Boundary-Layer Meteorology (2022) 185:199–228
https://doi.org/10.1007/s10546-022-00727-4

RESEARCH ARTICLE

Machine Learning for Improving Surface-Layer-Flux Estimates

Tyler McCandless[1] · David John Gagne[2] · Branko Kosović[2] · Sue Ellen Haupt[2] · Bai Yang[3,4] · Charlie Becker[2] · John Schreck[2]

Received: 23 August 2021 / Accepted: 8 June 2022 / Published online: 13 September 2022

Abstract
Flows in the atmospheric boundary layer are turbulent, characterized by a large Reynolds number, the existence of a roughness sublayer and the absence of a well-defined viscous layer. Exchanges with the surface are therefore dominated by turbulent fluxes. In numerical models for atmospheric flows, turbulent fluxes must be specified at the surface; however, surface fluxes are not known a priori and therefore must be parametrized. Atmospheric flow models, including global circulation, limited area models, and large-eddy simulation, employ Monin–Obukhov similarity theory (MOST) to parametrize surface fluxes. The MOST approach is a semi-empirical formulation that accounts for atmospheric stability effects through universal stability functions. The stability functions are determined based on limited observations using simple regression as a function of the non-dimensional stability parameter representing a ratio of distance from the surface and the Obukhov length scale (Obukhov in Trudy Inst Theor Geofiz AN SSSR 1:95–115, 1946), z/L. However, simple regression cannot capture the relationship between governing parameters and surface-layer structure under the wide range of conditions to which MOST is commonly applied. We therefore develop, train, and test two machine-learning models, an artificial neural network (ANN) and random forest (RF), to estimate surface fluxes of momentum, sensible heat, and moisture based on surface and near-surface observations. To train and test these machine-learning algorithms, we use several years of observations from the Cabauw mast in the Netherlands and from the National Oceanic and Atmospheric Administration's Field Research Division tower in Idaho. The RF and ANN models outperform MOST. Even when we train the RF and ANN on one set of data and apply them to the second set, they provide more accurate estimates of all of the fluxes compared to MOST. Estimates of sensible heat and moisture fluxes are significantly improved, and model interpretability techniques highlight the logical physical relationships we expect in surface-layer processes.

✉ Branko Kosović
branko@ucar.edu

1 E-Source, Boulder, CO, USA

2 National Center for Atmospheric Research, Boulder, CO 80303, USA

3 Earth Resource Technology, Laurel, MD 20707, USA

4 Air Resources Laboratory, National Oceanic and Atmospheric Administration, Idaho Falls, ID 83402, USA

Keywords Artificial neural networks · Machine learning · Monin–Obukhov similarity theory · Random Forest · Surface layer

1 Introduction

Flows in the atmospheric boundary layer (ABL) are turbulent, characterized by a large Reynolds number (Re), the existence of a roughness sublayer, and the absence of a well-defined viscous layer. The exchange of momentum, heat, and constituents between the land surface and atmosphere is mediated by the surface layer of an ABL. The surface layer is commonly considered to span approximately the lower 10% of an ABL in contact with the land surface. Exchanges between the land surface and the atmosphere through the surface layer are dominated by turbulent fluxes of momentum, heat, moisture, and other constituents.

At present it is not possible to fully resolve atmospheric flows in numerical models due to computational constraints. Resolving all turbulent motions in atmospheric flows characterized by Reynolds numbers on the order of 10^7 to 10^8 would require $\mathrm{Re}^{9/4}$ or up to 10^{18} grid points (Rogallo and Moin 1984). Instead, in numerical models of atmospheric flows the effects of turbulent stresses and fluxes on large-scale motions are parametrized, while turbulent fluxes at the surface must be specified. Since surface fluxes are not known a priori, they must be parametrized. When the surface heat flux and vertical virtual potential temperature gradient both vanish, the structure of a neutrally stratified ABL is represented well by a logarithmic velocity profile. While identically zero heat flux is rarely observed, mostly during transition periods between convective (e.g. daytime) and stably stratified (e.g. night-time) conditions, near-neutral ABLs occur under strong shear (i.e. higher wind) conditions. Due to diurnal variability and different atmospheric forcings, the ABL structure is commonly affected by surface fluxes of heat and moisture. Under such conditions, velocity profiles deviate from a logarithmic structure. Furthermore, baroclinicity due to non-uniform vertical profiles of horizontal pressure gradient can also result in deviation from a logarithmic profile.

Theoretical underpinnings of the surface exchanges with the atmosphere were laid out by Monin and Obukhov (1954). They developed a similarity theory linking measurements of wind speed and temperature at a level near the surface to the friction velocity and surface flux of sensible heat. Assuming that two relevant length scales, distance from the surface, z, and Obukhov length, L (Obukhov 1946), account for the effect of a land or water boundary and for the competing effects of shear and buoyancy, Monin and Obukhov defined a non-dimensional stability parameter z/L. The deviation from the shear associated with the logarithmic profile due to the effects of atmospheric stability can then be represented by a universal function that depends on the stability parameter and that must be determined empirically under stationary conditions with a flat, homogeneous upwind fetch. In this way a relationship between the wind shear in a surface layer and the friction velocity can be established. In a similar fashion, a relationship between virtual potential temperature scale and virtual potential temperature gradient is established. We can then use these relationships to compute the turbulent stress and the sensible heat flux at the surface. A similar relationship is extended and applied for moisture fluxes. Monin– Obukhov similarity theory (MOST) is currently used in virtually all atmospheric models to provide surface fluxes of momentum, heat, and moisture (e.g. Beljaars and Holtslag 1991; Jimenez et al. 2011).

A number of field studies under nearly homogeneous and stationary conditions were carried out to determine universal stability functions that modify velocity and temperature profiles under non-neutral conditions. These stability functions are determined as simple

linear and nonlinear regression fits for stably stratified and unstable conditions, respectively. The general forms of stability functions are commonly labelled Businger–Dyer functions (Dyer and Hicks 1970; Businger et al 1971; Dyer 1974). However, different regression parameters are obtained from different field studies. Even when extreme care is taken to control the quality of the data, the scatter is significant, in particular under stably stratified conditions. For example, Newman and Klein (2014) analysed surface observations from the Southern Great Plains site and found that under stably stratified conditions the coefficient of determination between direct observations of surface friction velocity and MOST estimation is only 0.5. Furthermore, MOST is based on single-point statistics, which implies that the local surface-layer eddies are responsible for the total turbulent flux at the surface. In a review marking 50 years of MOST, Foken (2006) noted that: "A better understanding of the limitations of the Monin–Obukhov similarity theory under non-ideal conditions depends upon an exact knowledge of all parameters of the similarity theory". Hicks et al. (2014) revisited MOST relationships using observations from an Ocotillo, Texas field study and pointed to limitations of MOST related to stably stratified boundary layers and non-stationary conditions associated with transitional boundary layers. Analysing large-eddy simulations, Khanna and Brasseur (1997) suggested a non-local dependence of surface fluxes on the stability parameter, z_i/L, accounting for the mixed-layer depth. Recently, Tong and Nguyen (2015) argued that MOST is an incomplete similarity theory because it does not account for the effects of non-local interactions on turbulent flux of momentum and heat. Li et al. (2018) used direct numerical simulations to analyse surface fluxes and demonstrated that eddies that originate from the outer layer contribute significantly to the heat transport in the surface layer, an effect not accounted for within MOST. To account for non-local interactions, Tong and Ding (2020, Ding and Tong 2021) proposed a multi-point MOST, while Salesky and Anderson (2020) propose extended similarity including a parameter related to large-scale motions. Sun et al. (2020) proposed a bulk parametrization of momentum flux based on a hockey-stick transition for weak wind conditions that accounts for the effect of non-local coherent turbulence eddies. The focus of the present work is on development of a surface-layer parametrization for atmospheric models that is not constrained by the assumptions inherent to MOST nor by single-point statistics. As such, this approach does not represent an alternative to the recent developments by Tong and Ding (2020), Salesky and Anderson (2020), and Sun et al. (2020), but complements these developments. The goal is to explore machine-learning approaches as an alternative to simple regression focusing on the homogeneous, flat-fetch boundary layers, the conditions under which MOST should perform well.

In practice, MOST stability functions determined under stationary conditions and flat homogeneous fetch are commonly used even when these conditions are not satisfied. Stiperski and Calaf (2018) proposed extending surface-layer similarity to more complex flows by accounting for turbulent stress anisotropy. In studies that followed, Stiperski et al. (2019, 2021) demonstrated that accounting for anisotropy can provide a more general surface-layer-similarity framework and improve estimates of velocity variances. However, under a wide range of conditions to which MOST is commonly applied, simple regression based on single-point measurements does not capture the relationship between governing parameters and surface-layer structure that, in addition to turbulence stresses, includes fluxes of sensible heat and moisture. Furthermore, empirically determined stability functions representing the non-dimensional shear and virtual potential temperature gradient are expressed as a function of stability parameter z/L so that there is implicit self-correlation because surface friction velocity and sensible heat flux figure in both stability functions, ϕ_m, ϕ_h, and the stability parameter, z/L (Hicks 1978; Klipp and Mahrt 2004). We have therefore developed a machine-learning model for an improved surface-layer parametrization using long-term surface-layer

observations. The ability of a neural network-based machine-learning model to estimate the Obukhov length and the mixed-layer height was previously demonstrated by Pelliccioni et al. (1999). Machine-learning models were also developed to estimate the index of refraction structure parameter, C_n, in a surface layer (Wang and Basu 2016) and develop a better understanding of the dependence of C_n on environmental parameters (Jellen et al. 2020).

To estimate surface fluxes of momentum, sensible heat, and moisture based on measurements of wind speed, temperature, and humidity as well as surface temperature and soil moisture, we developed, trained, and tested two machine-learning models. The machine-learning models are based on the artificial neural network and random forest algorithms. To train and test these machine-learning algorithms, we used several years of observations from the Cabauw mast in the Netherlands and from the National Oceanic and Atmospheric Administration's Field Research Division tower in Idaho. We use only directly observed quantities as predictors in machine-learning models.

In what follows we first review MOST, then we describe the data used for machine-learning-model training and validation, followed by development and testing of two machine-learning models based on random forest and neural network algorithms. Finally, we provide a summary of the results and outline potential further developments.

2 Monin–Obukhov Similarity Theory

Monin–Obukhov similarity theory states that under non-neutrally stratified atmospheric conditions the logarithmic profile is modified as a function of a stability parameter z/L, where L is the Obukhov length scale defined as (Obukhov 1946):

$$L = -\frac{u_*^3}{\frac{g}{T}\overline{w'\theta_v'}}, \tag{1}$$

where g is the acceleration due to gravity, T is the reference temperature, u_* is the surface friction velocity, and $\overline{w'\theta_v'}$ is the turbulent sensible heat flux. The non-dimensional wind shear in the surface layer (i.e. momentum stability function, Φ_M) can be expressed as:

$$\Phi_M\left(\frac{z}{L}\right) = \frac{\kappa z}{u_*}\frac{\partial U}{\partial z}, \tag{2}$$

where z is the distance from the surface, κ is the von Kármán constant, and U is the wind speed. The surface friction velocity u_* is defined as:

$$u_* = \sqrt[4]{\left(\overline{u'w'}\right)^2 + \left(\overline{v'w'}\right)^2}, \tag{3}$$

where $\overline{u'w'}$ and $\overline{v'w'}$ are the surface turbulent stress components. Similarly, the non-dimensional virtual potential temperature gradient in the surface layer (i.e. the virtual potential temperature stability function, Φ_H) can be experimentally determined for a virtual potential temperature profile:

$$\Phi_H\left(\frac{z}{L}\right) = \frac{\kappa z}{\theta_*}\frac{\partial \Theta_v}{\partial z}. \tag{4}$$

Here, Θ_v is mean virtual potential temperature and θ_* is a virtual potential temperature scale defined as:

$$\theta_* = -\frac{\overline{w'\theta'_v}}{u_*}. \tag{5}$$

The moisture lengths scale, q_*, is:

$$q_* = -\frac{\overline{w'q'}}{u_*}, \tag{6}$$

where $\overline{w'q'}$ is the surface moisture flux. An implicit assumption is that the surface friction velocity, the virtual potential temperature scale, and the moisture scale are constant throughout the surface layer. Equations 3, 5 and 6 are used to compute three scales based on observations.

In expressions for the non-dimensional shear and non-dimensional virtual potential temperature gradient, κ is the von Kármán constant. The constant is determined under neutrally stratified conditions when:

$$\Phi_M\left(\frac{z}{L}\right) = 1. \tag{7}$$

After integrating Eq. 2, it follows that:

$$U_1 = \frac{u_*}{\kappa}(\ln z_1 - \ln z_0). \tag{8}$$

Here, z_0 is the roughness length where $U = U_0 = 0$. Equation 8 represents a logarithmic wind profile. The wind speed, U_1, at the level z_1 above the surface and the surface friction velocity, u_*, are observed quantities. Equation 6 can be rewritten as:

$$U_1 = A\zeta - C, \tag{9}$$

where $\zeta = \ln z_1$, $A = u_*/\kappa$, and $C = u_*/\kappa \ln z_0$. The observations under neutral stratification are used to determine the von Kármán constant from the slope, A, of the line, while the roughness length is determined from the offset,C.

For the virtual potential temperature under non-neutral conditions, we can obtain the following relationship (Paulson 1970):

$$\Theta_{v1} - \Theta_{v0} = \frac{\theta_*}{\kappa}\left[\ln\frac{z_1}{z_{0T}} - \psi_H\left(\frac{z_1}{L}\right)\right]. \tag{10}$$

Here, z_{0T} is the heat flux roughness length (Owen and Thomson 1963) and ψ_H is the integral of the stability function Φ_H. The heat flux roughness length can be related to the momentum roughness length (Zilitinkevich 1995):

$$z_{0T} = z_0 \exp\left(-C_z\sqrt{\frac{u_* z_0}{\nu}}\right). \tag{11}$$

The non-dimensional constant C_z is commonly set to 0.1, although it may depend on land-cover characteristics (e.g. Chen et al. 1997; Chen and Zhang 2009), and $\sqrt{u_* z_0/\nu}$ is the roughness Reynolds number, where ν is the kinematic viscosity.

The surface roughness length depends on the characteristics of the upwind fetch and land cover. For a flat terrain covered by short grass or low crops, $0.008\text{m} \leq z_0 \leq 0.09\,\text{m}$ (Wieriga 1993). Kelly and Jørgensen (2017) showed that the uncertainty in estimation of surface roughness can be significant even under nearly ideal conditions of flow over flat, homogeneous terrain (cf. Fig. 1 in Kelly and Jørgensen 2017). For example, the Andreas

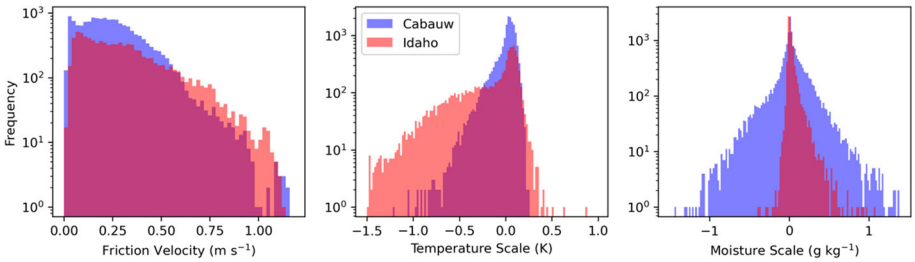

Fig. 1 Historical distributions of friction velocity, virtual potential temperature scale, and moisture scale derived from observations at Idaho (red) and Cabauw (blue) that show different distributions between sites

et al. (2006) estimates of roughness length over sea ice (a relatively consistent surface) during the Surface Heat Budget of the Arctic Ocean (SHEBA) field study show significant scatter (cf. Fig. 5 in Andreas et al. 2006). Even for the same value of observed surface friction velocity the estimated roughness lengths varied two orders of magnitude.

In addition, the universality of the von Kármán constant for different boundary-layer flows has been questioned (cf. Nagib and Chauhan 2008). Nagib and Chauhan argue that the value of the von Kármán constant depends on flow geometry (e.g. pipes, channels, or boundary layers) and the pressure gradient. Based on wind-tunnel measurements of zero pressure gradient boundary-layer flows they report a value of 0.384 ± 0.005. Similarly, Andreas et al. (2006) reviewed values of the von Kármán constant obtained from different atmospheric observations spanning the range from 0.35 to 0.41 for atmospheric flows. Based on atmospheric surface-layer observations during the SHEBA field study they concluded that for a weak pressure gradient ABL, the value of the von Kármán constant is 0.387 ± 0.003. Andreas (2009) presented an argument for a von Kármán constant value of 0.39 and showed that the stability function would not change the functional form, but only the constants in these functions would need to be modified. However, at present the two most commonly used stability functions determined by Dyer and Hicks (1970) and Businger et al (1971) are based on different estimates of the von Kármán constant, 0.41 and 0.35, respectively.

Even before universal stability functions can be determined from observations at a range of atmospheric stabilities, significant uncertainties exist in estimation of the roughness length and the von Kármán constant. These uncertainties compound uncertainty in determining universal stability functions is discussed by Salesky and Chamecki (2012).

Due to significant differences in the structure of velocity and virtual potential temperature profiles under convective and stably stratified conditions, the stability functions are estimated separately for two cases. However, stability functions derived by both Dyer and Hicks (1970) and Businger et al. (1971) have the same functional form. The functional form of similarity functions for stably stratified conditions is:

$$\Phi_M\left(\frac{z}{L}\right) = 1 + a\frac{z}{L}, \tag{12}$$

and

$$\Phi_H\left(\frac{z}{L}\right) = b + c\frac{z}{L}. \tag{13}$$

The functional form of similarity functions for convective conditions is:

$$\Phi_M\left(\frac{z}{L}\right) = \left(1 - d\frac{z}{L}\right)^p, \tag{14}$$

and

$$\Phi_H\left(\frac{z}{L}\right) = g\left(1 - h\frac{z}{L}\right)^q. \tag{15}$$

The constants $[a, b, c, d, p, g, h, q]$ are $[5, 1, 5, 16, -1/4, 1, 16, -1/2]$ and $[5, 0.74, 5, 15, -1/4, 0.74, 15, -1/2]$ based on Dyer and Hicks (1970) and Businger et al. (1971), respectively. Stability functions for moisture have the same form as those for virtual potential temperature.

The surface friction velocity, u_*, the virtual potential temperature scale, θ_*, and the moisture scale, q_*, are estimated using MOST by integrating Eqs. 2 and 3, and a corresponding equation for moisture using constants determined by Dyer and Hicks (1970). These equations are given in Appendix. Based on the Dyer and Hicks form of stability functions, the turbulent Prandtl number is implicitly equal to unity.

3 Data

Our focus is on developing a surface-layer parametrization for atmospheric models applicable at a wide range of conditions. While there are high-quality observations from a number of episodic field studies focused on specific physical processes (e.g. LeMone et al. 2000, Cooperative Atmosphere Surface Exchange Study 97 (CASES-97); Poulos et al. 2002, Cooperative Atmosphere Surface Exchange Study (CASES-99); Uttal et al. 2002, SHEBA; Lothon et al. 2014, Boundary-Layer Late Afternoon and Sunset Turbulence (BLLAST), etc.), we use observations from two locations that provide quality-controlled long-term observations to improve overall model skill by developing and test a machine-learning model that can be compared to MOST.

For this initial development of a machine-learning model for parametrization of surface fluxes, we use observations from two sites with flat homogeneous fetch to enable comparison with MOST. Using multiple sites allows us to compare a model trained with one site's data and applied to the other. One of the observational sites is the Royal Netherlands Meteorological Institute (KNMI) Cabauw Experimental Site for Atmospheric Research at Cabauw, Netherlands (Beljaars and Bosveld 1997), a site that has been used to validate surface-layer parametrizations and land-surface models since 1972. Cabauw is located in the western part of the Netherlands at 51.971°N, 4.927°E. The area is generally used for agriculture with minimal elevation changes within kilometres of the tower and generally covered with low brush. While the Cabauw observations are available starting from 26 February 2003, here we use data from 2013 through 2015 to match the three-year record period available from Idaho. The outgoing longwave radiation, which we convert to skin temperature, is derived from a measuring device approximately 200 m from the flux tower (Bosveld 2020). This dataset includes all of the variables needed for analysis at a 10-min temporal resolution.

Our second observational site is at the Idaho National Laboratory (INL) site, Idaho, USA, and managed by the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratory Field Research Division. The NOAA/INL eddy-covariance flux tower is located at 43.5959°N and 112.9288°W in a flat area with low (less than 1 m) brush. The outgoing longwave radiation measurement is collocated at 2 m height with the NOAA/INL flux tower. All other meteorological measurements are located at a tall tower that is about 900 m to the south-west of this flux tower (Finn et al. 2017). The dataset includes observations spanning 2015 to 2017. The variables measured match well with the ones from Cabauw, with a few exceptions. First, the Idaho dataset only includes relative humidity at one level

(2 m), which limits the moisture scale predictability at Idaho. Second, the Idaho dataset has slightly different levels for measuring wind speed and temperature (2 m, 10 m, 15 m, and 45 m) compared to Cabauw (2 m, 10 m, 20 m, and 40 m). In the comparison, we match the 15 m reading at Idaho to the 20 m reading at Cabauw and the 45 m reading at Idaho to the 40 m reading at Cabauw. Third, the Idaho dataset measures soil moisture content at 5 cm, while the Cabauw dataset measures soil moisture content at 3 cm. Additionally, we utilized measurements of wind velocity, temperature, and water vapour mixing ratio at their highest resolution, which was 10 min for Cabauw, and at Idaho we averaged 5-min resolution data to match the 10-min data from Cabauw. Flux measurements at Cabauw were also available at 10-min resolution, while at Idaho they were available at 30 min.

While the two locations are similar in terms of their horizontal homogeneity of grassland, they represent clearly different climatologies with different distributions of our target variables moisture scale, virtual potential temperature scale, and friction velocity, as shown in Fig. 1. Specifically, the moisture scale has a narrower distribution centred around zero for the arid Idaho site compared to a wider distribution for the Cabauw site. The Idaho site is also characterized by larger negative values of virtual potential temperature scales and higher frequency of larger values of friction velocity. Associated with the different distributions of target scaling parameters between the two sites are different distributions of atmospheric stability. We computed the bulk Richardson number (Ri_b) between the heights of 2 m and 10 m and displayed the number of cases of negative Ri_b (unstable) versus positive Ri_b (stable) conditions. While the Idaho data are characterized as 55% stable and 45% of unstably stratified conditions, the Cabauw data are characterized substantially higher ratio of stable to unstable instances, approximately 62.5% to 32.5%.

Figures 2 and 3 depict stability functions for momentum, ϕ_m, and heat, ϕ_h, respectively, computed using the data from the Cabauw and Idaho sites. It is clear that there is significant spread at both sites and potentially bias associated with determination of surface roughness for momentum and heat transfer as well as the von Kármán constant.

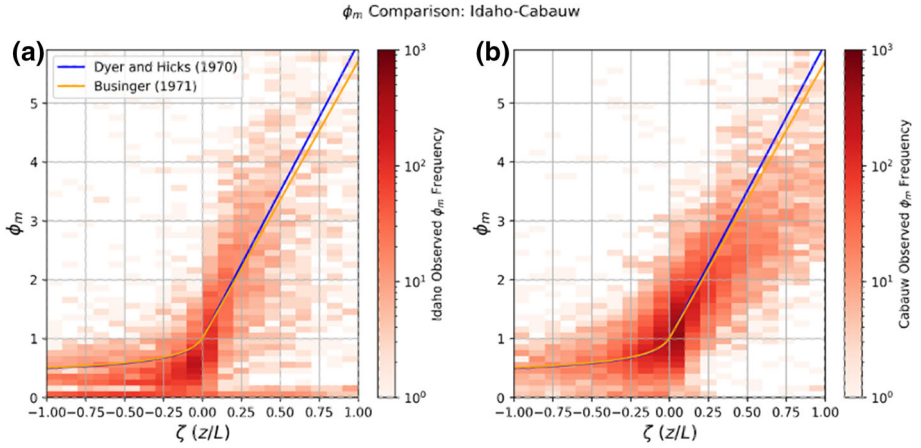

Fig. 2 Two-dimensional histograms of the momentum stability function, Φ_M, for Idaho (**a**) and Cabauw (**b**) as a function of the non-dimensional stability parameter z/L

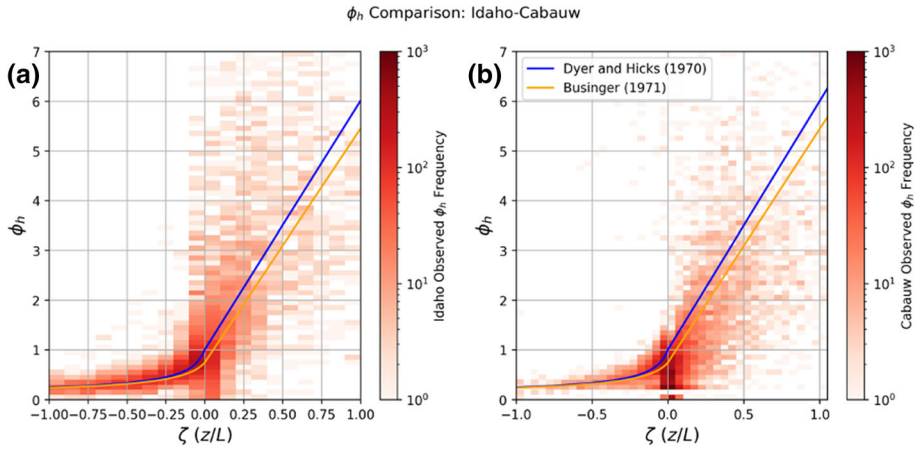

Fig. 3 Two-dimensional histograms of sensible heat stability function, Φ_H, for Idaho (**a**) and Cabauw (**b**) as a function of the non-dimensional stability parameter z/L

4 Machine-Learning Methods

We developed two machine-learning models for surface-layer parametrization based on the random forest (RF) and artificial neural network (ANN) methods. The two algorithms have different training requirements and they differ in complexity. The RF method requires less preprocessing and its training process is simpler. The ANN algorithm requires more experience in machine-learning model development because it includes a wider range of tunable hyperparameters. However, in comparison with RF-based models, ANN-based models produce a smoother prediction, and they are usually more compact models. For both model sets, we trained separate models to predict the friction velocity, virtual potential temperature scale, and moisture scale at each site. While we manually validated a select subset of model hyperparameters (settings governing model structure, such as the number of trees in a RF) configurations based on prior experience with machine-learning training, we did not perform an exhaustive hyperparameter search. Further incremental performance gains may be possible for these datasets with additional tuning, but those performance gains could come from overfitting to the validation set and as a result would not generalize.

4.1 Random Forest

The RF algorithm (Breiman 2001) has seen increased use across a wide range of meteorological applications in recent years (e.g. Gagne et al. 2017; Herman and Schumacher 2018; Yuval and O'Gorman 2020), because it provides a good balance of accuracy, robustness, and interpretability compared with many other machine-learning algorithms. The RF algorithm consists of an ensemble of classification and regression tree (CART; Breiman 1984) decision trees that have been diversified by incorporating random subsampling into an otherwise deterministic and greedy training process.

Classification and regression decision trees consist of decision nodes containing a yes-or-no question formatted as "Is x_n < threshold?" (i.e. "Is temperature < 280 K?"). If yes, the algorithm proceeds down the left branch, and if no, the algorithm proceeds down the right

branch. The x_n and threshold values are selected by exhaustively testing every combination of both and picking the one that minimizes the mean-squared error, for our regression RF, for the training examples at that node. Subsequent nodes in the tree can either be decision nodes or leaf nodes, where the final prediction from the tree is calculated from the training examples that fell into that node. For our RF, tree growth is stopped when a minimum number of examples in a node (2) are reached or a maximum number of leaf nodes (4096) have been created across all branches in order of the number of examples that proceed down each branch. A branch with more examples will be grown out before a branch with fewer examples. Limiting the size of the trees makes each tree less prone to overfitting by reducing prediction variance and also reduces the storage cost of the model. Fully growing the decision trees for the full Cabauw dataset resulted in a model that required multiple gigabytes of disk storage, whereas limiting leaf nodes resulted in a storage size of 22 MB.

For classification problems, the relative frequency of each class in the node determines the predicted probability. For regression problems, the mean of output values for the training examples in a leaf node determines the predicted output. In essence, a decision tree is a dynamic nearest-neighbours algorithm where the neighbourhood is determined by the decision thresholds of a particular branching path rather than a fixed distance metric.

The random forest builds on CART by incorporating random sampling into the training process with the combined goals of increasing tree diversity and reducing training time per tree. Each tree in the forest is trained on a bootstrap resampled subset of the original training data in which some examples are duplicated and others are ignored. The candidates for x_n at each node are selected by drawing a random subset of inputs of size square root of the total number of inputs and evaluating them under the same procedure as CART. The final prediction from the regression RF is the mean of the predictions from all the trees. Increasing the number of trees in the forest results in more stable predictions but also requires an increasing amount of computation, so we chose 100 trees as a reasonable compromise that has been robust across many problem domains.

Because the RF is a dynamic nearest-neighbours algorithm, it is efficient at interpolating within the space of the training data. Outside the range of the training data inputs, RF assumes a constant extrapolation value. As a result, RF will not predict any output values outside the range of the training distribution and will generally have a smaller prediction variance than the training output variance. The RF will not produce widely unphysical values even if given noisy input values, but it will underestimate extremes to a greater degree than other ML methods. We used the RF implementation from the *scikit-learn* package (Pedregosa et al 2011). For more information about the theory behind decisions trees and RF, see Hastie et al. (2009) and James et al. (2021).

4.2 Artificial Neural Network

An ANN is a flexible machine-learning method that can universally approximate smooth continuous functions (Hornik et al. 1989). Artificial neural networks consist of an input layer, a series of hidden layers that transforms the input vector into a latent vector, and an output layer that produces the final prediction. Each hidden layer consists of a matrix of perceptrons (Rosenblatt 1958; Reed and Marks 1998), or linear regression functions wrapped with a nonlinear transformation, the activation function. For this project, we used the scaled exponential linear unit (SELU) activation function (Klambauer et al. 2017), which encourages the neural network to self-normalize its signal and performed well in prior research on neural network parametrization (Gagne et al. 2020). The activation functions induce a sparser

representation of the latent vector, enabling the ANN to model different regimes within the data. Increasing the number of perceptrons in each hidden layer enables the model to learn a smoother and more detailed representation of the data at the potential risk of overfitting, so we chose 128 perceptrons per hidden layer as a reasonable compromise.

Artificial neural networks are trained or optimized through the process of stochastic gradient descent via backpropagation. The model is initialized with small random weights that are updated by selecting a random mini-batch of data, sending it forward through the model to generate a prediction, and calculating the prediction error, which is the mean-squared error for our implementation. The error gradient, or partial derivative of the error with respect to each model weight, is calculated in reverse through the model since the gradients for the first hidden layer depend on those for the second hidden layer, and so on. Increasing the number of hidden layers can reduce the magnitude of the gradient significantly, resulting in noisy gradients for early layers. Activation functions in the rectified linear unit family, including SELU, reduce this issue because the positive gradient is constant no matter how large the input value is. We chose two hidden layers for our model and found that additional hidden layers did not result in any performance improvements on validation data. The choice of optimizer determines how the gradient updates the weights by factoring in the learning rate, a constant multiplier to the gradient that determines the step size of the update. We chose the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.0001 because it promotes fast and stable convergence compared with other optimizer algorithms and has become the default choice for many neural network problems. The ANN training duration depends on the size of the mini-batch (128) and the number of epochs, or iterations over the training data. We found that 20 epochs were sufficient for the training and validation error decrease to begin levelling off without overfitting (see "Appendix 2").

Some ANNs require additional regularization constraints to fit their data robustly, especially if the data are noisy and high-dimensional, but our validation set performance did not indicate a need for regularization. Unlike decision trees, which include variable selection throughout the model building process, ANNs will use information from all variables. However, they will greatly reduce the weights assigned to less relevant inputs, leading to soft feature selection. The extrapolation pattern of neural networks depends on the choice of activation function. For SELU, the model extrapolates linearly based on the subset perceptrons activated by a given input. This extrapolation can increase the model variance but can also result in increasingly non-physical values the further away new predictions are from the training distribution. The model used here is built with the Keras–TensorFlow framework (Chollet 2015), which makes configuration of neural network settings and designs easy and scalable.

5 Comparing the Machine-Learning Surface-Layer-Parameter Estimates to Monin–Obukhov Similarity Theory

The predictands that must be estimated using a machine-learning model are surface friction velocity, u_*, virtual potential temperature scale, θ_*, and moisture scale, q_*. As the first step in development of machine-learning models for surface-layer parametrization, we need to determine the common set of predictors for the two observational sites. For effective development and best results in developing a reliable machine-learning model, the set of predictors should be based on consistent observations among different sites. However, the challenge is

that, in general, observations at different sites are not typically consistent. Frequently, observations of the same type are not made, or they may be made at different levels with respect to the ground level or at different frequencies. Surface fluxes of momentum, heat, and moisture can be and are frequently estimated with MOST using measurements of wind velocity, temperature, and specific humidity at only one level; however, as Panofsky (1963) indicated and Basu (2019) demonstrated, the Obukhov length, L, and therefore surface turbulent stress and sensible heat flux can be estimated using measurements of wind and temperature at three levels. Since our goal is to develop a surface-layer parametrization for atmospheric models and considering that they can provide prediction of atmospheric variables at multiple levels in a surface layer, we use observations at three levels above the ground.

The full list of the common set of predictors between Idaho and Cabauw sites appears in Table 1 with the corresponding height levels for each location. Several measurements were taken at different heights, at 15 m and 45 m at the Idaho site, while for Cabauw the measurements were taken at 20 m and 40 m above the ground. Additionally, the soil moisture and soil temperature were measured at 3 cm at Cabauw and at 5 cm at the Idaho site. For the purpose of this analysis, we assumed that the data were observed at the same level, but we would expect some small differences to result from these differences in levels.

We first trained ANNs and RFs for each site independently. We first applied the resulting machine-learning model to test datasets from the same site the training dataset was derived

Table 1 Observations from Idaho and Cabauw used as predictors in machine-learning models. Height levels at which observations are made are indicated in the second and the third column for Idaho and Cabauw, respectively

Observation (units)	Idaho	Cabauw
	Height level (m)	Height Level (m)
Potential temperature (K)	10	10
Potential temperature (K)	15	20
Potential temperature (K)	45	40
Low-level wind speed (ms^{-1})	10	10
Low-level wind direction (°)	10	10
Mid-level wind speed (ms^{-1})	15	20
Mid-level wind direction (°)	15	20
Top-level wind speed (ms^{-1})	45	40
Top-level wind direction (°)	45	40
Relative humidity (%)	2	2
Global horizontal irradiance (Wm^{-2})	0	0
Pressure (hPa)	2	2
Solar zenith angle (°)	0	0
Skin Temperature (K)	0	0
	Depth level (cm)	Depth level (cm)
Top-level soil water content (gm^{-3})	5	3
Top-level soil temperature (K)	5	4
	Difference between levels	
Bulk Richardson number	10–2 m	10–2 m

from. We then applied the models trained on the dataset from the first site to the test datasets from the second site to evaluate whether a model trained in one climate could perform in another climate, and thus determine whether the models can be generalized. Finally, we trained the machine-learning models on a training dataset that merged the Idaho and Cabauw training datasets. The Cabauw dataset was split into years 2013 to 2014 for training and year 2015 for testing, which resulted in 34,025 30-min averaged sets of observations in the training data and 16,553 sets in the testing data. For the Idaho dataset we used years 2016 to 2017 for training and year 2015 for testing, which included 27,787 30-min averaged sets of observations in the training data and 9,376 sets in the testing data. For consistency, the observations from the same year were used for testing from both locations. Any instances where any of the variables were missing were removed from the datasets. The mean absolute error (MAE) and the square of the Pearson correlation coefficient (R^2) were computed for the machine-learning model predictions and the MOST estimates with respect to observations of the friction velocity and virtual potential temperature scale. The MAE and R^2 results for the independent testing datasets are shown in Table 2 for the Idaho test dataset and in Table 3 for the Cabauw dataset. These results highlight the generally superior performance of both the ANN and the RF models over MOST with lower MAEs, and higher R^2 when models are trained and tested using data from the same site. Although forecast skill degrades when a machine-learning model trained in one climate is applied to the other, the models trained using the combined dataset consisting of merged Idaho and Cabauw datasets outperform MOST. In general, there are no major differences in the performance between the ANN and RF.

These results indicate that the additional data allow both the ANN and the RF model to learn the representative patterns and perform well. It would be expected that as more sites and data are added that both models would continue to generalize better to additional areas with minimal degradation compared to a site-specific model.

The distributions of the machine-learning model predictions compared to the surface flux variables predicted by MOST demonstrate the differences between the data-driven results and MOST. Figure 4 includes two-dimensional histograms (warmer colours indicate higher density of instances than cooler colours) that display the differences between observed and predicted surface friction velocity for Idaho (top) and Cabauw (bottom) from the RF (left),

Table 2 MAE and R^2 of the ANN and RF models trained on each dataset and applied to the Idaho test dataset using all common variables as predictors

Idaho test dataset (2015)						
	MAE			R^2		
	u_* (m s^{-1})	θ_* (K)	q_* (g kg^{-1})	u_*	θ_*	q_*
MOST	0.086	0.128	0.128	0.85	0.29	0.17
ANN trained on Idaho (2016–2017)	0.051	0.087	0.025	0.89	0.62	0.47
ANN trained on Cabauw (2013–2014)	0.087	0.199	0.161	0.87	0.58	0.22
ANN trained on both	0.049	0.081	0.027	0.90	0.66	0.46
RF trained on Idaho (2016–2017)	0.048	0.077	0.027	0.91	0.67	0.41
RF trained on Cabauw (2013–2014)	0.083	0.195	0.193	0.87	0.61	0.20
RF trained on both	0.048	0.078	0.027	0.91	0.67	0.42

Table 3 MAE and $R.^2$ of the ANN and RF models trained on each dataset and applied to the Cabauw test dataset using all common variables as predictors

Cabauw test dataset (2015)						
	MAE			R^2		
	u_* (m s^{-1})	θ_* (K)	q_* (g kg^{-1})	u_*	θ_*	q_*
MOST	0.038	0.043	0.127	0.92	0.68	0.23
ANN Trained on Idaho (2016–2017)	0.066	0.145	0.116	0.92	0.79	0.45
ANN Trained on Cabauw (2013–2014)	0.024	0.020	0.046	0.96	0.93	0.83
ANN trained on both	0.025	0.022	0.044	0.96	0.92	0.84
RF trained on Idaho (2016–2017)	0.070	0.078	0.118	0.93	0.78	0.57
RF trained on Cabauw (2013–2014)	0.022	0.021	0.043	0.96	0.92	0.84
RF trained on both	0.023	0.022	0.044	0.96	0.91	0.84

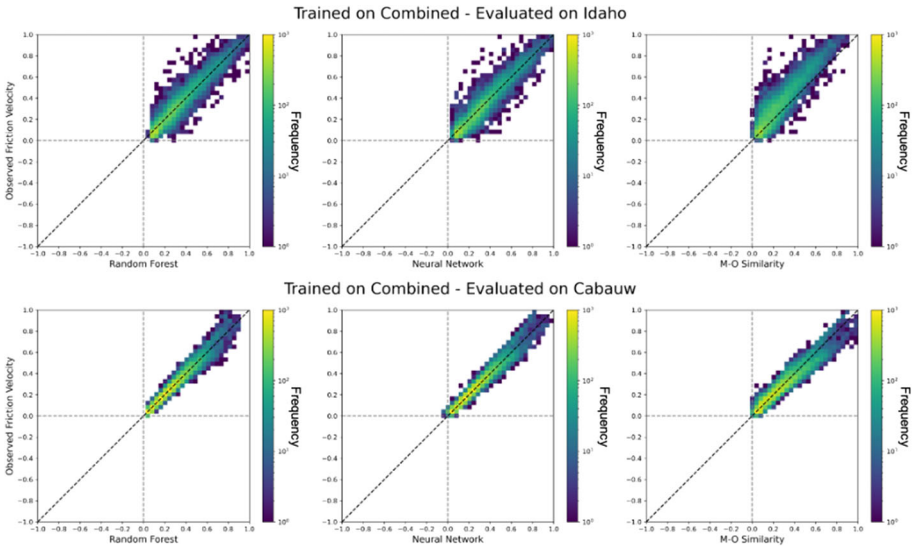

Fig. 4 Two-dimensional histograms comparing the observed and predicted friction velocity evaluated using Idaho (top row) and Cabauw (bottom row) data from the RF (left), ANN (centre), and MOST (right) with brighter colours indicating more instances and cooler colours indicating fewer instances

ANN (centre), and MOST (right). For the friction velocity predictions, the RF, ANN, and MOST produce generally similar distributions.

While two-dimensional histograms for the virtual potential temperature scale are also similar for RF, ANN, and MOST (Fig. 5), MOST histograms for both Idaho and Cabauw data have wider spread. The moisture scale distributions (Fig. 6) obtained by RF and ANN models are similar spanning a range of values from -0.5 to 0.8, MOST results in predominantly positive values. These results indicate that the machine-learning models are better capturing

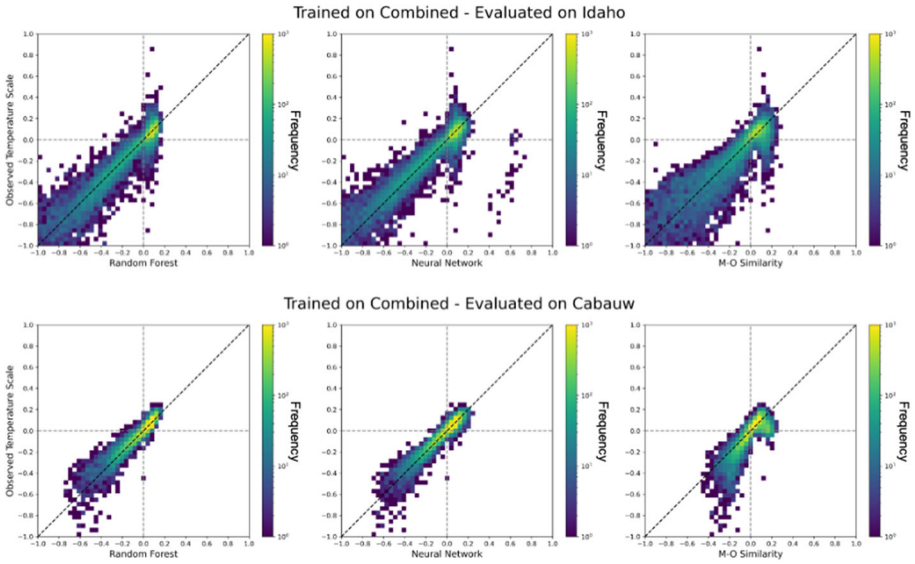

Fig. 5 Two-dimensional histograms comparing the observed and predicted virtual potential temperature scale evaluated using Idaho (top row) and Cabauw (bottom row) data from the RF (left), ANN (centre), and MOST (right) with brighter colours indicating more instances and cooler colours indicating fewer instances

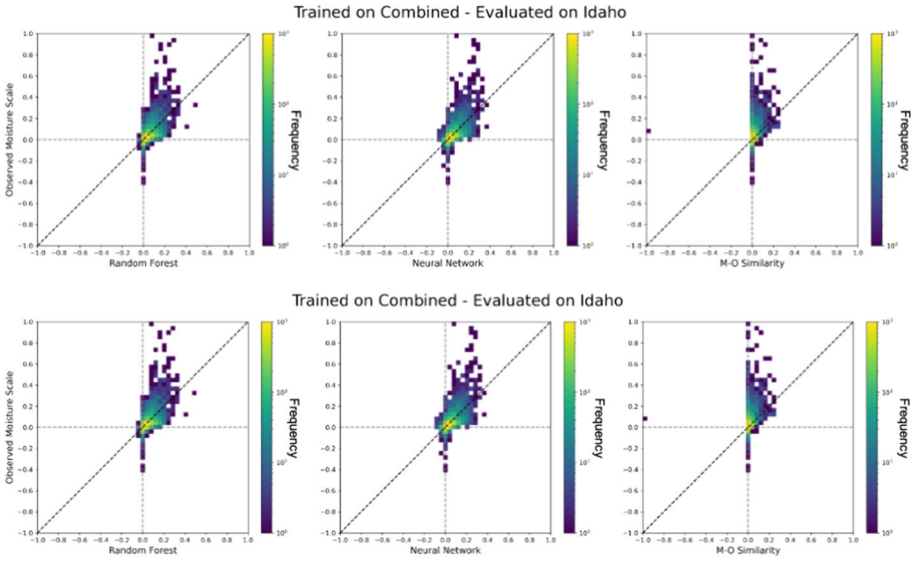

Fig. 6 Two-dimensional histograms comparing the observed and predicted moisture scale evaluated using Idaho (top row) and Cabauw (bottom row) data from the RF (left), and ANN (right) with brighter colours indicating more instances and cooler colours indicating fewer instances

the real distribution of the virtual potential temperature and moisture scale compared to the results computed from MOST.

6 Machine-Learning Interpretation

Explainable machine-learning methods can provide insights into what inputs a particular machine-learning model favours and how changes in those inputs affect the model predictions. Here, we perform and evaluate two machine-learning interpretability techniques: permutation feature importance and partial dependence plots (McGovern et al. 2019) for the dataset from Sect. 4 that utilizes all common predictors with 10-min average data. The predictor importance plots, Fig. 7 for Idaho data and Fig. 8 for Cabauw data, show the relative importance of each of the predictors by determining the increase in mean-squared error after permuting the values of each input among all examples and sending the permuted data through the model. For friction velocity the most important RF predictors are wind speeds at different levels encoding the level of shear followed by the bulk Richardson number, which encodes atmospheric stability. The ANN prefers the top-level potential temperature over bulk Richardson number for its indication of stability. For both the moisture and the virtual potential temperature scale, global horizontal irradiance (GHI) is the most important predictor for RF as it encodes the diurnal cycle, and therefore, indirectly the stability. A more direct measure of stability is the bulk Richardson number, the second most important predictor for these scales. The ANN again prefers the direct measures of both skin temperature and temperatures at different heights for its estimates of temperature and moisture scale. The next most important predictors for the moisture scale capture the heat and moisture content at and near the surface, which is

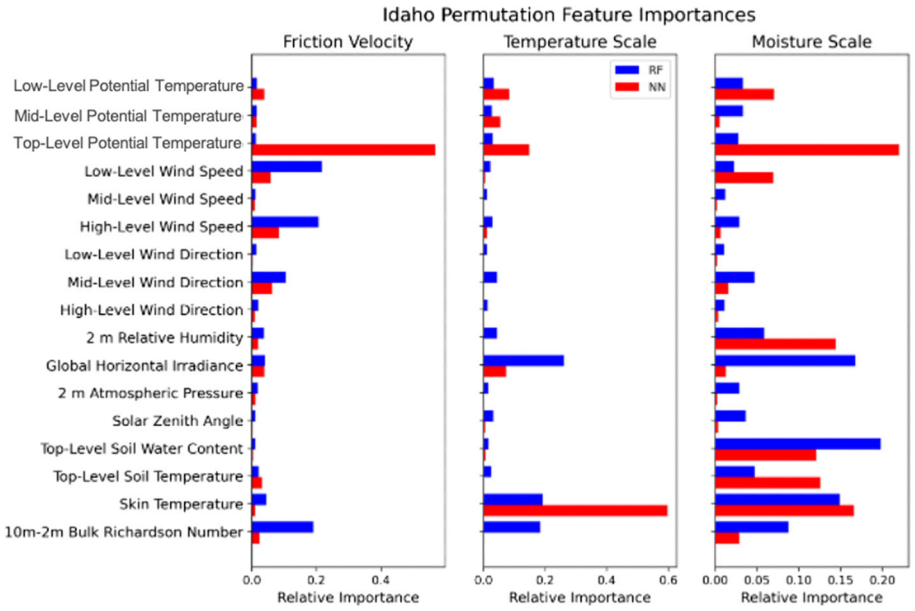

Fig. 7 Predictor importance rankings for the RF model on the Idaho dataset utilizing all common variables and the 10-min average fluxes

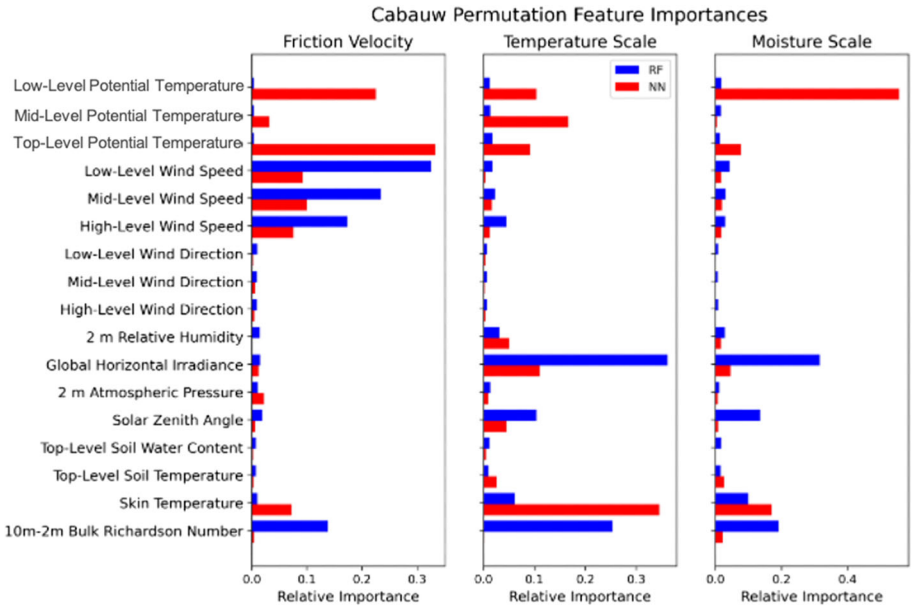

Fig. 8 Predictor importance rankings for the RF model on the Cabauw dataset utilizing all common variables and the 10-min average fluxes

expected given that the moisture scale quantifies the moisture flux from the surface of the earth to the surface layer of the atmosphere. In addition to GHI and the bulk Richardson number, relative humidity has a significant impact on the virtual potential temperature scale for the Idaho dataset, while solar zenith angle, which encodes seasonality, has a significant impact for the Cabauw dataset.

In addition to the predictor importance, we also analysed the partial dependence of predictors based on the Idaho and Cabauw data for both RF and ANN models. The partial dependence plots are shown for all common predictors for friction velocity (Figs. 9 and 10), virtual potential temperature scale (Figs. 11 and 12), and moisture scale (Figs. 13 and 14). Partial dependence plots illustrate the marginal effect a predictor has on the predicted outputs from a machine-learning model (Friedman 2001). The partial dependence plots hold a predictor variable constant at the low end of the range of the data and apply the trained machine-learning model to make a prediction. The average prediction of the predictand is then plotted against the predictor variable value that was held constant. Then, the predictor value is increased systematically while making predictions with the machine-learning model. For example, for the mid-level wind speed the value is given at 2.0 m s^{-1} and a prediction is made for all instances in the test data. Then the value is increased in ten consistent increments up to 10 m s^{-1} to show how the predicted variables (friction velocity, virtual potential temperature scale, and moisture scale) change as the mid-level wind speed changes.

The same model applied to the Idaho and Cabauw data produces similar partial dependence results. However, two machine-learning models yield significantly different partial dependences. These differences can be attributed to how the machine-learning algorithms handle correlated input fields and how they interpolate and extrapolate differently. In general, the RF is more likely to select variables with stronger signals and ignore others, while neural

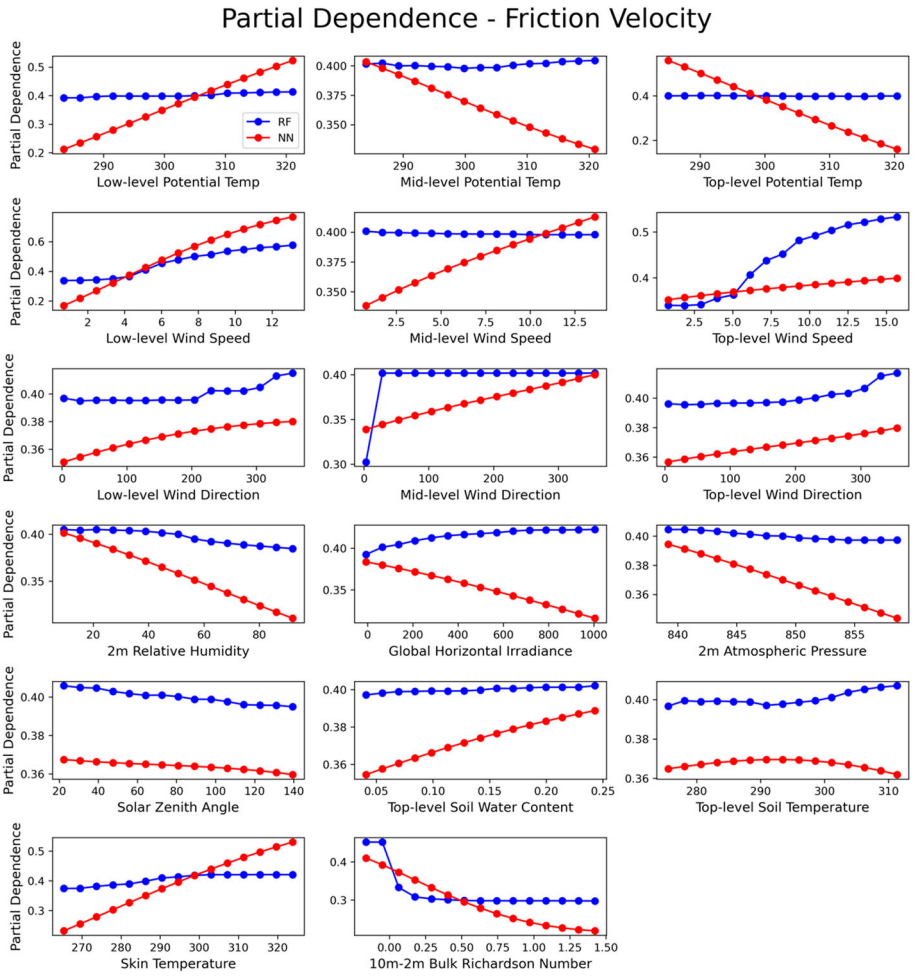

Fig. 9 Predictor partial dependence plots for the RF friction velocity predictions on the Idaho dataset utilizing all common variables and the 10-min average fluxes

networks provide similar weights to correlated variables that contain the same signal. The RF approach also shows flat dependence outside the range of the training data, while neural networks show a linear relationship expanding past the limits of the bulk of the training data.

Similar to the predictor importance analysis, the partial dependence results also highlight several physical behaviours that are expected given our knowledge of surface layer processes. The partial dependence analysis for friction velocity based on ANN and RF used on both datasets (Idaho, Fig. 9; Cabauw, Fig. 10) displays very different dependence on virtual potential temperature measurements. While the RF model estimates do not depend on the virtual potential temperature, the ANN model estimates exhibit linear dependence. Both machine-learning models display an expected positive increase for the low-level wind speed. However, ANN has a linear dependence, while the RF dependence is nonlinear. When the ANN model is used, the friction velocity also decreases as the bulk Richardson number increases, indicating that, as expected, stronger stability is associated with weaker momentum

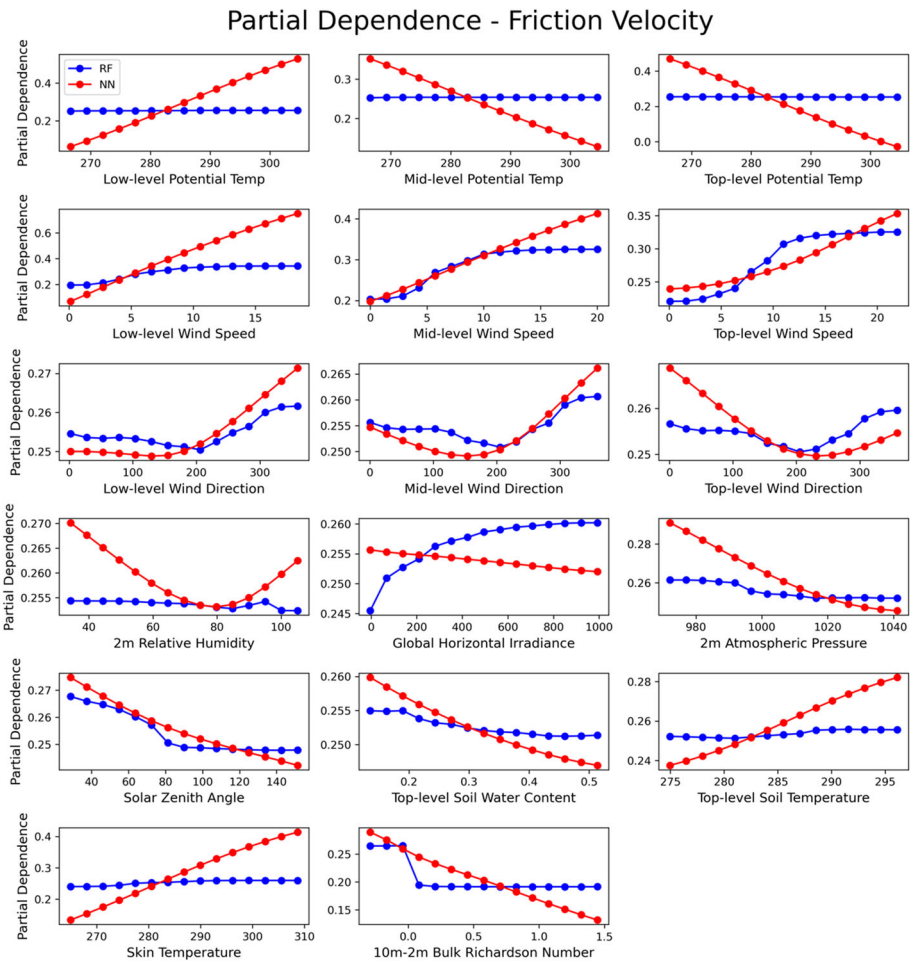

Fig. 10 Predictor partial dependence plots for the RF friction velocity predictions on the Cabauw dataset utilizing all common variables and the 10-min average fluxes

fluxes. However, due to an implicit classification, the RF model displays stepwise behaviour as the bulk Richardson changes sign, indicating transition from convective to stably stratified conditions. The virtual potential temperature scale partial dependence plots (Figs. 11 and 12) show that there is a strong negative dependence on GHI and a strong positive dependence for the ANN model on wind speeds and the lowest level virtual potential temperature. In contrast, for the RF model, the virtual potential temperature scale exhibits weak dependence on the wind speed and virtual potential temperature. It also exhibits positive linearly decreasing dependence for the ANN model. For the RF model, the virtual potential temperature scale shows significant positive and nearly stepwise dependence on a range of bulk Richardson number values close to zero, while it levels off for larger values corresponding to the stronger atmospheric stability. The moisture scale partial dependence plots (Figs. 13 and 14) show that there is a strong positive linear relationship with GHI and a negative linear dependence

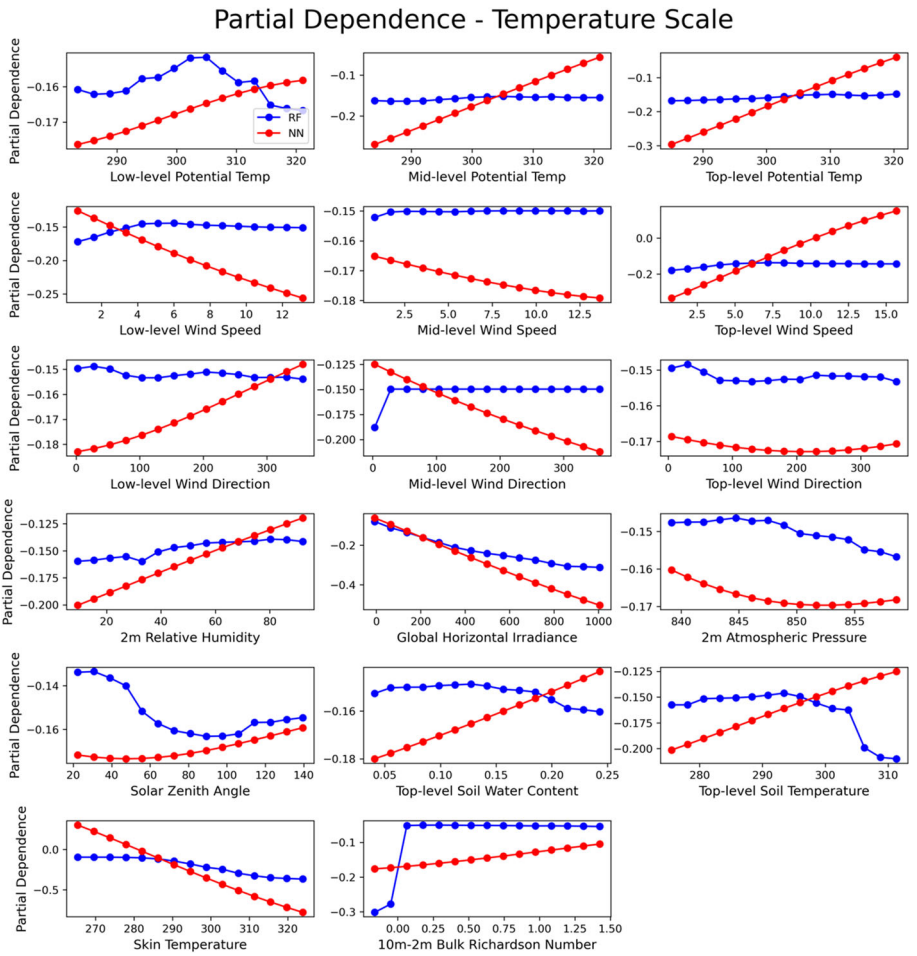

Fig. 11 Predictor partial dependence plots for the RF virtual potential temperature scale predictions on the Idaho dataset utilizing all common variables and the 10-min average fluxes

on the bulk Richardson number. It also indicates a weak dependence on calm to very light low-level wind speeds.

The analysis of the partial dependence of predictors demonstrates that the wind speed, i.e. wind shear, is the primary predictor controlling momentum exchange with the surface and that atmospheric stability is secondary. For heat and moisture exchanges, atmospheric stability encoded through GHI and the bulk Richardson number is the primary predictor, while the direct effect of wind speed or shear is a relatively distant second. Thus, the RF model displays an ability to effectively separate different stability conditions into implicit regimes as it grows regression trees.

The turbulent eddy structure of convective and stably stratified ABLs differs significantly due to buoyancy effects. Under unstable, convective conditions, the boundary layer fills with convective cells or helical rolls, or a combination thereof. In contrast, under stably stratified conditions, the boundary layer is dominated by a broader spectrum of shear production of

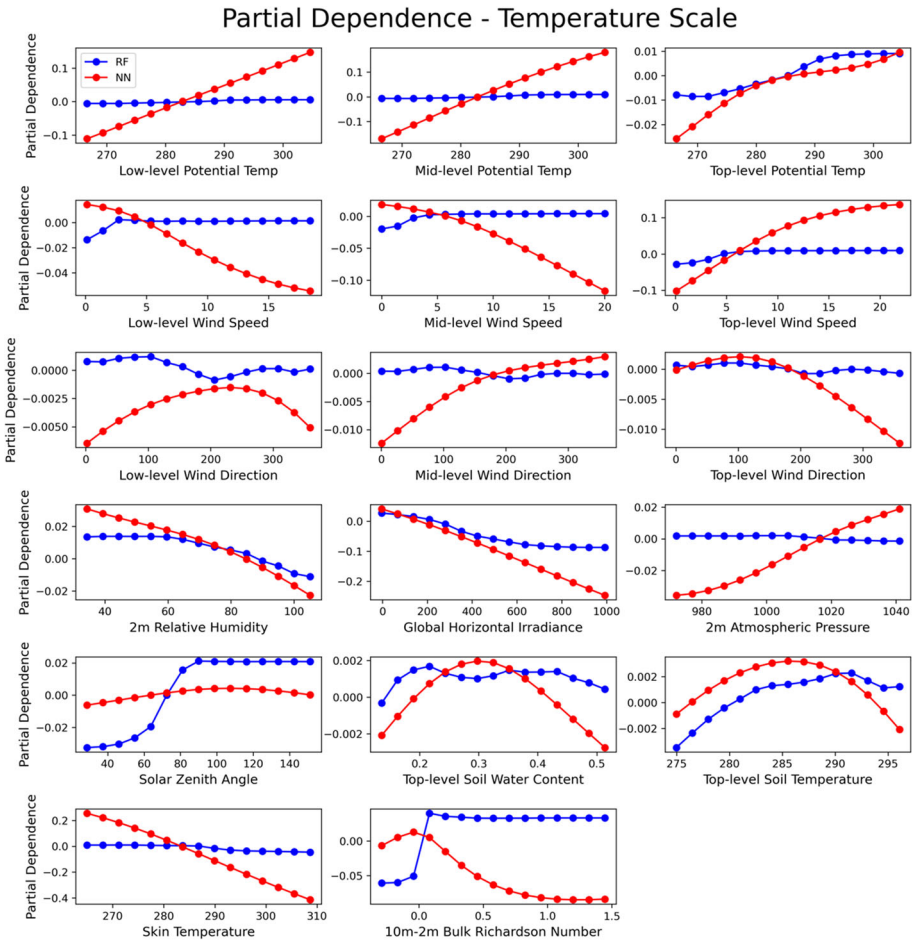

Fig. 12 Predictor partial dependence plots for the RF virtual potential temperature scale predictions on the Cabauw dataset utilizing all common variables and the 10-min average fluxes

turbulence kinetic energy. Therefore, surface exchanges differ significantly between these regimes as can be seen from the different forms of stability functions for convective and stably stratified conditions. To analyse the ability of the RF separately for each of these conditions, we divided the results into these two stability regimes based on the bulk Richardson number: stably stratified when the bulk Richardson number is positive or unstable when the bulk Richardson number is negative. For the friction velocity, there was minimal difference between the stability regimes, which is a consequence of its primary dependence on the wind speed, i.e. wind shear, and significantly less pronounced dependence on the stability of the atmosphere, and therefore, results are not shown here. For the moisture scale and virtual potential temperature scale, the RF produces substantially better results in the unstable regime compared to the stable regime, which is illustrated in Fig. 15 tested using the Idaho dataset and in Fig. 16 tested using the Cabauw dataset. For the virtual potential temperature scale tested on Idaho data, the value of R^2 for the stable regime is 0.43, while it is 0.42 for the unstable regime. Similarly, for the moisture scale analysis, the value of R^2 for the

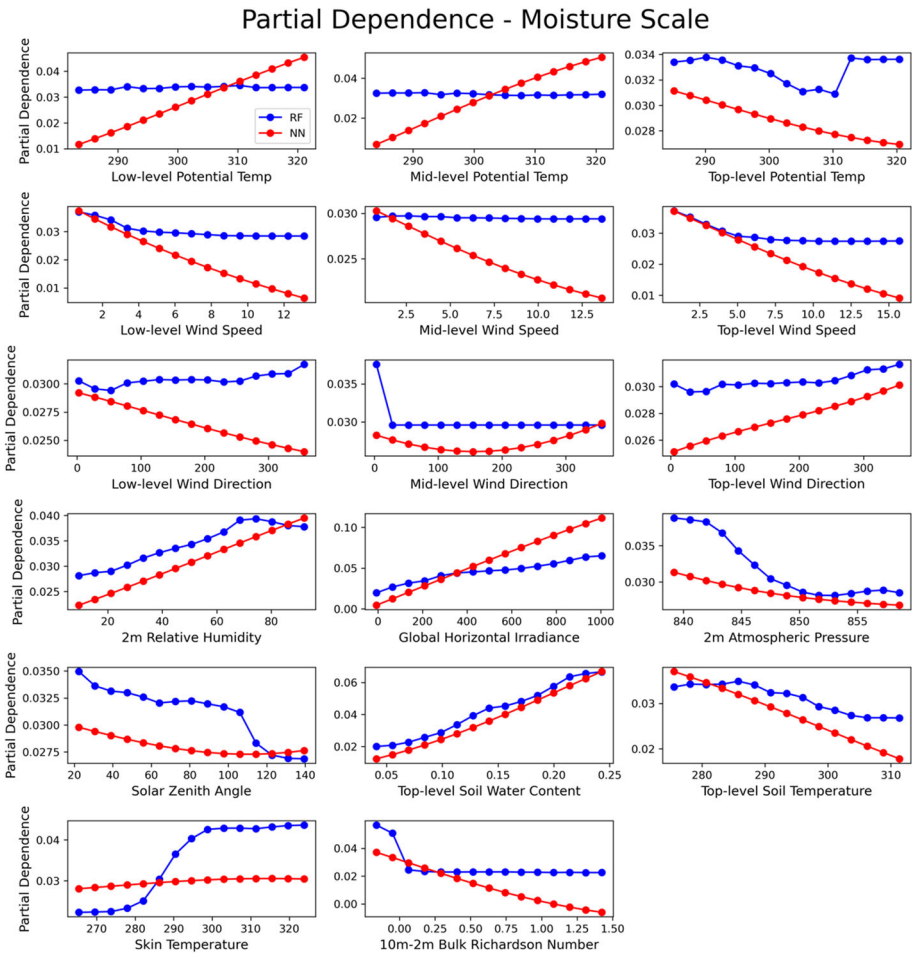

Fig. 13 Predictor partial dependence plots for the RF moisture scale predictions on the Idaho dataset utilizing all common variables and the 10-min average fluxes

stable regime is 0.249 and for the unstable regime is higher at 0.318. For the virtual potential temperature scale test on Cabauw data, the value of R^2 for the stable regime is 0.72, while it is much higher at 0.85 for the unstable regime. Similarly, for the moisture scale analysis, the value of R^2 for the stable regime is 0.703 and for the unstable regime is also much higher at 0.854. This analysis provides evidence that the machine-learning models have better ability to estimate the surface fluxes in the unstable regime. Similar conclusions can be made analysing ANN model prediction based on stability conditions. This agrees with the previously stated behaviour of MOST, which also struggles with stably stratified conditions (Hicks et al. 2014).

The datasets analysed in the current study are available from the corresponding author on reasonable request, and the source code used in analysis is available at https://github.com/NCAR/mlsurfacelayer.

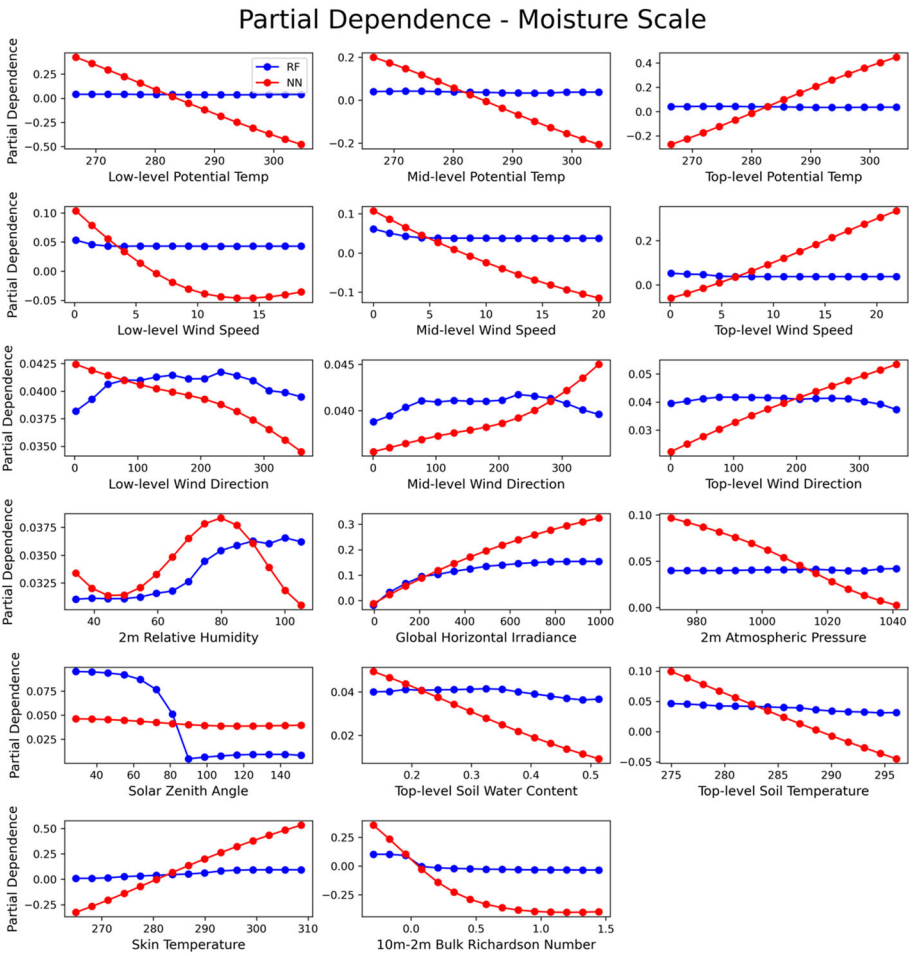

Fig. 14 Predictor partial dependence plots for the RF moisture scale predictions on the Cabauw dataset utilizing all common variables and the 10-min average fluxes

7 Conclusions and Future Work

Monin–Obukhov similarity theory is a semi-empirical theory relating wind and virtual potential temperature profiles to surface fluxes in ABLs, and it commonly provides lower boundary conditions in atmospheric flow models. Although MOST is based on the assumptions of horizontal homogeneity and stationarity, it is used over a wide range of atmospheric conditions. Recent proposals based on theoretical considerations, analysis of observations, and high-resolution simulations could extend MOST to a wider range of conditions. In this study we presented a complementary approach based on applying machine learning. We have shown that both RF and ANN machine-learning models have the potential to improve upon traditional similarity theory model, i.e. MOST. The advantage of the machine-learning approach is that, given appropriate training data, it can be extended to non-homogeneous and non-stationary conditions, since as recently stated by Hicks and Baldocchi (2020), "The matter

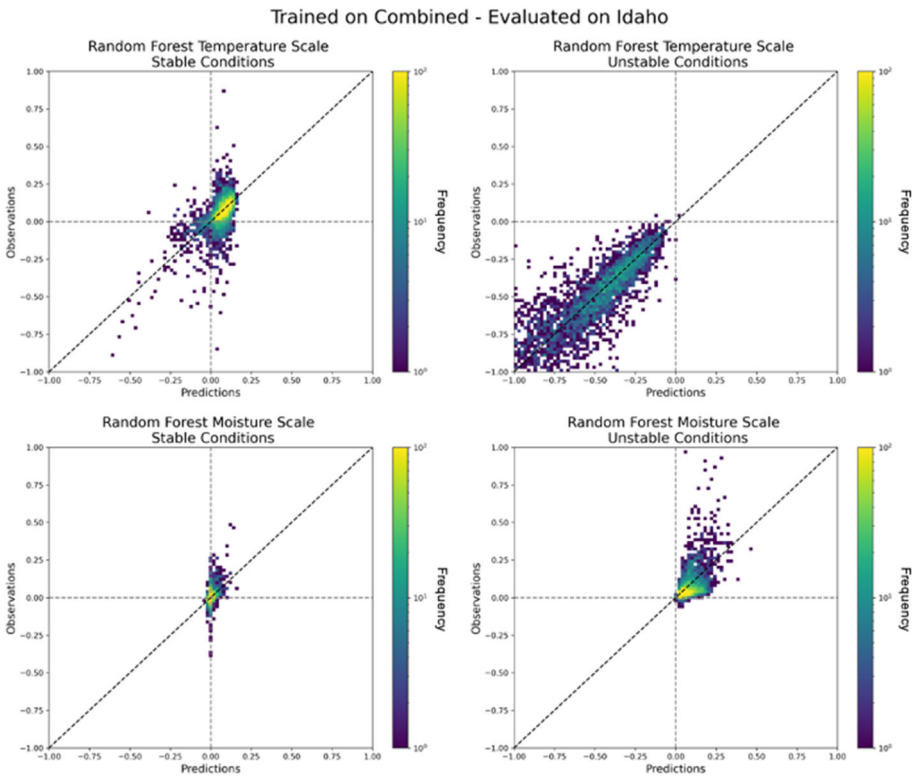

Fig. 15 Evaluation of results for the random forest on the Idaho dataset for the virtual potential temperature scale (top) and moisture scale (bottom) for stable conditions (left) and unstable conditions (right). The RF model performs better in the unstable cases than the stable cases

of fluxes over complex terrain remains unanswered". We have also shown that the machine-learning approaches have the ability to generalize well as highlighted by improved accuracy over MOST even when trained for one site and applied to another. The challenge is that a generalizable machine-learning model requires a collection of long-term, quality-controlled, complete, and consistent observations at diverse locations. While there are a number of excellent episodic field studies focused on specific processes, they do not provide sufficiently long records required for effective training of machine-learning models capable of capturing seasonal variability. Long-term observational networks such as AmeriFlux (Novick et al. 2018) or FLUXNET (Pastorello et al. 2020) focused on measurement of surface fluxes of CO2, methane, water, and energy may not include measurements of all the variables used to train machine-learning models in this study. Furthermore, the variables required to estimate surface fluxes in numerical weather prediction, such as skin temperature, are frequently not observed. The sites with sufficient data measured consistently are limited and even for the two sites that were used in this study, the differences in variables measured at the two sites have limited some of the analysis. For instance, it would have been beneficial to have all the measurements at the same levels with respect to the ground as well as measurements of relative humidity at multiple levels in the Idaho dataset. Furthermore, it is possible that using derived or estimated quantities, rather than only directly observed quantities, may be

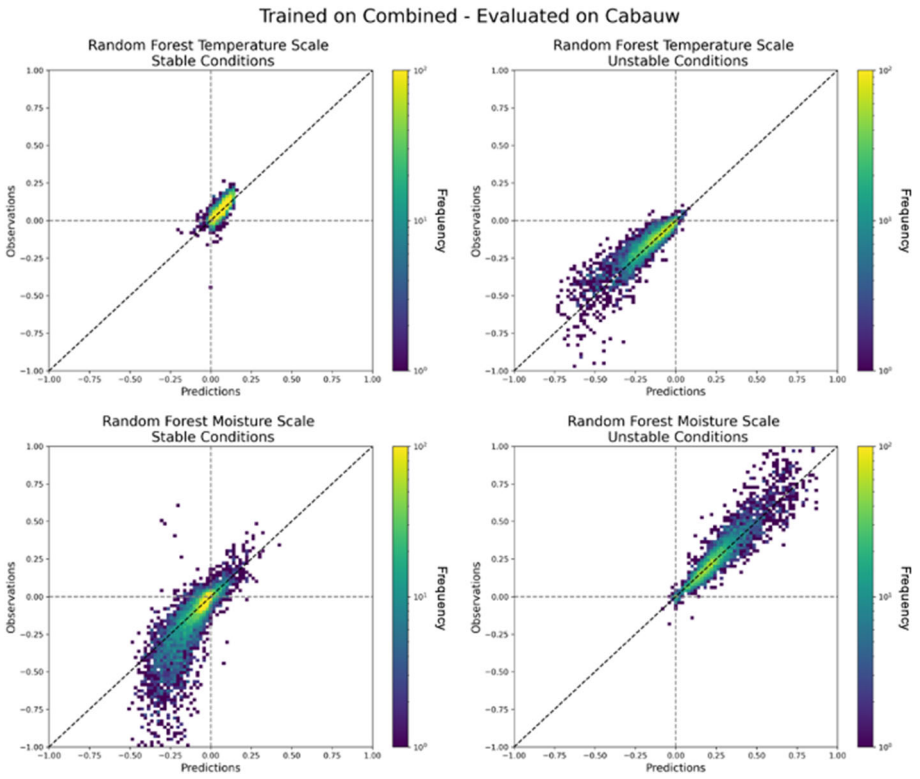

Fig. 16 Evaluation of results for the random forest on the Cabauw dataset for the temperature scale (top) and moisture scale (bottom) for stable conditions (left) and unstable conditions (right). The RF model performs better in the unstable cases than the stable cases

beneficial; however, this can also introduce additional uncertainties and errors. Extending the applicability of machine-learning models for surface parametrization to general topography and land use conditions in future studies would require determining the minimum set of required observations. Such a study was beyond current scope and it will be pursued in the future.

Model interpretability allows us to better understand the differences between the RF and ANN approaches at the two sites. The interpretability results show logical relationships among the predictors and surface friction velocity, virtual potential temperature scale, and moisture scale that align with our understanding of surface-layer processes. The results differ by stability (as determined by bulk Richardson number), indicating where MOST and machine-learning techniques (as applied here) have their strengths and weaknesses.

Ultimately, we can build machine-learning models to adapt to physics-based parametrizations implemented in a specific atmospheric model, such as the Weather Research and Forecasting (WRF) model. However, we should not be constrained by idiosyncrasies of a specific model's design. For example, implementing a machine-learning model for surface-layer parametrization in WRF does not guarantee that the benefits of improved surface flux estimation will affect the prediction of ABL structure. Boundary-layer parametrizations in the WRF model receive surface fluxes through the land-surface model, which relies on MOST. This

means that in order to benefit from the improved estimation of surface fluxes, the entire chain of parametrizations from land surface, through surface layer, to boundary-layer parametrization, would have to be modified to be internally consistent with the new parameterization. If atmospheric models were rewritten to take advantage of such machine-learning surface-layer approaches in a consistent manner, we would expect improvement in model flux predictions. We foresee future developments to improve the implementation of the machine-learning-model framework for surface-layer parametrization in numerical weather prediction and climate models that do not depend on the assumptions of MOST and are consistent across all the parametrizations.

Acknowledgements The authors would like to thank Peggy LeMone, Sukanta Basu, and an anonymous reviewer for their helpful comments and suggestions. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. This research was also funded in part by the U.S. Department of Energy through Pacific Northwest National Laboratory under Award No. 001657-00001. We acknowledge the Royal Netherlands Meteorological Institute for providing the Cabauw dataset and NOAA's Air Resources Laboratory for providing the Idaho data. We would also like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Appendix 1

We compute the surface friction velocity, virtual potential temperature scale, and moisture scale based on MOST using Dyer and Hicks (1970) stability functions.

The value of the von Kármán constant they used was $\kappa = 0.41$. After integrating Eq. 2, the wind speed difference between two levels, z_1 and z_0 for stably stratified boundary layers is

$$U_1 - U_0 = \frac{u_*}{\kappa}\left[\ln\frac{z_1}{z_0} + 5\frac{z_1}{L}\right], \tag{16}$$

while, for convective boundary layers, it is

$$U_1 - U_0 = \frac{u_*}{\kappa}\left[\ln\frac{z_1}{z_0} - 2\ln\left(\frac{1+\mathrm{x}}{2}\right) - \ln\left(\frac{1+x^2}{2}\right) + 2\tan^{-1}x - \frac{\pi}{2}\right], \tag{17}$$

where

$$x = \left(1 - 16\frac{z_1}{L}\right)^{\frac{1}{4}}. \tag{18}$$

For difference in virtual potential temperature at two levels we can obtain similar relationships for stably stratified conditions,

$$\Theta_{v1} - \Theta_{v0} = \frac{\theta_*}{\kappa}\left[\ln\frac{z_1}{z_{0T}} + 5\frac{z_1}{L}\right], \tag{19}$$

and for unstable conditions,

$$\Theta_{v1} - \Theta_{v0} = \frac{\theta_*}{\kappa}\left[\ln\frac{z_1}{z_{0T}} - 2\ln\left(\frac{1+x^2}{2}\right)\right]. \tag{20}$$

It should be pointed out that closing the surface energy budget under nearly ideal conditions characterized by stationarity and horizontal homogeneity remains an open problem. Sun et al. (2020) argued that the work by vertical density fluxes must be accounted for to close the budget, while in a recent review, Mauder et al. (2020) identify sub-mesoscale transport as the reason for non-closure.

Finally, the expression for the moisture mixing ratio difference has the same form as the one for virtual potential temperature. Therefore, for the stably stratified boundary layer the difference in moisture mixing ratio at two levels in a surface layer can be estimated using the following equation:

$$q_1 - q_0 = \frac{q_*}{\kappa}\left[\ln\frac{z_1}{z_0} - 5\frac{z_1}{L}\right], \tag{21}$$

while for the convective ABL the difference is

$$q_1 - q_0 = \frac{q_*}{\kappa}\left[\ln\frac{z_1}{z_0} - 2\ln\left(\frac{1+x^2}{2}\right)\right]. \tag{22}$$

Equations 16, 17, 18, 19, 20, 21 and 22 are used to compute surface friction velocity, u_*, virtual temperature scale, θ_*, and moisture scale, q_*, under stably stratified and convective conditions.

Appendix 2

We trained the ANN model using 20 epochs. Figure 17 depicts the training and validation loss functions mean-square error (MSE) for the friction velocity, the temperature scale, and the moisture scale. The loss functions for all three scales level off after 20 epochs. There is a

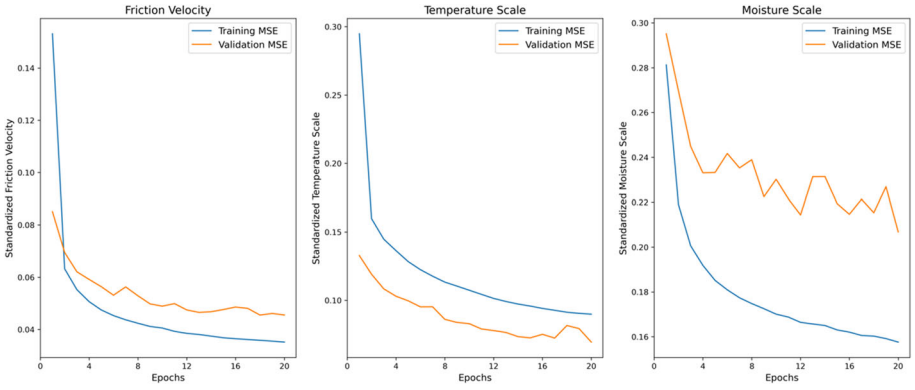

Fig. 17 The training (blue) and validation loss function (red) MSE for the friction velocity (left panel), the temperature scale (middle panel), and the moisture scale (right panel)

potential for further slight performance gains with longer training, but we do not expect the conclusions of the evaluation or interpretation to change significantly.

References

Andreas E, Claffey KJ, Jordan RE, Fairall CW, Guest PS, Persson OG, Grachev AA (2006) Evaluations of the von Kármán constant in the atmospheric surface layer. J Fluid Mech 559:117–149

Andreas E (2009) A new value of the von Kármán constant: implications and implementation. J App Meteorol Climatol 48:923–944

Basu S (2019) Hybrid profile-gradient approaches for the estimation of surface fluxes. Boundary-Layer Meteorol 170:29–44. https://doi.org/10.1007/s10546-018-0391-1

Beljaars ACM, Holtslag AAM (1991) Flux parameterization over land surfaces for atmospheric models. J Appl Meteorol 30(3):327–341

Breiman L (2001) Random forest. Mach Learn 45:5–32

Breiman L, Friedman JH, Stone CI, Olshen RA (1984) Classification and Regression Trees. Chapman and Hall/CRC p 368

Bosveld FC, Baas P, Beljaars AC, Holtslag AAM, Vilà-Guerau de Arellano J, van de Wiel BJH (2020) Fifty Years of Atmospheric Boundary-Layer Research at Cabauw ServingWeather, Air Quality and Climate. Boundary-Layer Meteorol 177:583–612. https://doi.org/10.1007/s10546-020-00541-w

Businger J, Wyngaard JC, Izumi Y, Bradley EF (1971) Flux-profile relationships in the atmospheric surface layer. J Atmos Sci 28:181–189

Chen F, Janjic Z, Mitchell K (1997) Impact of atmospheric surface layer parameterization in the new land-surface scheme of the NCEP Mesoscale Eta numerical model. Boundary-Layer Meteorol 185:391–421

Chen F, Zhang Y (2009) On the coupling strength between the land surface and the atmosphere: from viewpoint of surface exchange coefficients. Geophys Res Lett 36:L10404. https://doi.org/10.1029/2009GL037980

Chollet F (2015) Keras. https://github.com/fchollet/keras

Ding M, Tong C (2021) Multi-point monin-obukhov similarity of turbulence cospectra in the convective atmospheric boundary layer. Boundary-Layer Meteorol 178:185–199. https://doi.org/10.1007/s10546-020-00571-4

Dyer AJ, Hicks BB (1970) Flux-gradient relationships in the constant flux layer. Q J R Meteorol Soc 96:715–721

Dyer AJ (1974) A review of flux-profile relationships. Boundary-Layer Meteorol 7:363–372

Finn D, Clawson KL, Eckman RM, Carter RG, Rich JD, Reese BR, Beard SA, Brewer M, Davis D, Clinger D, Gao Z, Liu H (2017) Project sagebrush phase 2. In: NOAA technical memorandum OAR ARL-275, Air Resources Laboratory, Idaho Falls, Idaho. https://doi.org/10.7289/V5/TM-OAR-ARL-275

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Foken T (2006) 50 years of the Monin-Obukhov similarity theory. Boundary-Layer Meteorol 119:431–447. https://doi.org/10.1007/s10546-006-9048-6

Gagne DJ, Christensen HM, Subramanian AC, Monahan AH (2020) Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. J Adv Model Earth Syst. https://doi.org/10.1029/2019MS001896

Gagne DJ, McGovern A, Haupt SE, Williams JK (2017) Evaluation of statistical learning configurations for gridded solar irradiance forecasting. Sol Energy 150:383–393. https://doi.org/10.1016/j.solener.2017.04.031

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer-Verlag, New York

Hicks BB (1978) Some limitations of dimensional analysis and power laws. Boundary-Layer Meteorol 14:567–569

Hicks BB, Pendergrass WR III, Vogel CA, Keener RN, Leyton SM (2014) On the micrometeorology of the southern great plains 1: legacy Relationships Revisited. Boundary-Layer Meteorol 151:389–405. https://doi.org/10.1007/s10546-013-9902-2

Hicks BB, Baldocchi DD (2020) Measurement of fluxes over land: capabilities, origins, and remaining challenges. Boundary-Layer Meteorol 177:365–394. https://doi.org/10.1007/s10546-020-00531-y

Herman GR, Schumacher RS (2018) Money doesn't grow on trees, but forecasts do: forecasting extreme precipitation with random forests. Mon Weather Rev 146:1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8

James G, Witten D, Hastie T, Tibshirani R (2021) An introduction to statistical learning with applications in R. Springer, New York
Jellen C, Burkhardt J, Brownell C, Nelson C (2020) Machine learning informed predictor importance measures of environmental parameters in maritime optical turbulence. Appl Opt 59:6379–6389
Jimenez PA, Dudhia J, González-Rouco JF, Navarro J, Montávez JP, García-Bustamante E (2011) A revised scheme for the WRF surface layer formulation. Mon Wea Rev 140:848–918
Kelly M, Jørgensen HE (2017) Statistical characterization of roughness uncertainty and impact on wind resource estimation. Wind Energ Sci 2:189–209. https://doi.org/10.5194/wes-2-189-2017
Khanna S, Brasseur JG (1997) Analysis of Monin-Obukhov similarity from large-eddy simulation. J Fluid Mech 345:251–286
Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv [cs.LG]
Klambauer G, Unterthiner T, Mayr A, Hochreiter S (2017) Self-normalizing neural networks. arXiv [cs.LG]
Klipp C, Mahrt L (2004) Flux-gradient relationship, self-correlation and intermittency in the stable boundary layer. Q J R Meteorol Soc 130:2087–2103
LeMone MA, Grossman R, Coulter R, Wesely M, Klazura G, Poulos G, Blumen W, Lundquist J, Cuenca R, Kelly S, Brandes E, Oncley S, McMillen R, Hicks B (2000) Land-atmosphere interaction research, early results, and opportunities in the Walnut River Watershed in Southeast Kansas: CASES and ABLE. Bull Am Meteorol Soc 81:757–780
Li Q, Gentine P, Mellado JP, McColl KA (2018) Implications of nonlocal transport and conditionally averaged statistics on Monin-Obukhov similarity theory and townsend's attached eddy hypothesis. J Atmos Sci 75:3403–3431. https://doi.org/10.1175/JAS-D-17-0301.1
Lothon M, Lohou F, Pino D, Couvreux F, Pardyjak ER, Reuder J, Vilà-Guerau de Arellano J, Durand P, Hartogensis O, Legain D, Augustin P, Gioli B, Lenschow DH, Faloona I, Yagüe C, Alexander DC, Angevine WM, Bargain E, Barrié J, Bazile E, Bezombes Y, Blay-Carreras E, van de Boer A, Boichard JL, Bourdon A, Butet A, Campistron B, de Coster O, Cuxart J, Dabas A, Darbieu C, Deboudt K, Delbarre H, Derrien S, Flament P, Fourmentin M, Garai A, Gibert F, Graf A, Groebner J, Guichard F, Jiménez MA, Jonassen M, van den Kroonenberg A, Magliulo V, Martin S, Martinez D, Mastrorillo L, Moene AF, Molinos F, Moulin E, Pietersen HP, Piguet B, Pique E, Román-Cascón C, Rufin-Soler C, Saïd F, Sastre-Marugán M, Seity Y, Steeneveld GJ, Toscano P, Traullé O, Tzanos D, Wacker S, Wildmann N, Zaldei A (2014) The BLLAST field experiment: Boundary-Layer Late Afternoon and Sunset Turbulence. Atmos Chem Phys 14:10931–10960. https://doi.org/10.5194/acp-14-10931-2014
Mauder M, Foken T, Cuxart J (2020) Surface-energy-balance closure over land: a review. Boundary-Layer Meteorol 177:395–426. https://doi.org/10.1007/s10546-020-00529-6
McGovern A, Lagerquist R, John Gagne D, Jergensen GE, Elmore KL, Homeyer CR, Smith T (2019) Making the black box more transparent: Understanding the physical implications of machine learning. Bull Am Meteorol Soc 100:2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1
Newman JF, Klein PM (2014) The impacts of atmospheric stability on the accuracy of wind speed extrapolation methods. Resources 3(1):81–105. https://doi.org/10.3390/resources3010081
Novick KA, Biederman JA, Desai AR, Litvak ME, Moore DJP, Scott RL, Torn MS (2018) The AmeriFlux network: a coalition of the willing. Agri for Meteorol 249:444–456
Obukhov AM (1946) Turbulentnost v temperaturnoj – neodnorodnoj atmosfere (Turbulence in an Atmosphere with a non-uniform Temperature). Trudy Inst Theor Geofiz AN SSSR 1:95–115
Owen PR, Thomson WR (1963) Heat transfer across rough surfaces. J Fluid Mech 15:321–334
Panofsky HA (1963) Determination of stress from wind and temperature measurements. Q J R Meteorol Soc 89:85–94
Paulson CA (1970) The mathematical representation of wind speed and temperature profiles in the unstable surface layer. J Appl Meteorol 9:857–861
Pastorello G, Trotta C, Canfora E et al (2020) The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci Data 7:225. https://doi.org/10.1038/s41597-020-0534-3
Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Re 12:2825–2830
Pelliccioni A, Poli U, Agnello P, Coni A (1999) Application of neural networks to model the Monin-Obukhov length and the mixed-layer height from ground-based meteorological data. Tran Eco Environ 29(1055):1064
Poulos GS, Blumen W, Fritts DC, Lundquist JK, Sun J, Burns SP, Nappo C, Banta R, Newsom R, Cuxart J, Terradellas E, Balsley B, Jensen M (2002) CASES-99: A Comprehensive Investigation of the Stable Nocturnal Boundary Layer. Bull Am Meteorol Soc 83:555–581
Reed DR, Marks RJ (1998) Neural smithing: supervised learning in feedforward artificial neural networks. MIT Press, Cambridge
Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psych Rev 65(6):386–408

Rogallo RS, Moin P (1984) Numerical simulation of turbulent flows. Ann Rev Fluid Mech 16:99–137
Salesky ST, Anderson W (2020) Coherent structures modulate atmospheric surface layer flux-gradient relationships. Phys Rev Lett 125:124501. https://doi.org/10.1103/PhysRevLett.125.124501
Salesky ST, Chamecki M (2012) Random errors in turbulence measurements in the atmospheric surface layer: implications for Monin–Obukhov similarity theory. J Atmos Sci 69:3700–3714. https://doi.org/10.1175/JAS-D-12-096.1
Stiperski I, Calaf M (2018) Dependence of near-surface similarity scaling on the anisotropy of atmospheric turbulence. Q J R Meteorol Soc 144:641–657. https://doi.org/10.1002/qj.3224
Stiperski I, Calaf M, Rotach M (2019) Scaling, anisotropy, and complexity in near-surface atmospheric turbulence. J Geophys Res Atmos 124:1428–1448. https://doi.org/10.1029/2018JD029383
Stiperski I, Chamecki M, Calaf M (2021) Anisotropy of unstably stratified near-surface turbulence. Bound-Layer Meteorol 180:363–384. https://doi.org/10.1007/s10546-021-00634-0
Sun J, Takle ES, Acevedo OC (2020) Understanding physical processes represented by the Monin–Obukhov bulk formula for momentum transfer. Bound-Layer Meteorol 177:69–95. https://doi.org/10.1007/s10546-020-00546-5
Tong C, Nguyen KX (2015) Multipoint Monin-Obukhov similarity and its application to turbulence spectra in the convective atmospheric surface layer. J Atmos Sci 72:4337–4348. https://doi.org/10.1175/JAS-D-15-0134.1
Tong C, Ding M (2020) Velocity-defect laws, log law and logarithmic friction law in the convective atmospheric boundary layer. J Fluid Mech. https://doi.org/10.1017/jfm.2019.898
Uttal T, Curry JA, McPhee MG, Perovich DK, Moritz RE, Maslanik JA, Guest PS, Stern L, Moore JA, Turenne R, Heiberg A, Serreze C, Wylie DP, Persson OG, Paulson CA, Halle C, Morison JH, Wheeler PA, Makshtas A, Welch H, Shupe MD, Intrieri JM, Stamnes K, Lindsey RW, Pinkel R, Pegau WS, Stanton TP, Grenfeld TC (2002) Surface Heat Budget of the Arctic Ocean. Bull Am Meteorol Soc 83:255–276
Wang Y, Basu S (2016) Using an artificial neural network approach to estimate surface-layer optical turbulence at Mauna Loa. Hawaii Opt Lett 41:2334–2337
Wieriga J (1993) Representative Roughness Parameters for Homogeneous Terrain. Boundary-Layer Meteorol 63:323–363
Yuval J, O'Gorman PA (2020) Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. Nat Commun 11:3295. https://doi.org/10.1038/s41467-020-17142-3
Zilitinkevich SS (1995) Non-local turbulent transport: pollution dispersion aspects of coherent structure of convective flows. Trans Eco Environ 6:53–60