For examples with liquid cloud, predictions of F_{down}^{sfc} (panel a) and F_{net} (panel c) are significantly worse with the post-processing approach.

Figure 8 shows the model's performance by site. In attributes diagrams for the flux components (Figures 8a-c), reliability is nearly perfect, except that at a few sites, small positive predictions of F_{down}^{sfc} and F_{net} are up to $\sim$20 W m^{-2} too low. In the error profiles for heating rate (Figures 8d-f), all seven sites are similar to the whole testing set (Figures 6d-f), so there are no apparent outliers. Figures 8g-i show the reliability curves for heating rate at three randomly selected sites. Reliability is nearly perfect, except in the lower troposphere at the Perdido oil rig, where higher predictions are up to $\sim$0.5 K day^{-1} too low. This issue does not occur at the other four sites (not shown), whose reliability curves look similar to those for Bishop and Hilo.

Figure 9 shows the model's performance by zenith angle. For the sake of brevity, we show results for 1-km heating rate (lower troposphere), 10-km heating rate (upper troposphere in the testing data, which contain only tropical sites), 46-km heating rate (upper stratosphere; the height with the largest climatological heating rate), and F_{net}. Correlation is the Pearson correlation between predictions and observations, which ranges from $[-1, 1]$ and has an optimal value of 1. Kling-Gupta efficiency (KGE; Gupta et al. 2009) ranges from $(-\infty, 1]$, and the optimal value of 1 occurs

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

when the predictions and observations have perfect correlation, equal means, and equal variances. The unitless scores (left column of Figure 9) show that performance is worst at the extreme zenith angles, when the Sun is close to directly overhead or the horizon. However, except correlation and KGE for 46-km heating rate, unitless scores are close to their optimal values, even at local minima. Meanwhile, scores with units (MAE, RMSE, and bias) are shown in the right column of Figure 9. These scores are generally close to their optimum (0), except at zenith angles below 20°. At these zenith angles, the model has a negative bias for heating rate through most of the troposphere (including heights not shown) and negative bias for F_{net}, caused by a large negative bias for F_{down}^{sfc} and small negative bias for F_{up}^{TOA} (not shown). Zenith angles below 20° rarely occur in the training data (non-tropical sites only), so it is not surprising that the model has difficulty in generalizing to these scenarios.

c. Experiment 2

Again, results of the hyperparameter experiment are relegated to the Supplemental Material. The main conclusion to note here is the same as for Experiment 1: the U-net++ performs better with less regularization, which controls overfitting. Because models in Experiment 2 are trained with data from all latitudes, this is the model that would be used in NWP. Ultimately, we select the model with 4 fully connected layers, a dropout rate of 0.0, and L_2 weight of 10^{-7}. Results shown in the rest of this section, for the selected model only, are based on testing data rather than validation data.

Figure 10 shows the model's performance on the whole testing set (Assorted2 sites in 2017). For each flux component, the reliability is nearly perfect, as is the match between the observed and predicted histograms (Figures 10a-c). For heating rate, all heights have an absolute bias $\ll$ 0.1 K day^{-1}, including in the upper stratosphere (Figure 10d). As for the tropical testing data in

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Experiment 1, there is a spike in MAE at 46 km (Figure 10e), due to absorption by ozone, but the corresponding dip in MAE skill score is small (Figure 10f). In the reliability curve for heating rate (Figure 10g), all heights are nearly perfect, except in the lower troposphere, where higher predictions are up to $\sim$0.25 K day^{-1} too low.

Figure 11 shows the model's performance by cloud regime. For each flux component and each cloud regime, the reliability is nearly perfect (Figures 11a-c). For heating rate, all heights and all cloud regimes have an absolute bias $\ll 0.1$ K day^{-1} (Figure 11d), while values of MAE (Figure 11e) and MAE skill score (Figure 11f) are similar to the whole testing set. For all three cloud regimes, the reliability curves for heating rate (Figures 11g-i) are nearly perfect, with two exceptions: jagged curves for multi-layer cloud, due to small sample size, and the lower troposphere for single-layer cloud, where higher predictions are up $\sim$0.5 K day^{-1} too low.

Figure 12 shows the model's performance by site. For each flux component and each site, the reliability is nearly perfect (Figures 12a-c), except an underprediction of $\sim$50 W m^{-2} at the north pole for the lowest F_{up}^{TOA} bin (Figure 12b). In other words, the lowest predicted F_{up}^{TOA} values here tend to be 50 W m^{-2} too low. By inspection (not shown), we have found that this underprediction is associated with low albedos (< 0.7) at the north pole, which occur during the ice-free part of the year. Although the model correctly predicts that a lower albedo (less reflection from the surface) will lead to less upwelling radiation, it exaggerates this effect. For heating rates, all heights and sites have an absolute bias < 0.1 K day^{-1}, except at 46 km at the Perdido oil rig, where bias is $\sim$-0.11 K day^{-1} (Figure 12d). Profiles of MAE (Figure 12e) and MAE skill score (Figure 12f) are similar to the whole testing set, except for MAE at 46 km, where values are smaller at the Arctic sites (north pole and Tiksi) and larger at the tropical sites (Perdido and Bishop). This is because climatological 46-km heating rates are smaller at the Arctic sites (average of 29.7 K day^{-1} over the testing data) and larger at the tropical sites (36.3 K day^{-1}). Reliability for heating rate is nearly

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

perfect at the three sites shown (Figure 12g-i), except in the lower troposphere at Perdido, where higher predictions are up to $\sim$0.5 K day^{-1} too low. This issue also occurs for Perdido in Experiment 1 (Figure 8i). For the two sites not shown, reliability at Bishop is nearly perfect (similar to north pole and Lamont), while reliability at Tiksi has a similar issue to Perdido, except only at the lowest few heights and with an underprediction up to only $\sim$0.25 K day^{-1}.

Figure 13 shows the model's performance by zenith angle. The unitless scores (left column) show that performance is worst at the extreme zenith angles, but in general scores are better than for Experiment 1 (Figure 9), including at the lowest zenith angles. This is because the model from Experiment 2 is trained with more low zenith angles, due to the inclusion of tropical sites. Meanwhile, scores with units (right column of Figure 13) are very close to their optimum (0), especially bias, at all zenith angles. This contrasts starkly with the results for Experiment 1 (Figure 9), where every target variable has substantial bias for zenith angles $< 20^\circ$.

d. Additional analyses

Supplemental Section Ca presents a Kolmogorov-Smirnov and bias-variance analysis for both selected models (from Experiments 1 and 2). The main conclusions are (a) the models have more random variance than systematic bias; (b) although the difference between the predicted and observed distributions of heating rate are small, they are generally significant at the 99% level (as determined by the Kolmogorov-Smirnov p-value), because the sample sizes are large. Supplemental Section Cb shows results on training, validation, and testing data for both selected models. Although both models overfit to some extent, results on the testing data are highly skillful, as discussed in Sections 5b-c. Also, the model from Experiment 1 overfits more, because it performs more extreme spatial generalization (from non-tropical to tropical sites).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

e. Comparison of selected models

Overall, the model from Experiment 2 appears to outperform the model from Experiment 1 on testing data, consistent with our hypothesis. The comparison is not perfectly apples-to-apples, because the two testing sets contain different collections of sites, but they have two sites in common, both in the tropics: the Perdido oil rig and Bishop, Grenada. According to the site-specific reliability curves for heating rate (*c.f.* Figures 8g-i and 12g-i), there is no substantial difference between the two models. According to the site-specific attributes diagrams for flux components (*c.f.* Figures 8a-c and 12a-c), site-specific error profiles for heating rates (*c.f.* Figures 8d-f and 12d-f), and results for the lowest zenith angles (*c.f.* Figures 9 and 13) – seen primarily in the tropics – the model from Experiment 2 is significantly better.

Supplemental Section Cc compares the two models on a second testing set, containing non-tropical sites in 2017. The purpose of this analysis is to achieve a fairer comparison, using the same data. The model from Experiment 2 performs better on the second testing set as well, even though it was trained with only *some* tropical sites, while the model from Experiment 1 was trained with *all* tropical sites. We suspect that training on tropical sites allowed the model from Experiment 2 to learn additional relationships that improve its performance on non-tropical sites.

At inference time, both models (including the U-net++ and isotonic regression) can generate predictions for $\sim$500 000 profiles in one minute, while the shortwave RRTM can process $\sim$50 profiles in one minute. Thus, the ML models are $\sim 10^4$ times faster than the shortwave RRTM, which they emulate with impressive skill.

Lastly, Supplemental Section Cd compares the selected model (U-net++) from Experiment 1 to a traditional U-net and fully connected neural network (FCNN), developed via similar hyperparameter searches. The U-net and U-net++ clearly and significantly (at the 99% level) outperform the

FCNN, demonstrating the advantage of spatially aware layers (convolution and pooling). However, differences between the U-net and U-net++ are mixed, with the U-net performing better on some target variables and the U-net++ performing better on others. However, we believe that a major advantage of the U-net++ is superior performance on F_{net} in profiles with multi-layer cloud. F_{net} is arguably the single most important target variable (*i.e.*, more important than F_{down}^{sfc}, F_{up}^{TOA}, or heating rate at any individual height), and F_{net} errors are highest in profiles with multi-layer cloud, where radiative transfer is most complicated. Specifically, in profiles with multi-layer cloud, the U-net++ improves the absolute bias on F_{net} by $\sim$15 W m^{-2} compared to the U-net, and the difference is statistically significant. Also, the U-net++ significantly outperforms the U-net in predicting the other two flux components, F_{down}^{sfc} and F_{up}^{TOA}, with multi-layer cloud. We believe that this advantage of the U-net++ is due to more skip connections better preserving high-resolution information, which is crucial in profiles with multi-layer cloud and cloud in general (clouds create a discontinuity in the profile of extinction optical depth, so their exact boundaries matter). Since we do not train the U-net++ with deep supervision (another modification to U-nets proposed by Zhou et al. 2019, where intermediate feature maps, not only the output, are included in the loss function), this advantage of the U-net++ is not a result of deep supervision.

6. Model Interpretation

The permutation test measures the overall importance of each predictor variable, averaged over all grid points (*i.e.*, all heights for vector predictors) and testing examples. There are four versions of the permutation test – forward single-pass, forward multi-pass, backward single-pass, and backward multi-pass – which each handle correlated predictors differently. The backward multi-pass test begins with all predictors permuted – *i.e.*, randomly shuffled so that values are assigned to the wrong examples – and iteratively restores (puts back in the correct order) the most important

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

predictor still permuted, until all predictors have been restored. The k^{th} predictor to be restored is considered the k^{th}-most important. For more details on the permutation test, see McGovern et al. (2019). We run the permutation test with one of two loss functions – the dual-weighted MSE for heating rates (first term in Equation 3) or standard MSE for flux components (second term in Equation 3) – so that we can determine the most important predictors for each type of output. Figure 14 shows results for the backward multi-pass test, and Supplemental Figures S21-S23 show results for the other versions, which are very similar. We run the permutation test for both selected models, from Experiments 1 and 2.

With the heating-rate-only loss function, results for the two models (Figures 14a,c) agree on the top four predictors: zenith angle, LWC, downward LWP, and relative humidity. In other words, the most important factors for radiative heating are Sun angle, liquid water, and water vapour, with ice being much less important – likely because the dual-weighted MSE emphasizes large heating rates, which typically are not caused by ice clouds (Turner et al. 2018). With the flux-only loss function, results for the two models (Figures 14b,d) agree on the top four predictors: downward LWP, LWC, zenith angle, and surface albedo. Surface albedo is especially important for F_{up}^{TOA}, as higher-albedo surfaces reflect more radiation back to space. Surface albedo is much less important for heating rates (Figures 14a,c), because heating rates are measured at all 73 heights, which are generally far from the surface. All results discussed in this paragraph are significant at the 99% level, as indicated by the bold font in Figure 14.

7. Summary and Future Work

We developed U-net++ models, a type of deep learning, to emulate the shortwave RRTM. The U-net++ architecture contains more skip connections than the traditional U-net architecture, which improved our flux predictions in profiles with multi-layer cloud, while the inclusion of physical

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

constraints improved both flux and heating-rate predictions for multi-layer cloud. We bias-corrected the U-net++ models with isotonic regression, a simple ML method often used for this purpose. We conducted two hyperparameter experiments to find the best U-net++ configurations for predicting two output types: a heating-rate profile and three flux components (F_{down}^{sfc}, F_{up}^{TOA}, and F_{net}). In both experiments we found that the models perform best with minimal regularization, contrary to our prior experience with ML in atmospheric science. This result may generalize to other perfect-model experiments, where ML is used to emulate another model rather than fit observations.

We performed two experiments, with sites split among training and testing in different ways. In Experiment 1, we trained the models on non-tropical sites and tested on tropical sites, with the purpose of testing the models' spatial-generalization ability under extreme conditions (to a completely different region). In Experiment 2, we trained the models on assorted sites from all regions and tested on a different set of assorted sites from all regions, with the purpose of creating the best model possible for use as a parameterization in NWP. The selected model from Experiment 1 showed impressive skill on the testing set (tropical sites), but with four notable deficiencies. First, it has a large bias and MAE for heating rate in the upper stratosphere, where radiative heating is dominated by ozone absorption. Second, the lowest F_{up}^{TOA} predictions for examples with single-layer cloud have a large negative bias, of several hundred W m^{-2}. Third, the heating-rate bias for multi-layer cloud slightly exceeds 0.1 K day^{-1} (considered a threshold for stable integration into NWP) in the mid-troposphere. Fourth, at zenith angles below 20° (seldom seen in the training data), the model has a negative bias for the three flux components and for heating rates throughout the troposphere. With the exception of large MAE for heating rates in the upper stratosphere, none of these deficiencies appear in the testing data for the selected model from Experiment 2. According to the permutation test for both models, the most important factors for heating rate

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

(flux components) are zenith angle, liquid water, and water vapour (liquid water, zenith angle, and surface albedo).

The remainder of this section focuses on the model from Experiment 2, which outperforms the model from Experiment 1. In addition to closely emulating the shortwave RRTM, this model is $\sim 10^4$ times faster than the shortwave RRTM. In terms of heating rate, our performance is better than the emulator of Krasnopolsky et al. (2010, henceforth K10), which is a traditional (or fully connected) neural network. Their neural network achieves a PRMSE of 0.15 K day^{-1} (their Table 1), versus our 0.056 K day^{-1} on testing data[4]. In terms of F_{up}^{TOA} (K10 do not show results for the other flux components), our bias on the whole testing set is -2.2 W m^{-2}, with bias at individual testing sites ranging from -3.2 to -1.2 W m^{-2}. The overall bias of K10's emulator (their Figure 2, top right) is also negative, with zonal-mean bias ranging from approximately -3 to +1.25 W m^{-2}. Thus, our results for F_{up}^{TOA} are comparable with K10. However, the comparison is not apples-to-apples, because K10 evaluate on data from different times and locations; they emulate the RRTMG, rather than the RRTM; and they emulate the full RRTMG, including aerosols and non-climatological trace gases.

We attribute the success of our models to four factors. The first is the adoption of U-nets, which are specially designed to learn from gridded data and make pixelwise predictions. The second is the adoption of the U-net++ architecture, which outperforms the traditional U-net architecture in predicting fluxes with multi-layer cloud. The third factor is using isotonic regression for bias correction, and the fourth is knowledge-guided ML. We achieved knowledge-guided ML by incorporating a physical law (Equation 2) into the U-net++ models to ensure physically consistent and skillful F_{net} predictions, developing a custom loss function (Equation 3) to emphasize large heating

[4]Even for the model from Experiment 1, which is trained on non-tropical sites and tested on tropical sites, the testing PRMSE is 0.108 K day^{-1}.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

rates, and including custom predictors to allow vertical non-locality in heating-rate predictions, which is especially important for examples with multi-layer cloud.

We will continue this work along five lines. The first is developing models to emulate the full shortwave RRTM, including the effects of aerosols, precipitation, and non-climatological profiles of trace gases. Second, we will also emulate the longwave RRTM, using a similar framework. Third, we will make the models grid-agnostic (insensitive to exact heights in the profile), so that they can be applied to NWP models with different vertical grids. Fourth, we will experiment with other neural-network architectures, such as the U-net 3+ (Huang et al. 2020), which contains "full-scale" skip connections, combining data from all spatial resolutions at once, rather than just neighbouring resolutions as in the U-net++. Fifth, we will test the new models (emulating the full shortwave and longwave RRTM) online, *i.e.*, inside an NWP model as parameterizations. Since the models developed herein are orders of magnitude faster than the RRTM, if they were integrated stably into NWP, they could also be called at every atmospheric time step, which should improve the overall accuracy of the NWP model and free up computing time for other improvements to NWP.

Acknowledgments. We acknowledge Christina Kumler for ideological input during this project, as well as exploratory work during the preparation phase. 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. Input data (predictors and targets from 2017-2020) are available upon request from the authors, as well as the selected models (U-net++ and isotonic regression) for both Experiments 1 and 2. We used version 1.0.0 of ML4RT (Machine Learning for Radiative Transfer;

doi:10.5281/zenodo.4470077), a Python library managed by author Lagerquist, to train, evaluate, and interpret all ML models (both U-net++ and isotonic regression) in this work. Since U-net++ architecture is complicated, for each experiment we have included a script that creates the architecture for the selected U-net++. These can be found at `scripts/make_best_architecture_exp1.py` and `scripts/make_best_architecture_exp2.py`, respectively, in the Python library.

References

Barlow, R., and H. Brunk, 1972: The isotonic regression problem and its dual. *Journal of the American Statistical Association*, **67 (337)**, 140–147, URL http://doi.org/10.1080/01621459.1972.10481216.

Benjamin, S., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, **144 (4)**, 1669–1694.

Berthomier, L., and B. Pradel, 2021: Cloud cover nowcasting with deep learning. *Conference on Artificial Intelligence for Environmental Science*, New Orleans, Louisiana (became virtual due to COVID-19), American Meteorological Society, URL https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/380983.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2021: Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, **126 (9)**, 098 302, URL https://doi.org/10.1103/PhysRevLett.126.098302.

Bolton, T., and L. Zanna, 2019: Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, **11 (1)**, 376–399.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Brenowitz, N., T. Beucler, M. Pritchard, and C. Bretherton, 2020: Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, **77 (12)**, 4357–4375, URL https://doi.org/10.1175/JAS-D-20-0082.1.

Brenowitz, N., and C. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, **45 (12)**, 6289–6298, URL https://doi.org/10.1029/2018GL078510.

Chen, M., X. Shi, Y. Zhang, D. Wu, and M. Guizani, 2017: Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, URL https://doi.org/10.1109/TBDATA.2017.2717439.

Chen, Y., L. Bruzzone, L. Jiang, and Q. Sun, 2020: Aru-net: Reduction of atmospheric phase screen in sar interferometry using attention-based deep residual u-net. *IEEE Transactions on Geoscience and Remote Sensing*, **early online release**, 1–14, URL https://doi.org/10.1109/TGRS.2020.3021765.

Chollet, F., 2018: *Deep Learning with Python*. Manning, Shelter Island, New York.

Chollet, F., and Coauthors, 2020: Keras. GitHub, URL https://github.com/fchollet/keras.

Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, **online**, 1–49, URL https://doi.org/10.1175/BAMS-D-20-0097.1.

Felt, V., S. Samsi, and M. Veillette, 2021: A comprehensive evaluation of deep neural network architectures for precipitation nowcasting. *Conference on Artificial Intelligence for Environmental Science*, New Orleans, Louisiana (became virtual due to COVID-19), American Meteorological Society, URL https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/383115.

Fukushima, K., 1980: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36 (4)**, 193–202.

Fukushima, K., and S. Miyake, 1982: Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, **15 (6)**, 455–469.

Gagne, D., S. Haupt, D. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, **In Press**, URL https://doi.org/10.1175/MWR-D-18-0316.1.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, 2018: Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, **45 (11)**, 5742–5751, URL https://doi.org/10.1029/2018GL078202.

Gil, Y., and Coauthors, 2019: Intelligent systems for geosciences: An essential research agenda. *Communications of the Association for Computing Machinery*, **62 (1)**, 76–84.

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

Gupta, H., H. Kling, K. Yilmax, and G. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, **377 (1-2)**, 80–91, URL https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hayatbini, N., A. Badrinath, W. Chapman, L. D. Monache, F. Cannon, P. Gibson, A. Subramanian, and F. Ralph, 2021: Precipitation forecast with ConvLSTM using 3-dimensional numerical weather predictions. *Conference on Artificial Intelligence for Environmental Science*, New Orleans, Louisiana (became virtual due to COVID-19), American Meteorological Society, URL https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/381949.

Hinton, G., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2012: Improving neural networks by preventing co-adaptation of feature detectors. *arXiv e-prints*, **1207 (0580)**.

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. *International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, IEEE, URL Unet3+:Afull-scaleconnectedunetformedicalimagesegmentation.

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

Iacono, M., E. Mlawer, S. Clough, and J. Morcrette, 2000: Impact of an improved longwave radiation model, RRTM, on the energy budget and thermodynamic properties of the NCAR community climate model, CCM3. *Journal of Geophysical Research: Atmospheres*, **105 (D11)**, 14 873–14 890, URL https://doi.org/10.1029/2000JD900091.

Iacono, M., E. Mlawer, J. Delamere, S. Clough, J. Morcrette, and Y. Hou, 2005: Application of the Shortwave Radiative Transfer Model, RRTMG_SW, to the National Center for Atmospheric Research and National Centers for Environmental Prediction general circulation models. *Atmospheric Radiation Measurement (ARM) Science Team Meeting*, Daytona Beach, Florida, URL http://www.academia.edu/download/40132192/Application_of_the_Shortwave_Radiative_T20151118-2282-447csj.pdf.

Krasnopolsky, V., 2020: Using machine learning for model physics: An overview. *arXiv e-prints*, **2002 (00416)**, URL https://arxiv.org/abs/2002.00416.

Krasnopolsky, V., M. Fox-Rabinovitz, Y. Hou, S. Lord, and A. Belochitski, 2010: Accurate and fast neural network emulations of model radiation for the NCEP coupled climate forecast system: climate simulations and seasonal predictions. *Monthly Weather Review*, **138 (5)**, 1822–1842, URL https://doi.org/10.1175/2009MWR3149.1.

Kumler-Bonfanti, C., J. Stewart, D. Hall, and M. Govett, 2020: Tropical and extratropical cyclone detection using deep learning. *Journal of Applied Meteorology and Climatology*, **59 (12)**, 1971–1985, URL https://doi.org/10.1175/JAMC-D-20-0117.1.

Kurth, T., and Coauthors, 2018: Exascale deep learning for climate analytics. *International Conference for High Performance Computing, Networking, Storage, and Analysis*, Dallas, Texas, Institute of Electrical and Electronics Engineers (IEEE).

Lagerquist, R., J. Allen, and A. McGovern, 2020a: Climatology and variability of warm and cold fronts over north america from 1979 to 2018. *Journal of Climate*, **33 (15)**, 6531–6554, URL https://doi.org/10.1175/JCLI-D-19-0680.1.

Lagerquist, R., A. McGovern, and D. Gagne, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, **34 (4)**, 1137–1160.

Lagerquist, R., A. McGovern, C. Homeyer, D. Gagne, and T. Smith, 2020b: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, **148 (7)**, 2837–2861, URL https://doi.org/10.1175/MWR-D-19-0372.1.

Long, J., E. Shelhamer, and T. Darrell, 2015: Fully convolutional networks for semantic segmentation. *Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, Institute

of Electrical and Electronics Engineers (IEEE), URL https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

McGovern, A., R. Lagerquist, D. Gagne, G. Jergensen, K. Elmore, C. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175–2199, URL https://doi.org/10.1175/BAMS-D-18-0195.1.

Mlawer, E., S. Taubman, P. Brown, M. Iacono, and S. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlatedk 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.

Morrison, H., G. de Boer, G. Feingold, J. Harrington, M. Shupe, and K. Sulia, 2012: Resilience of persistent Arctic mixed-phase clouds. *Nature Geoscience*, **5 (1)**, 11–17, URL https://doi.org/10.1038/ngeo1332.

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.

Racah, E., C. Beckham, T. Maharaj, S. Kahou, Prabhat, and C. Pal, 2017: ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

of extreme weather events. *Advances in Neural Information Processing Systems*, Long Beach, California, Neural Information Processing Systems.

Reichstein, M., G. Camps-Balls, B. Stevens, M. Jung, J. Denzler, and N. Carvalhais, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204.

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.

Sadeghi, M., P. Nguyen, K. Hsu, and S. Sorooshian, 2020: Improving near real-time precipitation estimation using a u-net convolutional neural network and geographical information. *Environmental Modeling and Software*, **134**, URL https://doi.org/10.1016/j.envsoft.2020.104856.

Sha, Y., D. Gagne, G. West, and R. Stull, 2020a: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology*, **59 (12)**, 2057–2073, URL https://doi.org/10.1175/JAMC-D-20-0057.1.

Sha, Y., D. Gagne, G. West, and R. Stull, 2020b: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *Journal of Applied Meteorology and Climatology*, **early online release**, URL https://doi.org/10.1175/JAMC-D-20-0058.1.

Shanker, M., M. Hu, and M. Hung, 1996: Effect of data standardization on neural network training. *Omega*, **24 (4)**, 385–397, URL https://doi.org/10.1016/0305-0483(96)00010-2.

Stamnes, K., S. Tsay, W. Wiscombe, and K. Jayaweera, 1988: Numerically stable algorithm for discrete-ordinate-method radiative transfer in multiple scattering and emitting layered media. *Applied Optics*, **27 (12)**, 2502–2509, URL https://doi.org/10.1364/AO.27.002502.

Stewart, J., C. Kumler, D. Hall, and M. Govett, 2020: Deep learning approach for the detection of areas likely for convection initiation. *Conference on Artificial Intelligence for Environmental Science*, Boston, Massachusetts, American Meteorological Society, URL https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/365670.

Stone, P., 1978: Constraints on dynamical transports of energy on a spherical planet . *Dynamics of Atmospheres and Oceans*, **2 (2)**, 123–139, URL https://doi.org/10.1016/0377-0265(78)90006-4.

Turner, D., M. Shupe, and A. Zwink, 2018: Characteristic atmospheric radiative heating rate profiles in Arctic clouds as observed at Barrow, Alaska. *Journal of Applied Meteorology and Climatology*, **57 (4)**, 953–968, URL https://doi.org/10.1175/JAMC-D-17-0252.1.

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.

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

Wang, L., K. Scott, L. Xu, and D. Clausi, 2016: Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, **54 (8)**, 4524–4533.

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

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Wood, R., 2012: Stratocumulus clouds. *Monthly Weather Review*, **140 (8)**, 2373–2423, URL https://doi.org/10.1175/MWR-D-11-00121.1.

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.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

List of Tables

Table 1: Description of predictor variables. "Vector" means that the variable is defined at all 73 heights. If the cell does not contain a check mark, the variable is a scalar. Downward LWP at height z is LWC integrated from the top of the atmosphere down to z, and upward LWP at height z is LWC integrated from the surface up to z. The definitions of IWP and WVP are analogous.

Variable	**Units**	**Vector?**
Solar zenith angle	Degrees	
Surface albedo	Unitless	
Temperature	Kelvins	✓
Pressure	Pa	✓
Specific humidity	kg kg^{-1}	✓
Relative humidity	Unitless	✓
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}	✓

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Table 2: Training, validation, and testing data for each experiment. "Non-tropical" means both Arctic and mid-latitude. "Assorted1" contains sites from all regions; "Assorted2" also contains sites from all regions, which do not overlap with those in Assorted1. The validation and testing sets are used to evaluate bias-corrected U-net++ models (with isotonic regression).

Experiment 1			
Dataset	**Years**	**Sites**	**Number of Examples**
Training for U-net++	2019-2020	Non-tropical	1.50 million
Training for isotonic regression	2018, excluding last week	Non-tropical	0.89 million
Validation	2017, excluding last week	Non-tropical	0.42 million
Testing	2017, excluding last week	Tropical	0.26 million

Experiment 2			
Dataset	**Years**	**Sites**	**Number of Examples**
Training for U-net++	2019-2020	Assorted1	1.72 million
Training for isotonic regression	2018, excluding last week	Assorted1	0.99 million
Validation	2017, excluding last week	Assorted1	0.55 million
Testing	2017, excluding last week	Assorted2	0.13 million

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

Table 3: Normalization of predictor and target variables for U-net++ models.

Variable(s)	**Method**
Predictor variables	Transform to uniform distribution, then z-scores
F_{down}^{sfc} and F_{up}^{TOA}	Transform to uniform distribution over $[0, 1]$
Heating rate	No normalization (leave in units of K day^{-1})

Table 4: Experimental hyperparameters for U-net++ models. The loss function is Equation 3.

Hyperparameter	**Values Attempted**
Number of fully connected layers	2, 3, 4, 5
Dropout rate for fully connected layers	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
L_2 weight for convolutional layers	$10^{-7.0}$, $10^{-6.5}$, $10^{-6.0}$, $10^{-5.5}$, $10^{-5.0}$, $10^{-4.5}$, $10^{-4.0}$, $10^{-3.5}$, $10^{-3.0}$

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

List of Figures

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

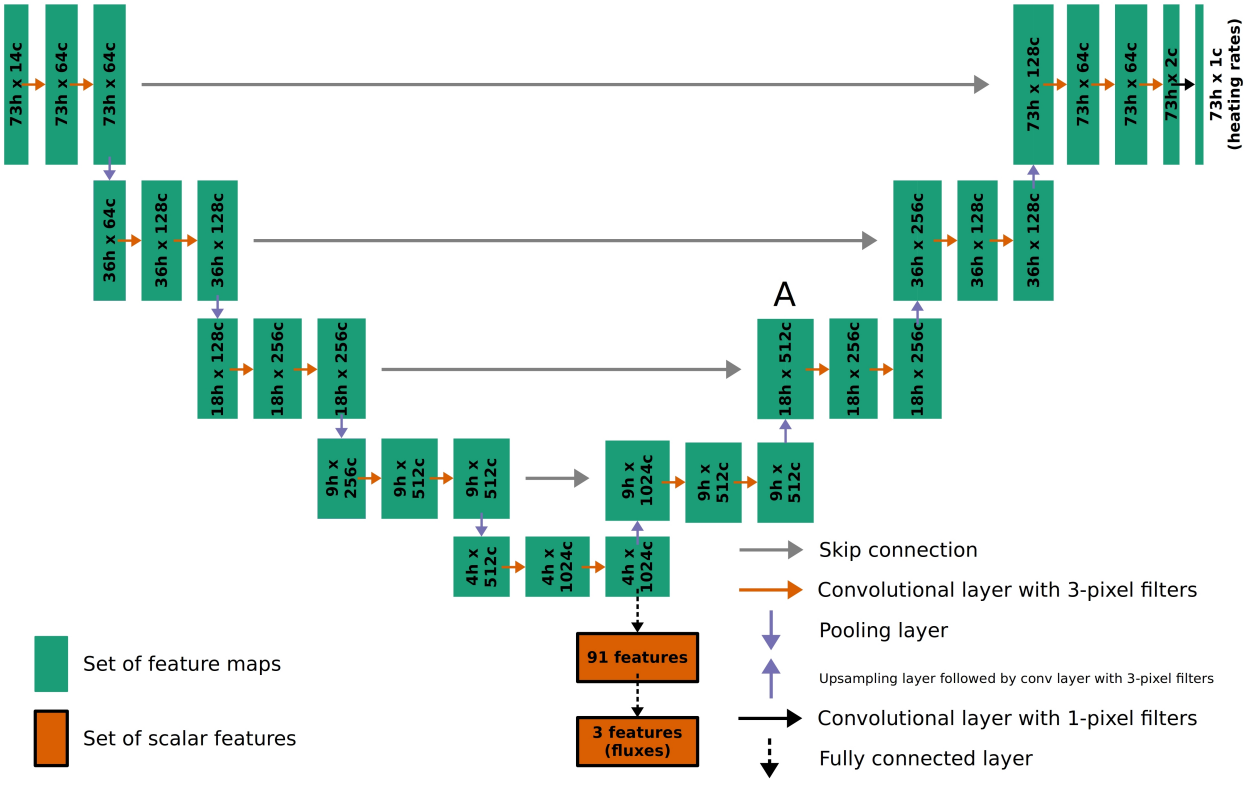

Figure 1: Architecture of traditional U-net with two fully connected layers. In each green box, "h" and "c" are the number of heights and channels (feature maps), respectively. The convolutional layer included wtih each upsampling layer (upward purple arrow), in addition to filling in spatial information, reduces the number of channels. For example, in the set of feature maps labeled "A," 256 channels come from the skip connection to the left and 512 channels come from the upsampling layer below. Thus, the convolutional layer included with this upsampling layer must reduce 768 channels to 512, which it achieves by having 512 filters. The shallowest layer is the convolutional layer at the top-left, and the deepest is the convolutional layer at the top-right. The top-left set of feature maps contains predictors: 14 variables at 73 heights. Although two predictor variables (albedo and zenith angle) are scalars, they are repeated over the 73 heights to create dummy grids, which are more easily input to the U-net. The outputs (predictions) are a length-73 grid of radiative-heating rates (top right) and three scalar fluxes (bottom).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

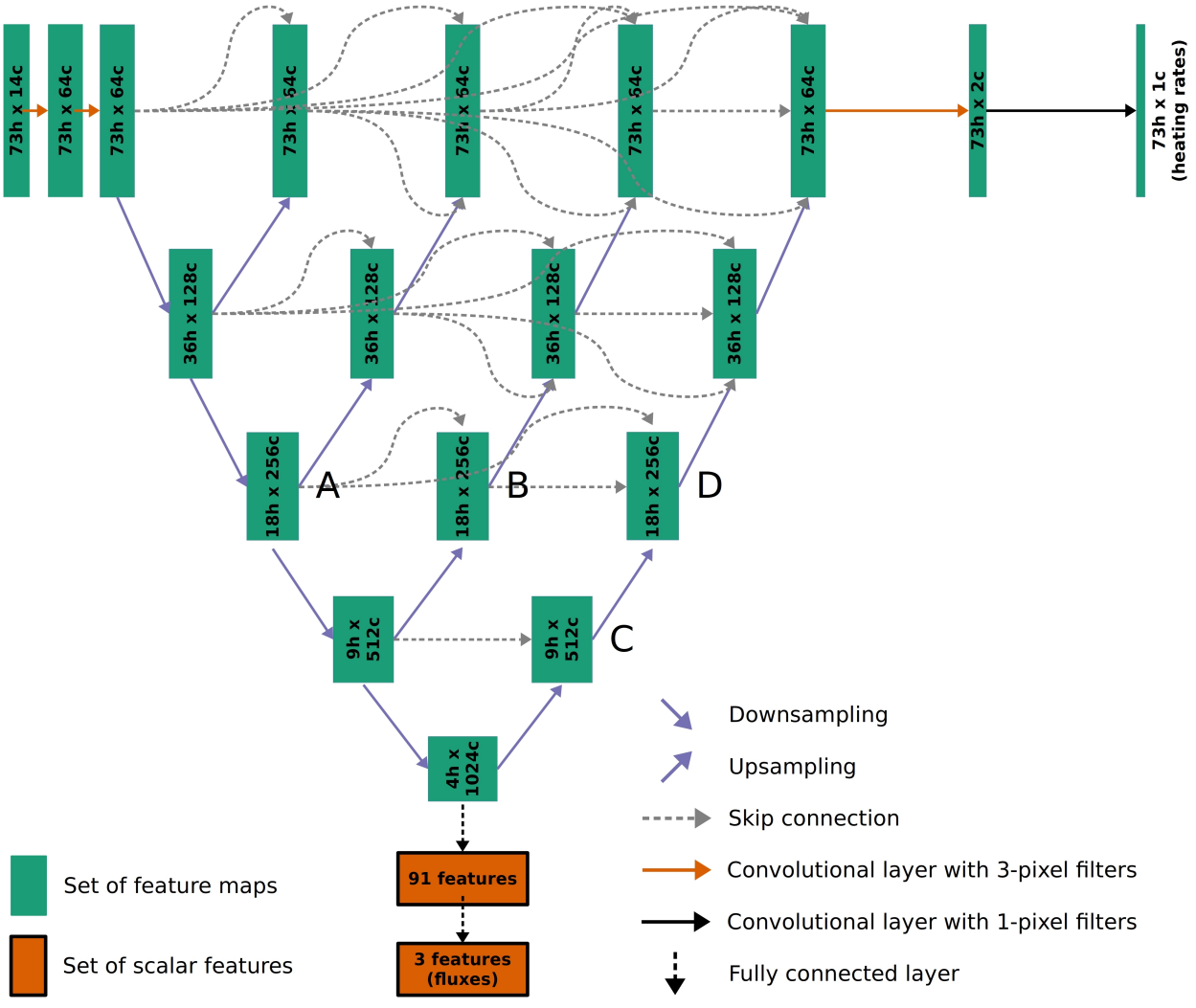

Figure 2: Architecture of U-net++ with two fully connected layers. Each "downsampling" arrow corresponds to a pooling layer followed by two convolutional layers with 3-pixel filters, as in one row of the downsampling side in Figure 1. Each "upsampling" arrow corresponds to an upsampling layer followed by two convolutional layers with 3-pixel filters, as in one row of the upsampling side in Figure 1. For each green box with multiple incoming arrows, feature maps are combined by concatenation (*i.e.*, stacking along the channel dimension), then convolution (with 3-pixel filters) to achieve the desired number of channels. For example, the set of feature maps labeled "D" is produced by concatenating A, B, and the upsampled version of C – which yields 1024 channels – then applying a convolutional layer that has 256 filters and therefore outputs 256 channels.

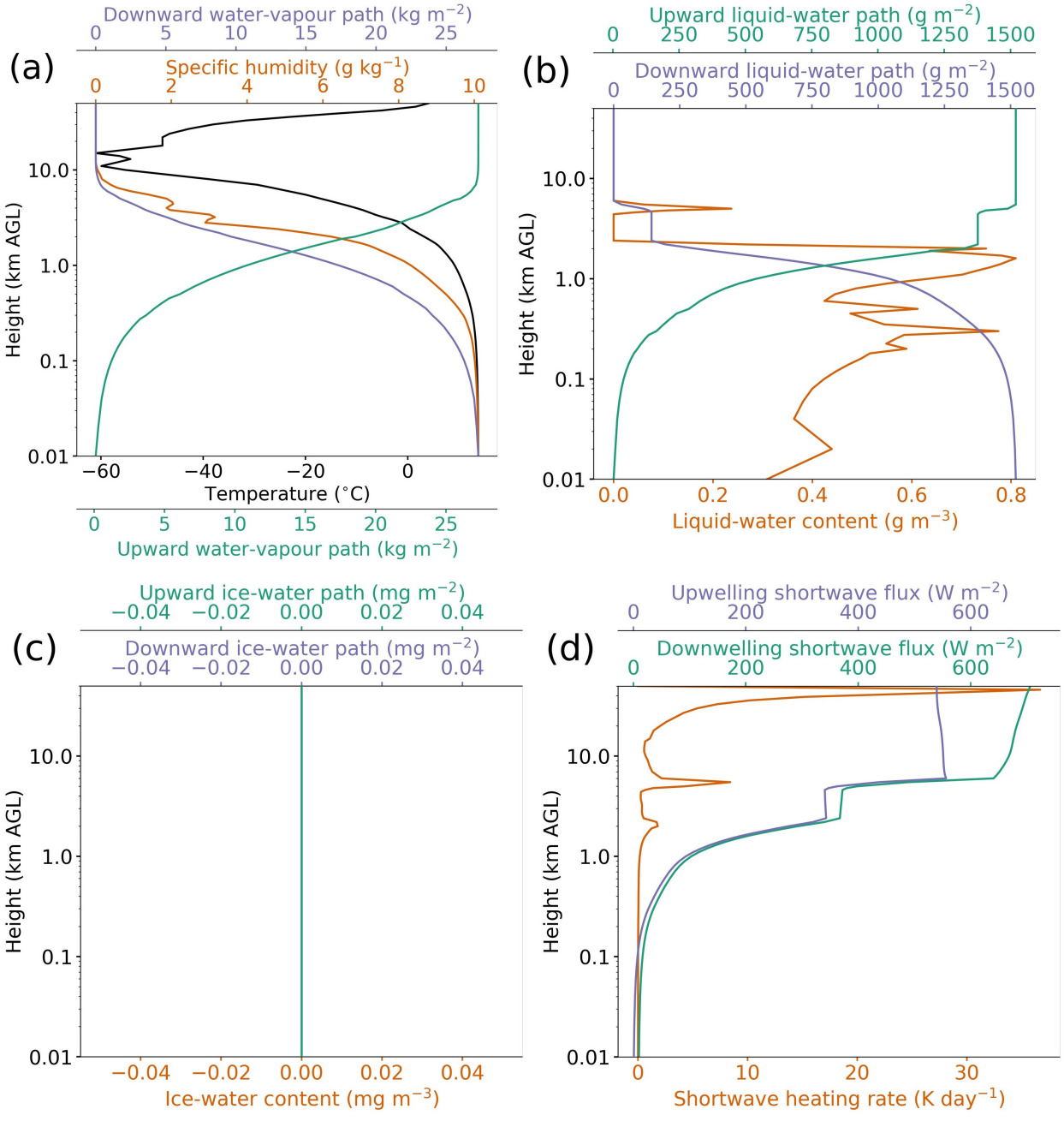

Figure 3: Predictor and target variables for one example: Santa Barbara, California, at 2200 UTC 16 Jan 2019. [a-c] All but four predictor variables: pressure, relative humidity, surface albedo, and solar zenith angle. [d] Target variables. Although the RRTM produces full profiles of downwelling and upwelling flux, the U-net++ models predict only the flux components required by an NWP model: F_{down}^{sfc} (the bottom value in the downwelling-flux profile), F_{up}^{TOA} (the top value in the upwelling-flux profile), and F_{net}, defined as $F_{down}^{sfc} - F_{up}^{TOA}$.

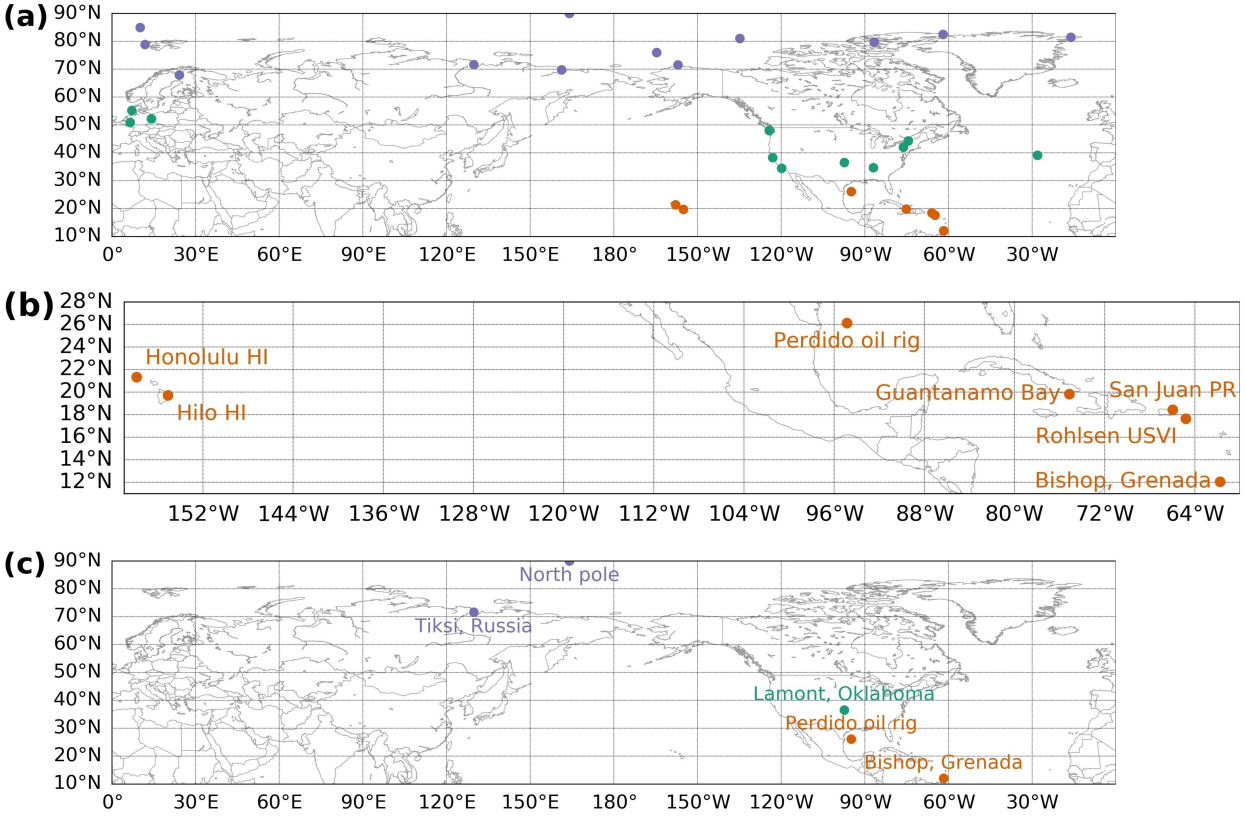

Figure 4: Sites used for model development (training, validation, and testing). Purple sites are in the Arctic; green sites are in the mid-latitudes; and orange sites are in the tropics. [a] All sites. [b] Testing sites for Experiment 1. [c] Testing sites for Experiment 2.

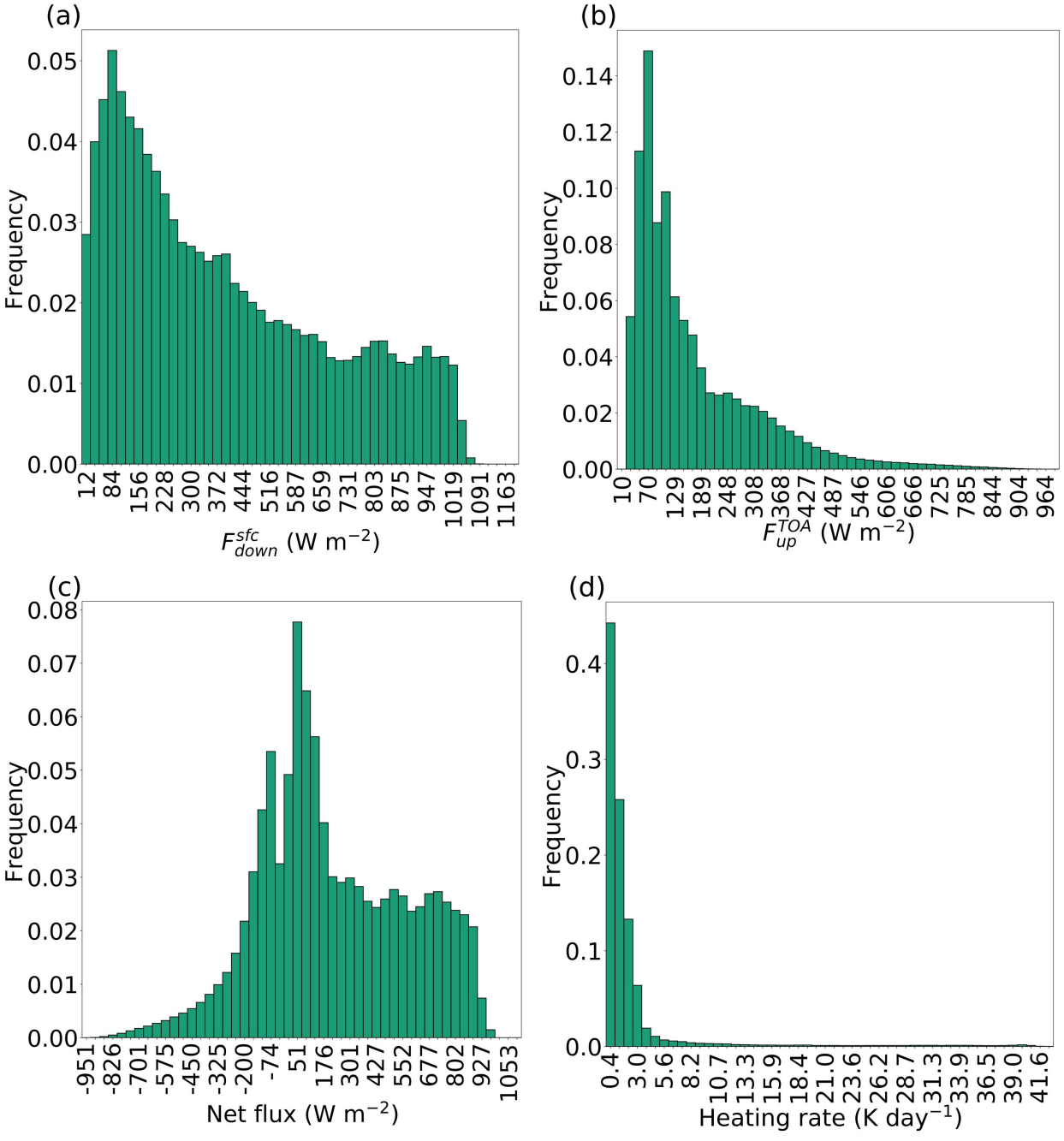

Figure 5: Distributions of target variables over the full dataset (all sites from 2017-2020).

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

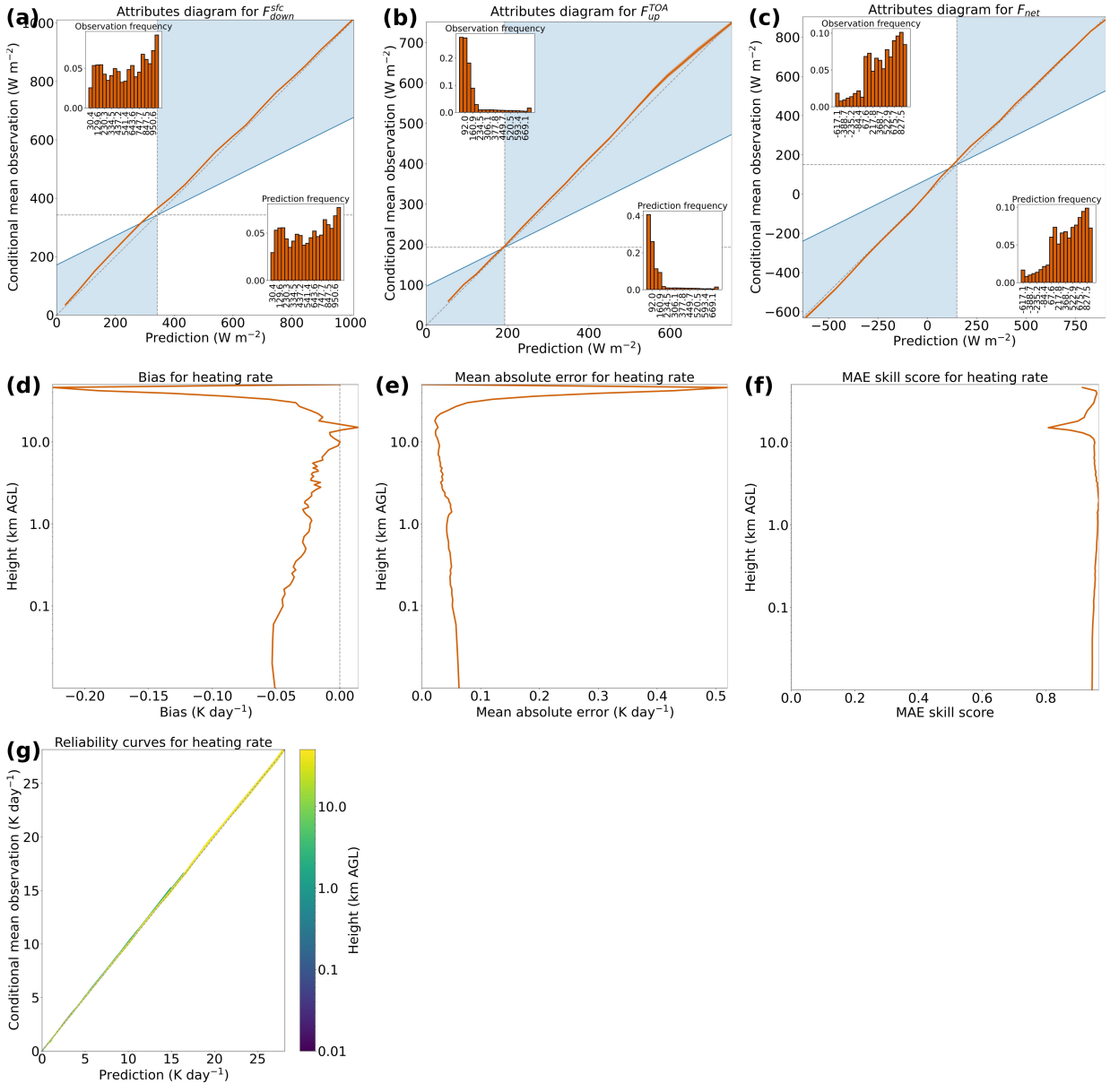

Figure 6: Performance of selected model from Experiment 1 on testing data. [a-c] Attributes diagram for each flux component. The orange curve is the reliability curve; the diagonal grey line is the perfect-reliability line; the vertical grey line is the climatology line; the horizontal grey line is the 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 heating rate. [g] Reliability curve at each height for heating rate. Each orange curve in panels a-f is the mean over 1000 bootstrap replicates. The 99% confidence interval is also plotted, but it is narrower than the line and thus invisible.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

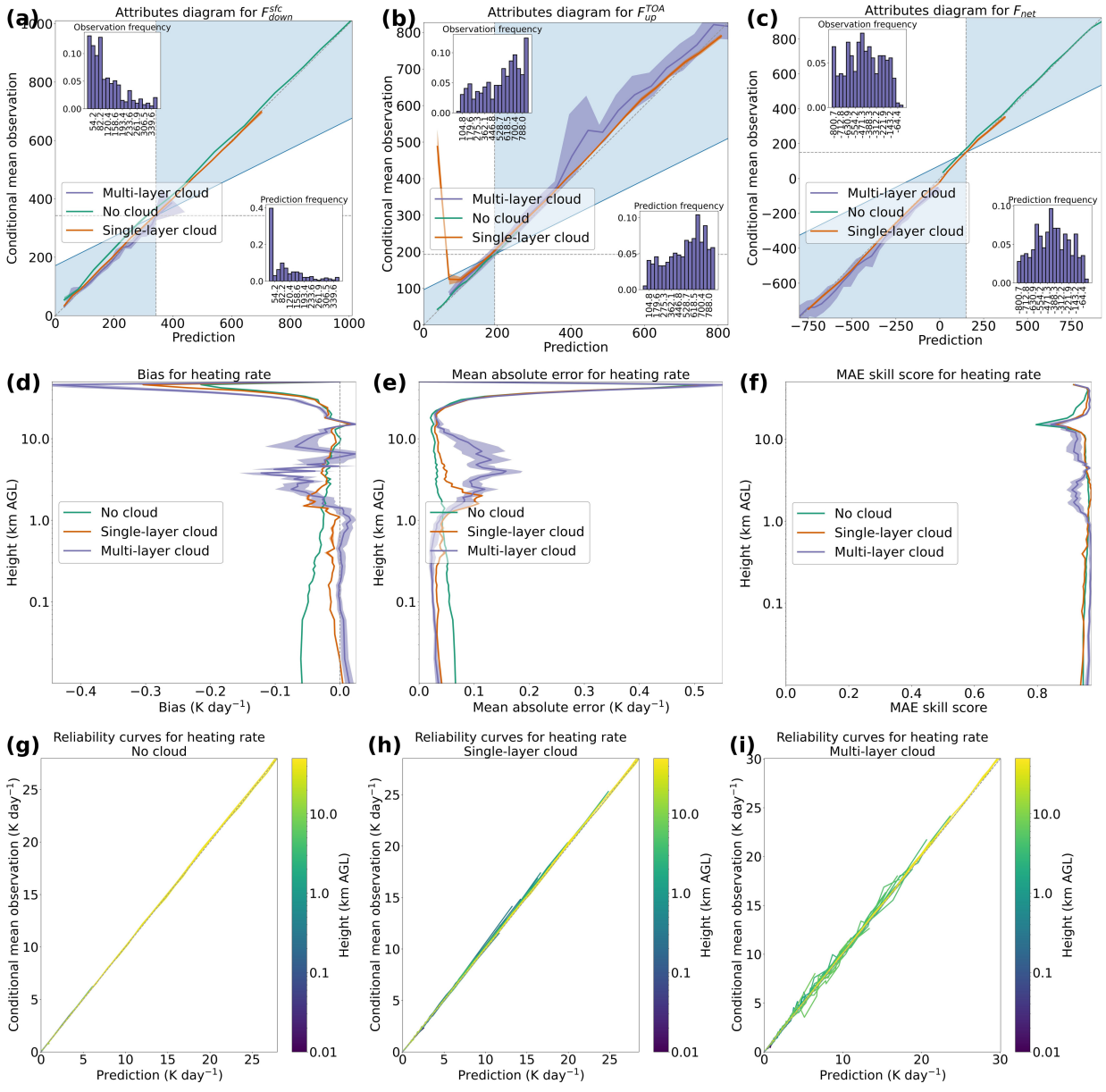

Figure 7: Performance of selected model from Experiment 1 on testing data, by cloud regime. [a-c] Attributes diagram (explained in the caption of Figure 6) for each flux component. The inset histograms and reference lines are based only on examples with multi-layer cloud. [d-f] Profiles of bias, MAE, and MAE skill score for heating rate. [g] Reliability curve at each height for heating rate, based only on examples with no cloud. [h] Same but for single-layer cloud. [i] Same but for multi-layer cloud. Each curve in panels a-f is the mean over 1000 bootstrapped replicates, and the surrounding shaded area is the 99% confidence interval.

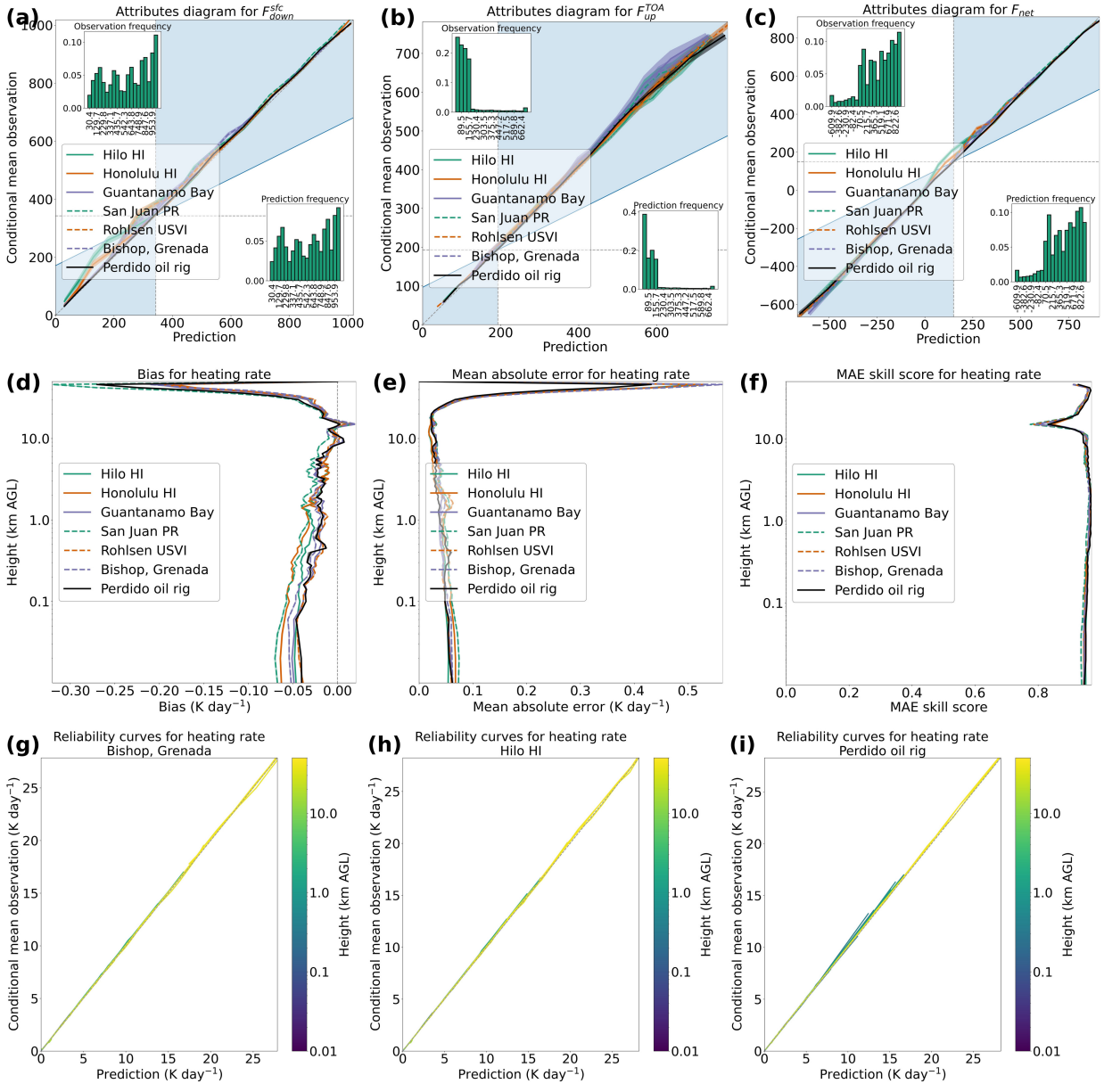

Figure 8: Performance of selected model from Experiment 1 on testing data, by site. [a-c] Attributes diagram (explained in the caption of Figure 6) for each flux component. In this case, The inset histograms and reference lines are based only on examples at Hilo, Hawaii. [d-f] Profiles of bias, MAE, and MAE skill score for heating rate. [g] Reliability curve at each height for heating rate, at Bishop, Grenada. [h] Same but for Hilo, Hawaii. [i] Same but for the Perdido oil rig. Each curve in panels a-f is the mean over 1000 bootstrapped replicates, and the surrounding shaded area is the 99% confidence interval.

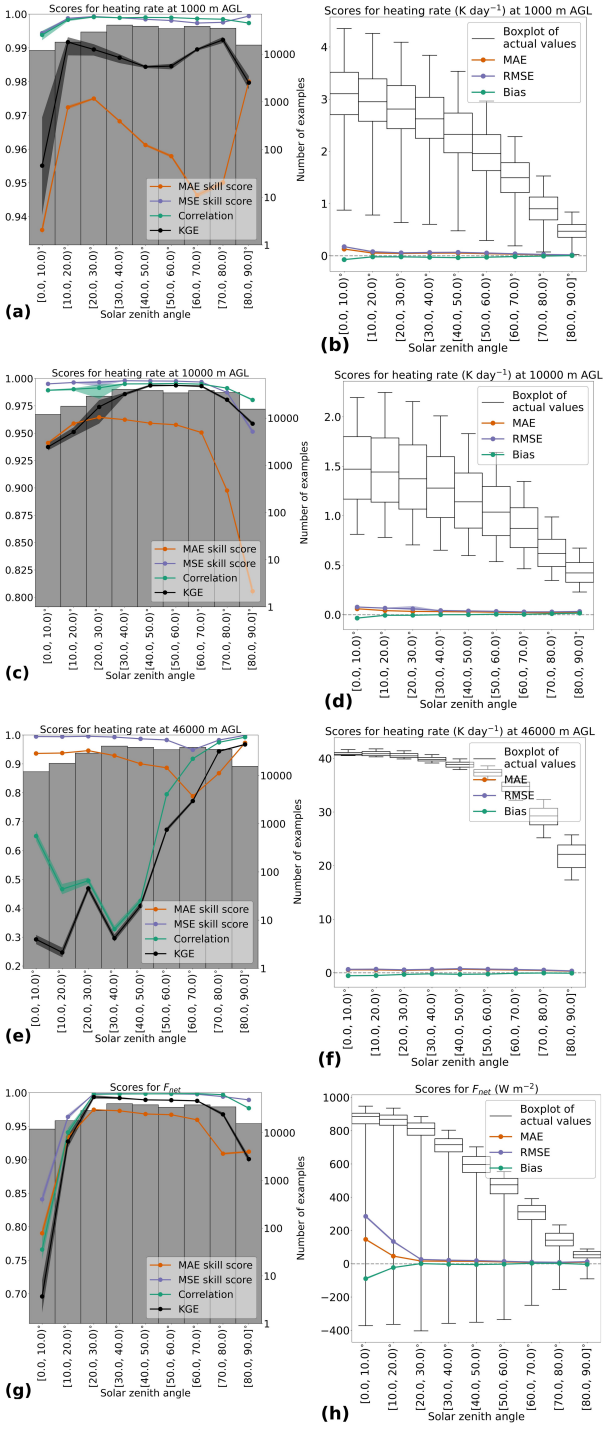

Figure 9: Performance of selected model from Experiment 1 on testing data, by solar zenith angle (0° means directly overhead, and 90° means on the horizon). [a-b] Scores without and with units, respectively, for heating rate at 1000 m AGL. [c-d] Same but for heating rate at 10 000 m AGL. [e-f] Same but for heating rate at 46 000 m AGL. [g-h] Same but for net flux. In each box plot, the center line is the median; the ends are the 25th and 75th percentiles; and the whiskers are the 5th and 95th percentiles. Each curve in panels a-h is the mean over 1000 bootstrapped replicates, and the surrounding shaded area is the 99% confidence interval.

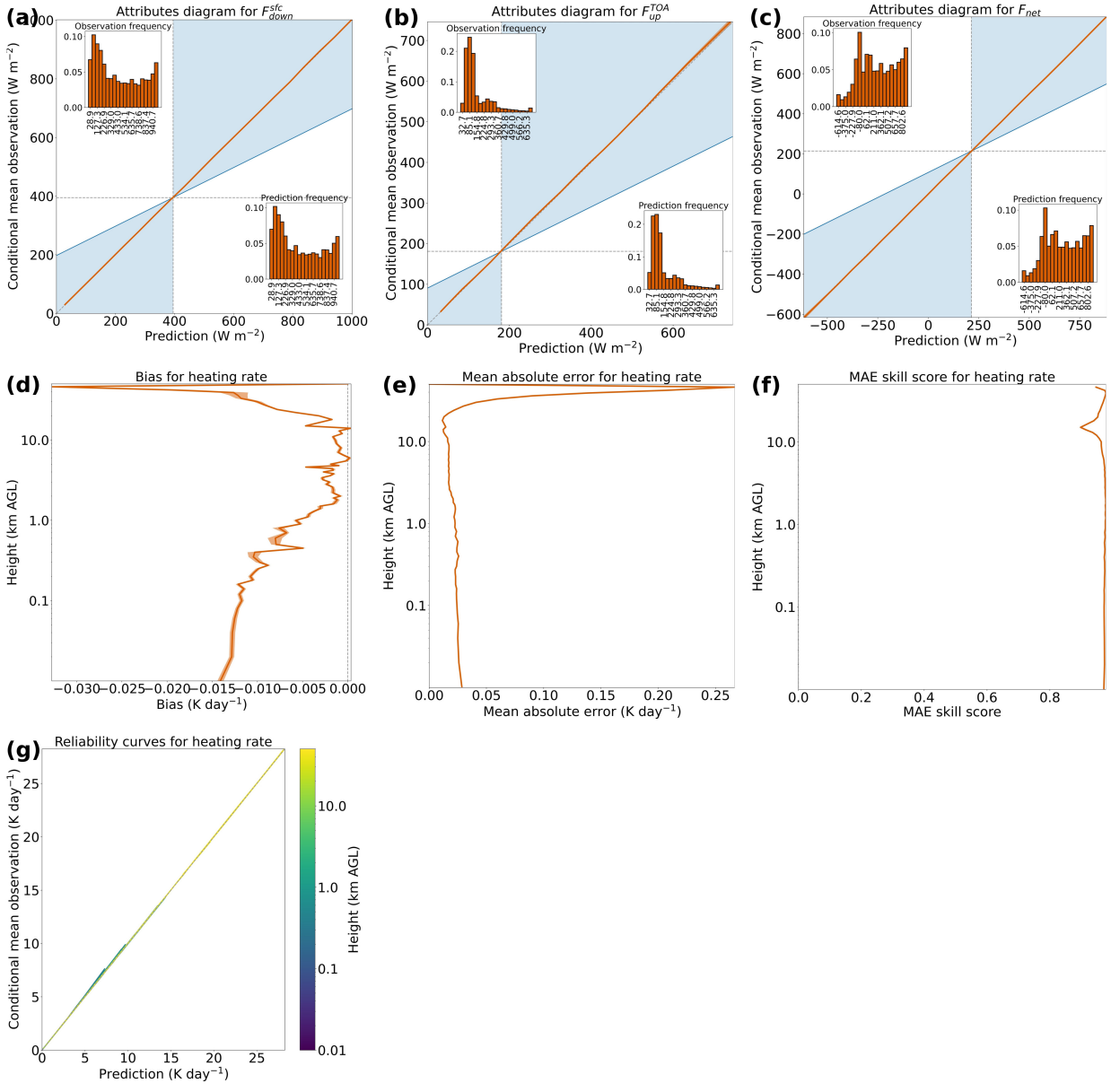

Figure 10: Performance of selected model from Experiment 2 on testing data. Formatting is explained in the caption of Figure 6, and each panel here is analogous to the same-letter panel in Figure 6. The x-axis ranges in panels d-e are markedly smaller here than in Figure 6.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

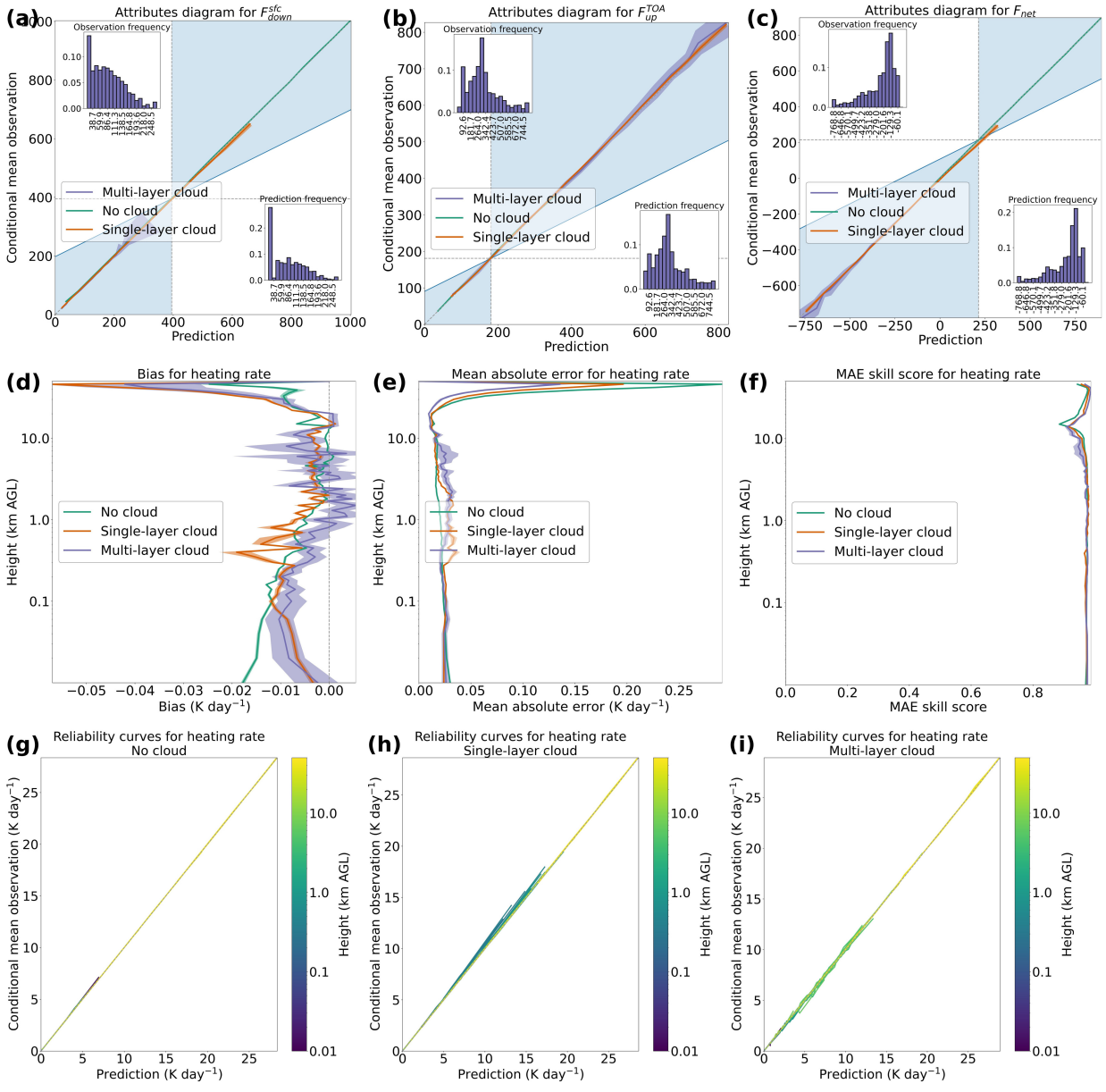

Figure 11: Performance of selected model from Experiment 2 on testing data, by cloud regime. In the attributes diagrams for flux components (a-c), the inset histograms and reference lines are based only on examples with multi-layer cloud. Formatting is explained in the caption of Figure 7, and each panel here is analogous to the same-letter panel in Figure 7. The x-axis ranges in panels d-e are markedly smaller here than in Figure 7.

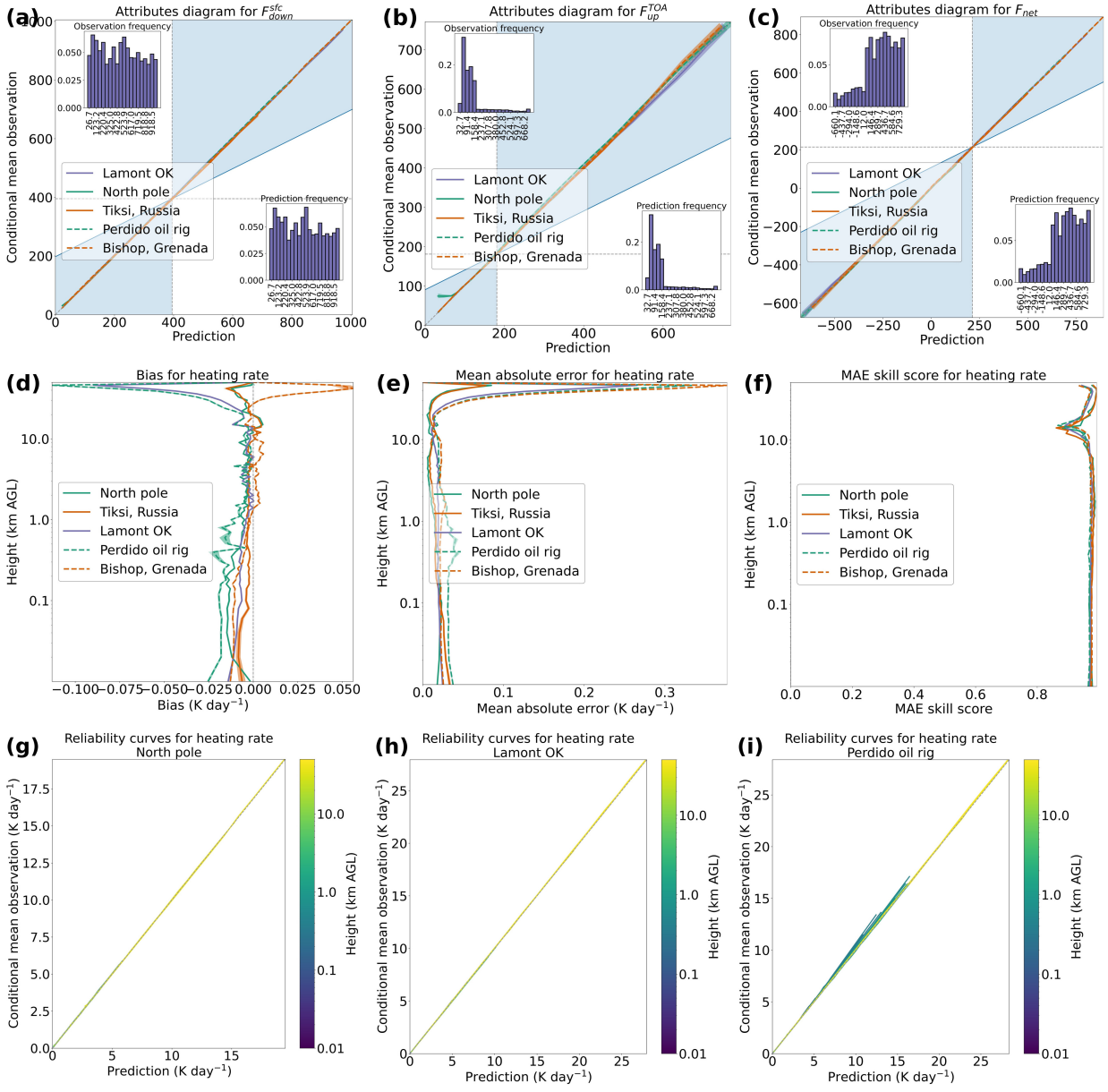

Figure 12: Performance of selected model from Experiment 2 on testing data, by site. In the attributes diagrams for flux components (a-c), the inset histograms and reference lines are based only on examples at Lamont, Oklahoma. Formatting is explained in the caption of Figure 8, and each panel here is analogous to the same-letter panel in Figure 8. The x-axis ranges in panels d-e are markedly smaller here than in Figure 8.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

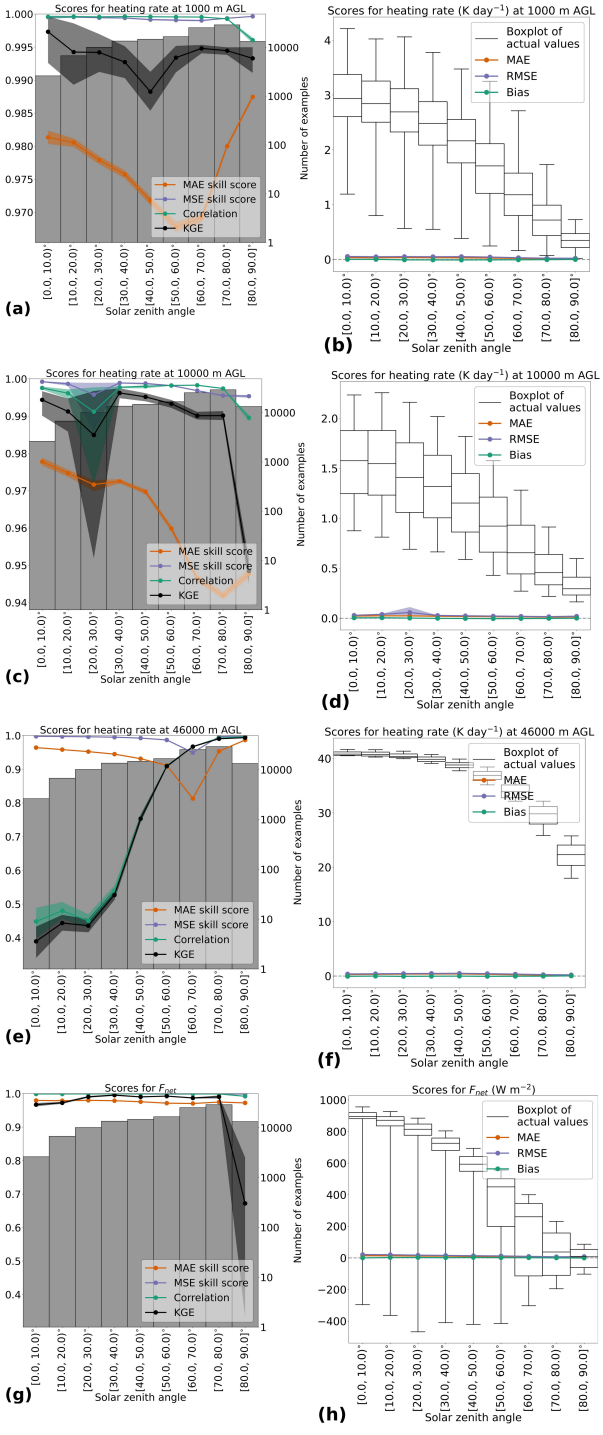

Figure 13: Performance of selected model from Experiment 2 on testing data, by solar zenith angle. Formatting is explained in the caption of Figure 9, and each panel here is analogous to the same-letter panel in Figure 9.

Accepted for publication in *Journal of Atmospheric and Oceanic Technology*. DOI 10.1175/JTECH-D-21-0007.1.

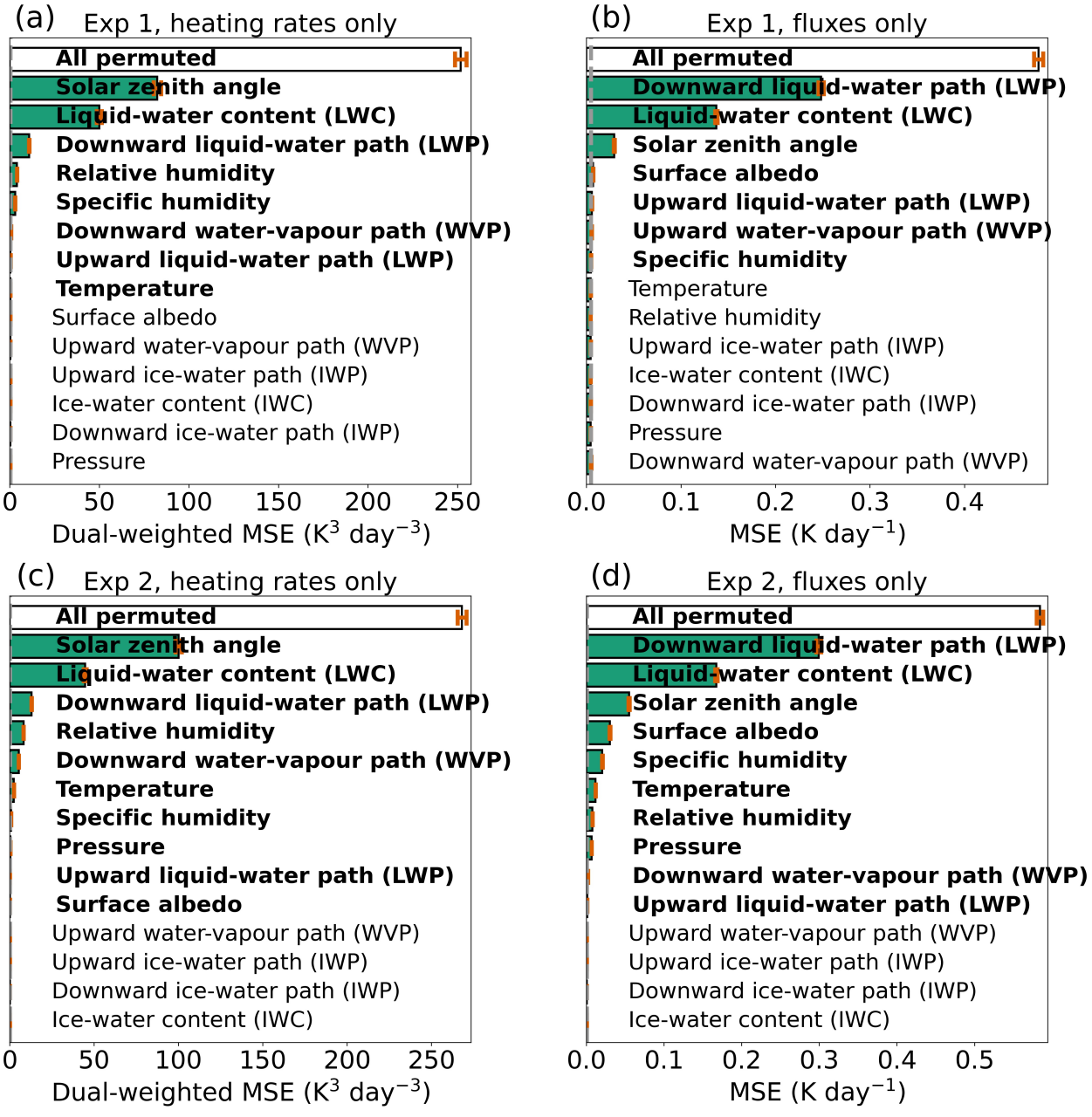

Figure 14: Results of backward multi-pass test on testing data for (a) best model from Experiment 1, with the heating-rate-only loss function; (b) best model from Experiment 1, with the flux-only loss function; (c) best model from Experiment 2, with the heating-rate-only loss function; (d) best model from Experiment 2, with the flux-only loss function. The value for the bar labeled "x_j" is the loss after restoring x_j and all predictors in the bars above x_j. The k^{th} predictor to be restored, and thus the k^{th}-most important, is k^{th} from the top. Orange error bars show the 99% confidence interval, based on bootstrapping 1000 times. If variable x_j is in bold font, this means that x_j is significantly more important than the variable below (at the 99% confidence level), based on a paired-bootstrapping test with 1000 replicates.