



# Estimation of the Surface Fluxes for Heat and Momentum in Unstable Conditions with Machine Learning and Similarity Approaches for the LAFE Data Set

Volker Wulfmeyer<sup>1</sup> · Juan Manuel Valencia Pineda<sup>1</sup> · Sebastian Otte<sup>2</sup> · Matthias Karlbauer<sup>2</sup> · Martin V. Butz<sup>2</sup> · Temple R. Lee<sup>3</sup> · Verena Rajtschan<sup>4</sup>

Received: 29 April 2022 / Accepted: 21 October 2022 / Published online: 22 November 2022  
© The Author(s) 2022

## Abstract

Measurements of three flux towers operated during the land atmosphere feedback experiment (LAFE) are used to investigate relationships between surface fluxes and variables of the land–atmosphere system. We study these relations by means of two machine learning (ML) techniques: multilayer perceptrons (MLP) and extreme gradient boosting (XGB). We compare their flux derivation performance with Monin–Obukhov similarity theory (MOST) and a similarity relationship using the bulk Richardson number (BRN). The ML approaches outperform MOST and BRN. Best agreement with the observations is achieved for the friction velocity. For the sensible heat flux and even more so for the latent heat flux, MOST and BRN deviate from the observations while MLP and XGB yield more accurate predictions. Using MOST and BRN for latent heat flux, the root mean square errors (RMSE) are  $107 \text{ Wm}^{-2}$  and  $121 \text{ Wm}^{-2}$ , respectively, as well as the intercepts of the regression lines are  $\approx 110 \text{ Wm}^{-2}$ . For the ML methods, the RMSEs reduce to  $31 \text{ Wm}^{-2}$  for MLP and  $33 \text{ Wm}^{-2}$  for XGB as well as the intercepts to just  $4 \text{ Wm}^{-2}$  for MLP and  $-1 \text{ Wm}^{-2}$  for XGB with slopes of the regression lines close to 1, respectively. These results indicate significant deficiencies of MOST and BRN, particularly for the derivation of the latent heat flux. In fact, in contrast to the established theories, feature importance weighting demonstrates that the ML methods base their improved derivations on net radiation, the incoming and outgoing shortwave radiations, the air temperature gradient, and the available water contents, but not on the water vapor gradient. The results imply that further studies of surface fluxes and other turbulent variables with ML techniques provide great promise for deriving advanced flux parameterizations and their implementation in land–atmosphere system models.

---

✉ Volker Wulfmeyer  
volker.wulfmeyer@uni-hohenheim.de

<sup>1</sup> Institute of Physics and Meteorology, University of Hohenheim, Stuttgart, Germany

<sup>2</sup> Computer Science & Psychology, Neuro-Cognitive Modeling Group, University of Tübingen, Tübingen, Germany

<sup>3</sup> NOAA/Air Resources Laboratory Atmospheric Turbulence and Diffusion Division, Oak Ridge, Tennessee, USA

<sup>4</sup> Institute of Soil Sciences and Land Evaluation, University of Hohenheim, Stuttgart, Germany

**Keywords** Monin–Obukhov theory · Bulk Richardson number · Machine learning · Multilayer perceptron · Extreme gradient boosting

## 1 Introduction

In numerical weather prediction (NWP), seasonal forecast, and climate models, unresolved atmospheric variables close to the Earth's surface and surface fluxes must be parameterized. Parameterizations of gradients and fluxes in the atmospheric surface layer (SL) are fundamental for an appropriate representation of land–atmosphere (L–A) or ocean–atmosphere interactions and, thus, accurate forecasts and simulations over all forecast ranges. The SL ranges up to approx. 10 to 100 m above the land surface and is generally stable during nighttime or over surfaces that are colder than the atmosphere. This leads to a reduction of surface fluxes and different mechanisms of vertical exchange such as intermittent turbulence induced by gravity waves. Over most ocean and land surfaces during daytime, in contrast, the SL is unstable resulting in an increase of vertical fluxes due to dynamical turbulence or turbulence due to surface heating. Typically, vertical resolutions of approx. 1 m for resolving gradients and temporal resolutions of 10 Hz are required to derive surface fluxes. Therefore, the resolutions of all the model systems mentioned above and not even the typical configurations of large eddy simulation (LES) are sufficient to resolve fluxes and gradients. Only special LES designs on the m-scale are appropriate, but these are computationally very demanding so that only a few case studies can be performed and evaluated (e.g., Maronga 2014; Maronga and Reuder 2017; Basu and Lacser 2017).

Corresponding research and understanding of flux–gradient relationships has a long tradition in atmospheric sciences. Pioneering studies were performed by Prandtl in the 1920s (Prandtl 1925). A huge step forward in the understanding of the parameterization of fluxes and gradients was the Monin–Obukhov similarity theory (MOST, Monin and Obukhov 1954). Here, an Obukhov length  $L$  was proposed, which is used as a scaling parameter in flux–gradient similarity functions. An overview of MOST is presented in Foken (2006). An alternative approach was proposed by Deardorff (1972), who suggested a scaling of similarity functions using the bulk Richardson number (BRN). This approach was adopted by Sorbjan (2006, 2010) and Mauritsen et al. (2007).

In almost all model systems, parameterizations based on MOST are implemented. This is due to the absence of alternatives that can easily be implemented in the model codes and measurement campaigns such as the Wangara (Hess et al. 1981) and the Kansas experiments performed in 1967 and 1968, respectively. During these experiments, measurements in the SL using on a combination of high-resolution sonic anemometers, hot wire thermometers, and Lyman Alpha water vapor sensors provided the basis for determination of the stability dependence of dimensionless gradient functions and indications of a universality of the results (Businger et al. 1971; Hicks 1976; Högström 1996). In addition, the experimental data provided estimations of some fundamental constants of turbulent flow in the ABL, such as the Von Kármán constant.

In spite of the apparent success of MOST and its implementations in models, recent research results indicate severe deviations with respect to observations. This can be due to self-correlations as well as incorrect assumptions and structural deficiencies of the proposed similarities. Additionally, as surface fluxes are the result of L–A feedback processes (Santanello et al. 2018), the scaling of surface fluxes should also depend on the heterogeneity of the land surface (Morrison et al. 2021), induced micro- and mesoscale circulations (Li et al.

2018; McNaughton and Brunet 2002), and the evolution of the atmospheric boundary layer (ABL) (van Heerwaarden et al. 2009). Therefore, it can be expected that the MOST-based scaling functions are not universal, or even incorrect, due to inappropriate assumptions of their shapes and scaling. Also, the omission of land surface heterogeneities as well as of reasonable ABL scaling variables such as the ABL depth  $z_i$  (van de Boer et al. 2014; Cheng et al. 2021) or entrainment fluxes (van Heerwaarden et al. 2009) can lead to errors in the determination of surface fluxes.

Therefore, it is essential to intensify research on flux derivations in the SL over various land cover types. Particularly, machine learning (ML) approaches hold great promise because of their capability to detect nonlinear relationships in large data sets without any constraints by the similarity relationships and self-correlations of variables prescribed in MOST and BRN. An overview of ML methods in L–A system research is given in Zhang (2008); Pal and Sharma (2019).

The application of ML methods for investigating surface fluxes is rapidly expanding. For instance, Qin et al. (2005a, b) applied variants of neural networks to study latent heat and CO<sub>2</sub> fluxes over cropland. They were able to train the ML methods to achieve good agreement and correlations with the observations. They also performed sensitivity analyses with respect to the observational inputs to the ML methods. However, they did not apply tower data and did not perform comparisons with MOST or BRN. Momentum and sensible heat fluxes were not studied either; the same limitations apply to the latent heat flux predictions by Wang et al. (2017); Xu et al. (2018); Wang et al. (2021). Safa et al. (2018) investigated sensible and latent heat fluxes with a multilayer perceptron (MLP) as the ML method including sensitivity analyses based on multiyear data sets of fluxes and L–A variables over maize. Momentum fluxes were not investigated. They demonstrated promising training results; however, Safa et al. (2018) neither investigated temperature nor moisture gradients in the SL nor compared the results with MOST or BRN. Leufen and Schädler (2019) used an extensive set of tower data in combination with MOST and an MLP to study surface momentum and sensible heat fluxes. After training the MLP using the data of one tower and application of the training results to an independent data set of another tower, Leufen and Schädler (2019) found comparable performance of MOST and MLP. Latent heat fluxes were not investigated.

In this study, we use data of three towers at two height levels operated during the land–atmosphere feedback experiment (LAFE) (Wulfmeyer et al. 2018) over vegetated surfaces for a duration of one month in August 2017. LAFE was performed at the Southern Great Plains (SGP) site of the US Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) program. We use the measurements of the surface friction velocity as well as the sensible and latent heat fluxes and compare these with MOST and BRN as well as two different machine learning (ML) methods without the consideration of additional ABL variables. For the ML methods, we chose multilayer perceptrons (MLP) (Goodfellow et al. 2016) and extreme gradient boosting (XGB) (Chen and Guestrin 2016).

We focus on the daytime unstable ABL over land. In contrast to the former well-established approaches, the ML models do not rely on explicit theory-driven formulations. Instead their predictions are primarily derived from the available data; they essentially learn relations and patterns within the available measurements of gradients and other variables in the L–A system.

In this work, based on our observations and the methods to evaluate similarity relationships, we focus on the following questions based on the LAFE data:

1. How accurate are the fluxes derived by MOST and BRN using appropriate data analyses?
2. How do these compare with the output of the ML techniques?

3. Can we use ML techniques to generate improved results and to identify of the most important drivers of the fluxes?

Please note that it is not the scope of this work to claim a universality of our results but to compare MOST, BRN, and ML for a specific location and time period. After a short introduction to the MOST and the BRN relationships, we present the ML approaches used in this work. Then, we give an overview of LAFE as well as the set up and operation of the three towers. We explain the training and the derivation of the results based on MOST, BRN, and the ML methods and compare their outputs. We summarize the results focusing on the future potential of ML with respect to the derivation of flux relationships, process understanding, and its implementation in earth system models.

## 2 Principles of the Parameterization of Surface Fluxes

### 2.1 General Principles of the Derivation of Surface Flux Similarity Relationships

Surface flux similarities are fundamental for the understanding of L–A exchange and form the basis for parameterizations in models. Typically, the most common form of first order closure is applied, which assumes that the surface fluxes are mainly dependent on the gradients of the transport variable. Within the SL, where the fluxes are determined and the gradients are derived, the fluxes are considered as constant with height (constant flux layer). Using Buckingham's  $\pi$  analyses, a set of surface scaling variables can be derived, which are expected to control this relationship completely. Flux–gradient relationships hold for typical temporal averaging times of 30–60 min so that enough eddies can be sampled to derive a reasonable flux with small error bars as well as meso- and microscale circulations can be separated from the turbulent fluctuations. It is assumed that the turbulence during this time period is quasi-stationary. As each flux value corresponds to a certain spatial footprint, the soil and the land cover properties are considered to be homogeneous over this area. We show that this is also the case for our measurements (see below for further discussions). Furthermore, we disregard potential dependencies of surface fluxes on ABL evolution. Therefore, we use the standard equations of the MOST and the BRN approach as well as we only apply soil, land cover, and SL variables for the ML approaches.

For applying MOST or BRN similarities, it turned out that flux relationships containing just a combination of surface scaling variables are not successful. It is necessary to combine these with similarity functions that consider the effects of the SL stability by additional scaling variables such as the Obukhov length  $L$  or the bulk Richardson number  $Ri_b$ . These additional relationships introduce self-correlations of the results, which may imply apparent artificial relations. Nevertheless, we disregard these effects in our study.

Currently, there is no general theory available to derive these similarity functions so that these are determined by observations or large eddy simulation (LES) with ultra-high resolution (Maronga and Reuder 2017). In the following, we apply the current state-of-the-art formulations of these similarity functions based on previous experiments derived under daytime unstable conditions.

### 2.2 Monin–Obukhov Similarity Theory

In this case, Monin and Obukhov proposed for the vertical gradient of the SL wind  $U(z)$ , potential temperature  $\theta(z)$ , and specific humidity  $q$  profiles:

$$\hat{U} := \frac{\partial U}{\partial z} \kappa (z - d) = u_* \phi_m \left( \frac{z - d}{L} \right), \quad (1)$$

$$\hat{\theta} := \frac{\partial \theta}{\partial z} \kappa (z - d) = -\frac{H_0}{u_*} \phi_h \left( \frac{z - d}{L} \right), \quad (2)$$

$$\hat{q} := \frac{\partial q}{\partial z} \kappa (z - d) = -\frac{Q_0}{u_*} \phi_q \left( \frac{z - d}{L} \right), \quad (3)$$

$$\text{with } u_* = \left\{ (\overline{u'w'})^2 + (\overline{v'w'})^2 \right\}^{1/4}, \quad (4)$$

$$\text{and } L = -\frac{\overline{\theta_V} u_*^3}{\kappa g w' \theta'_V}, \quad (5)$$

where  $U$  is the horizontal wind speed,  $\kappa \simeq 0.4$  is the von Karman constant,  $z$  is height above ground level (AGL),  $d = d_z + z_0$  is the displacement height, where  $d_z$  is the zero-plane displacement and  $z_0$  is the roughness length,  $u_*$  is the friction velocity,  $\theta$  is potential temperature,  $\theta_V$  is virtual potential temperature,  $H_0 = \overline{w'\theta'}$  is the surface kinematic heat flux,  $q$  is specific humidity, and  $Q_0 = \overline{w'q'}$  is the surface kinematic water vapor flux. The overline indicates the covariance of turbulent quantities averaged over the time period of interest.  $u'$ ,  $v'$ , and  $w'$  are the zonal, meridional, and vertical fluctuations of the three wind components.  $\theta'$  and  $q'$  are the fluctuations of potential temperature and specific humidity, respectively. Except the other variables introduced above,  $L$  is expressed by the mean  $\theta_V$  in the SL, the acceleration due to gravity  $g$ , and the surface virtual heat flux  $H_{0,V} = \overline{w'\theta'_V} \simeq H_0 + 0.61\overline{T}Q_0$  where  $\overline{T}$  is the mean temperature in the SL.

$\phi_i$  stand for the similarity functions that need to be derived to correct the flux–gradient relationships with respect to atmospheric stability and the strength of turbulence in the SL. In MOST, these functions scale with the dimensionless variable  $\zeta = (z - d)/L$ . The structure of the MOST equation shows a critical self-correlation due to the presence of  $u_*$  in both terms on the right side of Eq. 1 as well as due to the presence of fluxes and  $u_*$  in both terms on the right sides of Eqs. 2 and 3 (see Andreas and Hicks 2002).

As mentioned above, we disregard these effects here and use these relations to compare the observations with the theoretical expectations from MOST. For this purpose, the similarity functions in the following forms were used for the daytime unstable surface layer:

$$\frac{\hat{U}}{u_*} = \phi_m \left( \frac{z - d}{L} \right) = b_m \left( 1 - a_m \frac{z - d}{L} \right)^{-\alpha}, \quad (6)$$

$$-\frac{\hat{\theta} u_*}{H_0} = \phi_h \left( \frac{z - d}{L} \right) = b_h \left( 1 - a_h \frac{z - d}{L} \right)^{-\beta}, \quad (7)$$

$$-\frac{\hat{q} u_*}{Q_0} = \phi_q \left( \frac{z - d}{L} \right) = b_q \left( 1 - a_q \frac{z - d}{L} \right)^{-\gamma}. \quad (8)$$

Similar functional relationships are used in Dyer and Hicks (1970); Dyer (1974) and in the parameterization of surface fluxes in the Weather Research and Forecasting (WRF) model (Jiménez et al. 2012).

The challenge is to relate the measurements at the different heights either with gradients at specific heights or, in order to circumvent this, to use the integrated functions providing equations for the wind, temperature, and humidity profiles.

### 2.2.1 Using the Gradient Functions to Derive Fluxes

If the gradient functions are used directly to find relationships to the fluxes, analytical relationships can be derived with some approximations or the full set of the three equations is used by iterative methods. For instance, using these three similarity functions (Eqs. 6–8) with  $a := a_m \approx a_h \approx a_q$ , we can find the following relationships between gradient measurements and the fluxes:

$$u_* = \frac{\hat{U}}{b_m} \left( 1 - a \frac{z-d}{L} \right)^\alpha, \tag{9}$$

$$H_0 = -\frac{u_* \hat{\theta}}{b_h} \left( \frac{u_* b_m}{\hat{u}} \right)^{\beta/\alpha}, \tag{10}$$

$$Q_0 = -\frac{u_* \hat{q}}{b_q} \left( \frac{u_* b_m}{\hat{u}} \right)^{\gamma/\alpha}, \tag{11}$$

$$L = \frac{\bar{\theta}_v}{\kappa g \hat{\theta}} \frac{u_*^2}{\left\{ \frac{1}{b_h} \left( \frac{u_* b_m}{\hat{u}} \right)^{\beta/\alpha} + 0.61 \frac{\bar{T}}{\hat{\theta}} \frac{\hat{q}}{b_q} \left( \frac{u_* b_m}{\hat{u}} \right)^{\gamma/\alpha} \right\}}. \tag{12}$$

This system of equations can be solved implicitly by inserting Eq. 12 in 9 and proceeding consecutively to Eqs. 10 and 11.

Particularly, if we consider from previous studies that  $\beta \simeq \gamma \simeq 2\alpha$ , we achieve even an analytical solution of these equations starting with:

$$L = \frac{\bar{\theta}_v}{\kappa g \hat{\theta}} \frac{\hat{u}^2}{b_m^2 \left( \frac{1}{b_h} + 0.61 \frac{\bar{T}}{\hat{\theta}} \frac{\hat{q}}{b_q} \right)}. \tag{13}$$

However, there is a severe difficulty here, which makes the use of gradients challenging. If only tower measurements at two heights are available and their differences are used, the slope of these secants do not agree with the slope of the gradient function at the arithmetic mean between these two heights due to its strong nonlinearity. Of course, due to the mean value theorem of integral calculus, within the range of heights spanned by a secant, there is always a certain height (but not at the arithmetic mean) where the gradient function agrees with the secant. However, this height depends on  $L$  and can only be found, if the integrated gradient functions are derived and evaluated. Therefore, the gradients at the arithmetic mean between the two heights cannot be replaced by the secants, and if this was done, this would result in systematic errors of the fluxes. However, these solutions can be applied for scanning lidar systems and fiber-based sensors that are capable to resolve gradients of atmospheric variables (e.g., Späth et al. (2022)).

### 2.2.2 Using the Integrated Functions for the Derivation of Fluxes

This is the common approach, we also applied here for flux estimations. According to the overview in, e.g., Lee and Buban (2020), it is appropriate to set  $\alpha \approx 0.25$  and  $\beta \approx \gamma \approx 0.5$ .

Using these exponents, the integrals can be resolved analytically and we achieve:

$$\begin{aligned}
 U(z) = & -4 \frac{b_m u_*}{\kappa} \left[ 0.5 \arctan \frac{1}{\left(1 - \frac{a_m (z-d)}{L}\right)^{0.25}} \right. \\
 & + 0.25 \ln \left\{ \frac{1}{\left(1 - \frac{a_m (z-d)}{L}\right)^{0.25} + 1} \right\} \\
 & \left. - 0.25 \ln \left\{ 1 - \frac{1}{\left(1 - \frac{a_m (z-d)}{L}\right)^{0.25}} \right\} \right] + C, \tag{14}
 \end{aligned}$$

$$\begin{aligned}
 \theta(z) = & \frac{b_h H_0}{u_* \kappa} \left[ \ln \left( \sqrt{1 - \frac{a_h (z-d)}{L}} + 1 \right) \right. \\
 & \left. - \ln \left( \sqrt{1 - \frac{a_h (z-d)}{L}} - 1 \right) \right] + D, \tag{15}
 \end{aligned}$$

$$\begin{aligned}
 q(z) = & \frac{b_q Q_0}{u_* \kappa} \left[ \ln \left( \sqrt{1 - \frac{a_q (z-d)}{L}} + 1 \right) \right. \\
 & \left. - \ln \left( \sqrt{1 - \frac{a_q (z-d)}{L}} - 1 \right) \right] + E. \tag{16}
 \end{aligned}$$

for  $L < 0$  and  $z > d$ .

Therefore, the difference at two measurement heights can be expressed as:

$$\begin{aligned}
 U(z_2) - U(z_1) = & -4 \frac{b_m u_*}{\kappa} \left[ 0.5 \arctan \frac{1}{\left(1 - \frac{a_m (z_2-d)}{L}\right)^{0.25}} \right. \\
 & + 0.25 \ln \left\{ \frac{1}{\left(1 - \frac{a_m (z_2-d)}{L}\right)^{0.25} + 1} \right\} \\
 & - 0.25 \ln \left\{ 1 - \frac{1}{\left(1 - \frac{a_m (z_2-d)}{L}\right)^{0.25}} \right\} \\
 & \left. - 0.5 \arctan \frac{1}{\left(1 - \frac{a_m (z_1-d)}{L}\right)^{0.25}} \right]
 \end{aligned}$$

$$\begin{aligned}
 & -0.25 \ln \left\{ \frac{1}{\left(1 - \frac{a_m (z_1 - d)}{L}\right)^{0.25} + 1} \right\} \\
 & + 0.25 \ln \left\{ 1 - \frac{1}{\left(1 - \frac{a_m (z_1 - d)}{L}\right)^{0.25}} \right\} \Bigg], \tag{17}
 \end{aligned}$$

$$\begin{aligned}
 \theta(z_2) - \theta(z_1) = \frac{b_h H_0}{u_* \kappa} & \left[ \ln \left( \sqrt{1 - \frac{a_h (z_2 - d)}{L}} + 1 \right) \right. \\
 & - \ln \left( \sqrt{1 - \frac{a_h (z_2 - d)}{L}} - 1 \right) \\
 & - \ln \left( \sqrt{1 - \frac{a_h (z_1 - d)}{L}} + 1 \right) \\
 & \left. + \ln \left( \sqrt{1 - \frac{a_h (z_1 - d)}{L}} - 1 \right) \right], \tag{18}
 \end{aligned}$$

$$\begin{aligned}
 q(z_2) - q(z_1) = \frac{b_q Q_0}{u_* \kappa} & \left[ \ln \left( \sqrt{1 - \frac{a_q (z_2 - d)}{L}} + 1 \right) \right. \\
 & - \ln \left( \sqrt{1 - \frac{a_q (z_2 - d)}{L}} - 1 \right) \\
 & - \ln \left( \sqrt{1 - \frac{a_q (z_1 - d)}{L}} + 1 \right) \\
 & \left. + \ln \left( \sqrt{1 - \frac{a_q (z_1 - d)}{L}} - 1 \right) \right]. \tag{19}
 \end{aligned}$$

We used these equations for the derivations of the surface fluxes based on MOST. First of all, in order to optimize the comparisons, we fitted the combination of coefficients  $b_m$  and  $a_m$ ,  $b_h$  and  $a_h$  as well as  $b_q$  and  $a_q$  to the results. Afterward, the fluxes in Eqs. 14–16 were derived based on this adaptation of the similarity function coefficients.

In principle, these functions may also be used to find the respective heights for all variables where the gradients agree with the secants of the tower measurements at different heights but we keep these considerations for future efforts.

### 2.3 Bulk Richardson Number

Use of the BRN  $Ri_b$  has been proposed as an alternative to MOST (e.g., Sorbjan 2006, 2010; Mauritsen et al. 2007). Recent work has shown that using a  $Ri$ -based approach yields better predictions of near-surface gradients of wind, temperature, moisture, and heat fluxes than using the long-standing similarity relationships derived from MOST (Lee and Buban 2020; Lee et al. 2021; Lee and Meyers 2022). Furthermore, a  $Ri$ -based approach has the potential to avoid some of the long-known downsides associated with MOST such as a reduction of



self-correlation and also to be more computationally efficient. In the approach, local gradients present in the calculation of  $Ri$  are computed as bulk gradients such that the BRN,  $Ri_b$ , is computed as:

$$Ri_b = \frac{g \Delta \bar{\theta}_v \Delta z}{\bar{\theta}_v [(\Delta \bar{u})^2 + (\Delta \bar{v})^2]}. \quad (20)$$

In the above equation,  $\theta_v$  is the virtual potential temperature as well as  $u$  and  $v$  are the zonal and meridional wind components, respectively. Following Lee et al. (2021) and briefly summarized here, in the BRN approach,  $u_*$ ,  $H_0$ , and  $Q_0$  are computed, under unstable conditions, as (Panofsky et al. 1977):

$$u_* = U_z c_m [1 - d_m Ri_b]^{1/3}, \quad (21)$$

$$H_0 = -\Delta \theta u_* c_h [1 - d_h Ri_b]^{1/3}, \quad (22)$$

$$Q_0 = -\Delta q u_* c_q [1 - d_q Ri_b]^{1/3}. \quad (23)$$

In the above equations,  $U_z$  is the wind speed at height  $z$  as well as  $\Delta \theta$  and  $\Delta q$  are the differences in potential temperature and specific humidity between two sampling heights  $z_1$  and  $z_2$ , respectively.  $c_{u,h,q}$  and  $d_{u,h,q}$  are empirically determined coefficients that were found using the LAFE data set. More details appear in Lee et al. (2021).

### 3 Application of Machine Learning Analyses for Studying Flux Similarities

Data-driven ML algorithms can be understood as complementary, if not orthogonal, approaches to the theory-driven methods described earlier, i.e., MOST in Sect. 2.2 and BRN in Sect. 2.3. More precisely, an ML algorithm can learn a nonlinear mapping from a set of given input variables (input features), e.g., observational data, to predict a desired target variable. This mapping can be learned purely from data and without any prior knowledge about the real physical processes and the importance of individual input features. Once a model has been trained, it can be analyzed in terms of the importance that it attributes to the respective features.

These methods allow for studying possibly complex relationships between input and target variables without any prescribed relationships such as the similarity functions and the scaling variables for MOST and BRN. Moreover, we aim at using the ML results for improving the process understanding and advancing present theories.

In this work, we focus on predicting the following target variables:

$$u_* = F \left( U_{3m}, U_{10m}, \frac{U_{10m} - U_{3m}}{7 \text{ m}}, \theta_{2m}, \theta_{10m}, \frac{\theta_{10m} - \theta_{2m}}{8 \text{ m}}, \dots, R_{s,d}, A, R_{l,d}, R_{l,u}, R_{net}, T_s, \frac{T_{2m} - T_s}{2 \text{ m}} \right), \quad (24)$$

$$H_0 = F \left( U_{3m}, U_{10m}, \frac{U_{10m} - U_{3m}}{7 \text{ m}}, \theta_{2m}, \theta_{10m}, \frac{\theta_{10m} - \theta_{2m}}{8 \text{ m}}, \dots, R_{s,d}, A, R_{l,d}, R_{l,u}, R_{net}, T_s, \frac{T_{2m} - T_s}{2 \text{ m}}, T_{\text{soil},5 \text{ cm}}, \frac{T_s - T_{\text{soil},5 \text{ cm}}}{5 \text{ cm}}, AW \right), \quad (25)$$

$$Q_0 = F \left( U_{3m}, U_{10m}, \frac{U_{10m} - U_{3m}}{7 \text{ m}}, \theta_{2m}, \theta_{10m}, \frac{\theta_{10m} - \theta_{2m}}{8 \text{ m}}, \dots \right)$$

$$\begin{aligned}
 & q_{3m}, q_{10m}, \frac{q_{10m} - q_{3m}}{7 \text{ m}}, R_{s,d}, A, R_{l,d}, R_{l,u}, R_{net}, \dots \\
 & \left. T_s, \frac{T_{2m} - T_s}{2 \text{ m}}, T_{\text{soil},5 \text{ cm}}, \frac{T_s - T_{\text{soil},5 \text{ cm}}}{5 \text{ cm}}, AW \right). \tag{26}
 \end{aligned}$$

To facilitate the ML methods learning an effective input to output mapping, all non-target measurement variables from the LAFe were considered as inputs. More specifically, we used the following input features: wind speed at 3 m ( $U_{3m}$ ) and 10 m heights ( $U_{10m}$ ) along with the resulting gradients (or differences); wind direction at 10 m height; air and potential temperatures at 2 m ( $T_{2m}, \theta_{2m}$ ) and 10 m heights ( $T_{10m}, \theta_{10m}$ ) along with the gradients; five radiation terms (incoming short- and longwave radiations  $R_{s,d}$  and  $R_{l,d}$ , outgoing short- and longwave radiations  $R_{s,u}$  and  $R_{l,u}$  as well as the resulting net radiation  $R_n$  and the albedo  $A$ ); skin temperature  $T_s$ , soil temperatures in five cm depth ( $T_{\text{soil}}$ ), and the gradients from soil to skin and skin to air temperatures in 2 m height; specific humidity at 3 m ( $q_{3m}$ ) and 10 m height ( $q_{10m}$ ) along with the gradient; and the available water content  $AW$ . Separate models were trained to predict either friction velocity  $u_*$ , latent heat flux  $Q_0$ , or sensible heat flux  $H_0$ , as detailed in, e.g., Eqs. 24–26. More information about the recorded and used variables can be found in Sect. 4.

When choosing  $\Phi_p^y(\mathbf{x})$  as desired ML algorithm  $\Phi$  with learnable parameters  $p$  and input  $\mathbf{x}$  to predict the target variable  $y$ , the above equations transform into:

$$u_* \simeq \Phi_p^{u_*}(\mathbf{x}), \quad H_0 \simeq \Phi_p^{H_0}(\mathbf{x}), \quad \text{and} \quad Q_0 \simeq \Phi_p^{Q_0}(\mathbf{x}). \tag{27}$$

Usually,  $\mathbf{x}$  is an input vector that represents and holds all input features  $x_i$  for  $i \in \{1, \dots, n\}$ , where  $n$  is the number of input variables.

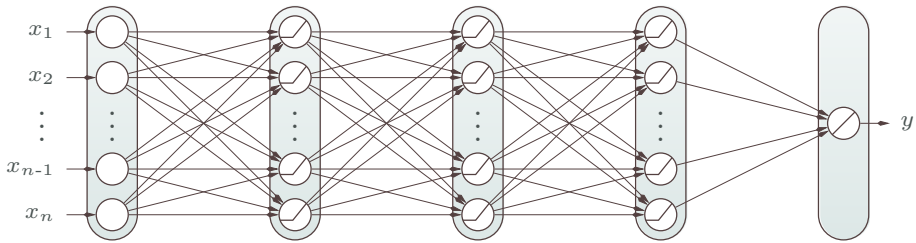
In this work, we train two ML approaches: a dedicated deep multilayer perceptron (MLP) model (Goodfellow et al. 2016) from the class of artificial neural networks and the extreme gradient boosting (XGB) method (Chen and Guestrin 2016). For both methods, we applied a threefold cross-validation to assess the model performance on unseen data. For this purpose, we used the data from two towers to train the ML models and tested these on the data from the respective third tower. Accordingly, to obtain predictions for all three towers, three models were trained for each target variable. Alternative data splits (e.g., 80% of data from all towers as train and the remaining 20% as test, or even training and testing on all data without split) were tested as well but not reported since the threefold cross-validation is considered the most robust method to prevent overfitting on the train data.

### 3.1 Deep Multilayer Perceptron

The first model we incorporated in our experiments is a deep fully connected MLP, see, e.g., Goodfellow et al. (2016). It consists of multiple layers of nonlinear neural processing units as shown in Fig. 1. In the following, we describe the computational scheme of the single neuron  $j$ :

$$x_j = \phi_j \left( \sum_i w_{ij} x_i + b_j \right) \quad \text{with, e.g.,} \quad \phi_j(u) \begin{cases} u & \text{if } u > 0 \\ 0.001u & \text{otherwise,} \end{cases} \tag{28}$$

where  $x_i$  refers to the input value or output of a neuron of the previous layer,  $w_{ij}$  refers to a trainable weight from neuron or input  $i$  to neuron  $j$ ,  $b_j$  is a trainable bias value, and  $\phi_j$  is the activation function, which computes the output of neuron  $j$ . In the overall network, an input pattern is propagated stage-wise from the input layer to the output layer.



**Fig. 1** Exemplary illustration of an MLP regression model with three hidden layers with ReLUs and one linear output unit.  $x_1, \dots, x_n$  represent the input features and  $y$  represents the respective output variable

Learning with such a model refers to adapting the free parameters  $w_{ij}$ 's and  $b_j$ 's (essentially the linear mappings from layer to layer) to approximate an unknown function, which is represented by a set of known input–target pairs.

For training the deep neural network, all 20 individually z-transformed (normalized using the mean and standard deviation from the respective training data) feature variables (of a particular time step) were fed into the network and an according prediction of the target variable ( $u_*$ ,  $Q_0$ , or  $H_0$ ) was generated. More precisely, 20 input neurons feed into three hidden layers 16, 32, 16 leaky rectified linear units (leaky ReLU, Maas et al. 2013) as shown in (28). Finally, the output is generated by a linear output layer (no activation function) consisting of a single output neuron, overall resulting in 1425 trainable parameters (including biases in each layer). In preliminary experiments, this particular architecture was identified as a good trade-off between generalization performance and model complexity. For a detailed introduction to optimization and supervised learning in neural networks, the reader is referred to Goodfellow et al. (2016).

### 3.2 Extreme Gradient Boosting

As an alternative machine learning approach, we evaluated a recent state-of-the-art decision tree-based method called extreme gradient boosting (XGB) proposed by Chen and Guestrin (2016). Gradient boosting is an ensemble method in which multiple weak predictors (here decision trees) are sequentially arranged to produce a joint output. First, an initial predictor  $f_0$  is trained to predict the target vectors  $z_i$ . Then the gradient of the loss function  $E$  with respect to the predictions  $y_i$  is calculated:

$$-\frac{\partial E}{\partial y_i} \text{ with, e.g., } E = \sum_i (x_i - z_i)^2. \quad (29)$$

This gradient is used to update the prediction:

$$y_i \leftarrow y_i - \frac{\partial E}{\partial y_i}, \quad (30)$$

and also as a learning target for the next predictor  $f_1$ . Thus, each new predictor  $f_{\tau+1}$  attempts to correct the error residual left by the previous predictor  $f_\tau$ . This is repeated until convergence is reached, e.g., until no more improvement is achieved by additional predictors. Finally, the overall ensemble prediction is the sum of all predictors of the above chain:

$$f(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{\tau=1}^n \eta f_\tau(\mathbf{x}), \quad (31)$$

where the learning rate  $\eta$  regulates the size of the residuals.

From a methodological perspective, XGB adds regularization (L1 and L2) to gradient boosting, but the name refers to its highly efficient and parallelized implementation, which makes it fast and thus effective in practice. Moreover, it has been shown to be particularly successful for both classification and regression on sparse and unstructured inputs (Chen and Guestrin 2016). For a detailed introduction to XGB, please refer to further literature, e.g., (Brownlee 2016).

### 3.3 Feature Importance Analysis

In order to assess the relative importance of each feature, as attributed by the models with respect to each target variable, we calculated the permutation feature importance as proposed by Breiman (2001): (1) Compute the prediction score of the trained model  $s$  (we used the root mean square error, RMSE). (2) Randomly permute the values of feature  $j$  over the data set (we used the respective test data sets) and recompute the score  $s'_{i,j}$  (repeating for  $i \in \{1, \dots, n\}$  times, we used  $n = 10$ ). (3) Calculate the average prediction score  $s'_j$  for feature  $j$  and subtract  $s$ , i.e.,  $s'_j = \left[ \frac{1}{n} \sum_i^n (s'_{i,j}) \right] - s$ . Additionally, we normalize  $s'_j$  to obtain the final and relative feature importance scores  $s_j = s'_j / \sum_j s'_j$ . We repeated this procedure for each target tower and target variable and computed the average importance scores with standard deviations for each feature over the three towers.

## 4 The Land–Atmosphere Feedback Experiment and its Measurements for Studying Flux Similarities

### 4.1 Overview of the Land–Atmosphere Feedback Experiment

The land–atmosphere feedback experiment (LAFE) took place in August 2017 at the DOE ARM program SGP site near Lamont, Oklahoma, USA. An overview of LAFE is presented in Wulfmeyer et al. (2018). The overarching goal of LAFE is the study of L–A feedback processes in the SGP region during summer time considering different vegetation types, which would have different soil moisture conditions, and the land surface heterogeneity. Specifically, LAFE has 4 scientific objectives:

1. Determine water vapor and vertical velocity, turbulence, and latent heat flux profiles, and investigate new similarity relationships for entrainment fluxes and variances
2. Map surface momentum, sensible heat, and latent heat fluxes using a synergy of range–height indicator (RHI) scanning wind, humidity, and temperature lidar systems
3. Characterize L–A feedback and the moisture budget at the SGP site by combining surface and ABL flux measurements as well as measurements of humidity advection in dependence of different soil moisture regimes.
4. Verify LES runs and improve turbulence parameterizations in mesoscale models

The LAFE measurements were complemented by the three tower measurements of NOAA for energy balance closure (EBC) measurements (EBC Towers 1, 2, 3 NOAA), which are subject of this work, and another EBC station of UHOH (Wizemann et al. 2015). These instruments were oriented along the main line of sight (LOS) of the LAFE remote sensing measurements in order to get a good overlap, enable comparisons, and taking advantage of the sensor synergy. For instance, this measurement configuration resulted in first measurements

**Table 1** Canopy parameters measured at the tower locations and derived based on Foken (2016)

Tower $i$	$h_i$ , m	$d_{z,i}$ , m	$z_{0,i}$ , m
1	0.5	0.4	0.05–0.1
2	0.6	0.5	0.05–0.1
3	0.6	0.5	0.05–0.1

of lidar-derived two-dimensional fields of LOS, moisture, and temperature. Their vertical structures over different towers were used for comparisons with MOST (Späth et al. 2022).

In this work, we focus on the analyses of the NOAA tower data. Our results have strong relationships to all LAFE objectives. The surface flux measurements will be used for complementing flux profiles measurements within LAFE objective 1, for the verification of scanning lidar-derived surface flux estimates (objective 2, see also Späth et al. 2022), for ABL budget studies (objective 3), and for the investigation of the parameterization of surface fluxes in mesoscale models (objective 4).

## 4.2 NOAA Tower Measurements

### 4.2.1 Data Sampling and Processing

Important for this work are the data sets from three 10-m towers that were installed within the LAFE domain. Tower 1 was installed in an early growth soybean crop field; Tower 2 in native grassland; and Tower 3 in a more mature soybean crop field. The towers were outfitted with an identical suite of measurements to sample wind, temperature, and moisture measurements as well as vegetation and soil variables such as albedo  $A$  and available water contents  $AW$ , which was calculated using a weighted average of the soil moisture measurements made at 5 cm ( $Smois_{05}$ ), 10 cm ( $Smois_{10}$ ), and 20 cm ( $Smois_{20}$ ):  $AW = 7.5 Smois_{05} + 7.5 Smois_{10} + 20 Smois_{20}$ . Temperature differences were sampled between 2 m and 10 m AGL, whereas the moisture and wind differences were determined between 3 m and 10 m AGL. Momentum, heat, and moisture fluxes were sampled at heights of 3 m and 10 m AGL. Bulk quantities were sampled at 1 Hz; turbulence and water vapor were sampled at 10 Hz using an CSAT3 sonic anemometer and EC155 closed path infrared gas analyzer, respectively. Additional details on the experimental setup during LAFE appear in Lee and Buban (2020); Lee et al. (2021); Lee and Meyers (2022). The whole data set during August 2017 was applied for our analyses. The diurnal cycles of the flux, gradient, and mean values during LAFE are shown in Lee and Buban (2020).

Table 1 summarizes the canopy heights  $h_i$  that we measured in the field, the derived zero-plane displacements heights  $d_{z,i}$ , and the roughness lengths  $z_{0,i}$  at each tower  $i$ . The zero-plane displacements were estimated using the approximation of  $0.8h$  and the roughness lengths from Table 2.7 in Foken (2016). The land cover and the terrain heights were rather homogeneous over the footprint areas and beyond. As the canopy heights and the roughness lengths were similar for all sites and the crop growth could be neglected, it was valid to use  $d \simeq 0.5$  m for all sites. As the roughness sublayer is less than three times the canopy height (Harman and Finnigan 2007), it was ensured that all measurements were taken above in the inertial sublayer, which is another requirement for the validity of MOST.

With respect to turbulence, we processed the data sets from the sonic anemometer and gas analyzer with the TK3 software by Mauder and Foken (2015), which applied cross-wind corrections, corrections of spectral loss, steady-state tests, and integral turbulence characteristics

tests consistently and reproducibly in one package. The differences in the results between the NOAA software and TK3 were in the range of a few  $10 \text{ Wm}^{-2}$  and we identified further 25 out of 1248 data points ( $\approx 2\%$ ) that did not pass the scientific level stationarity and integral turbulence characteristics test. Therefore, these two tests had negligible influence on our results.

As we were concerned about the unstable behavior of the turbulent fluxes, we applied further specific quality controls to reduce the impact of flow distortions on the measurements, which can happen at specific wind directions. First of all, as we dealt with unstable situations, we selected the data with a sensible heat flux of  $SH > 0 \text{ Wm}^{-2}$ , which corresponded also to  $LH > 0 \text{ Wm}^{-2}$ . Furthermore, to avoid indifferent situations during morning transitions and afternoon decays of turbulence, we chose only data with  $u_* > 0.1 \text{ ms}^{-1}$ . Flow distortion was detected and sorted out by the criterion  $0.9 < \sigma_w u_*^{-1} < 1.8$  according to Oliveira et al. (2021) where  $\sigma_w$  is the standard deviation of the vertical wind fluctuations. Also, values exceeding  $800 \text{ Wm}^{-2}$  for sensible and latent heat fluxes were considered as outliers. The rest of the data were applied for the comparisons of MOST, BRN, and the ML methods with the observations.

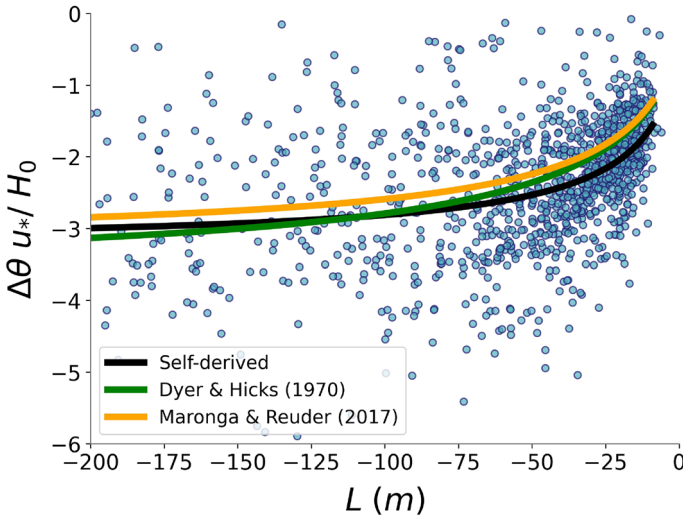
The following analyses gave us high confidence that MOST and BRN were applicable at all tower sites. As noted by Lee and Buban (2020) and briefly summarized here, there was nearly complete EBC at the early growth soybean and native grassland sites for the entire month-long data set. At the more mature soybean site, however, the EBC was around 95%, which was attributed to larger ground heat storage as compared with the other sites. As these errors are still quite small, we used the entire data set without any corrections for the EBC. Footprint analyses were performed for all towers using the tools provided in Kljun et al. (2015) (not shown). Both the footprints for the flux measurements at 3 m and 10 m showed rather homogeneous and consistent patterns after quality control of the data (see above). The 3 m flux footprints had a typical radius of 50 m so that all flux measurements corresponded to the soil and land cover over the fields of interest. Therefore, we focused on the flux measurements at this height for our flux retrievals and comparisons. Within the footprints, the terrains were very flat (variabilities of the order of 0.1 m) and the land covers were fairly homogeneous. The fluxes measured at 2 m, 3 m, and 10 m showed a reasonable correspondence without indications of inhomogeneous mesoscale flows. Furthermore, lidar scans analyzed so far over the sites did not indicate a strong inhomogeneous pattern either (Späth et al. 2022). We also emphasize that almost all mesoscale models apply MOST over these types of crop lands so that any verification of MOST such as in this work can be considered as very instructive to get insight into the expected performance of land–atmosphere model systems.

#### 4.2.2 Monin–Obukhov Similarity Theory

For the derivation of the fluxes, the integrated MOST functions were applied using the therein prescribed exponents of the similarity functions (see Eqs. 17–19) and  $d \simeq 0.5 \text{ m}$ . For improving the comparisons, we fitted first the parameters  $a_{m,h,q}$  and  $b_{m,h,q}$  using the entire tower 1–3 data set. Fits applied to the single tower data provided highly variable results with larger errors due to the reduction of the data points. A fit to all tower data (solid lines) is exemplarily demonstrated in Fig. 2 for potential temperature where the resulting MOST integrated functions are compared to the observations in dependence of  $L$ . This plot already demonstrates a structural problem of the MOST gradient and integrated functions. It is difficult to identify a clear relationship between the integrated function and its dependence of  $L$  and a large range of fit coefficients provide similar results with respect to the rms errors

**Table 2** Fit coefficients for the similarity or the integrated functions using  $\alpha = 0.25$  and  $\beta = \gamma = 0.5$ , respectively, for all three fluxes of interest based on the tower 1–3 data

Parameter	Result	Uncertainty
$a_m$	16	2
$b_m$	1.26	0.02
$a_h$	7	1
$b_h$	0.88	0.02
$a_q$	62.0	29
$b_q$	2.5	0.4



**Fig. 2** Fit of the scaled differences of the potential temperature measurements to the integrated MOST gradient function for potential temperature (see Eq. 18). The fits using the coefficients of Dyer and Hicks (1970) and Maronga and Reuder (2017) are also shown. The large scatter of the data and the resulting uncertain fits over the entire range of  $L$  already demonstrates fundamental problems of MOST

between the observations and the measurements of  $u_*$  and  $H_0$ . The results achieved with the coefficients of Dyer and Hicks (1970) and Maronga and Reuder (2017) are also shown, which led to significant biases in the results (as well as for  $H_0$  and  $Q_0$ , not shown). Therefore, it was reasonable to adapt these coefficients to our LAFE data. The results for the fit coefficients are presented in Table 2.

After fixing these coefficients, a simultaneous iteration between Eqs. 17–19 was performed to derive the three fluxes  $u_*$ ,  $H_0$ , and  $Q_0$  for each data point (three measurements of differences in three equations with three unknowns).

### 4.2.3 Bulk Richardson Number

In a similar manner as for MOST but now for the gradient functions, the coefficients for BRN were determined using the tower 1–3 data. In order to determine the BRN fitting coefficients, we followed the procedure discussed in Markowski et al. (2019); Lee and Meyers (2022), which is briefly summarized here. We first computed each observation’s uncertainty before performing a Levenberg–Marquardt least-squares fit that weights the uncertainties in each

**Table 3** BRN function fit coefficients used for the comparisons based on the tower 1–3 data

Parameter	Result	Uncertainty
$c_m$	0.11	0.01
$d_m$	2.4	1.6
$c_h$	0.28	0.04
$d_h$	16	12
$c_q$	0.18	0.01
$d_q$	28	7

**Table 4** XGB parameters that were identified by means of a systematic search over all three towers and used with respect to each target variable

XGB parameter	Target variable		
	$u_*$	$SH$	$LH$
Eta	0.19	0.16	0.11
Min child weight	0.70	0.02	0.44
Max depth	12	4	19
Gamma	0.00	0.01	1.92
Subsample	0.69	0.67	0.44
Colsample bytree	0.89	0.73	0.98
Colsample bylevel	0.87	0.47	0.77
Reg Lambda	0.29	0.65	0.62
$N$ estimators	21	26	70

$SH$ : sensible heat flux,  $LH$ : latent heat flux

observation to determine the fitting coefficients. Also here, the BRN similarity functions were analyzed specifically using all tower 1–3 data together. For this purpose, a fit of the experimental data was performed using Eqs. 21–23. The results are presented in Table 3. Please note that the differences between these coefficients with respect to each single tower were close considering the estimated error margins, which confirmed that it made sense to use the entire tower 1–3 data set for the fits and the retrievals.

#### 4.2.4 Machine Learning

For the MLP, we applied the Adam optimizer (Kingma and Ba 2015) to train the models' parameters using a learning rate of  $1 \times 10^{-3}$ , 1000 epochs of training with batch size 256, and default settings otherwise. Additionally, we used L2 regularization, i.e., weight decay, with a rate of 0.1 (Loshchilov and Hutter 2019). After cleaning (removing samples with NaN entries), the data amounted to 487, 519, and 450 samples for towers 1, 2, and 3, respectively. Training was performed on a GeForce GTX 1060 6GB graphics card using PyTorch 1.8.1.

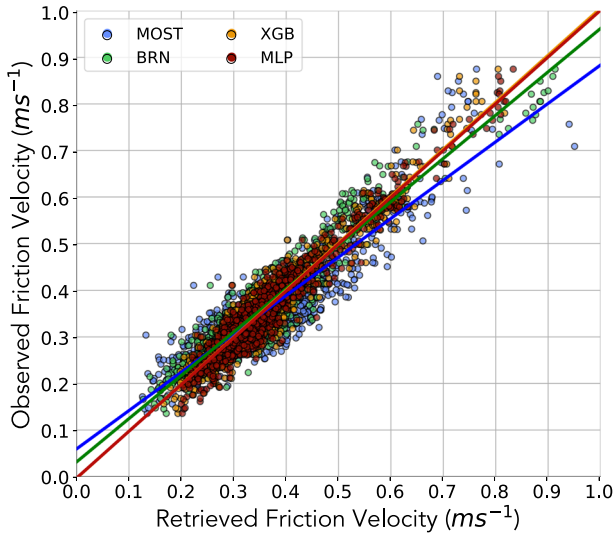
Using the XGB for learning the fluxes, we used the same data split between the tower data as reported in Sect. 3.1. A systematic search over all three towers revealed the best XGB parameters per target variable, which are reported in Table 4. We also tested the training using a subset of all tower data together (80:20 or 60:40) and found even improved results that indicated that our two-tower training strategy was reasonable.



**Table 5** Friction velocity evaluation metrics using linear regression analyses

Model	RMSE, $\text{ms}^{-1}$	R	OLS	Slope interval
BRN	0.04	0.94	$u_{*,o} = 0.9 u_{*,r} + 0.03 \text{ms}^{-1}$	[0.91 0.95]
MOST	0.05	0.9	$u_{*,o} = 0.8 u_{*,r} + 0.06 \text{ms}^{-1}$	[0.80 0.85]
MLP	0.03	0.96	$u_{*,o} = u_{*,r}$	[0.99 1.02]
XGB	0.03	0.96	$u_{*,o} = u_{*,r}$	[0.99 1.03]

RMSE: Root mean square error, R: regression coefficient, OLS: ordinary least square regression line coefficients, slope interval: 95% confidence interval,  $u_{*,o}$ : observed friction velocity,  $u_{*,r}$ : retrieved friction velocity



**Fig. 3** Scatter diagram of the retrieved friction velocities versus the observations. Blue: MOST, green: BRN, orange: XGB, red: MLP

## 5 Results

Using the derivation of the results for MOST, BRN, and ML, as described above, we compared their outputs with the observations using the eddy covariance measurements of the fluxes. For all comparisons, the kinematic heat fluxes were transformed in  $\text{Wm}^{-2}$  using  $SH = c_p \rho \overline{w'\theta'}$  and  $LH = l_q \rho \overline{w'q'}$  where  $c_p \approx 1005 \text{Jkg}^{-1}\text{K}^{-1}$  is the specific heat capacity of air at constant pressure,  $\rho$  is the surface air density which fluctuates here between  $1.08$  and  $1.2 \text{kgm}^{-3}$ , and  $l_q \approx 2500 \text{Jg}^{-1}$  is the specific heat of evaporation.

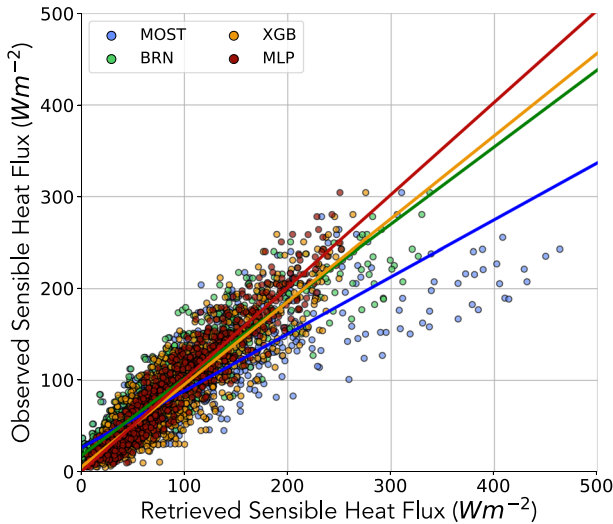
First of all, we visualized and analyzed the results using scatter diagrams. These scatter diagrams are presented in Figs. 3, 4 and 5 and are statistically evaluated in Tables 5, 6 and 7. For the regression analyses we used an ordinary least-squares model to obtain the slope and intercept of the line, followed by an error evaluation using a confidence interval of 95% based on a standard normal distribution.

Figure 3 presents the results for  $u_*$  and Table 5 provides the statistical comparison of the results based on regression analyses. The plot shows that MOST and BRN perform similarly; however, the correlation coefficient is slightly smaller for MOST. MOST rolls off at  $u_* \geq 0.6 \text{ms}^{-1}$  and often overestimates the friction velocity whereas this effect is reduced

**Table 6** Sensible heat evaluation metrics

Model	RMSE ( $\text{Wm}^{-2}$ )	R	OLS	Slope interval
BRN	24.8	0.92	$SH_o = 0.84 SH_r + 18 \text{ Wm}^{-2}$	[0.82 0.86]
MOST	42.9	0.86	$SH_o = 0.62 SH_r + 27 \text{ Wm}^{-2}$	[0.60 0.64]
MLP	17.0	0.95	$SH_o = SH_r - 1 \text{ Wm}^{-2}$	[0.99 1.03]
XGB	26.1	0.9	$SH_o = 0.9 SH_r + 5.5 \text{ Wm}^{-2}$	[0.87 0.93]

$SH_o$ : observed heat flux,  $SH_r$ : retrieved heat flux



**Fig. 4** Scatter diagram of the retrieved sensible heat flux versus the observations. Blue: MOST, green: BRN, orange: XGB, red: MLP

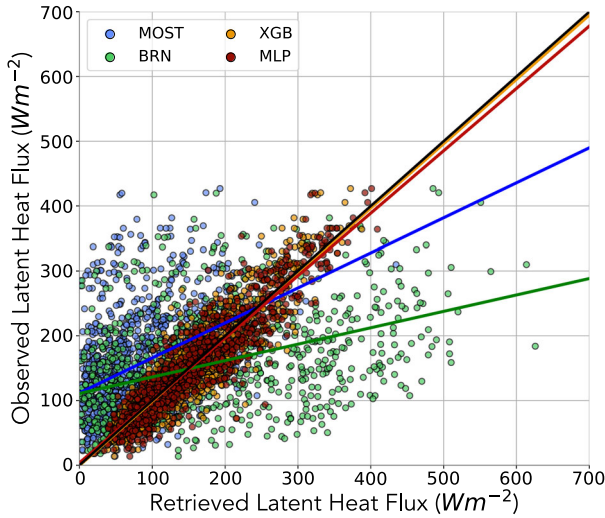
with BRN. The RMSEs for MOST and BRN are similar. The slight bias of these methods is also demonstrated by the deviating slopes and the significant intercepts of  $0.06 \text{ ms}^{-1}$  and  $0.03 \text{ ms}^{-1}$  of the regression lines, respectively. For MLP,  $u_*$  is slightly overestimated up to  $u_* \leq 0.25 \text{ ms}^{-1}$  whereas this bias is reduced for XGB. However, for the rest of the comparisons, these methods show a better correspondence to the observations than BRN or MOST with reduced RMSEs. Both ML methods perform very similar over the remaining range of data. This is also expressed in Table 5. A clearly better correspondence of the regression lines with slope 1 with respect to the observations is achieved showing nearly no intercept with respect to the bisecting line.

Figure 4 shows the results for  $SH$  and Table 6 provides the corresponding statistical evaluations. Here the behavior of MOST and BRN is considerably different. A large number of MOST retrievals overestimate  $SH$  nearly over the entire range, in spite of the optimization of the coefficients in the similarity function. Considering our careful outlier removal techniques, these deviations are significant. The number of strongly deviating data points with respect to the regression line is strongly reduced for BRN. Taking all these data into account, MOST has the highest  $RMSE \simeq 43 \text{ Wm}^{-2}$  of all methods and the smallest regression coefficient of 0.86. The regression line is far off the bisecting line. BRN has a similar tendency for an underestimation of  $SH$  indicated by a positive intercept of the regression line. However, in

**Table 7** Latent heat evaluation metrics

Model	RMSE ( $\text{Wm}^{-2}$ )	R	OLS	Slope interval
BRN	121	0.42	$LH_o = 0.3 LH_r + 111 \text{ Wm}^{-2}$	[0.22 0.29]
MOST	107	0.41	$LH_o = 0.5 LH_r + 112 \text{ Wm}^{-2}$	[0.47 0.61]
MLP	31	0.92	$LH_o = LH_r + 4 \text{ Wm}^{-2}$	[0.94 0.99]
XGB	33	0.91	$LH_o = LH_r - 1 \text{ Wm}^{-2}$	[0.96 1.02]

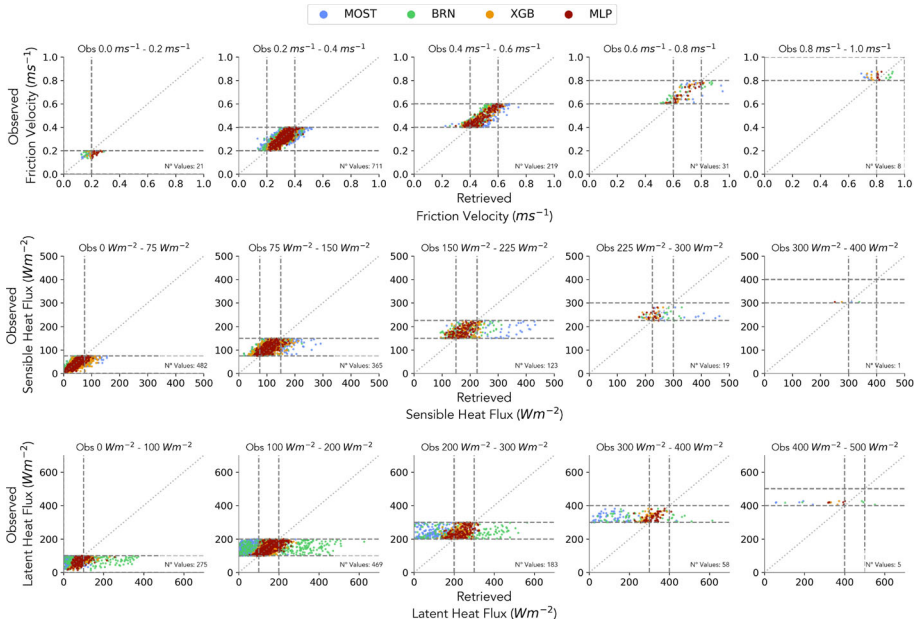
$LH_o$ : observed heat flux,  $LH_r$ : retrieved heat flux



**Fig. 5** Scatter diagram of the retrieved latent heat flux versus the observations. Blue: MOST, green: BRN, orange: XGB, red: MLP

contrast to MOST, its slope corresponds more to the bisecting line, the  $RMSE$  is significantly reduced to  $25 \text{ Wm}^{-2}$ , and the correlation coefficient is higher with  $R \simeq 0.92$  (see Table 6). Using MLP or XGB, a clearly better agreement with the observations is achieved. Here, MLP performs better than XGB because the latter is biased toward higher values for  $SH > 200 \text{ Wm}^{-2}$ . For XGB, the  $RMSE$  is slightly higher as for BRN; however, the slope is closer to the bisecting line and the intercept is reduced to  $5.5 \text{ Wm}^{-2}$ . The best performance is achieved with MLP. The  $RMSE$  is strongly reduced to  $17 \text{ Wm}^{-2}$ , the correlation coefficients is  $R \simeq 0.95$ , the slope is close to 1, and the intercept just  $-1 \text{ Wm}^{-2}$ .

The strongest improvement of the analysis of surface fluxes by the ML methods in comparison with BRN and MOST was achieved for the latent heat flux  $LH$ . Figure 5 presents the scatter diagram for  $LH$  and Table 7 the corresponding statistics. Here, MOST and BRN perform very suboptimal pointing to structural problems of the assumptions with respect to the similarity functions. The scatter is largest for BRN with an  $RMSE \simeq 121 \text{ Wm}^{-2}$  in comparison to MOST with  $RMSE \simeq 107 \text{ Wm}^{-2}$ . The regression coefficients are poor with  $R \leq 0.42$  and the intercepts are  $> 100 \text{ Wm}^{-2}$  in both cases with slopes of the regression lines far away from the bisecting line. In contrast, MLP and XGB still show a very good correlation and linearity almost over the entire range of the data set. Obviously, MLP and XGB are superior in their search for reasonable relationships between  $LH$  and the proposed



**Fig. 6** Scatter diagrams of the comparisons broken down with respect to different intervals of the observations. Beige: observations, Blue: MOST, green: BRN, orange: XGB, red: MLP

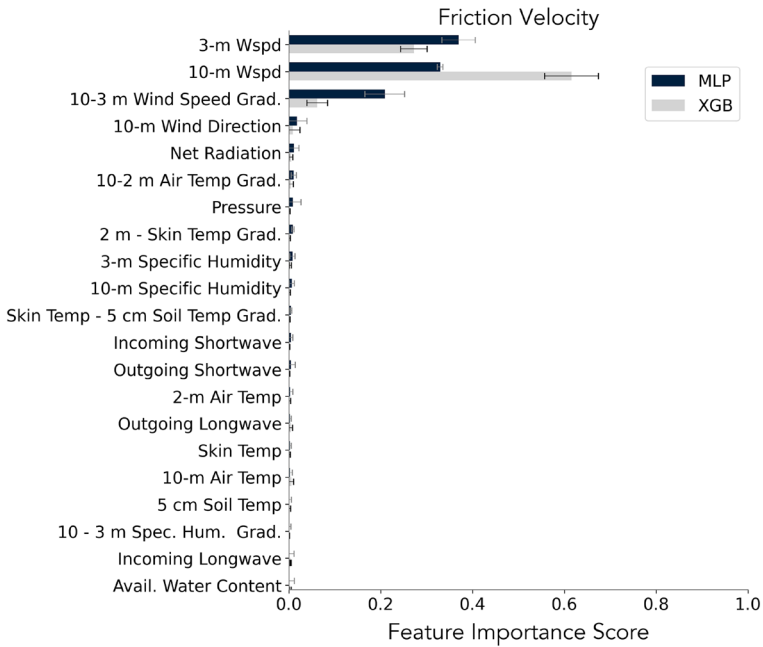
predictors. MLP and XGB perform similar with strongly reduced  $RMSE \approx 33 Wm^{-2}$  and high regression coefficients of  $R \approx 0.9$ . The slopes of the regression lines nearly correspond to the bisecting line and the absolute values of the intercepts are  $< 5 Wm^{-2}$ .

A more detailed breakdown of the performance of the various approaches for the flux similarities is presented in Fig. 6. Here, the scatter diagrams are presented with respect to intervals or classes of the observations for  $u_*$ ,  $SH$ , and  $LH$ . For all three variables, these refined scatter diagrams confirm the general reduced scatter of the ML methods and their reduced biases in comparison to MOST and BRN. Whereas the scatter and the bias for all these classes is acceptable for  $u_*$ , there are basically no intervals where MOST and BRN are bias-free for  $SH$  and  $LH$  with the most degraded performance for the latter.

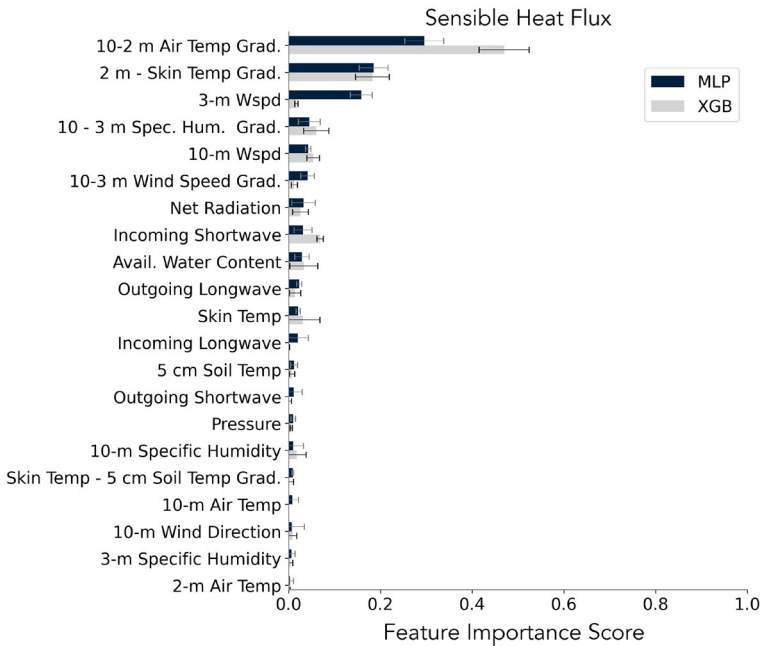
## 6 Discussion

Based on these results, it is very interesting to reconsider and to evaluate the key relationships that are prescribed in the MOST and the BRN similarity relationships as well as to search for more general and accurate ones. Particularly, the ML approaches permit the study of the relative importance of the respective features used during the training processes, as described above. The results are presented in Figs. 7, 8, and 9 for both ML methods ordered in relative importance for the MLP.

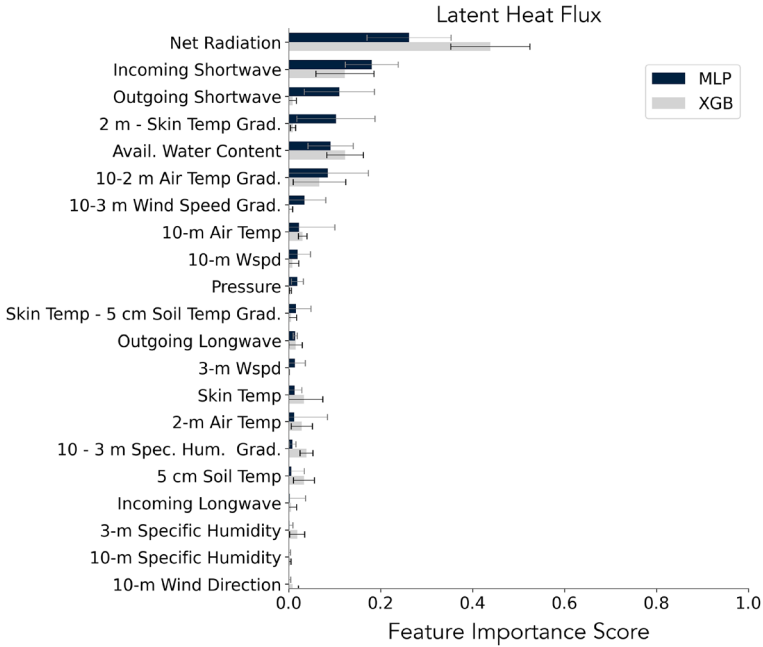
Figure 7 shows these weights for  $u_*$ . As expected from the relative good agreement of the MOST and BRN retrievals with the observations, indeed  $U_{3m}$  and  $U_{10m}$  are the most important features for MLP and XGB. Interestingly, the weight of  $U_{3m}$  and  $U_{10m}$  for MLP methods is similar but much higher with respect to  $U_{10m}$  for XGB. The third most important



**Fig. 7** Histogram of the importance scores for MLP and XGB with respect to the different features used during the training process for retrieving  $u_*$ . Black bars: MLP, gray bars: XGB



**Fig. 8** Histogram of the importance scores for MLP and XGB with respect to the different features used during the training process for  $SH$ . Black bars: MLP, gray bars: XGB



**Fig. 9** Histogram of the importance scores for MLP and XGB with respect to the different features used during the training process for *LH*. Black bars: MLP, gray bars: XGB

weight is the wind speed gradient between 3 m and 10 m. This result confirms that the main proposed structural relationships, the proportionality of  $u_*$  to  $U_{3m}$  or to  $U_{10m}$ , is largely valid. This dependence of the surface momentum flux on the strength of the SL wind is physically reasonable and understood, as the increase of the surface wind gradient (with wind at the displacement height equal to zero) must enhance the turbulence and dynamic exchange with the atmosphere. However, even in this case, the ML methods perform better for larger  $u_*$  in comparison to MOST and BRN. Based on Fig. 7, this is not due to missing relationships in MOST and BRN with respect to other variables because the remaining importance weighting of these is very low. Therefore, this deficiency of MOST and BRN is either due to a different nonlinear relationships of  $u_*$  to the wind speed that is not fully captured by the prescribed similarity functions or by a limited validity of the scaling variables  $L$  and  $Ri_b$  in the similarity functions. It will be subject of our future research to study and to separate these effects in more detail.

We investigated these most important relationships for  $u_*$  by plots of  $u_*$  versus  $U_{3m}$  or  $U_{10m}$ . The results are presented in Figs. 10 and 11. These figures confirm the major dependence of  $u_*$  on the surface winds. The main difference is the varying slope due to the different surface wind strengths between 3 m and 10 m. All retrieval methods show a slight under-dispersive behavior with respect to the scatter of the observations, which may be due to dependencies to other variables. Except at small  $u_*$ , where a small remaining bias may be present for MLP and XGB, the correspondence of all retrievals follow the observed relationships between  $u_*$  and the surface wind speeds  $U_{3m}$  and  $U_{10m}$ . A positive bias of MOST in the range of  $2.5 \text{ m s}^{-1} < U_{3m} < 6 \text{ m s}^{-1}$  or  $4 \text{ m s}^{-1} < U_{10m} < 8 \text{ m s}^{-1}$  is visible. In contrast, BRN shows a positive bias for  $U_{3m} > 6 \text{ m s}^{-1}$  and  $U_{10m} > 9 \text{ m s}^{-1}$ .

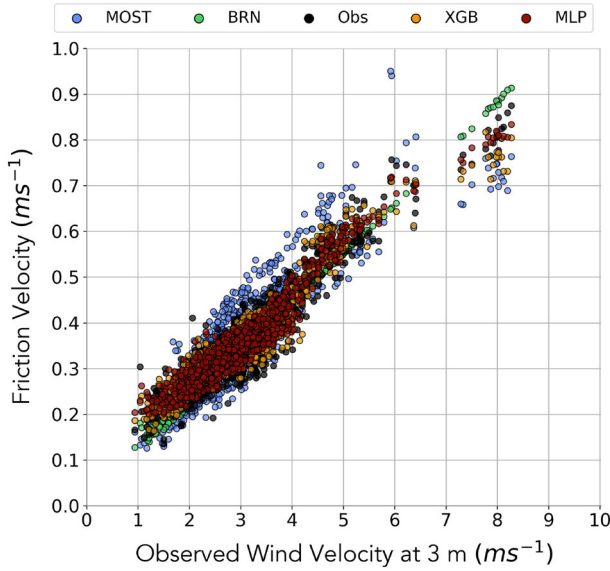


Fig. 10 Observed and retrieved  $u_*$  in dependence of  $U_{3m}$ . Black: Observations, blue: MOST, green: BRN, orange: XGB, red: MLP

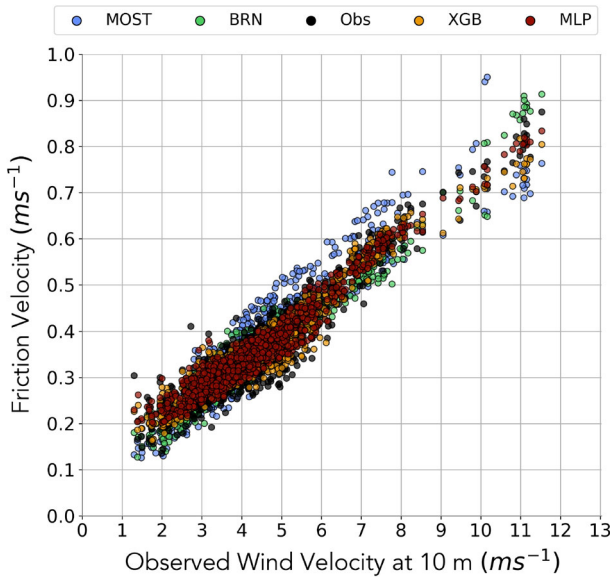
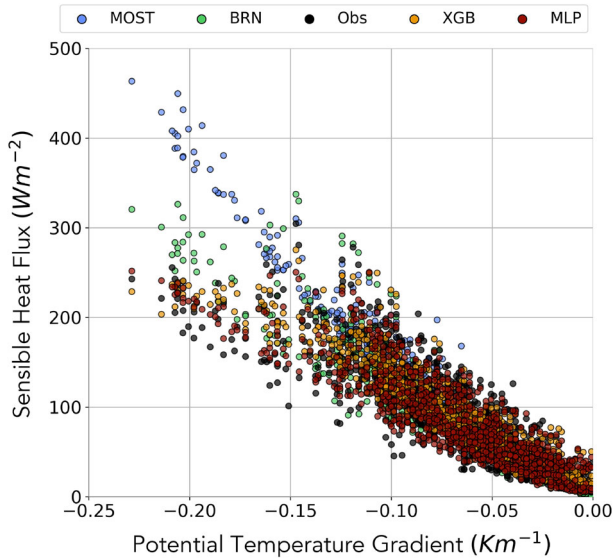


Fig. 11 Observed and retrieved  $u_*$  in dependence of  $U_{10m}$ . Black: Observations, blue: MOST, green: BRN, orange: XGB, red: MLP



**Fig. 12** Observed and retrieval  $SH$  in dependence of potential temperature gradient between 2 m and 10 m. Black: Observations, blue: MOST, green: BRN, orange: XGB, red: MLP

For  $SH$ , the importance weighting by MLP and XGB is presented in Fig. 8. As assumed in MOST and BRN, the most important drivers are indeed either the temperature differences between 10 m and 2 m or between 2 m and the land surface, respectively. Here, the weights are larger for XGB because the third most important feature is mainly present for MLP and is  $U_{3m}$ . This higher-order relationship will be studied more in detail in the future.

Therefore, it is reasonable to plot  $SH$  in dependence of the SL potential temperature gradient for the observations and the retrievals. The result for all overlapping data is presented in Fig. 12. Particularly, there seems to be a bifurcation of the data for potential temperature gradients  $d\theta/dz < -0.1 K m^{-1}$ . This figure demonstrates that the ML methods are able not only to reproduce this dependence very well but also these fine structures of the results. In order to allow for a better comparison of the results and the observations, the comparisons of the observations and the retrievals are presented in separate panels in Fig. 13. Except remaining slight deviations between the observations and MLP, the resulting clusters of retrievals perform similarly and show the best agreement with the measurements. XGB is more underdispersive and shows larger deviations at stronger temperature gradients. In contrast, for both BRN and MOST, the retrievals are largely underdispersive and do not follow the general shape of the observations. BRN and MOST represent only a subset of the data namely either the lower range of  $SH$  for BRN and a general incorrect slope dependence of the temperature gradient to  $SH$  on the temperature gradient. As the dependencies between retrievals and temperature gradients in MOST and BRN deviate strongly from the observations, these confirm structural limitations in their similarity relationships.

We can study this further by repeating the ML with a reduced set of variables, e.g., the three or two most important ones according to the importance weighting. Particularly interesting is  $SH$  because the importance weighting confirms that the temperature differences are the most powerful input variables as assumed in MOST and BRN (see Figs. 4, 8, and 12 as well as Table 6) but ML still shows a superior performance. As expected, using three input variables,



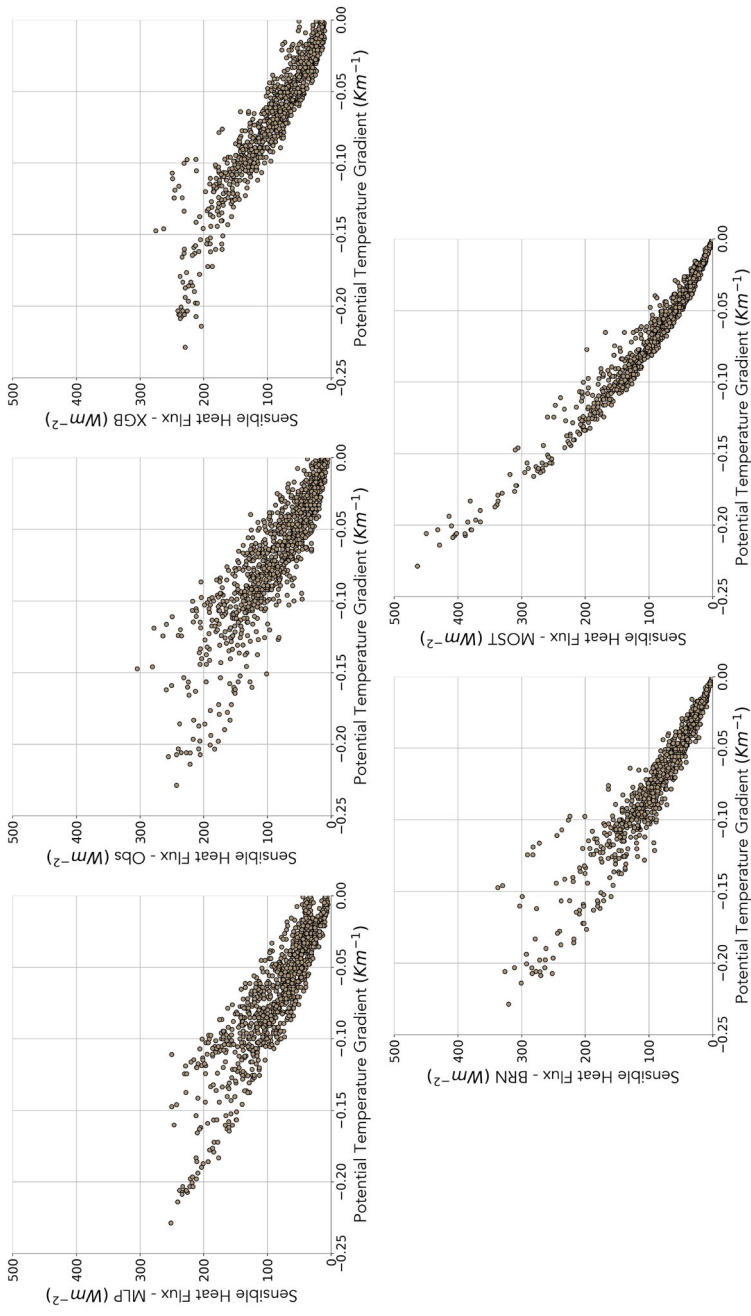


Fig. 13 Observed and retrieved  $SH$  in dependence of potential temperature gradient between 2 m and 10 m. Upper left panel: MLP; upper central panels: observations, upper right panel: XGB, bottom left panel: BRN, bottom right panel: MOST

**Table 8** Sensible heat evaluation metrics

Model	RMSE ( $\text{Wm}^{-2}$ )	R	OLS	Slope interval
Three input variables (two temperature differences and 3-m wind speed)				
MLP	19.8	0.94	$SH_o = 0.98 SH_r - 2.3 \text{ Wm}^{-2}$	[0.96 1.01]
XGB	23.3	0.91	$SH_o = SH_r - 1 \text{ Wm}^{-2}$	[0.97 1.03]
Two input variables (10-m and 2-m temperature difference and 3-m wind speed)				
MLP	22.5	0.92	$SH_o = 0.99 SH_r - 3.5 \text{ Wm}^{-2}$	[0.96 1.02]
XGB	34	0.83	$SH_o = 0.82 SH_r + 10.7 \text{ Wm}^{-2}$	[0.79 0.86]

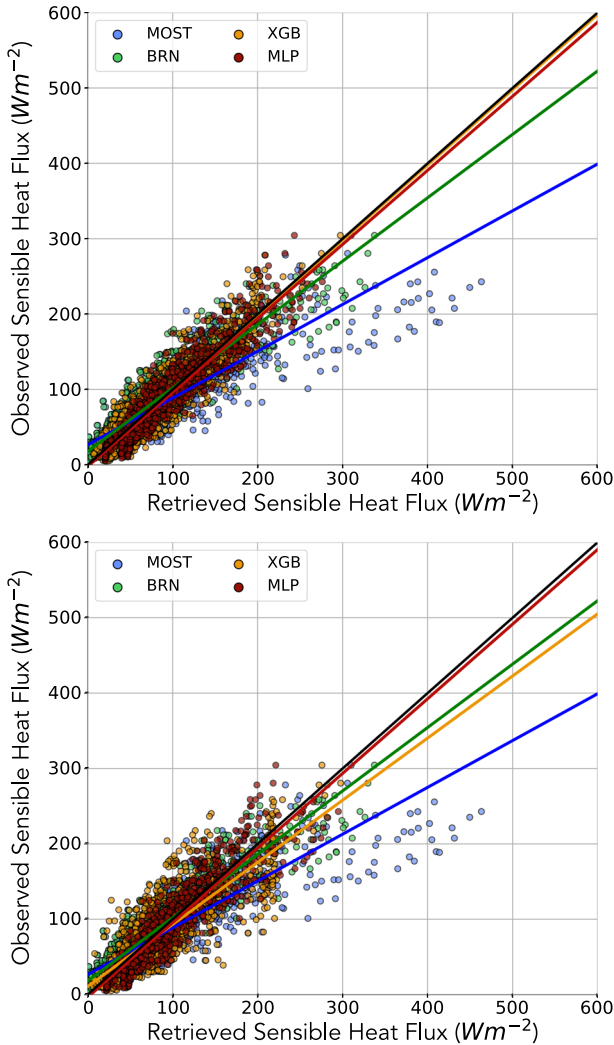
$SH_o$ : observed heat flux,  $SH_r$ : retrieved heat flux

the results are close to the performance using the full set of variables. If only the surface and the atmospheric temperature differences as well the 3-m wind speed are used for our threefold cross-validation and training strategy for ML, we still find similar results of the ML retrievals. This is demonstrated in Table 8 and Fig. 14, upper panel. The performance of MLP using the three input variables is almost indistinguishable from the results with the full training and XGB improves. As ML still achieves a better agreement with the observations, obviously ML overcomes structural limitations of the MOST and BRN similarity functions. If only two input variables are used (10-m, 2-m temperature difference as well as 3-m wind speed), this performance is basically maintained for MLP whereas XGB shows comparable results with BRN. Similar results are achieved for  $LH$  (not shown) and will be subject of future research as well as the reason why XGB performed worse than MLP just for  $SH$ .

Last but not least for  $LH$ , the importance weighting by MLP and XGB is presented in Fig. 9. It is striking that this result deviates strongly from the traditional expectations. There is basically no weight for the SL specific humidity gradient but mainly for the net and the incoming shortwave radiations for both XGB and MLP. Depending on the choice of XGB and MLP, the next important variables are the outgoing shortwave radiation, the air temperature gradient, and the available water contents.

In order to confirm the weak or even absent dependence of  $LH$  on the specific humidity gradient (in this case between 3 and 10 m), we plotted this dependence in Fig. 15 overlapping for the observations and all retrievals. This figure clearly shows the very weak relationship between  $LH$  and the moisture gradient between 10 m and 3 m. Additionally, this cluster of observations is hardly reproduced by BRN and MOST whereas the retrievals of MLP and XGB reproduce very well the distribution of the observations. This is seen in great detail in Fig. 16 where all results are presented in separate panels. The weak dependence of  $LH$  of the atmospheric moisture gradient is very well reproduced by MLP and XGB even with respect to two rather distinct clusters and some scattered data. There is a very slight underdispersive behavior of the entire range of the retrievals. In contrast, BRN and MOST propose relationships that deviate strongly from the observations with respect to their distributions and slopes. This slope dependence is far too large for BRN and also incorrect for MOST. Also, both MOST and BRN perform largely underdispersively in the region of the agreement of the data.

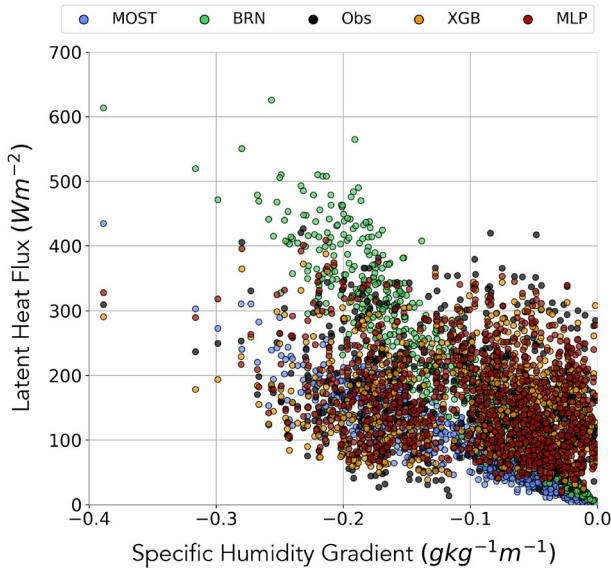
These results motivated us to plot the main dependencies that were predicted by the ML methods, namely the dependencies of  $LH$  on the net radiation and the incoming shortwave radiation. The results are presented in Fig. 17 for both variables (upper panel: net radiation, bottom panel: incoming shortwave radiation). Indeed, the correspondence between  $LH$  and the radiation terms is much clearer for MLP and XGB. The cluster of observations agrees



**Fig. 14** Upper panel: Scatter diagram of the retrieved sensible heat flux versus the observations using only three input variables for ML (two temperature differences, 3-m wind speed). Bottom panel: Scatter diagram of the retrieved sensible heat flux versus the observations using only two input variables for ML (10-m, 2-m temperature difference, 3-m wind speed). Blue: MOST, green: BRN, orange: XGB, red: MLP

very well with the retrievals of ML and XGB. In contrast, using BRN and MOST, a huge amount of scatter is produced in strong disagreement with the observations.

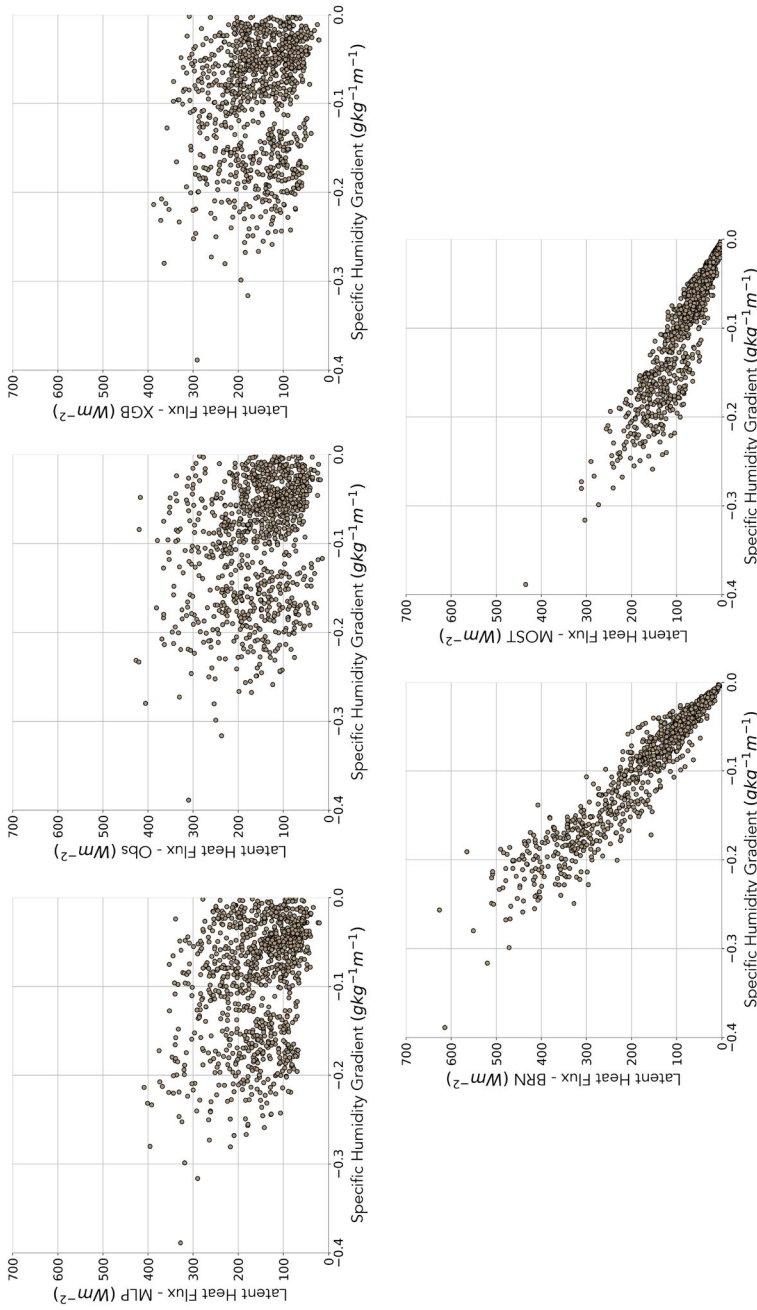
As we pointed out above, likely both MOST and BRN have further deficiencies such as dependencies from surface heterogeneities (Morrison et al. 2021), micro- and mesoscale circulations (Li et al. 2018; McNaughton and Brunet 2002), and ABL variables (van de Boer et al. 2014; Cheng et al. 2021; van Heerwaarden et al. 2009). In the MOST and BRN retrievals, we disregarded these effects and pointed out that the footprints of the measurements could be considered as homogeneous enough to apply these similarity theories. Also, our 2D lidar scans did not indicate any special conditions or microscale variability that may indicate that



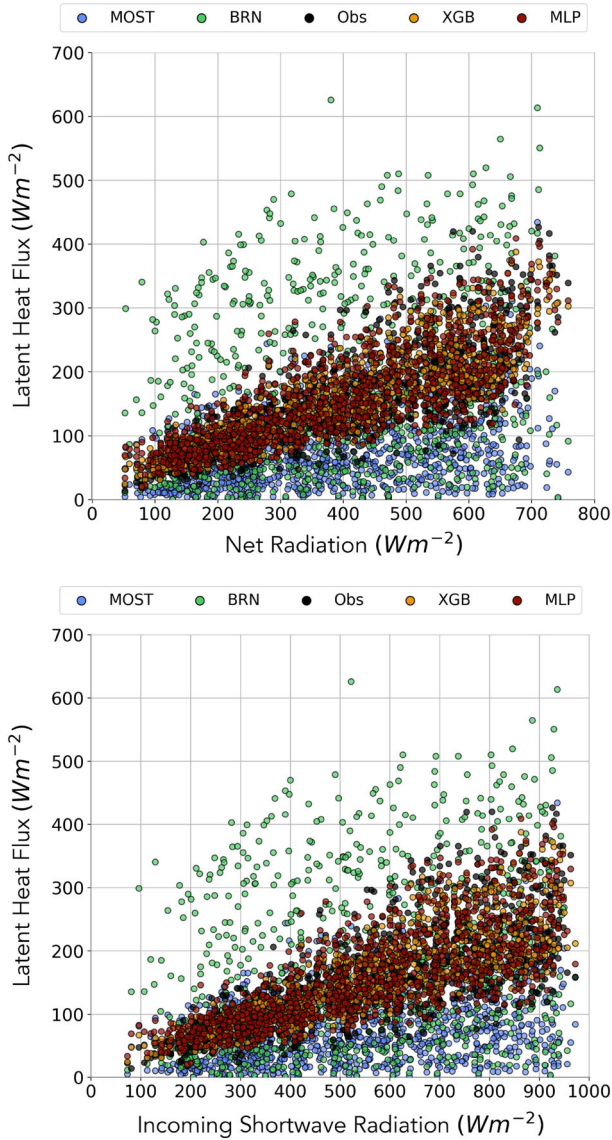
**Fig. 15** Observed and retrieval  $LH$  in dependence of specific humidity gradient between 3 m and 10 m. Blue: MOST, green: BRN, orange: XGB, red: MLP

these measurement sites were not suitable for the application of MOST and BRN relationships (Späth et al. 2022). Even without additional knowledge and ingesting this information in the ML algorithms, we find that the ML methods are capable of reproducing the observations very well. Therefore, we conclude that surface heterogeneities and ABL properties were not the main drivers for the erroneous behavior of MOST and BRN mainly with respect to  $SH$  and  $LH$ . The much better regression results of MLP and XGB indicate that the main problem of BRN and MOST are limitations of their relationships to temperature and moisture gradients, the prescribed but obviously incorrect shape and scaling of the similarity functions, and deficiencies in the choice of the scaling variables such as  $L$ .

It is interesting to compare our results with the work of Qin et al. (2005a, b, 2010), Safa et al. (2018), who used station but not tower data for the training of ML techniques over cropland, mainly above maize during different vegetation periods. Whereas Qin et al. (2005a, b) used mainly temperature, water vapor pressure deficit  $VPD$ , the soil water content in root zone  $W$ , the leaf area index  $LAI$ , and the photosynthetically active radiation  $PAR$  as input variables, Safa et al. (2018) also considered net radiation. Using a feed-forward back propagation neural network with an input layer, output layer as well as a hidden layer and output layer, Qin et al. (2005a) found a similar agreement of  $LH$  with the observations as in our study. In Qin et al. (2005b), they operated a least-squares support vector machine and achieve a similar performance with respect to  $LH$ . Qin et al. (2010) performed a sophisticated importance weighting based on the automatic relevance determination (ARD). ARD is another technique to determine the importance of input features by constructing a hyper-parameter accounting to the inverse variance of all weights that are connected to the respective input. A small value of the hyper-parameter means large influence of the input feature on the output. According to their analysis, the feature importance for  $LH$  turned out to be  $VPD > W > LAI > T > PAR$ . In our study, we also found a dependence on  $AW$ . However, we have not yet



**Fig. 16** Observed and retrieved  $LH$  in dependence of specific humidity gradient between 3 m and 10 m. Upper left panel: MLP, upper central panels: observations, upper right panel: XGB, bottom left panel: BRN, bottom right panel: MOST



**Fig. 17** Upper panel: Observed and retrieved  $LH$  in dependence on net radiation. Bottom panel: Observed and retrieved  $LH$  in dependence on shortwave incoming radiation. Blue: MOST, green: BRN, orange: XGB, red: MLP

investigated  $VPD$  and  $LAI$ . Obviously, this was also not necessary because MLP and XGB explain the driving variables mainly with net and shortwave incoming radiations.

Concerning  $VPD$ , we considered the moisture gradient as more important information. The very good prediction of  $LH$  using our input data set substantiates this approach. Also, our strong dependence on net radiation likely corresponds to the significant impact of  $PAR$ . In the future, we plan to enhance these importance weighting analyses and add further variables.

Safa et al. (2018) investigated sensible and latent heat fluxes with an MLP approach based on multiyear data sets of fluxes and L–A variables over maize. They also performed a sensitivity analysis, which was based on the computation of 5-% incremental changes of the input data with respect to the output (Nourani and Sayad Fard 2012). Momentum fluxes were not investigated. They demonstrated promising training results similar to Qin et al. (2010); however, they obtained a different importance weighting for  $LH$  with  $R_n > LAI > VPD > U$  and  $R_n > U > T > LAI > VPD$  for  $SH$ . Whereas the results for  $LH$  are in better agreement with our results, the dependence on net radiation for  $SH$  strongly disagree with our findings. A fundamental issue in these studies is the absence of the investigation of gradients with respect to  $T$  and  $q$ . Particularly, the temperature gradient turned out as most important input variables for  $SH$  in our study. Also, we suggest always including a comparison with the performance of MOST and BRN. Furthermore, the momentum flux should be studied.

Leufen and Schädler (2019) used an extensive set of tower data and compared these with retrievals of MOST and an MLP. They investigated the surface momentum and sensible heat fluxes, simultaneously. For the MOST retrievals, they also worked with the integrated similarity functions. After training the MLP using the data of one tower with six input variables and application of the training results to an independent data set of another tower, Leufen and Schädler (2019) found comparable performance of MOST and MLP. It is likely that the advanced results of our ML methods are due to the larger input data set such as the surface temperature gradient and radiation variables. Leufen and Schädler (2019) did not investigate latent heat fluxes and they did not perform a sensitivity study with respect to the MLP.

In summary, the comparison of our work shows correspondence with previous results although the input data sets for the various ML methods were different. In the future, it is important to compare the performance of ML methods at more sites and to use harmonized data sets as input variables in order to make the results comparable. In any case, our work demonstrates the importance of the incorporation of gradient information in combination with a large suite of potential driving variables as well as the structural deficits of MOST and BRN retrievals with respect to  $SH$  and  $LH$ .

## 7 Summary

In this work, we applied data of three towers operated during LAFE for one month in August 2017 in order to study retrievals of the friction velocity  $u_*$ , the surface heat flux  $SH$ , and the latent heat flux  $LH$  under unstable conditions during daytime. The retrievals were based on MOST and BRN relationships as well as two ML methods, the multilayer perceptron (MLP) and the extreme gradient boosting (XGB). For the determination of the fluxes using similarity relationships, we incorporated the observations in the MOST and BRN equations including fits of the specific parameters of the similarity functions as well as the integrated gradient functions for MOST.

Specifically, we are able to answer the scientific questions posed above:

1. How accurate are the fluxes derived by MOST and BRN using appropriate data analyses? In spite of deriving the best parameters of MOST and BRN to fit their similarity functions to the observations, their approximations of the fluxes  $u_*$ ,  $SH$ , and  $LH$  show increasingly significant deviations. These deviations were already visible for  $u_*$ , particularly in the slopes of the regression lines. The agreement with the observations of  $SH$



degraded further substantiated by an increase of RMSE, a reduction of the regression coefficient, a further deviation of the slopes from one, and an increase of the intercepts. Here BRN performed clearly better than MOST, particularly with respect to the slope. The strongest deviations in MOST and BRN retrievals were founded with respect to  $LH$ . Both MOST and BRN were not able to reproduce the weak dependence of  $LH$  on the moisture gradient. The slopes of the regression lines were far off the observations; the structural constraints due to the shapes of the similarity functions were not able to recover the pattern of the observations. Here MOST performed better than BRN, though still with overall unsatisfactory performance.

2. How do the derived fluxes from MOST and BRN compare with the output of the ML techniques? For all fluxes,  $u_*$ ,  $SH$ , and  $LH$ , the ML methods outperformed both MOST and BRN significantly. Whereas the improvement was rather weak with respect to  $u_*$ , it increased further for  $SH$  and  $LH$ . Particularly striking was that the MLP and the XGB were able to produce slopes of the regression lines that closely approximated one and yielded small intercepts. This is very promising for future studies. This performance was maintained, if only the first three input variables according to the importance weighting were taken for the ML training. The substantial improvements of  $SH$  and  $LH$ , in spite of the lack of additional ABL variables, point to structural deficits of MOST and BRN concerning the shape of the similarity functions and/or the definition of the scaling variables. Particularly, in the case of  $LH$ , the importance weighting indicates that flux–L–A variable relationship should be modified and should include mainly net and/or shortwave incoming radiation as well as soil moisture variables.
3. Can we use ML techniques to generate improved results and an identification of the most important drivers of the fluxes? Yes, the key is to incorporate and to apply techniques for importance weighting like the ones introduced in this work, or related techniques, such as automatic relevance determination. The results also highlight that ML techniques should not be considered to constitute pure black boxes. Importance weighting methods can provide key insights into the factors that critically influence L–A dynamics, such as the main drivers of fluxes. For instance, the importance weighting confirmed the expected main dependencies of  $u_*$  on the surface wind and its gradients as well as the major dependency of  $SH$  on the temperature gradient. However, also a significant dependency on  $U_{3m}$  was detected. For  $LH$ , in contrast to the basic assumptions of MOST and BRN, the main driving variables turned out to be the net radiation as well as the incoming and outgoing shortwave radiations.

At this stage, it is not possible yet to claim or to study a universality of our results. Therefore, in the future, we will test our ML methods on more tower data from other sites and include additional variables such as  $VPD$  and  $LAI$ . Furthermore, we suggest to harmonize these efforts internationally by the development and application of common data sets for the training of ML methods. A very important role will play the importance weighting analyses. Using these results at different sites all over the Earth, we are confident that ML methods, such as MLP and XGB, are fundamental tools for the derivation of advanced parameterizations of surface fluxes and their incorporation in L–A system models.

**Acknowledgements** LAFE was funded by the DOE, NASA, and NOAA in the US as well as the BMBF in Germany. Support concerning the preparation and performance of LAFE by the ARM program via Nicki Hickmon, Chris Martin, James Martin, Mike Ritsche, and John Schatz is highly appreciated. We acknowledge the support of Michael Buban concerning the collection of the tower data. SO, MK, and MVB acknowledge support from the DFG Cluster of Excellence “Machine Learning–New Perspectives for Science,” EXC 2064/1, project number 390727645 as well as from the Cyber Valley (CyVy–RF–2020–15). We also appreciate the suggestions of Prof. Dr. Thomas Foken and Prof. Dr. Matthias Mauder with respect to data quality control.



All LAFE data, particularly the tower data that are subject of this work, can be retrieved via the US ARM data archive at <https://www.arm.gov/data>.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Andreas EL, Hicks BB (2002) Comments on “critical test of the validity of Monin–Obukhov similarity during convective condition”. *J Atmos Sci* 59(17):2605–2607. <https://doi.org/10.1175/1520-0469>
- Basu S, Lacser A (2017) A cautionary note on the use of Monin–Obukhov similarity theory in very high-resolution large-eddy simulations. *Boundary-Layer Meteorol* 163:351–355
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Brownlee J (2016) XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. *Machine Learning Mastery*
- Businger JA, Wyngaard JC, Izumi Y, Bradley EF (1971) Flux-profile relationships in the atmospheric surface layer. *J Atmos Sci* 28(2):181–189. <https://doi.org/10.1175/1520-0469>
- Cheng Y, Li Q, Li D, Gentine P (2021) Logarithmic profile of temperature in sheared and unstably stratified atmospheric boundary layers. *Phys Rev Fluids* 6(034):606. <https://doi.org/10.1103/PhysRevFluids.6.034606>
- Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, association for computing machinery, New York, NY, USA*, pp 785–794
- Deardorff JW (1972) Parameterization of the planetary boundary layer for use in general circulation models. *Mon Weather Rev* 100(2):93–106
- Dyer AJ (1974) A review of flux-profile relationships. *Boundary-Layer Meteorol* 7(3):363–372. <https://doi.org/10.1007/BF00240838>
- Dyer AJ, Hicks BB (1970) Flux-gradient relationships in the constant flux layer. *Q J R Meteorol Soc* 96(410):715–721. <https://doi.org/10.1002/qj.49709641012>
- Foken T (2016) *Micrometeorology*, 2nd edn. Springer-Verlag, Berlin Heidelberg. ISBN 978-3-642-25439-0, <https://doi.org/10.1007/978-3-642-25440-6>
- 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>
- Goodfellow IJ, Bengio Y, Courville A (2016) *Deep learning*. MIT Press, Cambridge, MA, USA
- Harman IN, Finnigan JJ (2007) A simple unified theory for flow in the canopy and roughness sublayer. *Boundary-Layer Meteorol* 123:339–363. <https://doi.org/10.1007/s10546-006-9145-6>
- Hess G, Hicks B, Yamada T (1981) The impact of the Wangara experiment. *Boundary-Layer Meteorol* 20:135–174
- Hicks B (1976) Wind profile relationships from the “Wangara” experiment. *Q J R Meteorol Soc* 102:535–551. <https://doi.org/10.1002/qj.49710243304>
- Högström U (1996) Review of some basic characteristics of the atmospheric surface layer. *Boundary-Layer Meteorol* 78(3):215–246. <https://doi.org/10.1007/BF00120937>
- Jiménez PA, Dudhia J, González-Rouco JF, Navarro J, Montávez JP, García-Bustamante E (2012) A revised scheme for the WRF surface layer formulation. *Mon Weather Rev* 140:898–918
- Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. In: *3rd international conference for learning representations (ICLR)*
- Kljun N, Calanca P, Rotach MW, Schmid HP (2015) A simple two-dimensional parameterisation for flux footprint prediction (ffp). *Geosci Model Dev* 8(6):3695–3713. <https://doi.org/10.5194/gmd-8-3695-2015>
- Lee TR, Buban M (2020) Evaluation of Monin–Obukhov and bulk Richardson parameterizations for surface-atmosphere exchange. *J Appl Meteorol Clim* 59(6):1091–1107

- Lee TR, Buban M, Meyers TP (2021) Application of bulk Richardson parameterizations of surface fluxes to heterogeneous land surfaces. *Mon Weather Rev* 149(10):3243–3264
- Lee T, Meyers T (2022) New parameterizations of turbulence statistics for the atmospheric surface layer. *Mon Weather Rev* published online ahead of print 2022. <https://journals.ametsoc.org/view/journals/mwre/aop/MWR-D-22-0071.1/MWR-D-22-0071.1.xml>
- Leufen LH, Schädler G (2019) Calculating the turbulent fluxes in the atmospheric surface layer with neural networks. *Geosci Model Dev* 12(5):2033–2047
- 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(10):3403–3431. <https://doi.org/10.1175/JAS-D-17-0301.1>
- Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: 7th international conference on learning representations (ICLR)
- Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In: in ICML workshop on deep learning for audio, speech and language processing
- Markowski P, Lis N, Turner D, Lee T, Buban M (2019) Observations of near-surface vertical wind profiles and vertical momentum fluxes from vortex-southeast 2017: Comparisons to monin-obukhov similarity theory. *Mon Weather Rev* 147(10):3811–3824. <https://doi.org/10.1175/MWR-D-19-0091.1>
- Maronga B (2014) Monin–Obukhov similarity functions for the structure parameters of temperature and humidity in the unstable surface layer: Results from high-resolution large-eddy simulations. *J Atmos Sci* 71:716–733
- Maronga B, Reuder J (2017) On the formulation and universality of Monin–Obukhov similarity functions for mean gradients and standard deviations in the unstable surface layer: Results from surface-layer-resolving large-eddy simulations. *J Atmos Sci* 74:989–1010
- Mauder M, Foken T (2015) Documentation and instruction manual of the eddy-covariance software package tk3 (update). University of Bayreuth, Micrometeorology, Tech Rep Arbeitsergebnisse Nr, p 62
- Mauritsen T, Svensson G, Zilitinkevich SS, Esau I, Enger L, Grisogono B (2007) A total turbulent energy closure model for neutrally and stably stratified atmospheric boundary layers. *J Atmos Sci* 64(11):4113–4126
- McNaughton KG, Brunet Y (2002) Townsend’s hypothesis, coherent structures and Monin–Obukhov similarity. *Boundary-Layer Meteorol* 102(2):161–175. <https://doi.org/10.1023/A:1013171312407>
- Monin A, Obukhov A (1954) Osnovnye zakonomernosti turbulentnogo peremeshivaniya v prizemnom sloe atmosfery (basic laws of turbulent mixing in the atmosphere near the ground). *Tr Inst Teor Geofiz, Akad Nauk SSSR* 24(151):163–187
- Morrison T, Calaf M, Higgins CW, Drake SA, Perelet A, Pardyjak E (2021) The impact of surface temperature heterogeneity on near-surface heat transport. *Boundary-Layer Meteorol* 180(2):247–272. <https://doi.org/10.1007/s10546-021-00624-2>
- Nourani V, Sayad Fard M (2012) Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. *Adv Eng Softw* 47:127–146
- Oliveira BRF, Schaller C, Keizer JJ, Foken T (2021) Estimating immediate post-fire carbon fluxes using the eddy-covariance technique. *Biogeosciences* 18(1):285–302. <https://doi.org/10.5194/bg-18-285-2021>
- Pal S, Sharma P (2019) A review of machine learning applications in land surface modeling. *Earth* 2:174–190. <https://doi.org/10.3390/earth2010011>
- Panofsky H, Tennekes H, Lenschow D, Wyngaard J (1977) The characteristics of turbulent velocity components in the surface layer under convective conditions. *Boundary-Layer Meteorol* 11:355–361. <https://doi.org/10.1007/BF02186086>
- Prandtl L (1925) 7. bericht über untersuchungen zur ausgebildeten turbulenz. *ZAMM-Z Angew Math Me* 5(2):136–139. <https://doi.org/10.1002/zamm.19250050212>
- Qin Z, Su G, Yu Q, Hu B, Li J (2005) Modeling water and carbon fluxes above summer maize field in north china plain with back-propagation neural networks. *J Zhejiang Univ Sci B* 6(5):418–426. <https://doi.org/10.1631/jzus.2005.B0418>
- Qin Z, Yu Q, Li J, Wu Z, Hu B (2005) Application of least squares vector machines in modelling water vapor and carbon dioxide fluxes over a cropland. *J Zhejiang Univ Sci B* 6(6):491–495. <https://doi.org/10.1631/jzus.2005.B0491>
- Qin Z, li Su G, en Zhang J, Ouyang Y, Yu Q, Li J (2010) Identification of important factors for water vapor flux and co2 exchange in a cropland. *Ecol Model* 221(4):575–581. <https://doi.org/10.1016/j.ecolmodel.2009.11.007>
- Safa B, Arkebauer TJ, Zhu Q, Suyker A, Irmak S (2018) Latent heat and sensible heat flux simulation in maize using artificial neural networks. *Comput Electron Agric* 154:155–164. <https://doi.org/10.1016/j.compag.2018.08.038>

- Santanello JA, Dirmeyer P, Ferguson C, Findell K, Tawfik A, Berg A, Ek M, Gentine P, Guillod B, van Heerwaarden C, Roundy R, Wulfmeyer V (2018) loco perspective. *B Am Meteorol Soc* 99:1253–1272. <https://doi.org/10.1175/BAMS-D-17-0001.1>
- Sorbjan Z (2006) Statistics of scalar fields in the atmospheric boundary layer based on large-eddy simulations. Part ii: Forced convection. *Boundary-Layer Meteorol* 119:57–79
- Sorbjan Z (2010) Gradient-based scales and similarity laws in the stable boundary layer. *Q J R Meteorol Soc* 136(650):1243–1254
- Späth F, Behrendt A, Brewer A, Lange D, Senff S, Turner D, Wagner T, Wulfmeyer V (2022) Simultaneous observations of surface layer profiles of humidity, temperature and wind using scanning lidar instruments. *J Geophys Res-Atmos*. <https://doi.org/10.1029/2021JD035697>
- van de Boer A, Moene AF, Graf A, Schüttemeyer D, Simmer C (2014) Detection of entrainment influences on surface-layer measurements and extension of Monin–Obukhov similarity theory. *Boundary-Layer Meteorol* 152(1):19–44. <https://doi.org/10.1007/s10546-014-9920-8>
- van Heerwaarden CC, Vilà-Guerau de Arellano J, Moene AF, Holtslag AAM (2009) Interactions between dry-air entrainment, surface evaporation and convective boundary-layer development. *Q J R Meteorol Soc* 135(642):1277–1291. <https://doi.org/10.1002/qj.431>
- Wang X, Yao Y, Zhao S, Jia K, Zhang X, Zhang Y, Zhang L, Xu J, Chen X (2017) Modis-based estimation of terrestrial latent heat flux over north America using three machine learning algorithms. *Remote Sensing* 9(12):1326
- Wang L, Zhang Y, Yao Y, Xiao Z, Shang K, Guo X, Yang J, Xue S, Wang J (2021) Gbrt-based estimation of terrestrial latent heat flux in the Haihe river basin from satellite and reanalysis datasets. *Remote Sensing* 13(6):1054
- Wizemann HD, Ingwersen J, Högy P, Warrach-Sagi K, Streck T, Wulfmeyer V (2015) Three year observations of water vapor and energy fluxes over agricultural crops in two regional climates of southwest germany. *Meteorol Z* 24(1):39–59. <https://doi.org/10.1127/metz/2014/0618>
- Wulfmeyer V, Turner D, Baker B, Banta R, Behrendt A, Bonin T, Brewer W, Buban M, Choukulkar A, Dumas E, Hardesty R, Heus T, Ingwersen J, Lange D, Lee T, Metzendorf S, Muppa S, Meyers T, Newsom R, Osman M, Raasch S, Santanello J, Senff D, Späth F, Wagner T, Weckwerth T (2018) A new research approach for observing and characterizing land-atmosphere feedback. *B Am Meteorol Soc* 99:1639–1667. <https://doi.org/10.1175/BAMS-D-17-0009.1>
- Xu T, Guo Z, Liu S, He X, Meng Y, Xu Z, Xia Y, Xiao J, Zhang Y, Ma Y et al (2018) Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *J Geophys Res Atmosph* 123(16):8674–8690
- Zhang G (2008) Neural networks for data mining, In: Maimon O, Rokach L (eds) *Soft computing for knowledge discovery and data mining*. Springer US, Boston, MA, chap 21 pp 17–44, [https://doi.org/10.1007/978-0-387-69935-6\\_2](https://doi.org/10.1007/978-0-387-69935-6_2)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.