tex file

Generated using the official AMS LaTeX template v6.1

Effectiveness and Limitations of Machine Learning for Predicting Lower Atmosphere Temperature Profiles

Kip F. Nielsen,[a] David A. Rahn,[a] Nathaniel A. Brunsell,[a] and Erik S. Van Vleck[b]

[a] *Department of Geography and Atmospheric Science, University of Kansas, Lawrence, Kansas*

[b] *Department of Mathematics, University of Kansas, Lawrence, Kansas*

Corresponding author: David A. Rahn, darahn@ku.edu

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-26-0049.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2026 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

ABSTRACT: Temperature profiles in the lower atmosphere rapidly evolve during the morning transition, and predicting their details over complex urban environments is particularly important for pollutant dispersal. Temperature profiles can be predicted by operational weather models but are sensitive to model parameterizations. Machine learning models can be used, but require high-resolution profiles that are often unavailable. To overcome this, pre-dawn temperature and wind speed observations from aircraft over Dallas-Fort Worth, TX are used to predict the morning transition. Initially, only temperature is used as an input to assess several types and configurations of machine learning models. Each model's prediction is evaluated based on various methods because standard metrics, such as mean square error (MSE), may be minimized but miss important features, such as strong temperature gradients. A long short-term memory network with an AdamW optimizer does not result in the lowest MSE but does better capturing the temperature structure. The model tends toward the annual mean by underestimating the onset of mixing and near-surface temperature in summer and overestimating these in the winter. Including wind speed does not have a statistically significant improvement in MSE but performs better sequentially and with the inversion at the top of the boundary layer. Further refinements are likely to improve rapidly updated short-term forecasts, which may provide an additional forecast tool, but large gradients, extreme means, and external forcing remain an inherent challenge. These results highlight the importance of holistic evaluation of machine learning techniques to ensure adequate representation of predicted temperature profiles.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

SIGNIFICANCE STATEMENT: This work investigates how well machine learning models predict temperature changes in the lower atmosphere (<3,000 m) just after sunrise. This is important because pollution near the surface is impacted by the temperature profile. We used temperature and wind speed observations from aircraft arriving and departing airports in Dallas-Fort Worth to investigate the models' capability in capturing temperature changes. Basic statistics do not show large differences between models, but more complex models perform better in capturing rapid temperature changes in the lower atmosphere and retaining important features. The results of these machine learning models are promising, but there are inherent limitations, such as underestimating large temperature changes, the warmest and coldest days, and external factors such as frontal passages.

1. Introduction

Many processes in the planetary boundary layer (PBL) impact weather, climate, and air quality. The PBL can change rapidly and is influenced by local and remote processes, making it difficult to understand and predict. The nocturnal PBL is particularly challenging and is generally much less understood than the convective PBL (Seidel et al. 2012; Angevine et al. 2020). The morning transition (MT) is when the PBL transitions from stable to convective, and is typically defined as beginning at sunrise and ending at the onset of rapid growth of the well-mixed daytime PBL (Angevine et al. 2001; Lapworth 2006; Halios and Barlow 2018). The timing and duration of the MT can vary greatly and is difficult to predict, but are particularly important because changes in stability greatly influence the concentration of pollution near the surface (Kotthaus and Grimmond 2018; Nielsen and Rahn 2022).

The PBL is modified by the characteristics of the land surface and can be substantially different between urban and rural environments. For example, a greater sensible heat flux over cities at night can maintain a shallow convective or near-neutral layer, significantly changing the stability of the urban PBL compared to surrounding rural areas (Nunez and Oke 1977; Ryu and Baik 2012; Schrijvers et al. 2015). This increase in sensible heat contributes to the urban heat island effect, and these warmer temperatures can contribute to heat-related illnesses, such as heat stroke and heat exhaustion (Oke 1987). Urban canyons (the narrow space between tall buildings) reduce outgoing long-wave radiation from a city and there can be other non-local impacts, such as

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 01/16/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

warmer temperatures and air pollution transported to downwind rural communities (Cosgrove and Berkelhammer 2018). High concentrations of pollution can increase the incidence rate of acute or chronic respiratory illnesses (Brunekreef and Holgate 2002; Guan et al. 2016; Kim et al. 2018; Friedman et al. 2024). As urban sprawl and densification continue, the need to understand and predict the urban PBL will increase.

Numerical weather prediction using models such as the Weather Research and Forecasting modeling system is one approach to analyze the connection between these components and the PBL (Skamarock et al. 2005), but it is often sensitive to model parameterizations, including PBL schemes and urban canopy models. In addition, observations are necessary to assess the model (Barlow 2014; Hu et al. 2010).

Machine learning is rapidly gaining in popularity and can be used to predict complex components of the environment, including the urban PBL (McGovern et al. 2023; Molina et al. 2023; Peng et al. 2023). Linear regression is considered one of the most basic forms of machine learning, but sophisticated neural networks with many nodes (e.g., neurons) are more widely used. For a multi-layer neural network (i.e., deep neural network), the input for the second layer is the output from the first layer, and so forth (LeCun et al. 2015). This iteration continues for the entire training dataset trying to predict the user-defined output(s) from the user-defined input(s). Performance is assessed using the validation dataset to tune the internal hyperparameters. Finally, the resulting model is evaluated with the testing dataset, and the accuracy can be deduced using summary statistics and analyzing the individual predictions. The machine learning model gauges performance based on a single error metric rather than on underlying physical processes, so the user needs to balance minimizing global test statistics and how the model captures the PBL structure.

Even among the widespread uses of machine learning, its application is relatively limited in predicting the urban PBL. Past work often predicts only the PBL height (PBLH), but interpreting the results becomes a challenge. A major issue is that the results are largely dependent on how the PBLH is defined since there is such a wide variety of methods and no optimal PBLH formulation for every stability type (Vickers and Mahrt 2004; Zhang et al. 2014). An example using PBLH is Molero et al. (2022) who tested the performance of linear regression, regression trees, support vector machine, and Gaussian process regression to estimate PBLH using ground-level meteorological data and ceilometer profiles as ground truth. Temperature and relative humidity were the most

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/06/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

important, and the models generally performed the worst in predicting PBLH during the summer. Predicting daily PBLH using an extreme gradient boosting regression model performed better than reanalysis compared to spaceborne and airborne lidars and in situ aircraft observations (Ayazpour et al. 2023). An example using a sequential machine learning model is Kumar et al. (2021) who found that using observed values as the input of the next step instead of predicted values (i.e., non-autoregressive) in a long short-term memory network (LSTM) to predict PBLH decreased the mean absolute percentage error from 17.3% to 10.63% and the relative root mean square error decreased from 7.33% to 5.95%. The models performed similarly during the MT and the evening transition.

Machine learning can be used with numerical models to predict entire temperature profiles, which can lead to improvements in weather forecasts. For example, Haynes et al. (2024) showed that machine learning models greatly improved the dew point profiles and slightly improved temperature profiles from the Rapid Refresh model. A deep residual U-Net structure performed best by capturing spatial patterns between nearby levels, while complex neural networks tended to create noisy profiles that overfit observations. Retaining uncertainty estimates at each level of the atmosphere can be useful to the forecasting community in identifying potential errors. Fan et al. (2023) used a similar U-Net structure to predict a 3D temperature field up to 1,300 m above the surface with only 2-m air temperature as an input that was accurate to within 0.5 K compared to reanalysis. Adding physical constraints to the PBL, an initial guess, terrain height, and separating by time improved performance of this U-Net model. Fujita et al. (2024) used an algorithm that quantifies a profile's curve, the Signature method, to capture the vertical structure of atmospheric temperature and humidity to within 2.0 K and 2.0 g kg^{-1} on an annual basis using surface observations and satellite observations of water vapor column amounts and brightness temperature.

Using numerical models or only PBLH as input or output for machine learning models is a common approach (Krishnamurthy et al. 2021; Molero et al. 2022), but another approach is to use observed profiles directly, if available. More complicated structures of the PBL could be included to train a model by using temperature profiles instead of just PBLH. This makes it easier to interpret which parts of the PBL the machine learning model is capable of representing, which is challenging to do with a single test statistic output. Capturing temporal changes in the PBL is a important, and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.
Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

a natural fit is sequential machine learning models, such as a recurrent neural network (RNN) or LSTM. The applications of machine learning of the PBL are broad due to the importance of the PBL, but the application is currently developing with some limitations, which we highlight in this work.

One of the fundamental limitations is that having a finely resolved vertical structure of the rapidly evolving lower atmosphere is necessary to produce effective predictions. The lack of observations at the surface and within the PBL greatly limits our fundamental understanding of physical processes occurring at the land-atmosphere interface (Helbig et al. 2021). Urban environments are especially challenging because there is generally a lack of information in the upper 90% of the urban PBL (Grimmond 2006; Barlow 2014; Li et al. 2018).

Here, we leverage the vast network of observations from the Aircraft Meteorological Data Relay (AMDAR) Program to overcome the paucity of data common in PBL studies (Petersen 2016; Zhang et al. 2020). Given the importance of the morning temperature profile over urban areas, we are using the high-resolution AMDAR temperature profiles to construct empirical predictions based on several machine learning models. We create daily batches of temperature profiles that are available every half hour. The soundings in each batch are split into pre-dawn soundings and morning soundings. The pre-dawn soundings are used as input to predict the evolution of the temperature profile during the first four hours after sunrise. We examine several approaches and highlight strengths and weaknesses in predicting the temperature profile throughout the morning. This is done using a combination of summary statistics and inspecting all predicted profiles to determine the effectiveness and limitations of using machine learning.

To examine the potential for predicting temperature profiles using machine learning and AMDAR data, the three main goals are to: (1) find the model type, configuration, and optimizer that provides relatively low error while working well in a variety of PBL stability types, (2) diagnose under what conditions even the best model struggles compared to observations, and (3) determine whether the addition of wind speed profiles noticeably improves accuracy.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

2. Methods

a. Data

AMDAR observations are high-quality with root mean square errors compared to radiosondes of 1.16-1.52 K for temperature and 2.00 - 2.26 m s^{-1} for wind speed at pressure levels greater than 850 hPa for 54 U.S. airports (Zhang et al. 2019). These observations are used for a variety of applications (Drüe et al. 2010; Dai et al. 2014; World Meteorological Organization 2014, 2015, 2018). Only one city at a time should be selected to avoid training a single model based on two distinct local climates, such as a coastal location that has a well-defined sea breeze and a location in the interior of the continent that does not. Dallas-Fort Worth, TX is selected as the test location because it has ample observations from flights taking off and landing at Dallas-Fort Worth International Airport (DFW) and Dallas Love Field Airport (DAL), which includes many pre-dawn soundings (World Meteorological Organization 2017). There are no major topological or coastal influences. DFW and DAL are located 18 km apart in fairly flat urban terrain with very little difference in temperature profiles (Nielsen and Rahn 2022). Observed profiles below 3,000 m and from 2010-2023 for DFW and DAL are combined to improve vertical and temporal resolution and ultimately better represent the profile over the city, which can lead to better overall performance. Similar to Rahn and Mitchell (2016), there is no ascent/descent bias. Although only one city is selected, the methods, conclusions, and limitations are generalizable and should be considered when training at other locations.

The use of aircraft observations over traditional soundings yields an unprecedented analysis because of the much larger amount of data available to capture the diurnal evolution of the PBL under a variety of stability types. Dallas-Fort Worth often has weak to moderate mixing in the PBL at sunrise because it is a large urban location, but temperature inversions are also noticeably evident (Nielsen and Rahn 2022).

Due to the inconsistency of flight times, profiles are constructed from all point observations from DFW and DAL within a half-hour window that is normalized relative to sunrise. Height is with respect to sea level. Nielsen and Rahn (2022) have shown that combining both airports in one dataset leads to better representation and fidelity of the PBL structure. They used a simple linear interpretation to construct a single representative sounding, but the present work uses locally

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

weighted scatterplot smoothing to better retain structures within the profile, which is essentially a moving window regression done throughout the profile.

Temperature profiles are separated into daily batches, where one batch consists of 13 half-hour profiles. The first five are from two hours before sunrise to sunrise and are used as model input. Observations earlier than two hours before sunrise quickly become sparse due to fewer flights, particularly in the summer. The next eight half-hour time steps begin a half-hour after sunrise and continue to four hours after sunrise. These are used as the model output (predictions). The PBL is often well-mixed four hours after sunrise. This analysis is designed to assess the performance of the model during the most challenging time of the day when the PBL changes rapidly during the MT. Days with missing profiles are excluded from the analysis because some models, such as LSTMs, rely on continuous time steps for the best accuracy. This results in a total of 1,547 batches (days) that are shuffled before being separated into datasets of 70% for training, 20% for validation, and 10% for testing. The same seed from the random generator is used in all models, so each model has the same training, validation, and testing datasets. Shuffling the data is important so that there is no known relationship between consecutive batches.

b. Models

The best performing model type is found before testing various configurations, learning rates, and optimizers. Initially, only temperature is used as input, then the wind speed is added to the input. The goal is to start simple, so reducing the number of inputs makes it easier to deduce how well a model operates on the temperature profiles, its limitations in performance, and what it is able to capture. Including additional inputs and complexity will likely improve the machine learning model output, but here the focus is on fundamental model performance.

The model types tested are a linear neural network (LNN), dense neural network (DNN), convolutional neural network (CNN), RNN, and LSTM. RNNs are optimized to work with sequential information and CNNs are optimized to work with images or two-dimensional data. One challenge with traditional RNNs is that the error may explode or vanish as the model progresses, which is overcome by the more advanced LSTM that has four interacting layers in each repeating module (Hochreiter and Schmidhuber 1997). LSTMs build upon RNNs by being designed to work with time-series data, but LSTMs are trained to learn what information to retain and what to forget from

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

earlier in the sequence. More information on these models, including neural network diagrams, can be found in LeCun et al. (2015).

Choosing and applying the desired model is not a simple task. Some components, such as the number of nodes, are easier to select due to previous studies in this area. Other components can be selected using various techniques, such as early stopping or hyperparameter optimization via methods like random search and Bayesian optimization (Liashchynskyi and Liashchynskyi 2019; Alibrahim and Ludwig 2021). However, some are more challenging because there are not many PBL-related studies to use as a starting point. Because of this, each component of the model is tested and presented to reveal the optimal model type, configuration, learning rate, and optimizer to predict temperature profiles over a built environment using a large dataset. A vital component of finding the optimal components is inspecting the predictions to understand the strengths and weaknesses inside each model and between the models. Looking beyond the summary statistics provides invaluable information to determine the capability and usefulness of these models. For example, the existence of lapse rates greater than the dry adiabatic lapse rate would indicate a non-physical solution.

An important part of model convergence is based on the optimizer, number of epochs, and learning rate. An epoch is one iteration of training, validating, and testing the model, and the evaluation metric used is mean square error (MSE), which is a common evaluation metric useful for training the model to reduce errors throughout the profile (Fan et al. 2023). However, inspecting each of the predicted profiles can show whether the model is capable of capturing the dynamic and complex temperature structure inside the PBL, which is not usually captured in a single test statistic. The reduction in MSE for each iteration depends on the learning rate and optimizer. If the learning rate is too high, it may overshoot the desired minimum. If it is too small and the number of epochs is too small, it may never reach the minimum. If the number of epochs is too large, the model may start performing worse (overfitting), or it may make little to no improvement.

The optimizer controls how the model should change based on recent changes in MSE. Some optimizers have momentum, which is when the model makes larger changes if there is a large decrease in MSE, allowing it to converge faster than if it had made the same repeated small steps. The optimizers tested are Adadelta, adaptive gradient algorithm (Adagrad), adaptive moment estimation (Adam), Adam with decoupled weight decay (AdamW), Adamax, Nadam (Adam with

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Nesterov momentum), root mean square propagation (RMSprop), and stochastic gradient descent (SGD, with momentum). The descriptions of the optimizers and their benefits are provided in Table A2.

Only temperature is used as an initial input to keep the models relatively simple. Predicting temperature profiles near the surface using AMDAR is a novel application, and the focus on the evolution of the profile removes issues related to how the PBLH is defined. The mean profile is removed and divided by the standard deviation before being split into their respective datasets to normalize the data. The mean profile is added back to the profiles and multiplied by the standard deviation after the model is finished. The first step is to determine which type of machine learning model is most appropriate based on the assumption that the type of model has the largest influence on performance. The five model types tested consist of a single layer neural network with 500 nodes, 100 epochs, a learning rate of 0.0001, and use the Adam optimizer, which generally has a good balance of low error and fast computation time. Each model is trained using MSE as the loss metric and stops once the MSE stops improving after 5 epochs. The lowest MSE is then compared with all the models to help determine the optimal one. No single training metric can best capture all parts of the profile, especially sharp gradients. However, MSE is selected here because of its widespread use, so it will be used in conjunction with inspecting each predicted profile.

After finding the model with the lowest MSE that performs well under a variety of stability conditions, the model configuration and the learning rate are changed to find the best combination. This is done by changing the order of the layers, changing the number of nodes in each layer between 1,000 and 2,000, all while also testing learning rates of 0.01, 0.001, 0.0001, 0.00001, and 0.00001 with 200 epochs. The lowest MSE from these model runs is used to find the optimizer that balances a low MSE while visually performing well. This is crucial to ensure that it reasonably captures the complex thermal structure of the PBL. Individual evolutions and seasonal averages are then analyzed before adding wind speed as input to determine if it improves the output.

Brought to you by | UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Dataset	LNN	DNN	CNN	RNN	LSTM
MSE of training	0.54	0.46	0.52	**0.35**	0.38
MSE of validation	0.60	0.63	**0.60**	0.64	0.62
MSE of testing	**0.74**	0.76	0.75	0.78	**0.74**

TABLE 1. Mean square error (MSE; °C) of the training, validation and testing datasets from each of the models tested. In bold is the lowest MSE for each dataset.

3. Results

a. Only temperature as input

1) MODEL TYPE

A LNN, DNN, CNN, RNN, and LSTM are compared using MSE and also identifying differences in the shapes of the projected profiles to determine the best type of machine learning model. The lowest MSE for the training dataset is the RNN, the lowest for the validation dataset is the CNN, and the LNN and LSTM have the nearly identical lowest MSE of 0.74°C (Table 1). The spread in MSE of the testing datasets is less than 0.04°C. The errors increase between the training, validation, and testing of the model, because the three datasets cover different non-sequential dates and therefore different conditions (e.g., different initial surface temperature inversions, residual layers, large-scale subsidence, etc.). The RNN and LSTM have the largest difference in MSE in the three datasets of 0.43°C and 0.36°C, respectively. These models could be interpreted as less generalizable since they work noticeably better on the training dataset. Small differences in performance between the RNN and LSTM could be due to either the limited information retained throughout each batch or to the total length of the window not being large enough to fully utilize the functionality of the LSTM.

The RNN and LSTM have a relatively low MSE in the training dataset and better capture local minima and maxima (Fig. 1). This example reflects the general trend that the LSTM best captures the base of the inversion, the top of the inversion, and the near-surface temperature. The LNN profile does not have extrema as large as the observation, and there is noticeably higher small-scale variance in the profiles. This profile is explicitly chosen to illustrate the increased variance. The more well-mixed profiles have less variance or a smoother profile in their prediction, which results in a smaller difference between the models. More nodes also lead to a noisier prediction, which

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

is consistent with previous studies (Haynes et al. 2024). However, having more nodes tends to increase performance across PBL conditions, such as a strong mid-level temperature inversion. For example, the DNN is up to 1°C better than the LNN in capturing the minimum and maximum temperature of the inversion in Fig. 1. Although the MSE of the testing dataset is nearly identical among the five models, all profiles show that the LSTM is more aggressive in capturing a wide range of PBL structures. This makes it easier to learn from the LSTM than an LNN or other model types that have predictions that trend more quickly towards a mean profile by weakening stable layers too rapidly. The LSTM also has the lowest average MSE in all three datasets. For these reasons, the LSTM is selected as the model type for further analysis.

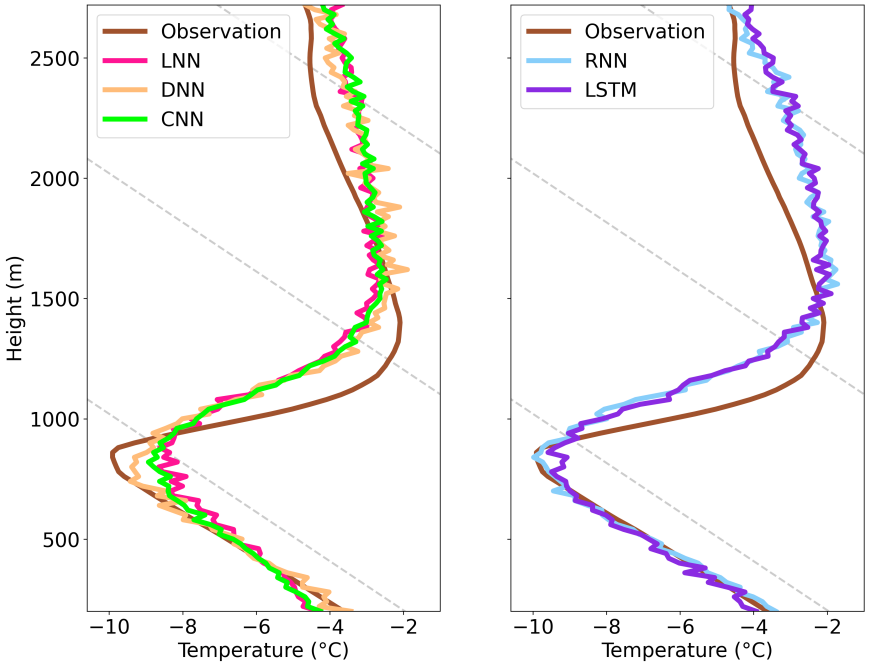


FIG. 1. Example predicted temperature (°C) profiles of the LNN (pink), DNN (peach), CNN (green), RNN (blue), and LSTM (purple) compared with the observation (maroon) on 20 January 2022 at 3 hours after sunrise. The dashed gray lines represent the dry adiabatic lapse rate (9.8°C km^{-1}).

2) Model configuration

Various configurations are assessed by changing the order of layers, the number of nodes, and the learning rate to find the lowest MSE for further analysis. Individual predictions are not shown because of the large number of models compared. The comparison of the MSE testing dataset across multiple LSTM configurations is shown in Table 2. Configurations were chosen to determine

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/06/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Model configuration/learning rate	0.01	0.001	0.0001	0.00001	0.00001*
LSTM(1,000)/Dense(1,000)/Dense(1,000)	4.81	0.94	0.91	**0.90**	1.02
LSTM(2,000)/Dense(1,000)/Dense(1,000)	7.05	0.96	**0.87**	0.94	0.99
LSTM(2,000)/Dense(2,000)/Dense(2,000)	7.33	0.94	**0.90**	0.98	0.95
Dense(2,000)/LSTM(2,000)/Dense(2,000)	3.95	1.27	0.97	0.98	**0.93**
Dense(2,000)/Dense(2,000)/LSTM(2,000)	16.56	4.83	0.96	**0.91**	0.94
CNN(2,000)/LSTM(2,000)/Dense(2,000)	4.37	1.96	1.10	**0.95**	1.03

TABLE 2. Testing dataset mean square error (MSE; °C) as a function of model configuration, with number of nodes in each layer in parentheses, and learning rate. In bold is the lowest MSE for each configuration. *200 epochs.

the role of doubling the number of nodes, changing the order of each layer, and quantifying the sensitivity of these changes with respect to the learning rate. The number of nodes chosen is based on preliminary analysis where fewer nodes converged too quickly and were not accurate, and too many nodes took much longer to converge. Decreasing the learning rate generally reduces the MSE but there is an inflection point for some configurations where the smallest learning rate with 200 epochs does not improve the model. This is likely due to overfitting or the model not generalizing well to different datasets (e.g., between training and testing). Larger differences would arise between 0.0001, 0.00001, and 0.00001 with 200 epochs if fewer epochs were used because the model may never converge to the solution. Learning rates of 0.01 and 0.001 are too high and repeatedly overshoot the solution, but demonstrate the role of the learning rate. Decreasing the learning rate generally reduces the MSE more than altering the model configuration. With the exception of the 0.01 learning rate, having the LSTM first yields the lowest MSE. Having a CNN layer instead of a dense layer does not noticeably decrease the MSE (with the exception of 0.00001 when they are relatively close together). Now that the model with the LSTM layer with 2,000 nodes followed by two dense layers with 1,000 nodes and a learning rate of 0.0001 is found to have the lowest MSE of 0.87°C, various optimizers with that configuration are tested.

3) OPTIMIZERS

Table 3 shows Nadam with the smallest testing MSE of 0.87°C and SGD completing first in 39 min. Nadam has a MSE 0.05°C lower than Adam, which highlights that the Nesterov momentum in Nadam is useful. Adamax has a 0.05°C larger MSE than Adam, suggesting that the usage of the infinity norm does not reduce the error in this case. The first and second order moments in

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Optimizer	MSE	Time to run (min.)
Adadelta	1.96	149
Adagrad	0.99	53
Adam	0.92	60
Adamax	0.97	66
AdamW	0.88	78
Nadam	**0.87**	89
RMSprop	0.88	57
SGD	1.37	**39**

TABLE 3. Testing dataset mean square error (MSE; °C) and time to run (min.) for the eight optimizers tested, with the lowest MSE in bold for each. Each model is run using 48 CPU cores with 4 gb of memory per core (CPU efficiency of 80-97%).

Adam are likely what cause it to have noticeably lower MSE and allow it to better capture a larger temperature range in the PBL than SGD. AdamW has a 0.04°C lower MSE than Adam, highlighting the usefulness of the additional method to decay the weights through L_2 regularization. By these metrics, the desired optimizer comes down to what is needed by the model, whether that is a fast converging model, a highly accurate model, or a balance between the two. In practice, initially creating the model takes the most amount of time, but the inference time will be more useful for making the machine learning model operational.

An example predicted profile comparing the eight optimizers represents characteristics seen on many other days (Fig. 2). A single statistic does not adequately quantify the model performance throughout the temperature profile, during all weather conditions, and at all times. One of the first notable differences between the optimizers is the amount of variance in the predicted profile. For example, Adadelta and SGD have a large variance throughout the profile, whereas Adam and AdamW have much smoother curves. Another difference is how each optimizer captures the base and top of the inversion. Adam, AdamW, and Adamax all capture the base of the inversion and the well-mixed lapse rate below it, while Adagrad, Adam, AdamW, and Nadam perform the best at capturing the top of the inversion. Most do a reasonable job of predicting the near-surface temperature, especially since it is 3 hours after sunrise. In other cases not shown here, Adadelta, SGD and Adagrad do not capture the minima and maxima throughout the profiles as well as the rest of the optimizers. SGD has large temperature variations throughout the profile when there is an inversion in the profile. AdamW performs the best at predicting the near-surface temperature

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

of approximately 7°C, the base of the inversion around 4°C, and is within 1.5°C at the top of the inversion.

The other optimizers perform worse at capturing the minimum and maximum value of the temperature inversion, which means that the stability of the PBL is not well represented. Between Adam and AdamW, the predicted profiles are nearly identical, but AdamW has less of a bias above 1,500 m, which can largely influence the MSE. This conclusion also holds when examining profiles on other days. Occasionally, there is a pronounced bias in the profile, leading to a larger MSE. However, the shape of the profile is still captured, despite an offset of the profile's mean temperature. This highlights the importance of analyzing individual predictions instead of relying on a single summary statistic. Although three of the optimizers essentially have the same relatively low MSE of 0.88°C, the AdamW optimizer is selected since the profile is smoother and tends to capture sharp gradients better than the other optimizers. This configuration is used to conduct a further examination of the overall strengths and weaknesses of this model.

4) Performance of the optimal configuration

The remainder of the analysis uses the optimal model configuration found in Sections 1 - 3, which is an LSTM layer with 2,000 nodes followed by two dense layers with 1,000 nodes, a learning rate of 0.0001, the AdamW optimizer, and 100 epochs of training. This configuration has a global MSE of 0.193, and the example in Fig. 2 demonstrates how it generally captures the profile relative to other configurations. The overall performance of this configuration is examined as a function of height, and two examples are used to highlight the performance.

Probability density functions (PDFs) of the temperature difference between predictions and observations are created every 200 m for each of the half-hour predictions (Fig. 3). As expected, the PDFs at all levels are initially narrow and broaden at longer prediction times. Higher heights have the narrowest distribution that broadens towards the surface. There is no abrupt shift around the end of the MT. This suggests that the model either adequately captures the rapid transition at the end of the MT, or the shift is simply smoothed out by the variability of MT durations. As prediction times increase, lower levels develop a slight cold bias, and higher levels develop a slight warm bias. This suggests a slight decrease in average predicted stability between the surface and 2,000 m.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

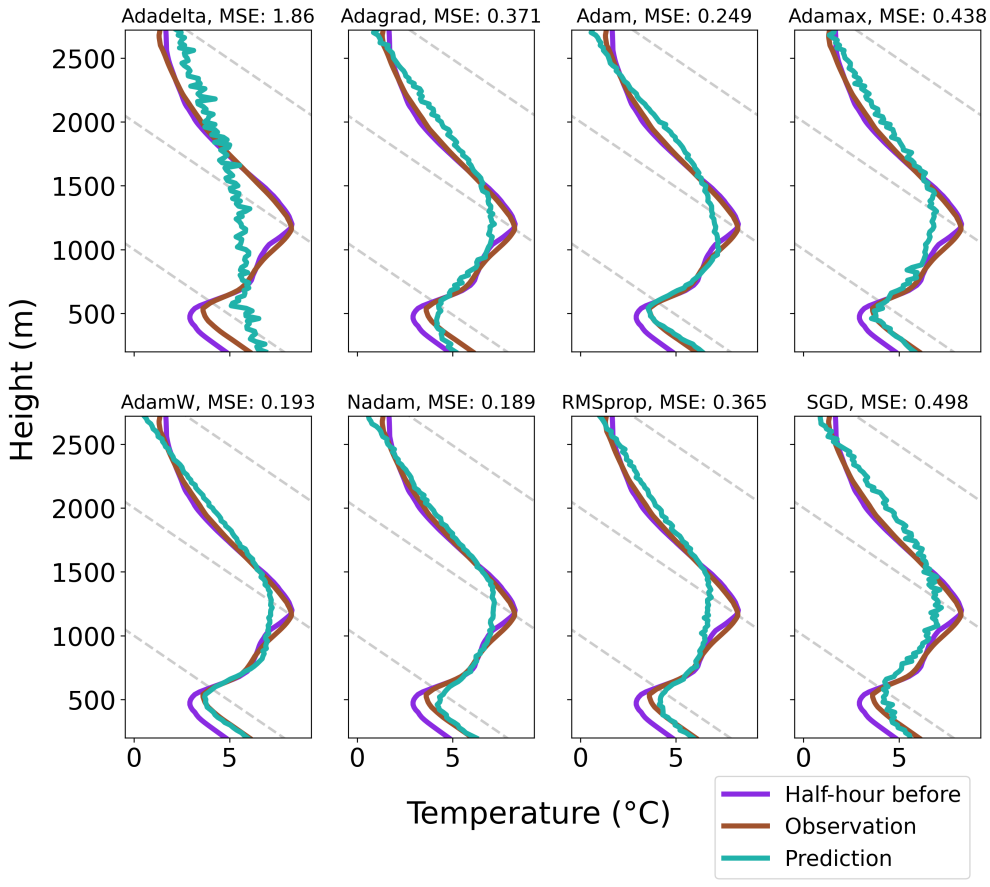


FIG. 2. Example predicted temperature profile (°C; cyan) on 30 January 2018 at 3 hours after sunrise for the eight optimizers tested. The observation at that time (maroon) and from half-hour before (purple) are also shown. The dashed gray lines represent the dry adiabatic lapse rate (9.8°C km^{-1}).

Evaluating individual profiles can yield more informative results than simply using the overall MSE of PDFs. The evolution of the first sample case is characterized by a mildly stable to well-mixed PBL (Fig. 4). The model is able to capture the shift from stable to isothermal to well-mixed, even though there is a complicated structure at sunrise. The model is about one timestep behind, starting at 2.5 hours, but does well in retaining the residual layer characteristics above 1,200 m. In the last timestep, the predicted lapse rate is consistent with observations below 700 m but with a 1°C cold bias. There is a 1°C warm bias at the base of the temperature inversion at 1,000 m. We note that using the Adam optimizer instead yields similar results but performs poorly with the lapse rate above the PBL (not shown). In this example, using the additional method to decay the weights with AdamW appears to primarily impact the top of the profile.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

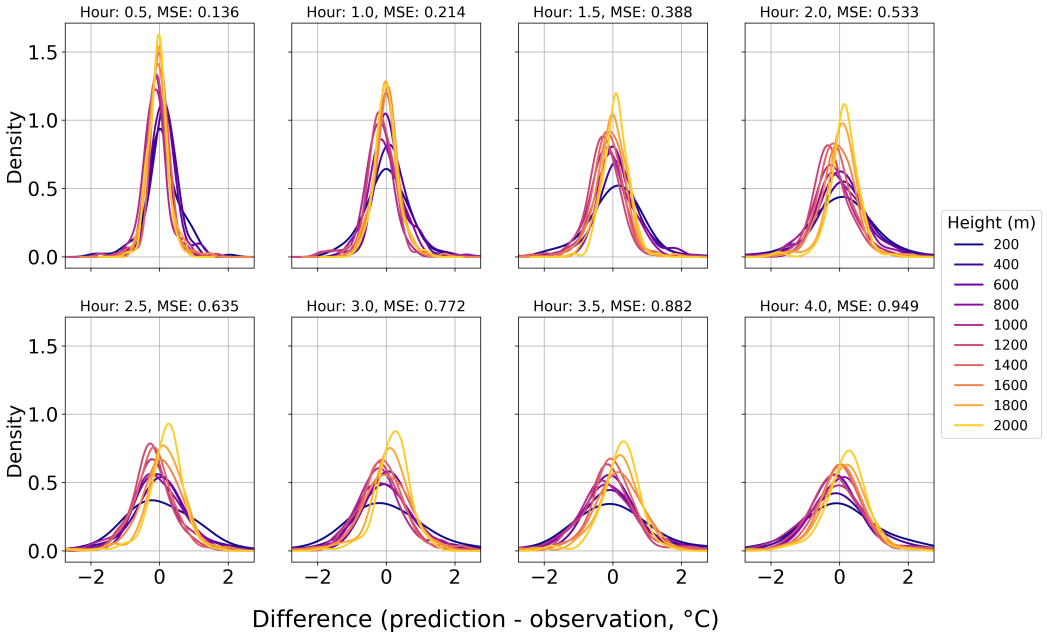


FIG. 3. Probability density functions (PDFs) of the difference between the prediction and observation (°C) as a function of hour then height (m). The PDF at each height is the average between 100 m below and 100 m above (e.g., the 1,000 m PDF is the average PDF from 900 m to 1,100 m). The average mean square error (MSE; °C) across all heights at that time is also given.

Well-mixed profiles do reasonably well with a slight bias in the PBL, or they struggle only near the top of the profile. A nocturnal surface inversion can exist over urban areas, and an example is shown in Fig. 5. Understanding and predicting the evolution of stable boundary layers remains a challenge (Lemone et al. 2018), especially compared to the growth of well-mixed boundary layers. The model captures the warming and transition from stable to isothermal during the first two hours. The model warms the near-surface temperature and reduces the temperature inversion too quickly, starting 2.5 hours after sunrise and leading to a warm bias of 1°C. This is the opposite problem of Fig. 4. The substantial differences at the end of the forecast period are not entirely unexpected, as this case starts with one of the most stable profiles. There are fewer stable cases to train on than well-mixed cases, and extreme cases tend not to be captured as well, which is a common limitation of machine learning.

The rapid weakening of the predicted inversion layer at the top of the mixed layer in Fig. 5 is consistent with other days. This is an inherent challenge for several reasons. As highlighted above, extreme features tend to be diminished. Deviations aloft can arise from external forcing,

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

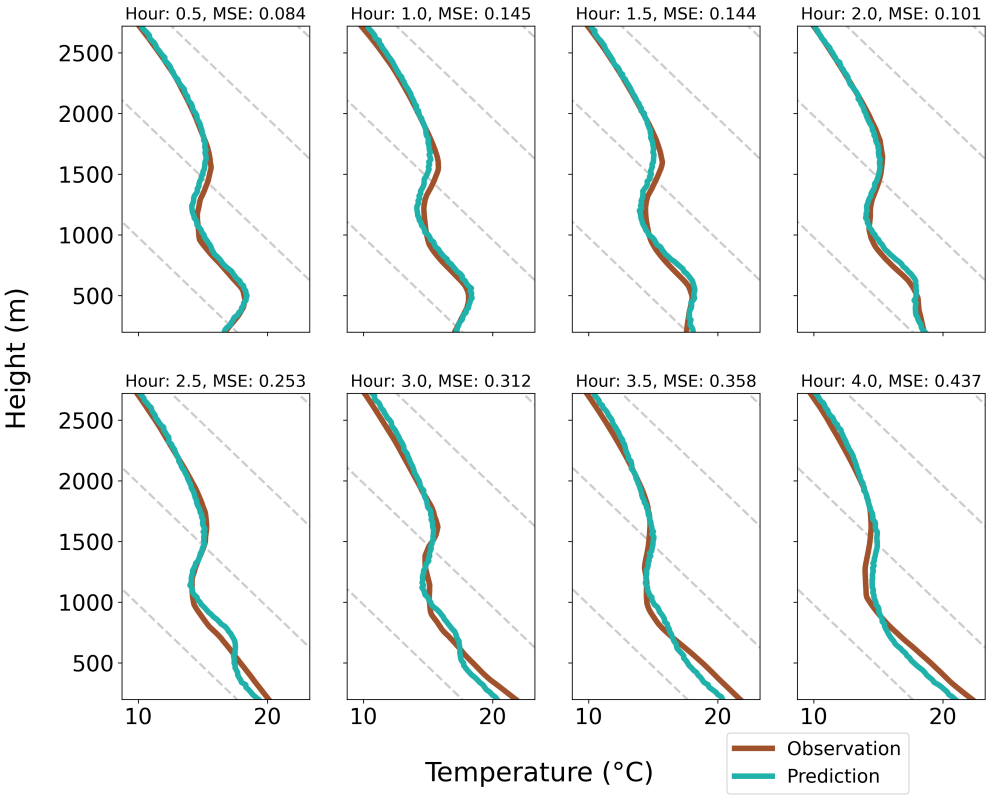


FIG. 4. An example evolution on 01 October 2022 of predicted temperature profiles (°C; cyan) in the testing dataset compared with observations (maroon). The dashed gray lines represent the dry adiabatic lapse rate (9.8°C km^{-1}).

such as the amount of subsidence. Fortunately, changes in the lapse rates aloft tend to be much slower since it is related to synoptic forcing, which is in contrast to rapid changes in the lower atmosphere driven by surface processes. The goal of this work is to focus on these rapid changes in the lower atmosphere using machine learning because of their inherent complexity, especially over cities. In most cases, the prediction only misses the MT by about half an hour and generally captures the approximate shape of the profile above the PBL. These are just two examples from the testing dataset that highlight the broader model performance. The bias across the entire dataset as a function of season is next examined to illustrate overall differences.

In the two examples shown, the model predictions eventually tend toward the annual means in that they underestimate the near-surface temperature and the onset of mixing when it is warmer (Fig. 4) and overestimate both when it is cooler (Fig. 5). This near-surface temperature difference is consistent with the bias as a function of height when separated by the predicted season and hour (Fig. 6a). The strongest signal occurs when the model has a mean warm bias of 0.8°C during

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

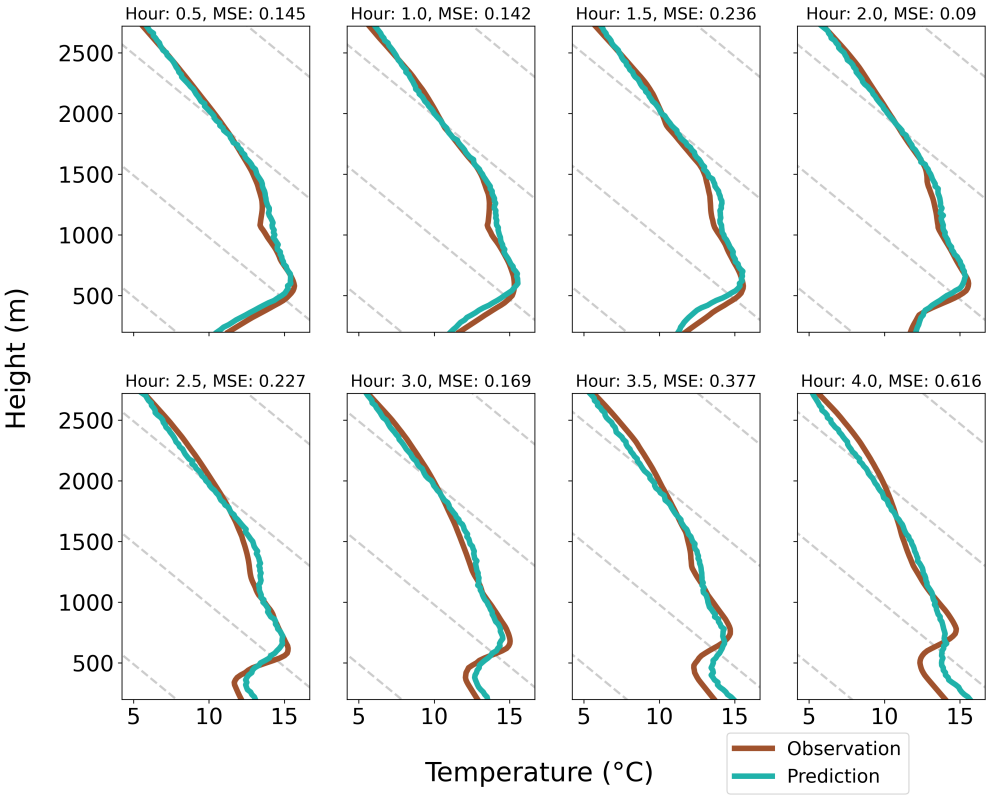


FIG. 5. An example evolution on 26 December 2019 of predicted temperature profiles (°C; cyan) in the testing dataset compared with observations (maroon). The dashed gray lines represent the dry adiabatic lapse rate (9.8°C km^{-1}).

winter at the end of the prediction sequence. This is also when the standard deviation reaches its highest value of 2.5°C (Fig. 6b). However, the model has a cold bias of less than 0.5°C below 1,000 m during the summer. This is because the model is not aggressive enough to capture seasonal extremes. The standard deviation is greater than 2°C below 500 m at 1.5 hours after sunrise but the depth of the relatively high standard deviation increases to 750 m by 4 hours after sunrise, which is mainly associated with a deeper PBL. The transition seasons, which are closer to the annual mean temperature profile, do not have as strong an effect, but there is a slight overestimation of the model below 1,200 m. The model also greatly underestimates the temperature above 2,000 m in winter. Making the model more complex (e.g., adding more nodes, increasing the number of layers, allowing it to train more) does not noticeably impact these results. Increasing complexity exponentially increases the computation time while only providing marginal improvements in MSE. Next, we investigate the possible improvement of including wind speed profiles as an additional

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

input, which could serve as a proxy for mechanically-generated turbulence that would influence the MT.

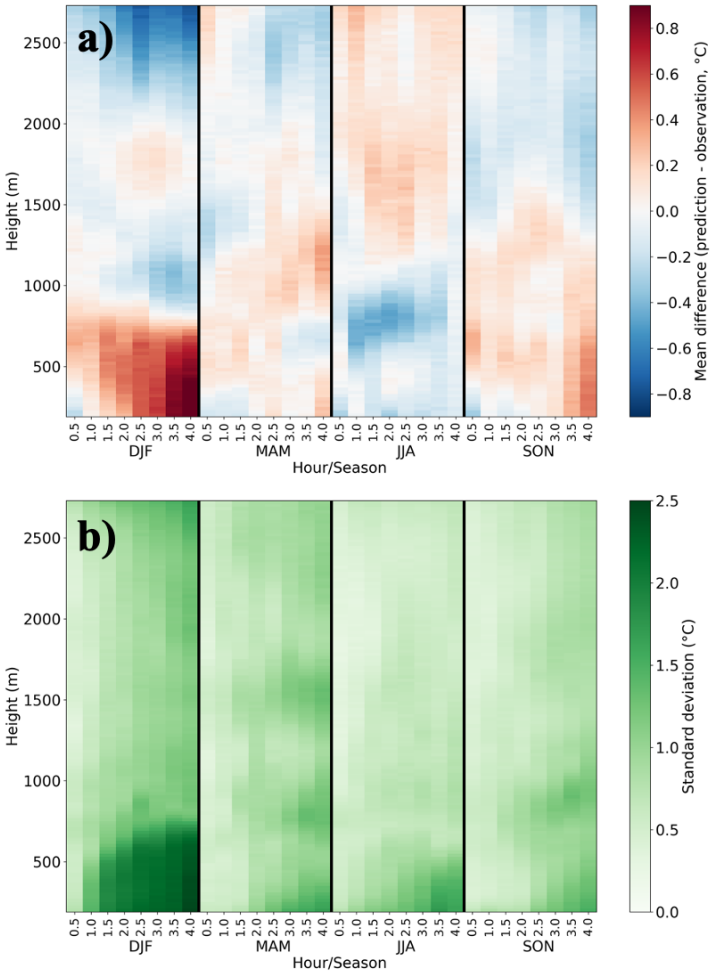


FIG. 6. The (a) mean (°C) and (b) standard deviation (°C) of the temperature difference between the prediction and observation as a function of height (m), separated by season then hour predicted.

b. Temperature and wind speed as inputs

AMDAR wind speed profiles are easily added to the model, but there is a smaller sample size due to missing data. There are 1,547 days with temperature data, but 1,188 days with both temperature and wind data. The subset of 1,188 days is used for all subsequent tests. The same basic model configuration and method described at the beginning of Section 4 is used. Three approaches are tested: (1) using temperature only (TO) as input, (2) using temperature and wind (TW) as input,

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

and (3) using temperature and wind as input, but with 200 epochs and twice the number of nodes in the LSTM and dense layers (i.e., 4,000 in the LSTM and 2,000 in the dense layers, TW2). Doubling the epochs and nodes is done to see if a deeper model is needed because more data is used.

Predictions on a day with an initially well-mixed PBL highlight the results from these three configurations (Fig. 7). All predictions are nearly indistinguishable for the first time step, besides the TO profile, which has a slightly more accurate near-surface temperature. The TO profile relaxes the lapse rate of the inversion too quickly but performs best above 1,200 m, resulting in a noticeably lower MSE than the other two scenarios. By hour 3.5, the TW profile better captures the minimum and maximum of the inversion, while the TO profile predicts too large a temperature inversion. The mean MSE for this sequence is 0.29°C for TO, 0.50°C for TW, and 0.41°C for TW2. This case highlights the importance of analyzing individual profiles, as each model has its strengths and weaknesses throughout the profile that the mean MSE alone cannot represent.

Seasonal plots demonstrate the general differences between TO and TW (Fig. 8). Including wind speed reduces bias throughout the entire profile up to 0.4°C in all seasons and times, especially near the surface in winter and above 1,000 m in summer. Spring and fall have the same location of warm and cold biases, but are lower when including the wind. The warm bias is about 0.1°C weaker but 400 m higher in the atmosphere when including the winds. The bias has less of an abrupt increase after 1 hour after sunrise when the winds are included. Including the winds helps predict the change in convective state during the MT, leading to less bias throughout the rest of the sequence. In general, the addition of wind speed reduces the MSE by better capturing the minima and maxima throughout the profile.

TW is often better at capturing the evolution of the PBL than TO. TW better captures the base and top of the inversion layer above the PBL, whereas TO is often too warm at the base of the inversion. TW often performs better by reducing the greater MSE throughout the profile by the end of the prediction window, with near-surface temperature in the summer, and at the top of the profile. Less improvement occurs when increasing the number of nodes or training time. This demonstrates that changing the inputs plays a larger role than making the model more complex. This is consistent with the findings above, where changing the number of nodes in each layer or increasing the number of epochs only marginally changes the performance compared to altering the

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

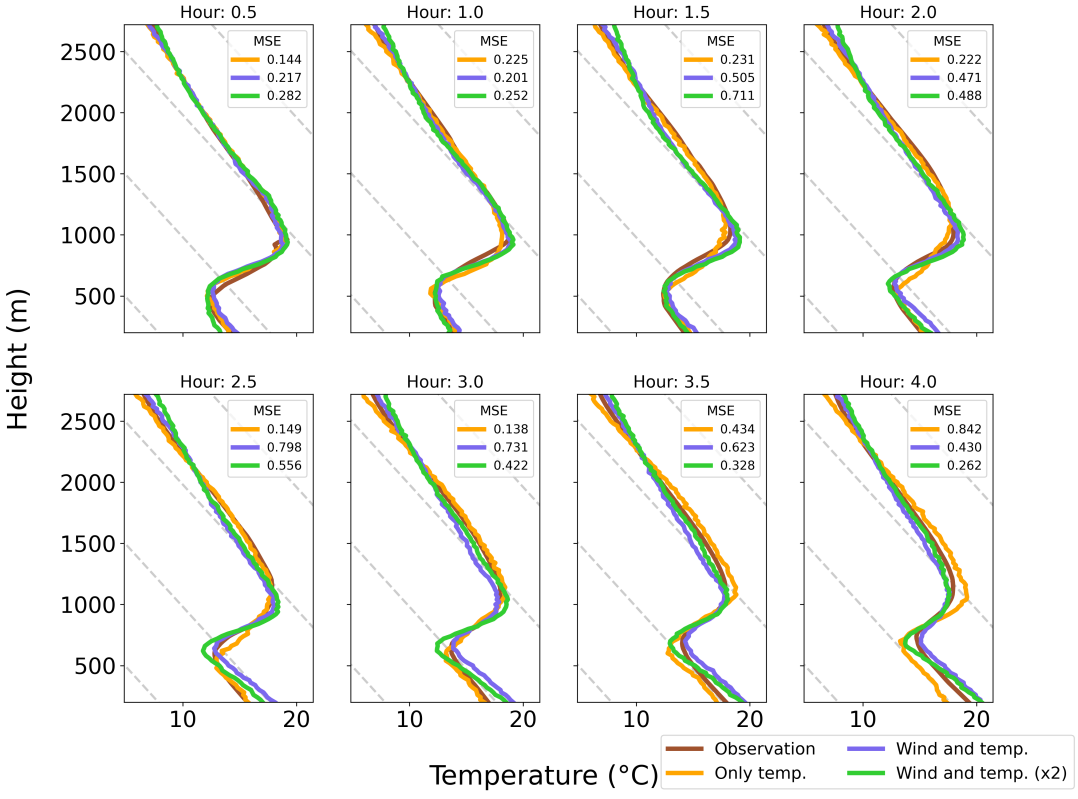


FIG. 7. An example evolution on 23 December 2021 of predicted temperature (°C) profiles in the testing dataset using only temperature as an input (TO; orange), wind and temperature as inputs (TW; purple), and wind and temperature as inputs but with double the number of epochs and nodes (TW2; green) compared with observations (maroon). The dashed gray lines represent the dry adiabatic lapse rate (9.8°C km^{-1}).

configuration (Table 2). The overall MSE across the testing dataset is 0.56°C for TO, 0.62°C for TW, and 0.59°C for TW2. The higher MSE for TW is likely due to a slightly larger bias throughout the profile on average. However, a t-test does not show a statistically significant difference between TO and TW (p-value 0.92) or the MSE between them (p-value 0.57). Although including winds better captures the characteristics of the PBL (e.g., closer to predicting the minimum and maximum temperatures of inversions at the top of the PBL), there is no statistically significant increase in global performance using MSE by including the winds.

Increasing complexity and training time could yield additional improvement, but such changes marginally change the overall MSE and the errors are at the same general height. Using only temperature as input, the MSE changes from 0.56°C to 0.54°C after using five times the number of

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

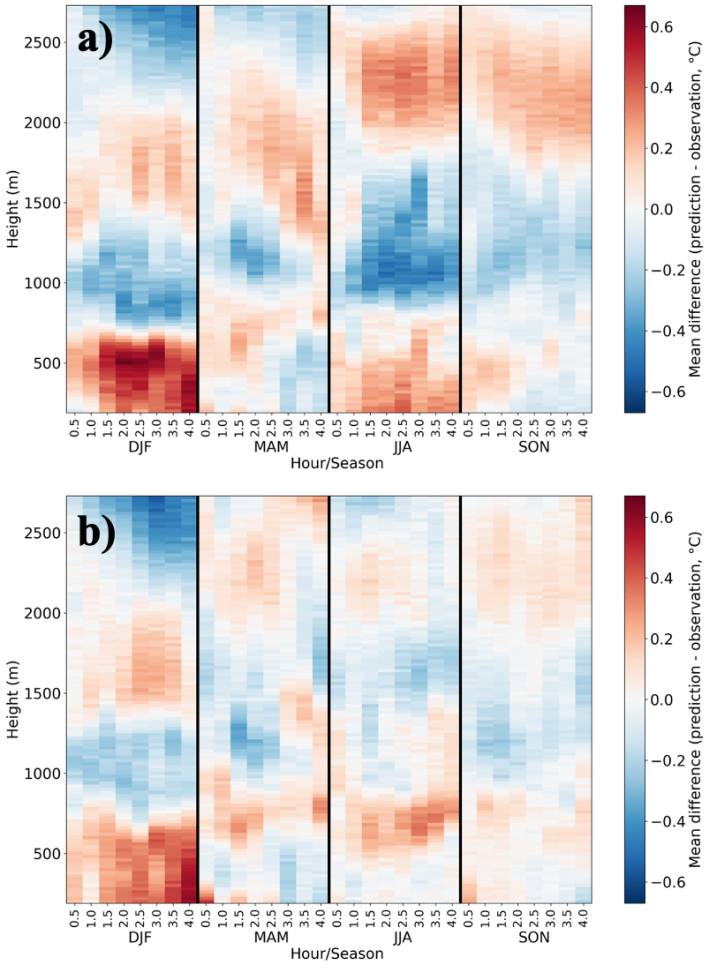


FIG. 8. The mean temperature difference (°C) between the prediction and observation as a function of height (m), separated by season then hour predicted for (a) only using temperature as the input (TO), and (b) using both temperature and wind speed as inputs (TW).

epochs and nodes in each layer and including an additional layer. Doing the same with temperature and wind speed as input increases the MSE from 0.62°C to 0.65°C. The larger error is likely due to either overfitting or the model performing too well on the training dataset and is therefore less generalizable to the validation and testing datasets. These small changes suggest that the model is not gaining any additional nuances from the training dataset. Altering the learning rate does not noticeably change these conclusions either. However, having a larger dataset as the number of AMDAR observations increase over time is still likely to be beneficial.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/16/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

4. Discussion and Conclusions

The PBL is a vital component of our atmosphere, especially over cities because it impacts the dispersion of pollution and the urban heat island. Machine learning is increasingly being used to understand and predict the PBL, however, most work examines only the PBLH, which is sensitive to how the PBLH is calculated and does not capture other important features like stability. Capturing the evolution of the nocturnal surface inversion through the MT is important but often difficult, particularly in numerical weather prediction. We demonstrate the potential use of machine learning approaches to short-term predictions of the entire temperature profile below 3,000 m, but emphasize that the strengths and limitations of using an empirical machine learning approach for any related application should be considered and identified.

Machine learning can be a powerful tool and we demonstrate that it can perform well, but there are several fundamental limitations when applied to temperature profiles in the lower atmosphere. First, data availability and quality for machine learning is a concern for any application, but especially for the urban PBL, where measurements are often infrequent and sparse (Grimmond 2006; Barlow 2014; Li et al. 2018; Helbig et al. 2021). However, infrequent observations are overcome here using high-resolution half-hourly profiles of temperature and wind speed from AMDAR (Nielsen and Rahn 2022). Limited work in this area means that there is no single type or configuration of a machine learning model that is already optimized to accomplish this task. This is true for any application that uses empirical relationships with limited data. This approach should generalize to other regions with appropriate adaptation; however, retraining using local profile data is critical. For example, if temperature profiles from the coast or from a location near complex terrain are used to train the evolution of the lower atmosphere, then the training would pick up on diurnal phenomena such as a sea breeze or upslope flow that would not exist over a flat location away from water.

Finding the optimal version of any machine learning model is a daunting task given the variety of base models and optimizers, as well as large degrees of freedom for aspects such as the number of nodes or layers. Haynes et al. (2024) found that increasing the number of nodes can reduce the training error, but there is an inflection point where the predicted profiles start to become noisier. They also concluded that linear regression models can sometimes perform better than fully connected neural networks based on the data being trained on. Since many components of the

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

model can be altered, it is difficult to objectively determine the optimal configuration based solely on model error due to confounding effects. Predicting the same state of the atmosphere using different datasets may lead to different optimal configurations. To remove the sensitivity of the optimal configuration to the dataset being trained on, it is necessary to interrogate each prediction to determine how well it captures important features of the PBL. If not, the conclusions would be too specific and not generalizable to future studies.

Defining metrics that quantify the optimal version of a machine learning model to predict a temperature profile is also problematic because a single test statistic, such as MSE, does not necessarily capture important features, such as the strength and location of temperature inversions. Machine learning models minimize MSE without necessarily considering the retention of these important features. An advantage of operational numerical weather prediction is that the parameterizations are driven by some representation of physical processes, although the way that the parameterizations are formulated can lead to large sensitivity of the predicted temperature profile, especially for stable boundary layers or over complex land surfaces such as urban areas.

It is beneficial to examine both standard summary statistics and individual profiles to assess more nuanced aspects of performance as they relate to the intended application. Striking a balance between the physical features picked up in each profile and global metrics such as the MSE is a more complete way to determine which performs the best. Balancing these components highlights the usefulness of LSTMs given a sequential dataset and the benefit of having an LSTM layer coming before deep layers to better retain time-dependent information. The selected optimizer also impacts the retention of features in the profile where some, such as AdamW, are more aggressive in capturing a wider range of stability conditions (Fig. 2). Simpler models, such as the LNN, produced a more linear profile that sometimes reduced the MSE but did not capture extrema throughout the profile (Fig. 1). This once again underscores that training towards a minimum MSE can produce soundings that have globally low MSE, but can return a more linear profile that fails to capture important features such as strong temperature gradients.

Another concern is how the machine learning models trend towards the mean for extreme conditions with respect to both individual profiles and seasonal means. There is a bias towards the annual mean such that in the winter (summer), the model warms the near-surface temperature too quickly (slowly), causing the predicted profile to be too warm (cool). This is consistent with Pasche

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.
Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

et al. (2025), who found that several machine learning models are too cold when predicting the heat index in heat wave events and too warm when predicting wind chills in a winter storm. Machine learning models cannot respond quickly or with sufficient intensity to approaching extreme events as a physics-based model would.

Despite limitations, the LSTM predictions reasonably capture the evolution of the temperature profile during the MT on most days. However, machine learning methods are at a clear disadvantage when transient weather occurs, such as the passage of a cold front. It is not possible to capture this without including additional input. More subtle influences on the growth of the PBL are also not directly present. This includes enhanced synoptic subsidence aloft, which may evolve relatively slowly but is indirectly reflected in the initial lapse rates aloft. In contrast, this approach performs well if the goal is to predict the evolution of the temperature profile during the MT to better anticipate the surface pollution concentration. High pollution often occurs in the morning when stability is high during clear, quiescent conditions. With rapidly evolving systems, such as mid-latitude systems, temperature profiles change quickly, but pollution is generally low given the low-level mixing and precipitation scavenging. Therefore, predictions are best when conditions are most conducive to high pollution in the morning, and predictions are worst when pollution is most likely to be low anyway. Past work has predicted vertical profiles of the entire atmosphere using machine learning (Haynes et al. 2024), but a high vertical and spatial resolution of the PBL is necessary to capture important details near the surface.

Strengths and weaknesses are easier to understand by sifting through all the outputs of each machine learning model. Even when the predicted profile has a bias from the actual profile, the lapse rates are still physically realistic. This would not be discovered if the individual profiles were not inspected or otherwise queried. These results are consistent with other studies where the machine learning model tends toward the mean or a base state (geostrophic), when the models need to be reworked to emphasize extremes, or when the input data needs to be subset to test the detection of extreme events (Bonavita 2024; Lopez-Gomez et al. 2023; Nevo et al. 2022). The public tends to focus on the ability of a model solely on its performance with extreme events, which makes it a challenge to create a skillful model for all conditions (Lerch et al. 2017).

All too often, the typical approach with machine learning is to give a model all available data and let it find the dominant patterns. However, this makes it harder to analyze the inputs, outputs,

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

and the connection between the two to approximate how the model is working. In this work, the inclusion of wind speed profiles as an additional input did not significantly improve the global MSE of the LSTM (Figs. 7 and 8). However, the predicted profiles that included wind speed appeared to better capture the minimum and maximum temperature for temperature inversions at the top of the PBL. If compute time is a concern for near-real-time forecasts, wind could be omitted at the cost of slightly lower accuracy. Increasing the complexity of the model may reduce the error, but it often plateaus at the expense of much longer computational times.

To fully appreciate the strengths of machine learning, the limitations must also be held with equal priority. Applying a one-size-fits-all approach to a problem with machine learning greatly limits the usefulness of the tool. The same can be said for only finding the lowest MSE without any other assessment of the output, which could lead to unsatisfactory results.

Acknowledgments. This work was supported by the Madison and Lila Self Graduate Fellowship at the University of Kansas and by the high performance computing facilities operated by the Center for Research Computing at the University of Kansas. We thank the reviewers for their constructive comments.

Data availability statement. Observations from surface stations and aircraft that are older than 48 h are publicly available on the Meteorological Assimilation Data Ingest System (MADIS; https://madis-data.ncep.noaa.gov). Data are available for real-time use by federal agencies, including NOAA.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

APPENDIX

A1. Acronyms

Acronym	Full name
Adagrad	Adaptive gradient algorithm
Adam	Adaptive moment estimation
AdamW	Adam with decoupled weight decay
AMDAR	Aircraft Meteorological Data Relay
CNN	Convolutional neural network
DAL	Dallas Love Field Airport
DFW	Dallas-Fort Worth International Airport
DNN	Dense neural network
LNN	Linear neural network
LSTM	Long short-term memory network
Nadam	Adam with Nesterov momentum
MSE	Mean square error
MT	Morning transition
NOAA	National Oceanic and Atmospheric Administration
PBL	Planetary boundary layer
PBLH	Planetary boundary layer height
PDF	Probability density function
RMSprop	Root mean square propagation
RNN	Recurrent neural network
SGD	Stochastic gradient descent
TO	Temperature only used as the input
TW	Temperature and wind speed as inputs
TW2	TW with double the epochs and number of nodes in the LSTM and dense layers

TABLE A1. List of acronyms and abbreviations.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

A2. Optimizers tested

Optimizer	Description	Benefit
SGD	Finds the gradient of the cost function with the goal of finding the optimal solution. This is done stochastically by calculating the gradient at a single training example and updating the model's parameters accordingly.	Desired for large datasets because it does not require iterating through the entire dataset.
Adagrad (Duchi et al. 2012)	Designed to adapt the subgradient methods to the geometry of the problem at hand. Utilizes diagonal approximation for the matrix by using the outer products of subgradients computed in the algorithm.	Works well with both sparse and dense solutions. Has parameter-specific learning rates.
Adadelta (Zeiler 2012)	A first order refined version of Adagrad that uses a moving window of gradient updates instead of using all previous gradients.	Less computational demands compared to SGD and has a per-dimension learning rate so it is less of a need to manually specify the learning rate ahead of time. Hyperparameters do not need to be tuned.
RMSprop	Uses momentum to maintain the direction of the moving average of the gradients, which are then used to approximate the variance.	Performs better than SGD because of the adaptive learning rates, which make it more stable while converging faster.
Adam (Kingma and Lei Ba 2015)	Similar to SGD but uses adaptive estimation of the first-order and second-order moments. Combines how Adagrad can work with sparse gradients and how RMSprop can work with non-stationary objects.	Computationally efficient, does not require a lot of memory, the hyperparameters do not require a lot of tuning, and is optimized for large datasets or those with high-dimensional parameter spaces.
Adamax (Kingma and Lei Ba 2015)	Similar to Adam but with a parameter for the exponentially weighted infinity norm, which is updated based on the maximum between the second exponential decay rate multiplied by the previous exponentially weighted infinity norm and the gradients.	Designed to work with sequential datasets.
AdamW (Loshchilov and Hutter 2019)	Similar to Adam but with another method to decay the weights to perform better than the Adam algorithm with L_2 regularization.	Often performs better than Adam because of its more separable hyperparameter search space and it decouples the weight decay and learning rate. Compared to Adam, AdamW has lower training loss and better generalizability performance.
Nadam (Dozat 2016)	Similar to Adam but with the addition of Nesterov momentum, which includes a velocity term to update the gradient.	Often converges slightly faster than Adam and is more stable. Better to use than using RMSprop with momentum to train the model.

TABLE A2. Descriptions and benefits of the eight optimizers tested.

References

Alibrahim, H., and S. A. Ludwig, 2021: Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. *2021 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 1551–1559, https://doi.org/10.1109/CEC45853.2021.9504761, URL https://ieeexplore.ieee.org/document/9504761/.

Angevine, W. M., H. K. Baltink, and F. C. Bosveld, 2001: Observations of the morning transition of the convective boundary layer. *Boundary-Layer Meteorology*, **101 (2)**, 209–227, https://doi.org/

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.
Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

10.1023/A:1019264716195.

Angevine, W. M., J. M. Edwards, M. Lothon, M. A. LeMone, and S. R. Osborne, 2020: Transition periods in the diurnally-varying atmospheric boundary layer over land. *Boundary-Layer Meteorology*, **177**, 205–223, https://doi.org/10.1007/s10546-020-00515-y.

Ayazpour, Z., S. Tao, D. Li, A. J. Scarino, R. E. Kuehn, and K. Sun, 2023: Estimates of the spatially complete, observational-data-driven planetary boundary layer height over the contiguous united states. *Atmospheric Measurement Techniques*, **16 (2)**, 563–580, https://doi.org/10.5194/amt-16-563-2023.

Barlow, J. F., 2014: Progress in observing and modelling the urban boundary layer. *Urban Climate*, **10 (P2)**, 216–240, https://doi.org/10.1016/j.uclim.2014.03.011.

Bonavita, M., 2024: On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, **51**, https://doi.org/10.1029/2023GL107377.

Brunekreef, B., and S. T. Holgate, 2002: Air pollution and health. *The Lancet*, **360 (9341)**, 1233–1242, https://doi.org/10.1016/S0140-6736(02)11274-8.

Cosgrove, A., and M. Berkelhammer, 2018: Downwind footprint of an urban heat island on air and lake temperatures. *npj Climate and Atmospheric Science*, **1 (1)**, 1–10, https://doi.org/10.1038/s41612-018-0055-3.

Dai, C., Q. Wang, J. A. Kalogiros, D. H. Lenschow, Z. Gao, and M. Zhou, 2014: Determining Boundary-Layer Height from Aircraft Measurements. *Boundary-Layer Meteorology*, **152 (3)**, 277–302, https://doi.org/10.1007/s10546-014-9929-z.

Dozat, T., 2016: Incorporating Nesterov Momentum into Adam. *International Conference on Learning Representations Workshop*, **(1)**, 2013–2016.

Drüe, C., T. Hauf, and A. Hoff, 2010: Comparison of boundary-layer profiles and layer detection by AMDAR and WTR/RASS at Frankfurt airport. *Boundary-Layer Meteorology*, **135 (3)**, 407–432, https://doi.org/10.1007/s10546-010-9485-0.

Brought to you by | UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 04/14/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Duchi, J. C., P. L. Bartlett, and M. J. Wainwright, 2012: Randomized smoothing for (parallel) stochastic optimization. *Proceedings of the Institute of Electrical and Electronics Engineers Conference on Decision and Control*, **12**, 5442–5444, https://doi.org/10.1109/CDC.2012.6426698.

Fan, H., Y. Liu, Y. Li, Y. Liu, J. Duan, L. Li, and Z. Huo, 2023: A deep learning method for predicting lower troposphere temperature using surface reanalysis. *Atmospheric Research*, **283**, https://doi.org/10.1016/j.atmosres.2022.106542.

Friedman, E., B. R. Lee, D. Rahn, B. L. Martinez, and A. Mena, 2024: Assessing environmental injustice in kansas city by linking paediatric asthma to local sources of pollution: a cross-sectional study. *BMJ open*, **14 (7)**, e080 915.

Fujita, M., N. Sugiura, and S. Kouketsu, 2024: Prediction of atmospheric profiles with machine learning using the signature method. *Geophysical Research Letters*, **51 (6)**, e2023GL106 403.

Grimmond, C. S., 2006: Progress in measuring and observing the urban atmosphere. *Theoretical and Applied Climatology*, **84 (1-3)**, 3–22, https://doi.org/10.1007/s00704-005-0140-5.

Guan, W.-J., X.-Y. Zheng, K. F. Chung, and N.-S. Zhong, 2016: Impact of air pollution on the burden of chronic respiratory diseases in china: time for urgent action. *The Lancet*, **388**, 1939–1951, https://doi.org/10.1016/S0140-6736(16)31597-5.

Halios, C. H., and J. F. Barlow, 2018: Observations of the Morning Development of the Urban Boundary Layer Over London, UK, Taken During the ACTUAL Project. *Boundary-Layer Meteorology*, **166 (3)**, 395–422, https://doi.org/10.1007/s10546-017-0300-z.

Haynes, K., J. Stock, J. Dostalek, C. Anderson, and I. Ebert-Uphoff, 2024: Exploring the use of machine learning to improve vertical profiles of temperature and moisture. *Artificial Intelligence for the Earth Systems*, **3**, https://doi.org/10.1175/AIES-D-22-0090.1.

Helbig, M., and Coauthors, 2021: Integrating continuous atmospheric boundary layer and tower-based flux measurements to advance understanding of land-atmosphere interactions. *Agricultural and Forest Meteorology*, **307**, 108 509, https://doi.org/https://doi.org/10.1016/j.agrformet.2021.108509.

Hochreiter, S., and J. Schmidhuber, 1997: Long Short-Term Memory. *Neural Computation*, **9 (8)**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/19/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Hu, X. M., J. W. Nielsen-Gammon, and F. Zhang, 2010: Evaluation of three planetary boundary layer schemes in the wrf model. *Journal of Applied Meteorology and Climatology*, **49**, 1831–1844, https://doi.org/10.1175/2010JAMC2432.1.

Kim, D., Z. Chen, L. F. Zhou, and S. X. Huang, 2018: Air pollutants and early origins of respiratory diseases. *Chronic Diseases and Translational Medicine*, **4**, 75–94, https://doi.org/10.1016/j.cdtm.2018.03.003.

Kingma, D. P., and J. Lei Ba, 2015: Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 1–15, https://doi.org/https://doi.org/10.48550/arXiv.1412.6980.

Kotthaus, S., and C. S. B. Grimmond, 2018: Atmospheric boundary-layer characteristics from ceilometer measurements. Part 2: Application to London's urban boundary layer. *Quarterly Journal of the Royal Meteorological Society*, **144 (714)**, 1511–1524, https://doi.org/10.1002/qj.3298.

Krishnamurthy, R., R. K. Newsom, L. K. Berg, H. Xiao, P. L. Ma, and D. D. Turner, 2021: On the estimation of boundary layer heights: A machine learning approach. *Atmospheric Measurement Techniques*, **14**, 4403–4424, https://doi.org/10.5194/amt-14-4403-2021.

Kumar, N., K. Soni, and R. Agarwal, 2021: Prediction of temporal atmospheric boundary layer height using long short-term memory network. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **73 (1)**, 1–14, https://doi.org/10.1080/16000870.2021.1926132.

Lapworth, A., 2006: The morning transition of the nocturnal boundary layer. *Boundary-Layer Meteorology*, **119 (3)**, 501–526, https://doi.org/10.1007/s10546-005-9046-0.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Lemone, M. A., and Coauthors, 2018: 100 years of progress in boundary layer meteorology. *Meteorological Monographs*, **59**, 9.1–9.85, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0013.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/06/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, **32**, 106–127, https://doi.org/10.1214/16-STS588.

Li, H., M. Wolter, X. Wang, and S. Sodoudi, 2018: Impact of land cover data on the simulation of urban heat island for berlin using wrf coupled with bulk approach of noah-lsm. *Theoretical and Applied Climatology*, **134**, 67–81, https://doi.org/10.1007/s00704-017-2253-z.

Liashchynskyi, P., and P. Liashchynskyi, 2019: Grid search, random search, genetic algorithm: A big comparison for nas. URL http://arxiv.org/abs/1912.06059.

Lopez-Gomez, I., A. McGovern, S. Agrawal, and J. Hickey, 2023: Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, **2**, https://doi.org/10.1175/AIES-D-22-0035.1.

Loshchilov, I., and F. Hutter, 2019: Decoupled weight decay regularization. *7th International Conference on Learning Representations*, 1711.05101.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artificial Intelligence for the Earth Systems*, **2**, https://doi.org/10.1175/aies-d-22-0077.1.

Molero, F., R. Barragán, and B. Artíñano, 2022: Estimation of the atmospheric boundary layer height by means of machine learning techniques using ground-level meteorological data. *Atmospheric Research*, **279**, https://doi.org/10.1016/j.atmosres.2022.106401.

Molina, M. J., and Coauthors, 2023: A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artificial Intelligence for the Earth Systems*, **2**, https://doi.org/10.1175/aies-d-22-0086.1.

Nevo, S., and Coauthors, 2022: Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, **26**, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022.

Nielsen, K. F., and D. A. Rahn, 2022: Morning Transition of the Boundary Layer over Dallas–Fort Worth. *Journal of Applied Meteorology and Climatology*, **61 (10)**, 1433–1448, https://doi.org/10.1175/JAMC-D-21-0169.1.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/10/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Nunez, M., and T. R. Oke, 1977: The energy balance of an urban canyon. *Journal of Applied Meteorology and Climatology*, **16 (1)**, 11–19, https://doi.org/10.1175/1520-0450(1977)016⟨0011:TEBOAU⟩2.0.CO;2.

Oke, T. R., 1987: *Boundary Layer Climates*. 2nd ed., Routledge: Taylor Francis Group, 288–294 pp.

Pasche, O. C., J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, 2025: Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, **4**, https://doi.org/10.1175/AIES-D-24-0033.1.

Peng, K., and Coauthors, 2023: Machine learning model to accurately estimate the planetary boundary layer height of beijing urban area with era5 data. *Atmospheric Research*, **293**, https://doi.org/10.1016/j.atmosres.2023.106925.

Petersen, R. A., 2016: On the Impact and Benefits of AMDAR Observations in Operational Forecasting—Part I: A Review of the Impact of Automated Aircraft Wind and Temperature Reports. *Bulletin of the American Meteorological Society*, **97 (4)**, 585–602, https://doi.org/10.1175/BAMS-D-14-00055.1.

Rahn, D. A., and C. J. Mitchell, 2016: Diurnal climatology of the boundary layer in Southern California using AMDAR temperature and wind profiles. *Journal of Applied Meteorology and Climatology*, **55 (5)**, 1123–1137, https://doi.org/10.1175/JAMC-D-15-0234.1.

Ryu, Y.-H., and J.-J. Baik, 2012: Quantitative analysis of factors contributing to urban heat island intensity. *Journal of Applied Meteorology and Climatology*, **51 (5)**, 842–854.

Schrijvers, P. J., H. J. Jonker, S. Kenjereš, and S. R. de Roode, 2015: Breakdown of the night time urban heat island energy budget. *Building and Environment*, **83**, 50–64, https://doi.org/10.1016/j.buildenv.2014.08.012.

Seidel, D. J., Y. Zhang, A. Beljaars, J. C. Golaz, A. R. Jacobson, and B. Medeiros, 2012: Climatology of the planetary boundary layer over the continental United States and Europe. *Journal of Geophysical Research Atmospheres*, **117 (17)**, 1–15, https://doi.org/10.1029/2012JD018143.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/10/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research wrf version 2. Tech. rep., National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div.

Vickers, D., and L. Mahrt, 2004: Evaluating formulations of stable boundary layer height. *Journal of applied meteorology*, **43 (11)**, 1736–1749, https://doi.org/10.1175/JAM2160.1.

World Meteorological Organization, 2014: The Benefits of AMDAR Data to Meteorology and Aviation. Tech. rep., World Meteorological Organization.

World Meteorological Organization, 2015: Impact and Benefits of AMDAR Temperature, Wind and Moisture Observations in Operational Weather Forecasting. Tech. rep., World Meteorological Organization.

World Meteorological Organization, 2017: Guide to Aircraft-based Observations. Tech. Rep. 1200, World Meteorological Organization.

World Meteorological Organization, 2018: WIGOS, WMO Integrated Global Observing System - Benefits to the Environment and Society from the Availability and Use of AMDAR Data. Tech. rep., World Meteorological Organization.

Zeiler, M. D., 2012: ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints*, 1212.5701.

Zhang, Y., Z. Gao, D. Li, Y. Li, N. Zhang, X. Zhao, and J. Chen, 2014: On the computation of planetary boundary-layer height using the bulk Richardson number method. *Geoscientific Model Development*, **7**, 2599–2611, https://doi.org/10.5194/gmd-7-2599-2014.

Zhang, Y., D. Li, Z. Lin, J. A. Santanello, and Z. Gao, 2019: Development and Evaluation of a Long-Term Data Record of Planetary Boundary Layer Profiles From Aircraft Meteorological Reports. *Journal of Geophysical Research: Atmospheres*, **124 (4)**, 2008–2030, https://doi.org/10.1029/2018JD029529.

Zhang, Y., K. Sun, Z. Gao, Z. Pan, M. A. Shook, and D. Li, 2020: Diurnal Climatology of Planetary Boundary Layer Height Over the Contiguous United States Derived From AMDAR and Reanalysis Data. *Journal of Geophysical Research: Atmospheres*, **125 (20)**, 1–21, https://doi.org/10.1029/2020JD032803.

Brought to you by UNIVERSITY OF KANSAS | Unauthenticated | Downloaded 11/10/26 03:34 PM UTC

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-26-0049.1.