Exploring the Use of Machine Learning to Improve Vertical Profiles of Temperature and Moisture

KATHERINE HAYNES,[a] JASON STOCK,[b] JACK DOSTALEK,[a] CHARLES ANDERSON,[b] AND IMME EBERT-UPHOFF[a,c]

[a] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[b] *Department of Computer Science, Colorado State University, Fort Collins, Colorado*
[c] *Department of Electrical and Computer Engineering Colorado State University, Fort Collins, Colorado*

(Manuscript received 7 December 2022, in final form 19 September 2023, accepted 23 October 2023)

ABSTRACT: Vertical profiles of temperature and dewpoint are useful in predicting deep convection that leads to severe weather, which threatens property and lives. Currently, forecasters rely on observations from radiosonde launches and numerical weather prediction (NWP) models. Radiosonde observations are, however, temporally and spatially sparse, and NWP models contain inherent errors that influence short-term predictions of high impact events. This work explores using machine learning (ML) to postprocess NWP model forecasts, combining them with satellite data to improve vertical profiles of temperature and dewpoint. We focus on different ML architectures, loss functions, and input features to optimize predictions. Because we are predicting vertical profiles at 256 levels in the atmosphere, this work provides a unique perspective at using ML for 1D tasks. Compared to baseline profiles from the Rapid Refresh (RAP), ML predictions offer the largest improvement for dewpoint, particularly in the middle and upper atmosphere. Temperature improvements are modest, but CAPE values are improved by up to 40%. Feature importance analyses indicate that the ML models are primarily improving incoming RAP biases. While additional model and satellite data offer some improvement to the predictions, architecture choice is more important than feature selection in fine-tuning the results. Our proposed deep residual U-Net performs the best by leveraging spatial context from the input RAP profiles; however, the results are remarkably robust across model architecture. Further, uncertainty estimates for every level are well calibrated and can provide useful information to forecasters.

KEYWORDS: Convective storms/systems; CAPE; Forecasting; Deep learning; Machine learning; Neural networks

1. Introduction

Atmospheric instability is a key ingredient in the formation of thunderstorms, which can produce severe weather. To accurately forecast these potentially hazardous events and to conduct near-term threat assessments, it is crucial to have frequent and accurate vertical profiles of temperature and dewpoint because they affect the instability of the atmosphere, which is often measured by the convective available potential energy (CAPE). The likelihood that the instability will be realized, measured by the convective inhibition (CIN), is also strongly dependent on the thermodynamic structure, particularly in the lowest layers in the atmosphere.

Radiosondes produce high quality thermodynamical profiles with high vertical resolution while ascending in the atmosphere; however, balloon releases are conducted routinely only twice a day with an average of several hundred kilometers separating the stations of the irregularly spaced network. This sparsity remains a leading cause for errors, and the cost and environmental impact of launching enough radiosondes to overcome this limitation hampers our ability to resolve this sparsity issue (Hurlbut and Cohen 2014). As a result, forecasters leverage vertical profiles generated by numerical weather prediction (NWP) models because they have a high temporal resolution and are produced on a dense, regularly spaced grid. These models are based on observational data that are assimilated to predict three-dimensional temperature, moisture, and other atmospheric variables. Generally, NWP models produce excellent short-term forecasts, but uncertain initial conditions, the application of necessary assumptions, and the mathematics of prognosis produce errors in a model's output.

Using statistical models to postprocess NWP model forecasts in order to improve severe weather prediction has been a topic of research for decades (Vannitsem et al. 2018). To improve NWP model predictions, work has been ongoing to develop different postprocessing statistical methods aiming to remove systematic biases, correct errors, incorporate local-scale adjustments, and produce finer-scale end-use products (Schultz et al. 2021; Rojas-Campos et al. 2023). Originally this work focused on uni- and multivariate corrections of horizontal atmospheric variables (Schoenach et al. 2020). Recently, postprocessing methods have extended to employ deep learning neural networks, and machine learning implementations exist for optimizing temperature (e.g., Rasp and Lerch 2018; Peng et al. 2020), wind speed (e.g., Candido et al. 2020; Veldkamp et al. 2021, and precipitation (e.g., Rojas-Campos et al. 2023).

Despite the importance of vertical profiles of temperature and dewpoint in weather forecasting, little work has been

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-22-0090.s1.

Corresponding author: Katherine Haynes, katherine.haynes@colostate.edu

DOI: 10.1175/AIES-D-22-0090.1 e220090

done to improve vertical profile prediction with two notable exceptions. First, Renkl (2013) postprocessed ensemble NWP predictions to approximate the vertical profiles using vertical normal modes and kriging. Recently, Schoenach et al. (2020) utilized a two-step strategy to correct ensemble NWP forecasts, first using univariate distributional regression to correct the probability distribution separately at each vertical level, then using the forecast rank order structure to reinstall vertical dependence among neighboring levels.

In this work, we investigate using deep learning as a postprocessing tool to correct NWP vertical profiles of temperature and dewpoint. Specifically, we use machine learning (ML) to postprocess NWP profiles from the Rapid Refresh (RAP) model and combine them with surface analyses from the Real-Time Mesoscale Analysis (RTMA) and satellite radiances from the Geostationary Operational Environmental Satellite (GOES) to optimize temperature and dewpoint profiles. We perform a comprehensive analysis over various neural network architectures and physically meaningful training methodologies. Our contributions can be summarized as follows:

- We apply ML to 1D vectors and show the limitations of fully connected and convolutional neural networks for predicting vertical profiles. As also shown in Lagerquist et al. (2021), a U-Net structure leverages relationships between vertical levels for accurate profiles with low variance.
- We present three physically inspired loss functions that are built on the mean absolute error (MAE) or the mean-square error (MSE). These functions adjust the profile weights with height, yielding more accurate near-surface predictions than using the traditional metrics.
- We show that ML is able to improve existing biases in the RAP, regardless of cloud cover, month, or location, using six physically based metrics.
- We demonstrate that ML models are able to provide useful, well-calibrated uncertainty estimates at every level in the vertical column.

2. Methods

a. Data

Data are collected from January 2017 through May 2020 over 18 sounding locations that span between North Dakota and Texas. This region, known as "Tornado Alley," constitutes a significant number of severe weather events each spring and summer. The target values in this study are 1D vertical profiles of temperature and dewpoint from the radiosonde observations (raob). The input features (predictors) are 1D vertical profiles from the RAP model, spatially averaged surface data from the RTMA, and spatially averaged satellite data from *GOES-16*. A sample is a single 1D profile in space and time, where all the input features are collocated with the 1D target profile using the raob release location and time. Because we are focused on convective activity that has the potential to result in severe weather, we restrict the samples to those occurring from April through August (except for the seasonal and regional analyses where we use all the data). To reduce bias to any given launch location and release time, we use a spatiotemporally stratified sampling approach to split the data into 0.75, 0.1, and 0.15 partitions for training, validation, and testing, respectively, ensuring that each site occurs in all three partitions for all months. By temporally ordering the sites, the testing partition has no overlaps in time and includes the latest 15% of the data. This results in 12 929 training, 1727 validation, and 2574 test samples. We use the validation dataset for the architecture search and hyperparameter tuning, and all reported statistics for the analyses shown use the test dataset.

For training and testing, we standardize the input features and target values to have a mean of zero and unit variance (z-score normalization). Every vertical level and each observational variable are standardized independently by subtracting the mean and dividing by the standard deviation from the training data. To convert predictions from a model back to their original units, we simply multiply by the standard deviation and add the mean using the statistical values from the training dataset.

1) RAOB

Data are from the National Oceanic and Atmospheric Administration's Earth System Research Laboratories radiosonde archive for the locations of interest (Schwartz and Govett 1992). The majority of these samples are from daily NWS launches that occur shortly before 0000 and 1200 UTC. Less common, but still prevalent, are radiosondes launched when atmospheric conditions are of interest (i.e., during severe weather events). These launches usually occur between 1800 and 2100 UTC. Every observation in the archive undergoes extensive quality assurance analysis and correction procedures to resolve erroneous data and to check for various hydrostatic consistencies. Thus, only minimal preprocessing is needed. Dewpoint depression is converted to dewpoint temperature (herein referred to as dewpoint) and profiles with missing values are removed. Profiles are linearly interpolated to regularly spaced intervals of geopotential height coordinates that are consistent across geographic regions; this establishes 256 fixed vertical levels extending up to 17 km above the surface.

2) RAP

The Rapid Refresh is an operational assimilation and modeling system for North America. Designed primarily for NWP guidance, the RAP provides hourly updated short-range weather forecasts out to 18 h (Benjamin et al. 2016). The community-driven Advanced Research Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) underpins the NWP, and the Gridpoint Statistical Interpolation analysis system (Wu et al. 2002; Whitaker et al. 2008; Kleist et al. 2009) is used for data assimilation and initializing the model. The RAP has a horizontal grid spacing of 13 km and a hybrid sigma vertical coordinate system with 50 levels.

The time and location of the raob launch is used to locate the nearest RAP forecast, and we assume that the raob does not drift outside of the 169-km^2 grid cell. We extract the total pressure, temperature, specific humidity, and geopotential height at every vertical level. To better align with the raobs, we convert

specific humidity to dewpoint (see section S1 in the online supplemental material for details). The four 1D profile components (i.e., total pressure, temperature, dewpoint, and geopotential height) are linearly interpolated with respect to geopotential height to align with the raobs, having 256 levels up to a top boundary layer of 17 km above the surface.

3) RTMA

The RTMA provides accurate near-surface weather conditions at a high spatial resolution (De Pondeca et al. 2011). Observations centered ±12 min around the analysis time are assimilated following the Gridpoint Statistical Interpolation system (Wu et al. 2002). The result is a 2.5-km grid over the conterminous United States (CONUS) with analyses of 2-m temperature, 2-m dewpoint, 10-m wind components, and surface pressure. Several quality-control steps are done during analysis to remove erroneous data, verify threshold constraints, and resolve static and dynamic blacklisted data, which eliminates the need for additional variable corrections.

The temperature, dewpoint, and pressure from the RTMA are temporally aligned with the release time of individual raobs. RTMA samples from the nearest hour, prior or in the future, are used for irregular or late-release radiosondes. Although rare, data samples may be discarded when there are no RTMA data within an hour window. Once aligned, the samples are cropped to the neighborhood around each launch location. Specifically, we extract a 3×3 or 56.25-km^2 patch from each RTMA variable centered closest to the latitude and longitude of a given raob. Note that radiosondes take roughly 30 min to ascend, and with wind speeds of 20 m s^{-1}, neither their exact time nor location is fixed. Thus, we take the spatial mean of each patch to generally represent the area with a single scalar value (per variable).

4) *GOES-16*

Satellite data are from a multichannel passive imaging radiometer, the Advanced Baseline Imager (ABI), onboard *GOES-16* (Schmit et al. 2017). The ABI captures imagery every 5 min over CONUS with a nadir footprint of 0.5–2 km in 16 spectral wavelengths covering the visible, near-infrared, and infrared spectrum. Each channel is centered on specific wavelengths to highlight certain atmospheric properties. In this study, we use the radiances from a subset of infrared wavelengths commonly used to retrieve vertical profiles of temperature and moisture (Schmit et al. 2019; Hilburn 2020). The specific channels chosen are the 6.2-, 6.9-, and 7.3-μm water vapor channels, the 8.4-, 10.3-, 11.2-, and 12.3-μm infrared window channels, and the 13.3-μm CO_2 channel. As with the RTMA data, we use a 3×3 grouping of ABI pixels for the satellite input. The channels used have nadir footprints of 2 km, resulting in a 36-km^2 (nominal) area of consideration. The cadence of the ABI instrument ensures at most 5 min of separation between the satellite data and the release time of the radiosonde.

b. *Evaluation metrics*

The goal of this work is to predict profiles of both temperature and dewpoint. Because the mechanisms driving the variability in these profiles are different, we utilize traditional metrics and atmospheric quantities to analyze each separately as well as their interactions. First, we use the root-mean-square errors (RMSEs) of the entire profile of temperature (T) and dewpoint (TD) separately. Second, as the boundary layer is the most important for forecasting convection and severe weather, we evaluate the RMSEs for both in the lowest 25 layers (SFC T, TD), which corresponds to ~2 km above ground level.

In addition to evaluating the profiles, we use two important atmospheric quantities derived from the profiles. The first is mixed-layer CAPE, which is a measure of atmospheric instability and indicates the energy available for thunderstorm development. The second is CIN, which corresponds to the amount of energy that prevents air parcels from rising buoyantly, suppressing convection. Since both of these metrics require the pressure profile (which is not predicted), we use the corresponding RAP pressure profiles.

Because the different evaluation metrics have different scales, to allow for comparison among them we use two derived scores. The first is a normalized RMSE score (NRS), which is given by

$$\mathrm{NRS}(x) = 1 - \frac{\mathrm{ML}_{\mathrm{RMSE}(x)} - \mathrm{ML}^{\min}_{\mathrm{RMSE}(x)}}{\mathrm{ML}^{\max}_{\mathrm{RMSE}(x)} - \mathrm{ML}^{\min}_{\mathrm{RMSE}(x)}}, \tag{1}$$

where for each metric x, $\mathrm{ML}_{\mathrm{RMSE}}$ is the RMSE for each ML model, $\mathrm{ML}^{\min}_{\mathrm{RMSE}}$ is the minimum RMSE across all models, and $\mathrm{ML}^{\max}_{\mathrm{RMSE}}$ is the maximum RMSE across all models. The second is a normalized improvement score (NIS), which is calculated as

$$\mathrm{NIS}(x) = \frac{I(x)}{I_{\max}}, \text{ and} \tag{2}$$

$$I(x) = \frac{\mathrm{RAP}_{\mathrm{RMSE}(x)} - \mathrm{ML}_{\mathrm{RMSE}(x)}}{\mathrm{RAP}_{\mathrm{RMSE}(x)}}, \tag{3}$$

where for metric x, $I(x)$ is the improvement, $\mathrm{RAP}_{\mathrm{RMSE}(x)}$ is the RAP RMSE, and $I_{\max}$ is the maximum improvement across all models. For both scores, 1 is the highest value and skill decreases with decreasing values.

c. *Machine learning models*

1) MODEL ARCHITECTURES

We focus on neural networks to learn nonlinear mappings between initial guess RAP profiles and ground truth raobs. In addition to learning existing RAP model biases, we provide the ML models with the GOES and RTMA data. We explore four different ML model architectures, each with added complexity: linear regression (LIN), fully connected neural networks (NN), convolutional neural networks (CNN), and deep residual U-Nets. For all architectures, the target temperature and dewpoint profiles are flattened and concatenated to 512 output features (256 levels × 2 variables).

For the LIN and NN models, we flattened the input features into a 1D vector for each observation, where all the input levels are treated independently. Thus, for experiments using all possible inputs, the data are concatenated together

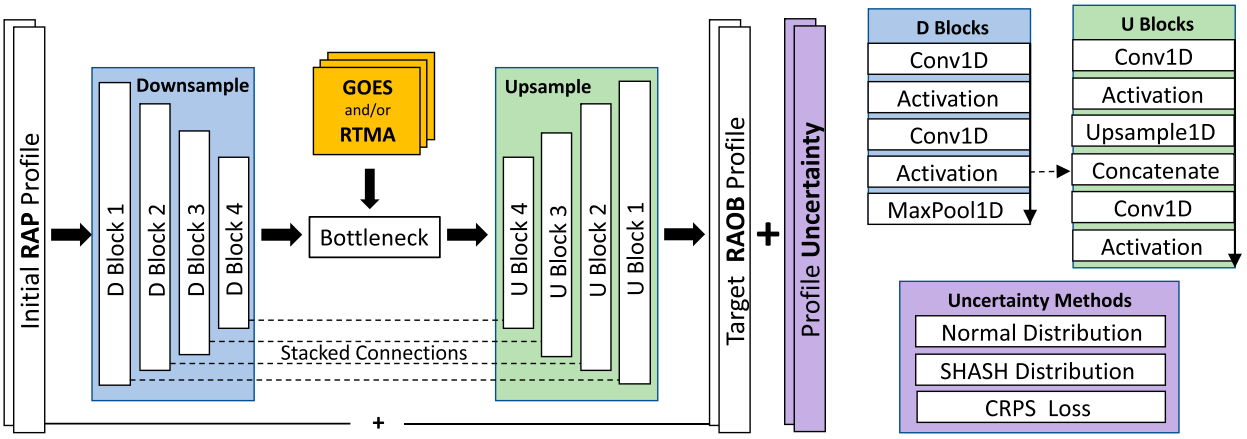


FIG. 1. U-Net architecture using the RAP temperature and moisture profiles as inputs and concatenating the GOES ABI and RTMA data at the bottleneck of the network. After completing the stacked connections, the architecture includes an additive connection between the input and output profiles, as indicated by the "+" symbol. The outputs are the reconstructed temperature and dewpoint profiles, which can be directly compared to the radiosonde observations. Additionally, uncertainty information is predicted for each layer in the profiles. Adapted from S21, which did not include uncertainty estimates.

to create 1035 input features (4 RAP Profiles × 256 levels + 8 GOES + 3 RTMA). For the LIN and NN models, we performed an ablation study to reduce the number of input features. We tested averaging from 1 to 25 levels and evaluated the results using the normalized RMSE score from the six metrics (RMSEs of T and TD, SFC T and TD, CAPE, CIN). We found that for both the LIN and NN models, the highest score used every RAP level in the lowest 25 levels and an average of 8 levels above. This decreases the total number of possible input features down to 226.

CNNs and U-Nets take spatial information into account by considering patterns in neighboring levels, thus these ML models may be able to detect patterns and relationships in the vertical profile. For both of these architectures, we performed convolutions over the input RAP profiles and ended with a linear output layer matching the size of the raob profiles (256 levels × 2 variables). For the CNN, we evaluated a sequence of convolutional layers followed by pooling layers, concatenating the GOES and RTMA data with the output of the last pooling layer.

U-Nets are a type of CNN that encodes input through downsampling layers and then decode the compressed representation of the input through a series of upsampling layers via U-shaped stacked connections (Ronneberger et al. 2015). Figure 1 shows the U-Net architecture, which includes downsampling blocks (blue), a bottleneck layer, and upsampling blocks (green) with stacked connections to the downsampling layers. The GOES and RTMA data are concatenated in the bottleneck layer, and we include an additive connection between the input profile with the final output of the upsampling layer, forcing the model to learn the residuals between the input and target profiles.

To find the highest-performing NN, CNN, and U-Net model architectures, we performed hyperparameter searches to optimize the number of layers, number of weights, and kernel size/stride (for the CNN and U-Net). Full details are in Stock (2021, chapter 6, hereafter S21). For the NN, 2 layers both with 1024 weights was the best-performing model (S21, Fig. 6.1). For the CNN, the model with the smallest profile RMSEs contained five convolutional blocks of size [32, 64, 128, 256, 512] each using a filter size of 3 × 1 with a stride size of one, followed by two fully connected layers of size [512, 256], and a final output layer (S21, Fig. 6.3). For the U-Net, a network structure of size [32, 64, 128, 256] with a mirrored design around the bottleneck and no additional layers after the U-Net had the lowest mean near-surface error (S21, Fig. 6.4). S21 also showed that maintaining the structural integrity of the RAP profile using symmetrical skip connections resulted in the U-Net outperforming the traditional CNN; thus, in this study we will not show any CNN results. For all architectures, we used the Adam optimizer, the ReLU activation function, and batch size of 128. For the deterministic loss functions, we used a learning rate of 0.001; for the models predicting uncertainty, we used a learning rate of 0.0001.

To avoid overfitting, we performed hyperparameter searches of the final model architectures to optimize the dropout (final layer, internal dense layers, and convolutional layers) and kernel regularization. Using the validation data, we performed both gridded and guided searches using Hyperopt with the tree of Parzen estimators algorithm (Bergstra et al. 2013). We found kernel regularization was not helpful for this application. The top NN model uses a dropout of 0.05 in the last layer and the top U-Net uses 0.1 dropout in the last layer and 0.05 dropout in the convolutional layers. Further, we employed early stopping to quit training when the model validation score did not continue to improve, and plots of the validation and training loss ensured that during training the validation score remained similar to the training score.

2) LOSS FUNCTIONS AND UNCERTAINTY ESTIMATES

To optimize performance, we tested seven loss functions, of which four are deterministic. Because it is essential to have accurate profiles near the surface in order to accurately predict the convective activity, we developed three custom loss functions. While all of these loss functions are designed to

more heavily weight the near-surface levels, they vary in the way the weights decrease with height. The four deterministic loss functions are as follows:

- MSE:

$$\text{MSE} = \frac{1}{512n}\sum_{i=1}^{n}\sum_{v=1}^{256}\sum_{t=1}^{2}(y_{i,v,t} - \hat{y}_{i,v,t})^2, \tag{4}$$

where y are the raob observations and $\hat{y}$ are the predictions. The average square difference is taken over n samples across $v = 256$ vertical levels and $t = 2$ variables (temperature and dewpoint).

- Surface-weighted mean absolute error (MAES): MAE where the performance of the output temperature and dewpoint in the lowest 25 levels accounts for 80% of the error and the remaining upper profile accounts for 20% of the error:

$$\text{MAES} = \frac{0.8}{50n}\sum_{i=1}^{n}\sum_{v=1}^{25}\sum_{t=1}^{2}\left|y_{i,v,t} - \hat{y}_{i,v,t}\right| + \frac{0.2}{462n}\sum_{i=1}^{n}\sum_{v=26}^{256}\sum_{t=1}^{2}\left|y_{i,v,t} - \hat{y}_{i,v,t}\right|. \tag{5}$$

- Exponentially weighted mean absolute error (MAEW): MAE where the weight of the temperature and dewpoint in the profile decreases exponentially with height:

$$\text{MAEW} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}\left[\sum_{v=1}^{256}(\alpha e^{-\lambda v} + \beta)|y_{i,t,v} - \hat{y}_{i,t,v}|\right], \tag{6}$$

where α is an initial value, λ is a decay constant, and β is an offset value for an exponential decay function. The absolute difference of the profile is multiplied by an exponential decay function so that the weights applied per vertical level decrease with altitude. Through initial experiments, we found $\alpha = 3.75$, $\lambda = 0.01$, and $\beta = 0.25$ to be appropriate values for the data.

- Pressure-weighted mean square error (MSEW): MSE where the weight of the temperature and dewpoint decreases exponentially with pressure:

$$\text{MSEW} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}\left[\sum_{v=1}^{256}\frac{p_v}{p_{\text{Tot}}}(y_{i,t,v} - \hat{y}_{i,t,v})^2\right], \text{ and} \tag{7}$$

$$p_{\text{Tot}} = \sum_{v=1}^{256} p_v, \tag{8}$$

where p is the pressure in the profile level and p_{Tot} is the sum of the pressure in the vertical profile. Since the ML models do not predict pressure, we use the RAP pressure.

In addition to deterministic prediction, we also tested three probabilistic approaches to obtain a forecast distribution that predicts not only the vertical profiles of temperature and dewpoint, but also the associated uncertainty of each per vertical level. The approaches are the following:

- Gaussian parametric distribution prediction (NORM): The loss function trains the NN to predict the parameters of the normal distribution by maximizing the Bayesian likelihood between the true and observed distribution (e.g., Rasp and Lerch 2018; Schoenach et al. 2020; Veldkamp et al. 2021). In this case, the mean is the central prediction, and the standard deviation of the prediction is used to calculate the uncertainty.
- Sinh–arcsinh parametric distribution prediction (SHASH): The loss function trains the NN to predict the parameters of a sinh–arcsinh–normal distribution, with four parameters to represent the shape of the distribution (e.g., Barnes et al. 2023, 2021; Haynes et al. 2023). The predictions can then be drawn from the full distribution, where the mean of the distribution is used as the central predicted value and the 95% range of the distribution is used for the uncertainty.
- Ensemble prediction with the continuous-ranked probability score (CRPS): Ensemble-based approach that minimizes the cumulative distribution function (CDF) of the ensemble members (e.g., Matheson and Winkler 1976; Hersbach 2000; Gneiting et al. 2005), which has been used as an NN loss function for numerous postprocessing applications (e.g., Dai and Hemri 2021; Ghazvinian et al. 2021; Scheuerer et al. 2020; Schulz and Lerch 2022). The CRPS is a generalization of the MAE for probabilistic forecasts:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty}[F(\hat{y}) - \mathscr{H}(\hat{y} - y)]^2 d\hat{y}, \tag{9}$$

where y is the single observed value; F is the CDF of the predicted distribution; $\hat{y}$, the variable of integration, is one value in the predicted distribution; and $\mathscr{H}$ is the Heaviside step function, evaluating to 1 if $\hat{y} \geq y$ and 0 otherwise. Thus, Eq. (9) is the error between the predicted and observed CDF. The CRPS can be modified to be used as a loss function with uncertainty quantification for NNs, as shown in Haynes et al. (2023), where the central prediction is the median of the ensemble members, and the uncertainty is calculated from their spread.

Both the NORM and SHASH loss functions use the maximum-likelihood approach with a log-loss formulation to optimize the distribution parameters. For CRPS prediction, we use 60 ensemble members, which was optimized during the hyperparameter search. All three of these methods are probabilistic approaches to predict a temperature and dewpoint distribution at every level in the vertical profile.

3. Model comparisons

a. *Architectures and loss functions*

We ran each of the model architectures (LIN, NN, U-Net) with the seven loss functions. The results are shown in Fig. 2a, which compares the model architectures and loss functions using the NRS score for the six metrics. The RAP is shown in the top line in Fig. 2a. Postprocessing using ML improves all dewpoint, CAPE, and CIN predictions. Overall, the U-Net architecture has the highest performances. Comparing the loss functions, the U-Net MAEW predicts the best surface dewpoint, the U-Net MSE predicts the best dewpoint profile, and the U-Net MAEW predicts the best surface temperature and dewpoint. While most

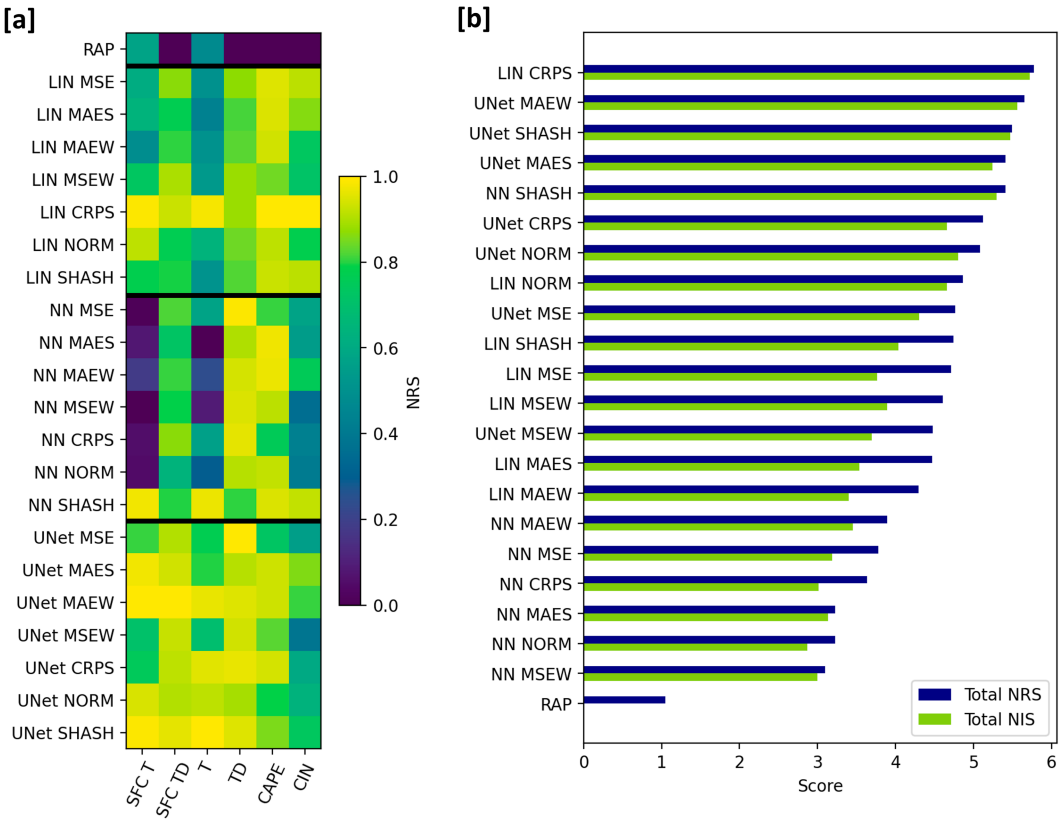


FIG. 2. Performance comparisons between model architectures and loss functions, with the RAP included for reference. (a) NRS scores per metric. (b) Model performances ranked by total NRS and NIS scores. The totals are the score sums for the six evaluation metrics.

of the LIN models have an average performance compared with the other architectures, LIN CRPS is the most consistent across all evaluation metrics, with the best predictions for the temperature profile, CAPE and CIN. The NN models do not perform well, particularly for temperature; however, the dewpoint, CAPE, and CIN improve over the RAP. Of the different loss functions with the NN architecture, NN SHASH has the most consistent performance and is the only NN model to improve over the RAP across all metrics.

To rank the overall model performance, we added the NRS and NIS scores for the six metrics (Fig. 2b). LIN CRPS performs the best overall, followed by the U-Net models. The NN architecture performs the worst, except NN SHASH, which has scores comparable to the U-Nets. Each architecture takes advantage of different loss functions. The LIN model takes advantage of the CRPS ensemble approach, whereas the NN performs best with the SHASH loss function, which focuses on predicting a nonsymmetric distribution associated with each prediction. The U-Net is able to take advantage of the custom loss functions, with the weighted and surface MAE losses being top performers.

b. Feature importance via ablation

To test feature importance, we performed an ablation study where we removed the GOES and RTMA data (always keeping RAP profiles as inputs). This results in four different combinations per model: 1) using only the RAP input; 2) using RAP and RTMA data ("R"); 3) using RAP and GOES data ("G"); 4) using RAP, RTMA, and GOES data ("R + G"). We tested this for all model combinations (3 architectures × 7 loss functions) and then calculated the number of times each combination had the lowest RMSE per evaluation metric (Fig. 3a). For all evaluation metrics, adding RTMA data to the RAP profiles improves performance. Adding GOES data does not improve profile prediction; however, GOES does improve the CAPE and CIN performances, indicating that they are providing information on the profile structure even if they do not directly reduce the profile errors. Overall, adding RTMA data results in the highest profile performances, but adding RTMA and GOES data yields the highest CAPE and CIN performances.

Comparing the model performances, Fig. 3b shows the NRS scores for the ablation experiments with the top LIN, NN, and U-Net models. For these models, adding both RTMA and GOES data improves performances, particularly for dewpoint, CAPE and CIN; and these models are the top performers. In contrast, the ML models using only RAP data perform the worst, although they all improve over the RAP (Fig. 3c).

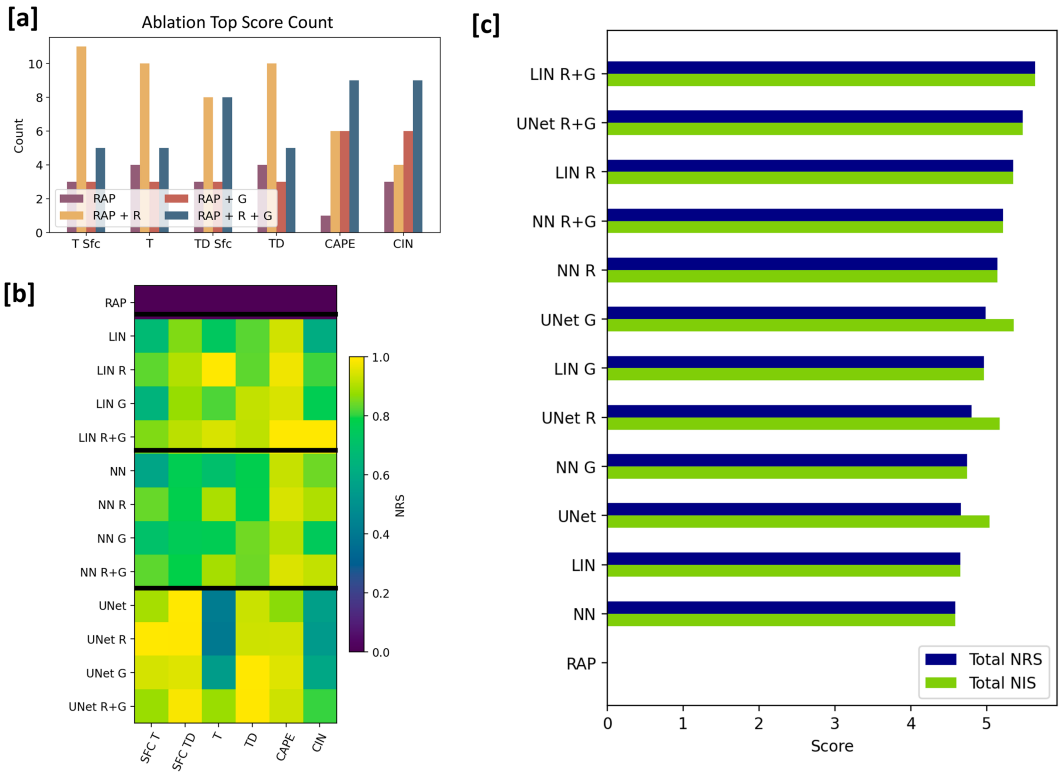


FIG. 3. (a) Counts of the feature combinations resulting in the top-performing model per architecture/loss combination. All simulations use the RAP data; "R" indicates RTMA data are added; "G" indicates GOES data are added; and "R + G" indicates both RTMA and GOES data are added. (b) NRS scores per metric comparing feature ablation performances for the top LIN, NN, and U-Net models. (c) Ranked NRS and NIS scores from (b). Note that for the combinations that are the same as in Fig. 2 (e.g., models with R + G), the exact values may be different because these scores depend on the population statistics.

4. Evaluation results: How well do the best ML models perform?

a. Metrics evaluation

Until now, we have been looking at normalized errors to compare the different models. In Fig. 4, we evaluate actual performance for the six metrics. We selected the top four performing U-Net models along with the top-performing LIN and NN models. Starting with temperature (Figs. 4a,c), the performance for near-surface and total column is similar, and all ML models improve over the RAP baseline performance. Looking at the magnitude of the errors, the temperature errors are ~1°C, which is much smaller than the dewpoint errors, and the improvement is <10%. For dewpoint (Figs. 4b,d), all models perform similarly and improve substantially over the RAP baseline. Near the surface, the errors are ~2.4°C, which is ~13% improvement; over the total column the dewpoint errors are ~5°C, a ~25% improvement compared to the RAP. Substantial improvements are also seen in the CAPE predictions, where the ML models reduce the CAPE errors by ~40%. The CIN improvement is modest, with the best ML model reducing the error by ~14%.

Overall, differences in performance between specific architectures are small compared to the improvement against the RAP, revealing that the ML models are robust. As expected with postprocessing techniques, the ML models are primarily learning and fixing RAP biases. The improvements are greatest for dewpoint and CAPE, with more modest temperature and CIN improvements.

b. Evaluation across convective potential

Because CAPE is an important indicator of convective activity, we looked at model predictions across the range of mixed-layer CAPE values. Figure 5 shows the RMSE in temperature and dewpoint binned by observed CAPE. The gray histograms show the number of profiles per CAPE bin, revealing that most of the cases have <50 J kg^{-1} of CAPE, and the number of samples decreases with increasing CAPE (which makes sense because intense storms are associated with high CAPE and are rare events).

All ML models improve upon the RAP across CAPE bins for both temperature and dewpoint, with the U-Net performing the best (Figs. 5a,d). The errors in the ML models follow the shape of the RAP errors, and there are relatively constant offsets of ~0.1° and ~1.5°C for temperature and dewpoint, respectively, across all CAPE bins. This indicates the ML models are improving the profiles for both convectively active and nonconvectively active scenarios.

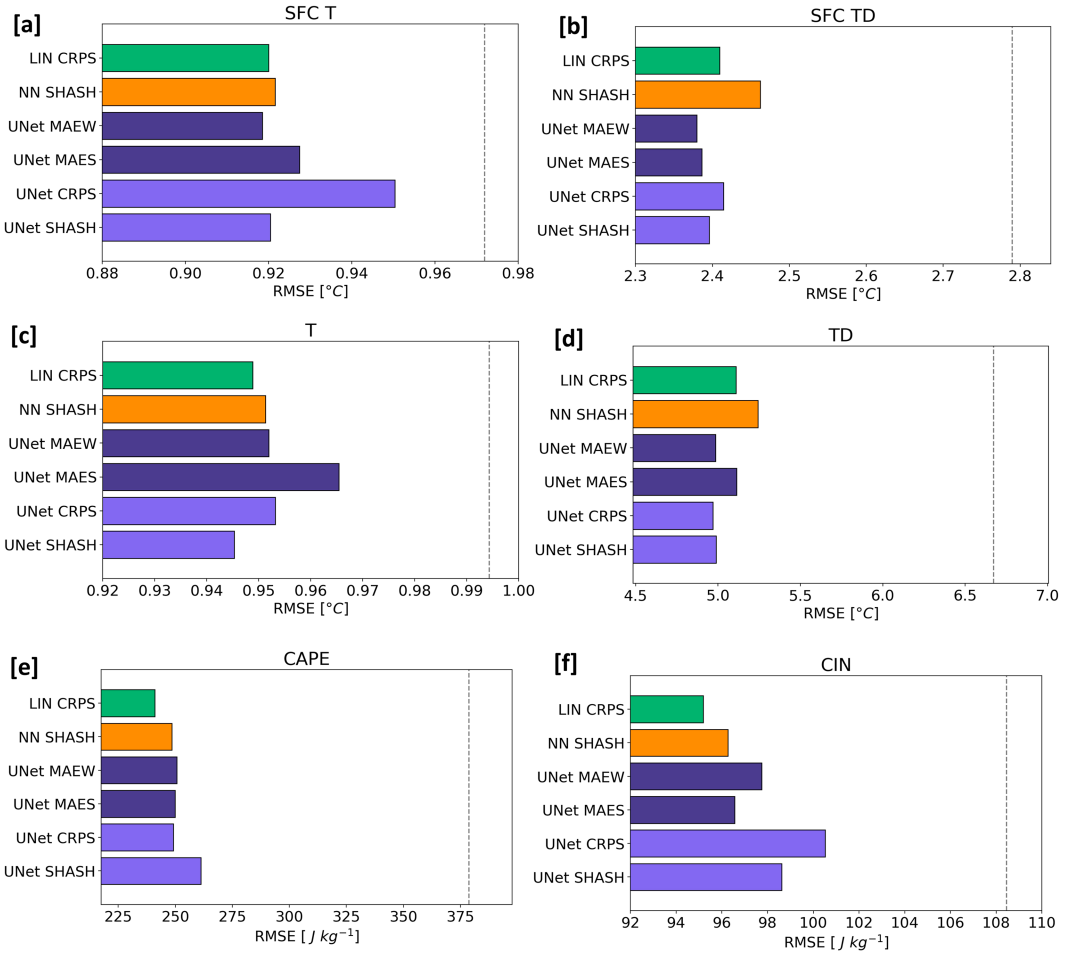


FIG. 4. Top model evaluation for the six metrics. The dashed line indicates the RAP baseline performance. (a) Near surface temperature, (b) near surface dewpoint, (c) temperature, (d) dewpoint, (e) CAPE, and (f) CIN.

To determine how often the temperature profiles are improved, Fig. 5b shows the fraction of samples where the ML RMSE is lower than the RAP RMSE by CAPE bin. The NN model has the lowest number of samples improved, and the U-Net has the highest fraction of improved cases, with an average of 71% of the profiles improving over the RAP. Since the near surface levels are the most influential as to whether storms will develop, Fig. 5c shows the fraction of improved temperature samples near the surface. None of the ML models offers substantial, consistent improvement in near-surface temperature over the RAP model.

Looking at dewpoint, all ML models improve the dewpoint profiles across all CAPE bins, improving ~82% of the profiles (Fig. 5e). The improvement is less near the surface, where on average the ML models improve 69% of the profiles, with slightly less improvement for profiles with higher CAPE than for profiles with lower CAPE. The performances near the surface for both temperature and dewpoint vary between bins, indicating that specific cases may be driving the errors and that the models could likely benefit from additional training data for convectively active conditions.

c. *Gradient and gradient rate of change*

The vertical thermal gradient is an important quantity in weather forecasting: it determines how easily an air parcel rises, which drives exchange processes. In addition to preserving the temperature and dewpoint profile values, it is essential for postprocessing methods to capture the vertical gradients of these profiles (Schoenach et al. 2020). Mean absolute value observed and modeled temperature gradient profiles are shown in Fig. 6a. Positive gradients up to 300 mb (1 mb = 1 hPa) indicate the temperature decreases with height on average, as expected. The observed and modeled profiles all show a sharp gradient in the lowest level of the profile. This feature is particularly prominent when using the absolute value of the gradients to calculate the mean profile, because this formulation does not allow for cancellation due to the sign of the gradient. All models systematically underestimate the gradient in the surface layer from 900 to 750 mb, indicating that there are large temperature swings through this layer that are not fully predicted. Above the surface layer, the models capture the thermal gradient well.

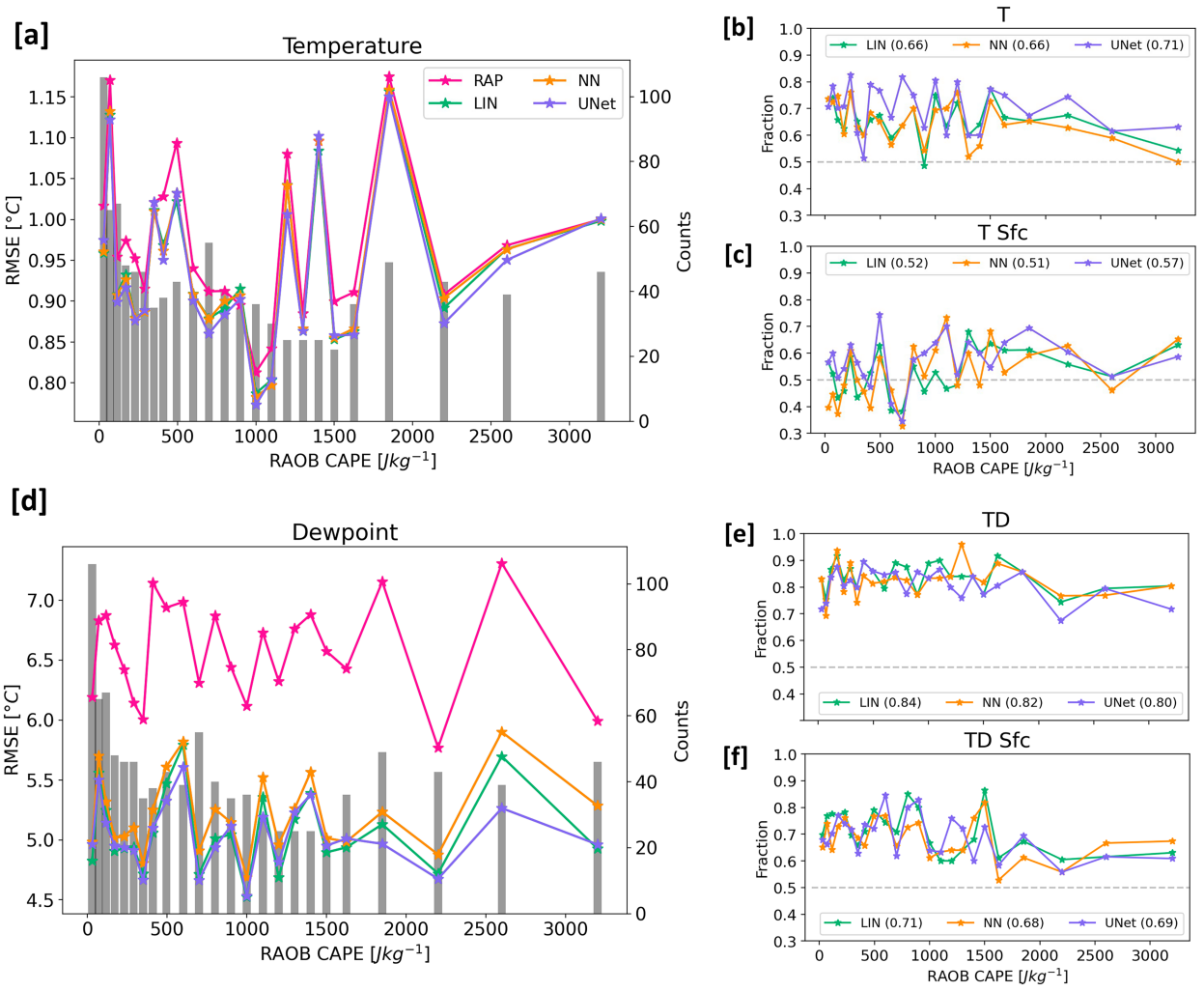


FIG. 5. Model performance binned by observed CAPE. (a) Temperature errors per CAPE bin (colors). The counts per bin are shown in the gray bars. (b) Fraction of temperature samples that improve over the RAP per CAPE bin. (c) As in (b), but for the surface temperature. (d) Dewpoint errors per CAPE bin. (e) Fraction of dewpoint samples that improves over the RAP per CAPE bin. (f) As in (e), but for surface dewpoint. For each model, the mean across CAPE bins is shown in parentheses.

Looking at the dewpoint in Fig. 6b, it also decreases with height on average, and the models consistently underestimate the dewpoint gradient throughout the atmosphere. This phenomenon is seen only in the absolute value formulation when taking the mean gradient across all cases, indicating that the models underestimate the change of signs in the dewpoint gradient, which cancel out when using signed values. Interestingly, the LIN and NN models do better than the U-Net at capturing the dewpoint gradient.

In addition to preserving vertical profile gradients, postprocessing methods should also preserve realistic smoothness in the profiles, which can be seen via the profile second derivative, or the rate of change of the gradients. We used the second derivative as a proxy for the curvature to provide an estimate as to how much the temperatures and dewpoints change between each level in the profile. This is important because changes that occur between neighboring levels impact forecaster confidence: noisier profiles with alternating jumps in temperature between adjacent levels are not only less pleasing to look at but are less physical and can lower forecaster confidence.

The mean temperature second derivatives are shown in Fig. 6c, which shows that on average the observations have a mean second derivative of 0.02°C mb^{-2}, with a maximum of ~0.05°C mb^{-2}. The RAP underestimates this, indicating that the RAP profiles are too smooth compared to the observations. In contrast, using the CRPS and SHASH loss functions overestimates the second derivative, indicating that these profiles are too jumpy. The U-Net models with the physical-based loss functions do the best job of matching the profile smoothness, which is not surprising given that they take into account vertical profile information.

The dewpoint second derivatives (Fig. 6b) show similar characteristics to temperature: the RAP profiles are smoother than the radiosondes and the LIN, NN, and U-Net models

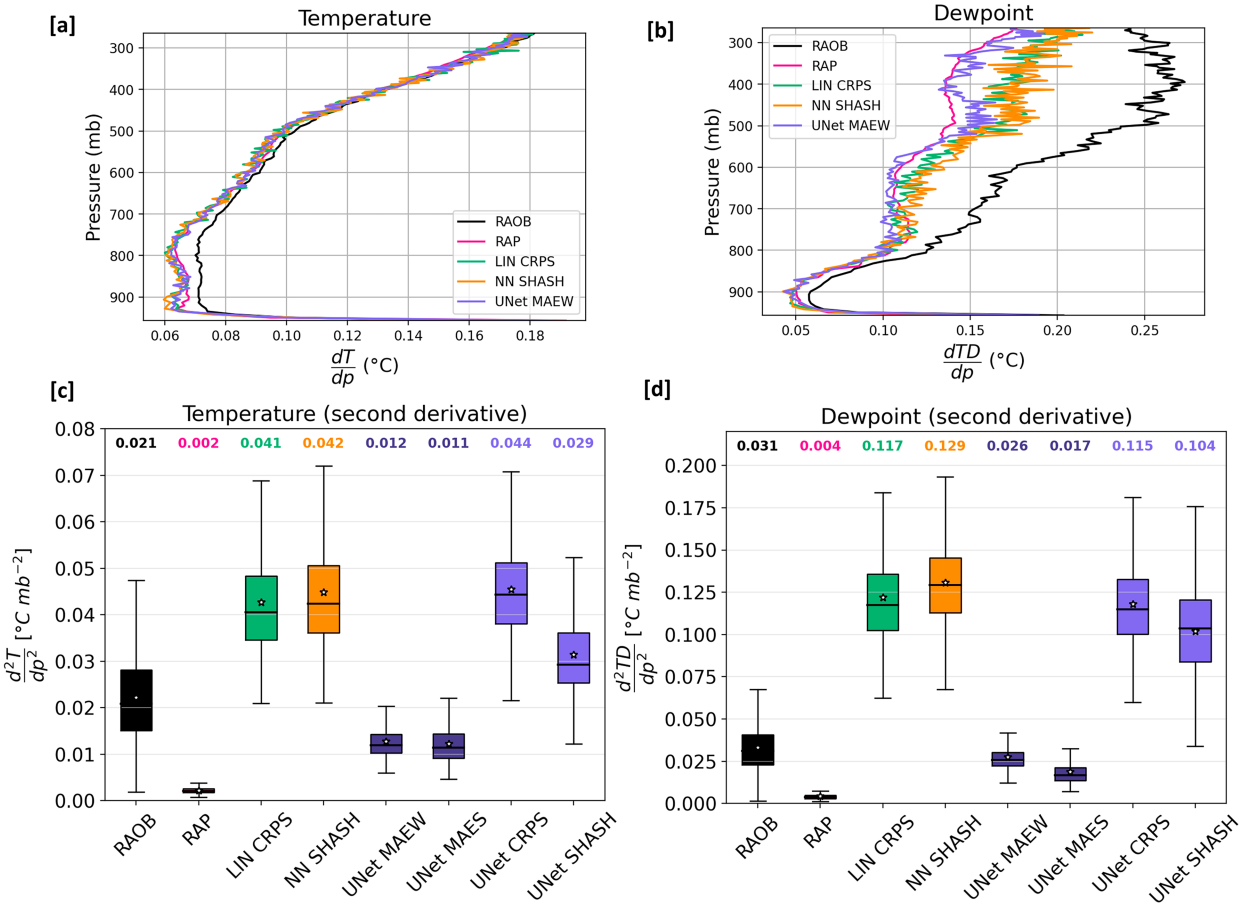


FIG. 6. (a) Mean vertical temperature gradient profile. (b) Mean vertical dewpoint gradient profile. (c) Mean vertical temperature gradient rate of change (T second derivative). (d) Mean vertical dewpoint gradient rate of change (TD second derivative). Note that for all plots, we used the absolute values to calculate the mean across all cases.

with uncertainty contain more noise than the radiosondes, while the U-Net profiles with physics-based loss have second derivatives similar to the radiosondes. This analysis suggests that while the LIN and NN models yield similar error statistics, the U-Net model may be more desirable in practice because the profiles have less variance between levels, with curvature characteristics that are more similar to the observations.

d. Profile analysis

1) PROFILE ERRORS

Figure 7 shows the mean profile RMSEs for RAP and the top ML models, one per architecture type. For temperature, the errors range from 0.6° to 1.5°C, while for dewpoint the errors are larger and range from ~1°C near the surface up to almost 8°C in the upper atmosphere for the RAP. For both temperature and dewpoint, all ML models have lower errors than the RAP over the entire profile. The U-Net has the lowest RMSEs; however, all models perform similarly throughout both profiles, except briefly from 800 to 700 mb when the LIN and NN models have slightly higher temperature errors. For temperature, the ML models reduce the RAP errors from ~15% near the surface to ~2% in the upper atmosphere, while for dewpoint the error reduction is modest at the surface (~5%–8%) but increases up to ~40% in the upper atmosphere.

2) BEST PROFILES

To provide an example of individual soundings, the best and most improved profiles are shown in Fig. 8. Differences between the RAP and ML models for the temperature profiles are difficult to see, highlighting the small existing errors in the RAP and the minimal improvements from the ML models. For the best temperature profile (Fig. 8a), the RAP already has low errors and improvements by the ML models are minimal. The ML improvements can best be seen in the most improved T profile shown in Fig. 8c, where the ML models lower the near-surface temperatures (below 850 mb) to more closely match the observations.

Improvements to the dewpoint profiles can be seen in all the selected soundings. In all panels, the ML models come closer to matching the lower dewpoints seen above 400 mb. For the best dewpoint sounding (Fig. 8b), the U-Net lowers

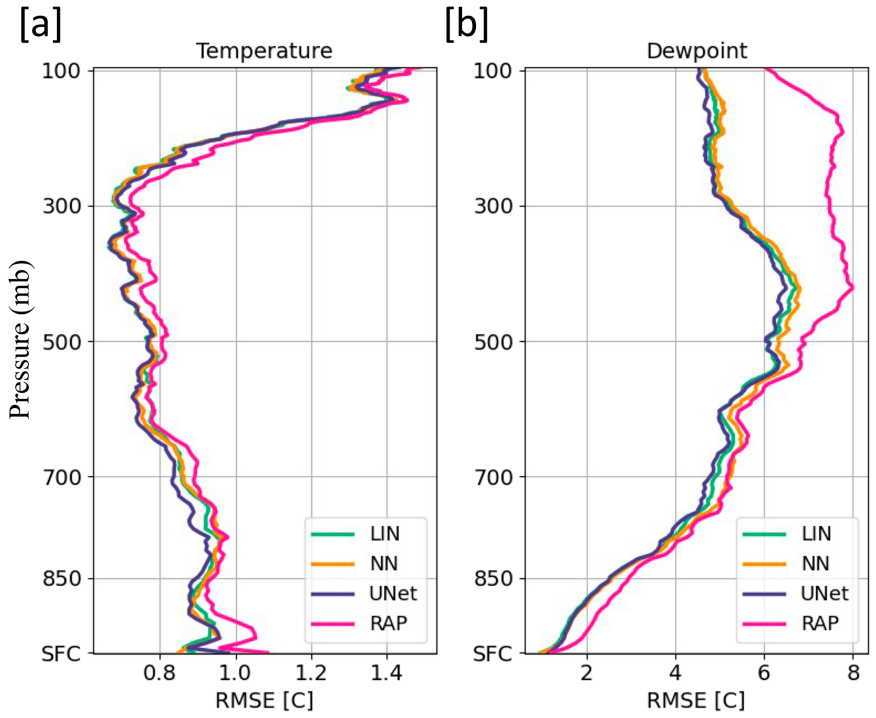


FIG. 7. Mean profile RMSEs for (a) temperature and (b) dewpoint.

the errors by improving the profile primarily above 650 mb. For the sounding with the most improved dewpoint profile (Fig. 8d), the RAP has two major difficulties: 1) predicting too low temperatures lower than 500 mb and 2) predicting too high temperatures above 400 mb. The ML models address both problems, reducing the low temperature bias near the surface and more closely matching the observations in the upper atmosphere.

3) WORST PROFILES

The worst and most degraded near-surface profiles are shown in Fig. 9. These are shown for the top-performing U-Net model with uncertainty estimates, which uses a CRPS loss function with 60 ensemble members. As throughout, the central line shown for the U-Net CRPS is the ensemble median and the 95% error bar is the corresponding range in ensemble members per vertical level. We demonstrate in section S2 in the online supplemental material that the predicted uncertainty estimates are well calibrated using spread–skill diagrams, PIT histograms, and discard tests.

For the worst near-surface temperature profile (Fig. 9a), both the RAP and the U-Net miss a strong surface temperature inversion. While the uncertainty estimates increase over the poorly performing region, they do not cover the discrepancies seen between the predictions and the observations. For the most degraded temperature sounding (Fig. 9c), while the differences are difficult to see, the U-Net shifts the RAP profile further from the observations throughout the majority of the profile. Lower than 500 mb, the U-Net lowers the RAP temperatures below the observations, and above 500 mb the U-Net increases the temperatures above the observations. The U-Net degrades the RAP performance by 22%; however, the degradation is in the uncertainty range predicted by the model.

For the worst near-surface dewpoint profile (Fig. 9b), both the RAP and the U-Net miss a region of low dewpoints around 800 mb, which is also outside of the expected uncertainty. This supports the gradient results (Fig. 6b), where the models underestimate the dewpoint gradients compared to the radiosondes. The U-Net improvements in the upper atmosphere compensate for the higher errors lower in the atmosphere, causing the U-Net to have lower overall errors than the RAP. This is not true for the degraded profile (Fig. 9d), where the U-Net has a higher RMSE than the RAP because the U-Net shifts the dewpoints colder both lower than 700 mb and above 400 mb, both of which shifts the dewpoints away from the observations. Utilizing the probabilistic predictions, the U-Net predicts higher uncertainties for this case, which mostly cover the model errors.

4) CONVECTIVELY ACTIVE PROFILES

Figure 10 shows soundings for two cases with high mixed-layer CAPE, indicating they have the potential to be convectively active. In Fig. 10a, this sounding occurred on a day when there were severe weather reports throughout the region, including tornadoes. For this case, both the LIN CRPS and U-Net MAEW models most closely match the observations, improving upon the RAP by ~45%. While the RAP

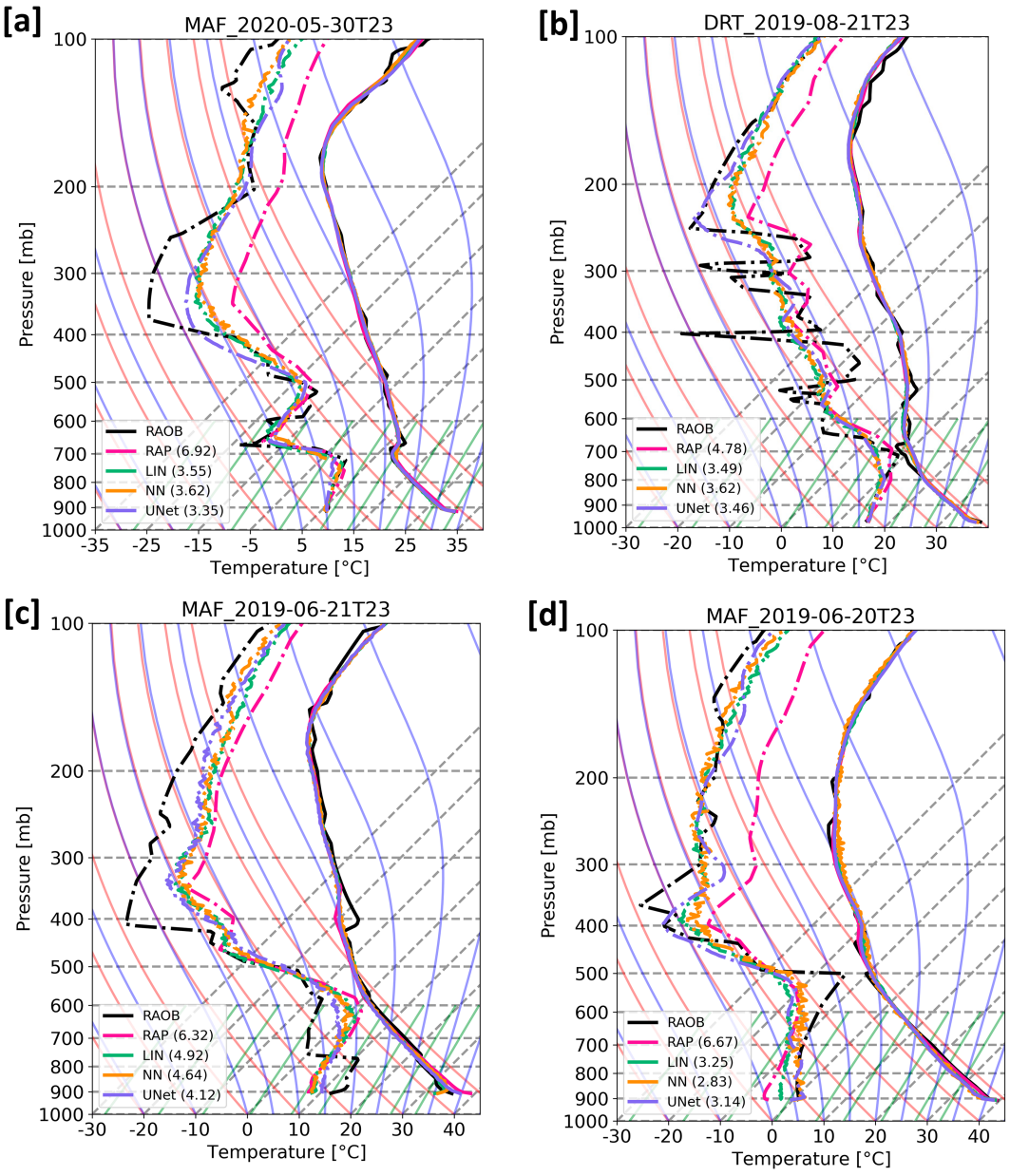


FIG. 8. Example soundings. Temperature profiles are solid lines; dewpoint profiles are dashed lines. The different colors show the observations (black), RAP (pink), and ML model results. Best (a) temperature and (b) dewpoint; most improved (c) temperature and (d) dewpoint. The RMSE shown in parentheses is the combined errors for both temperature and dewpoint.

predicts the temperature profile pretty well, it overestimates the dewpoints near the surface, particularly at 800 mb. The ML models lower the dewpoints at this altitude, making them more in line with the observations, which shifts the CAPE toward the lower observations as well.

In Fig. 10b, the sounding shows significant CAPE above a region of CIN, and on this day there were reports of wind and hail throughout the region, but no tornadoes. In this case, all the models have large errors in near-surface temperature and dewpoint, where they miss strong gradients at ~880 mb. Because of this, all models underestimate the CIN; however, slight shifts in the temperature and dewpoint profiles near the surface help the ML models have lower CAPE values closer to the observations.

e. CAPE/CIN direct predictions

We tested directly predicting CAPE and CIN, rather than deriving these from the predicted profiles. To do this, we used random forests (RF), LIN, and fully connected NN architectures. We used the same inputs as for the profile predictions, including

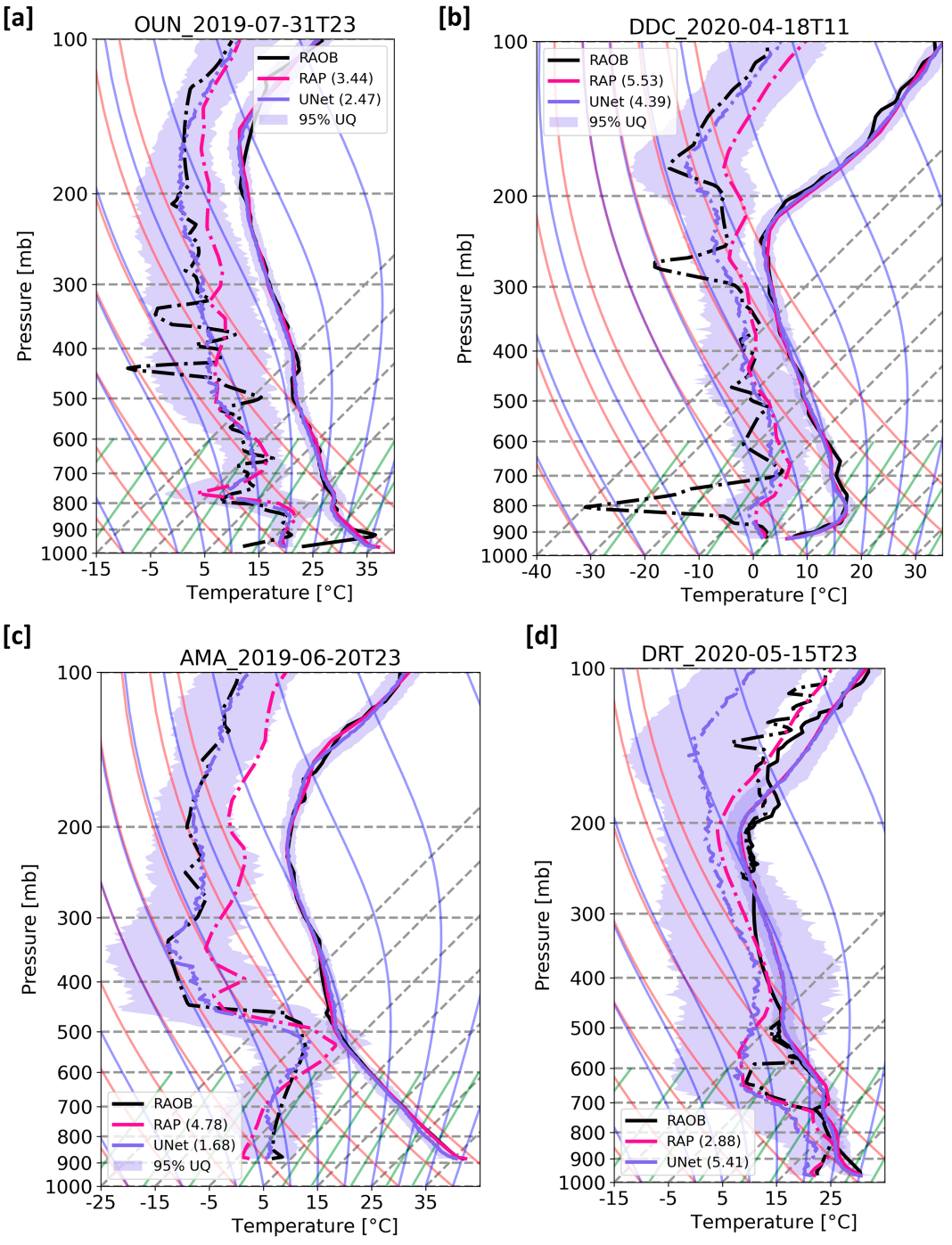
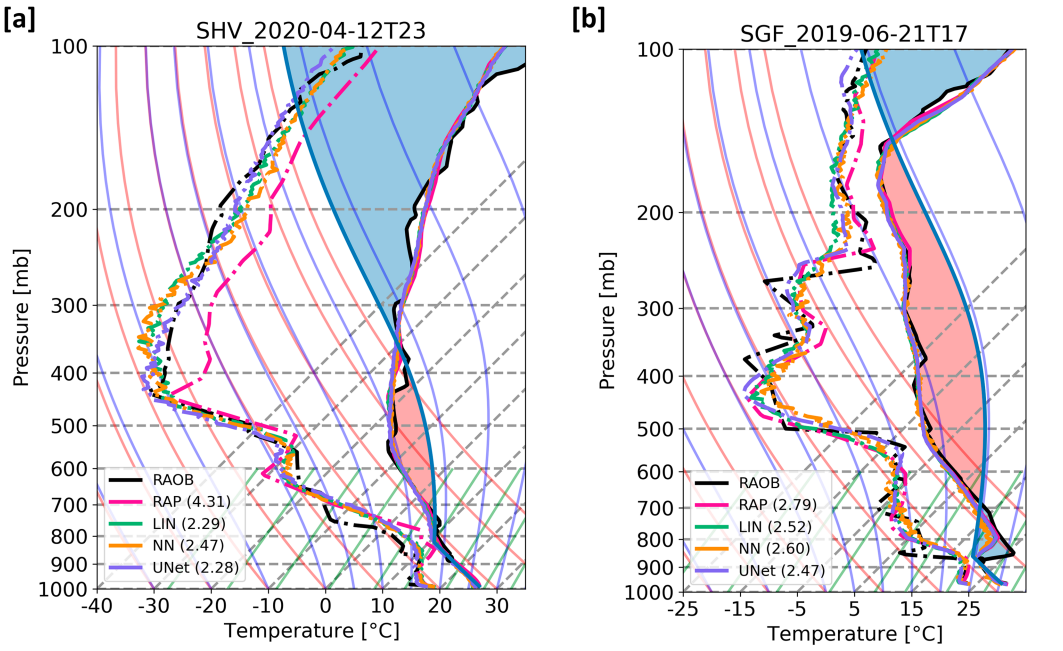


FIG. 9. Temperature (solid) and dewpoint (dashed) profiles for the (top) worst and (bottom) most degraded performance near the surface (first 25 levels up to 800 mb). Worst (a) temperature and (b) dewpoint; most degraded (c) temperature and (d) dewpoint. As in Fig. 8, the total RMSE for both temperature and dewpoint is shown in parentheses.

the RAP, GOES, and RTMA data, and all models are optimized using a hyperparameter search. Rather than predicting profiles, the outputs are now scalars for CAPE and CIN, and these are evaluated against CAPE and CIN computed from the raob observations. The results are shown in Table 1.

Predicting CAPE and CIN directly improves the performance over the RAP for all ML models, and all models improve the CAPE predictions more than the CIN predictions, decreasing the prior RAP CAPE and CIN errors by ~32% and ~13%, respectively. For CAPE, the models that predict the profiles perform



	CAPE	CIN	T RMSE	T Sfc RMSE	TD RMSE	TD Sfc RMSE
RAOB	742	33				
RAP	1226	0	1.83	0.65	5.81	3.26
LIN	1162	2	1.74	0.47	2.73	2.26
NN	1092	5	1.76	0.53	3.01	2.03
UNet	985	0	1.77	0.55	2.70	1.64

	CAPE	CIN	T RMSE	T Sfc RMSE	TD RMSE	TD Sfc RMSE
RAOB	3377	324				
RAP	3910	113	1.12	2.61	3.78	3.91
LIN	3123	157	1.31	2.59	3.31	3.45
NN	3427	94	1.42	2.87	3.39	3.88
UNet	3190	153	1.25	2.66	3.27	3.53

FIG. 10. Soundings for atmospheric conditions supporting high CAPE. Red (blue) shading has been added for CAPE (CIN), and the total combined temperature and dewpoint RMSEs are in parentheses. (a) Example of U-Net improving over RAP. (b) Example of already well-predicted RAP.

the best. For CIN, direct prediction using an NN has the lowest error; however, the profile models have only slightly worse performances with the advantage of providing additional information regarding the temperature and dewpoint profiles.

5. Discussion and conclusions

This study investigates the potential to use ML to post-process NWP output for improving vertical profiles of temperature and dewpoint. We explored the use of different ML model architectures, including linear regression, fully connected neural networks, convolutional neural networks, and deep residual U-Nets. Since the target data for this project are 1D vertical profiles, we outlined methods to incorporate 1D profiles with 2D image data in unified networks. To improve model training, we introduced physically inspired custom loss functions designed specifically for this task in order to accentuate different levels in the vertical profile. Finally, we utilized six physically based metrics to evaluate model performance. All the ML models improved the dewpoint profiles by up to 45%, particularly in the middle and upper atmosphere, and these improvements occur ~82% of the time. The improvements to the temperature profiles are more modest and less robust, with improvements of ~10% that occur 68% of the time. The modest temperature improvements are likely due to the RAP already having significantly lower temperature errors compared to dewpoint, making improvements more difficult.

While the overall results are relatively robust among the different model architectures, model performance differences illuminate important science and ML aspects. First, the U-Net

TABLE 1. Results for predicting CAPE/CIN directly vs calculating these from predicted temperature and dewpoint profiles. The values shown are the RMSEs (J kg^{-1}). The RF, LIN, and NN models predict CAPE/CIN directly. The LINP, NNP, and U-NetP models are the best models predicting temperature/dewpoint profiles. Predicting the CAPE/CIN from the ML-predicted temperature and dewpoint profiles uses the RAP pressure. The best performing models for CAPE and CIN are shown in bold.

	RAP	RF	LIN	NN	LINP	NNP	U-NetP
CAPE	378	267	278	243	**240**	241	244
CIN	108	96	98	**93**	95	95	96

models perform the best overall. This indicates that using vertical spatial information and skip connections helps improve the model, particularly for the temperature profiles, and it also allows the U-Net-predicted profiles to have similar gradients and smoothness as the observations. Since the ML models are able to improve both the temperature and dewpoint profiles, resulting predictions of both CAPE and CIN substantially outperform the RAP predictions by up to 36%. These improvements are greater than using ML models to predict CAPE/CIN directly.

Second, the NN models perform worse than the linear regression models, indicating that adding model complexity does not improve performance, especially given the cost of reduced understanding for these large models. While these models perform comparably in obtaining the CAPE and CIN, the vertical profiles of dewpoint and temperature are lower performing than their linear and U-Net counterparts. Additionally, these models create profiles that are noisier than the observations due to not taking spatial context into account, evident from comparing second order derivatives.

Third, the majority of the information learned by the models is obtained from the RAP profiles, indicating that these models are primarily fixing existing biases in the RAP. While adding the RTMA and GOES data improves the results, particularly for CAPE and CIN, these improvements are modest. Investigating the performance of the models for various cloud cover conditions (see supplemental section S3a) reveals that the ML improvement is not sensitive to cloud cover, indicating that any potential consequences from cloud contamination to the data are more than offset by the increase in sample size. Further, the ML predictions do not exhibit any monthly or regional biases (see supplemental sections S3b and S3c), indicating the improvements to the RAP model are robust and reliable, regardless of location or time of year.

Fourth, the ML models are able to provide well-calibrated uncertainty estimates that may be useful to the forecasting community (see supplemental section S2). The ML models predict uncertainty estimates that match the corresponding error for the majority of the predictions without suffering any degradation to the temperature and dewpoint predictions. Having uncertainty estimates at all vertical levels helps identify where the ML model may have the largest errors, providing forecasters not only with a heads-up for potential errors, but also with information regarding where in the profile these errors are occurring so that they may use their domain expertise to help overcome them. Further, since removing the most uncertain samples results in improved model performance, aggregating uncertainty estimates over the entire profile provides an estimate of how uncertain the entire profiles are. If specific applications require more accurate results, thresholds can be established to utilize only the most-certain profiles with reasonable confidence. And since both temperature and dewpoint profiles are predicted separately, users can adjust their quality requirements to temperature and dewpoint separately.

A few limitations of this work are to be noted. This study focused on the central United States; thus, the results are not expected to generalize to sites outside of this region. In our study region, while the overall statistics indicate the ML models are an improvement over the RAP, this is sounding specific and is not the case for all profiles. As seen in the worst and most degraded soundings, cases exist where the ML model degrades the RAP performance, adjusting the profiles incorrectly. Further, the improvement to the RAP is not uniform across all samples, with the ML models doing the best job at correcting large errors in the dewpoint profiles, particularly in the middle to upper atmosphere levels. However, this work suggests that the postprocessing the RAP using ML offers improvements to the temperature and dewpoint profiles, as well as derived CAPE and CIN values, and that when taken with the knowledge of the model's limitations, the ML-predicted profiles can provide useful information. Last, we note that this work was performed with RAPv4. Due to data limitations, we cannot test this proposed approach on RAPv5; however, we hypothesize that it could be extended to other versions and NWP models with proper training data.

Although this study focused on improving temperature and moisture profiles for severe weather forecasting, adjusting a first guess model profile with satellite and other data can be used in other applications as well. For example, moisture throughout the entire atmospheric column is of interest to the climate community, and the improvements in upper-level water vapor demonstrated here may benefit research and forecasting on processes at longer time scales.

Finally, this study suggests several topics for future work. First is to include skin temperature as an input feature, which could be derived from the satellite data and may help the predictions in the lower level. Second is to include the winds from the RAP data, which may provide more information to the ML models. Third is to incorporate CAPE and CIN in the loss functions, which we currently found to be too time-consuming for training; however, as computer resources improve this could be a possible avenue for improvement in the future. Finally, since one of the current weaknesses is in predicting the temperature and dewpoint in the boundary layers, including more information regarding boundary layer height could be helpful. For example, estimates derived from the temperature gradients and wind profile could improve performance.

Acknowledgments. We thank Dr. Louie Grasso, who provided valuable domain expertise, helpful resources, and continuous feedback. We thank John Dandy as well as Dr. Grasso for their assistance in acquiring the data, which accelerated our research efforts. We thank John M. Haynes for his help in analysis, and we thank Charles H. White and two anonymous reviewers for their insightful and helpful comments on our manuscript. This project was funded by NOAA Grant NA19OAR4320073.

Data availability statement. The data used in this study are available via Dryad at https://doi.org/10.5061/dryad.h70rxwdqn. The code for this project is available on Zenodo at https://doi.org/10.5281/zenodo.10041772.

REFERENCES

Barnes, E. A., R. J. Barnes, and N. Gordillo, 2021: Adding uncertainty to neural network regression tasks in the geosciences. arXiv, 2109.07250v1, https://doi.org/10.48550/arXiv.2109.07250.

——, ——, and M. DeMaria, 2023: Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environ. Data Sci.*, **2**, e15, https://doi.org/10.1017/eds.2023.7.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Bergstra, J., D. Yamins, and D. Cox, 2013: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *ICML'13: Proc. 30th Int. Conf. on Int. Conf. on Machine Learning*, Atlanta, GA, PMLR, 115–123, https://proceedings.mlr.press/v28/bergstra13.html.

Candido, S., A. Singh, and L. D. Monache, 2020: Improving wind forecasts in the lower stratosphere by distilling an analog ensemble into a deep neural network. *Geophys. Res. Lett.*, **47**, e2020GL089098, https://doi.org/10.1029/2020GL089098.

Dai, Y., and S. Hemri, 2021: Spatially coherent postprocessing of cloud cover forecasts using generative adversarial networks. *EGU General Assembly 2021*, Online, European Geosciences Union, Abstract EGU21-4374, https://doi.org/10.5194/egusphere-egu21-4374.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's national centers for environmental prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Ghazvinian, M., Y. Zhang, D.-J. Seo, M. He, and N. Fernando, 2021: A novel hybrid artificial neural network-parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour.*, **151**, 103907, https://doi.org/10.1016/j.advwatres.2021.103907.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, 220061, https://doi.org/10.1175/AIES-D-22-0061.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hilburn, K., 2020: Inferring airmass properties from GOES-R ABI observations. *2020 Fall Meeting*, Online, Amer. Geophys. Union, Abstract A008-0009, https://doi.org/10.1002/essoar.10504854.1.

Hurlbut, M. M., and A. E. Cohen, 2014: Environments of northeast U.S. severe thunderstorm events from 1999 to 2009. *Wea. Forecasting*, **29**, 3–22, https://doi.org/10.1175/WAF-D-12-00042.1.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting*, **24**, 1691–1705, https://doi.org/10.1175/2009WAF2222201.1.

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

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087.

Peng, T., X. Zhi, Y. Ji, L. Ji, and Y. Tian, 2020: Prediction skill of extended range 2-m maximum air temperature probabilistic forecasts using machine learning post-processing methods. *Atmosphere*, **11**, 823, https://doi.org/10.3390/atmos11080823.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Renkl, C., 2013: The vertical structure of the atmosphere in COSMO-DE-EPS: Multivariate ensemble postprocessing in the space of vertical normal modes. Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn.

Rojas-Campos, A., M. Wittenbrink, P. Nieters, E. J. Schaffernicht, J. D. Keller, and G. Pipa, 2023: Postprocessing of NWP precipitation forecasts using deep learning. *Wea. Forecasting*, **38**, 487–497, https://doi.org/10.1175/WAF-D-21-0207.1.

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—MICCAI 2015*, N. Navab et al., Eds., Lecture Notes in Computer Science, Vol. 9351, Springer, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

——, and Coauthors, 2019: Legacy atmospheric profiles and derived products from GOES-16: Validation and applications. *Earth Space Sci.*, **6**, 1730–1748, https://doi.org/10.1029/2019EA000729.

Schoenach, D., T. Simon, and G. J. Mayr, 2020: Postprocessing ensemble forecasts of vertical temperature profiles. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **6**, 45–60, https://doi.org/10.5194/ascmo-6-45-2020.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philos. Trans. Roy. Soc.*, **A379**, 20200097, https://doi.org/10.1098/rsta.2020.0097.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

Schwartz, B., and M. Govett, 1992: A hydrostatically consistent North American radiosonde data base at the Forecast System Laboratory, 1946–present. NOAA Tech. Memo. ERL FSL-4, 81 pp., https://repository.library.noaa.gov/view/noaa/32850.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Stock, J. D., 2021: Using machine learning to improve vertical profiles of temperature and moisture for severe weather

nowcasting. M.S. thesis, Dept. of Computer Science, Colorado State University, 95 pp..

Vannitsem, S., D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier Science, 347 pp., https://doi.org/10.1016/C2016-0-03244-8.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using covolutional neural networks. *Mon. Wea. Rev.*, **149**, 1141–1152, https://doi.org/10.1175/MWR-D-20-0219.1.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482, https://doi.org/10.1175/2007MWR2018.1.

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916, https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2.