Gap-Filling of Turbulent Heat Fluxes over Rice–Wheat-Rotation Croplands Using the Random Forest Model

Jianbin Zhang[1], Zexia Duan[1], Shaohui Zhou[1], Yubin Li[1], Zhiqiu Gao[1]

[1] School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing, 210044, China

Correspondence to: Dr. Yubin Li (liyubin@nuist.edu.cn)

Abstract. This study investigated the accuracy of the Random Forest (RF) model in gap-filling the sensible (H) and latent heat (LE) fluxes, by using the observation data collected at a site over rice–wheat-rotation croplands in Shouxian County of eastern China from 15 July 2015 to 24 April 2019. Firstly, the variable significances of the machine learning (ML) model's five input variables, including the net radiation (Rn), winds speed (WS), temperature (T), relative humidity (RH), and air pressure (P), were examined, and it was found that Rn accounted for 78% and 76% of the total variable significance in H and LE calculating, respectively, showing that it was the most important input variable. Secondly, the RF model's accuracy with the five-variable (Rn, WS, T, RH, P) input combination was evaluated, and the results showed that the RF model could reliably gap-fill the H and LE with mean absolute errors (MAEs) of 5.88 Wm^{-2} and 20.97 Wm^{-2}, and root mean square errors (RMSEs) of 10.67 Wm^{-2} and 29.46 Wm^{-2}, respectively. Thirdly, 4-variable input combinations were tested, and it was found that the best input combination was (Rn, WS, T, P) by removing RH from the input list, and its MAE values of H and LE were reduced by 12.65% and 7.12%, respectively. At last, through the Taylor diagram, H and LE gap-filling accuracies of the RF model, the support vector machine (SVM) model, the k-nearest neighbor (KNN) model, and the gradient boosting decision tree (GBDT) model were inter-compared, and the statistical metrics showed that RF was the most accurate for both H and LE gap-filling, while the LR and KNN model performed the worst for H and LE gap-filling, respectively.

1 Introduction

The turbulent fluxes between the atmosphere and the ground play a crucial role in global climate change and atmospheric circulation, and the inaccuracy of long-term observations of surface turbulent fluxes is a major factor in erroneous weather predictions and climate projections. Research on the ecological effects of urban green spaces, agricultural ecosystems, and forests all use surface turbulent fluxes as key indicators. Currently, the eddy covariance (EC) technique can be used to directly measure the turbulent fluxes (Wilson et al., 2001; Jiang et al., 2021; Wang et al., 2021). However, due to sensor failure and adverse meteorological factors (such as rainfall and frost), these high-frequency turbulence data are subject to errors (Khan et al., 2018). As a result, it is difficult to obtain a continuous time series of ground-based turbulent fluxes. Furthermore, quality assurance methods lead to unavailable sections of flux datasets (Nisa et al., 2021). Based on the above reasons, gap-filling is in need to retrieve continuous datasets of EC-based fluxes. Researchers have developed approaches based on existing meteorological information to fill up the gaps in atmospheric databases, such as interpolation, nonlinear regression, mean diurnal method, and sampling techniques from the marginal distribution (Falge et al., 2001; Hui et al., 2004; Stauch et al., 2006; Foltnov et al., 2020). Further, the ML technique has also become an effective method to be used in the calculation of turbulent fluxes (McCandless et al., 2022).

As a result of recent developments in high computing technology, machine-learning-based algorithms have been developed and successfully used in various areas, such as natural language processing, data mining, biometrics, computer vision, search engines, clinical applications, video games, robots, etc. To address the missing data issue, machine-learning-based models have recently been used to fill data gaps in meteorological elements and turbulent fluxes (Bianco et al., 2019; Yu et al., 2020). As a result of their reliable and repeatable results, these models are now regarded as a standard gap-filling algorithm (Beringer et al., 2017; Isaac et al., 2017). ML algorithms have several deficiencies even if they perform well in some areas. For instance, over-fitting is a major concern that can occur when the training window is too short or the training dataset's quality is poor. That's because the present ML approaches are not sufficiently adaptable to work in extreme situations with large values (Kunwor et al., 2017; Moffat et al., 2007). Furthermore, even with the best technique, the model uncertainty of gap-filling still plays a role, particularly when the gaps are relatively large. Numerous novel ML and optimization algorithms have been created and put to use in numerous scientific domains since the 2000s, and their superiority has been demonstrated, either singly or as a component of a hybrid or ensemble model (e.g. Gani et al., 2016).

Based on the need for fluxes dataset gap-filling, and the effectivity of the ML technique, this paper aims, firstly, to investigate the performance of the RF machine learning algorithm trained from a dataset obtained over rice-wheat-rotation croplands in Shouxian County, eastern China, in gap-filling the sensible and latent heat fluxes; and secondly, to analyze the RF model's accuracy with various meteorological input combinations during training; and thirdly, to compare the performance of RF model with other four typical ML models.

2 Materials and Methods

2.1 Study area

This observation was conducted at a site in Shouxian County in the eastern Chinese province of Anhui (32.42 °N, 116.76 °E) (Figure 1). The altitude of the site is 27 meters, and the annual mean air temperature and annual cumulative precipitation here are 16 ℃ and 1115 mm, respectively. Summer (from June to September) precipitation accounts for nearly 60% of the annual precipitation amount, which meets the high water demand of rice. Drought sometimes occurs due to lack of precipitation in the growing season of wheat. This observation site is rather flat, with farmland accounting for more than 90% of the area. Winter wheat is grown here from November until late May, while from June to November the field is flooded, plowed, and harrowed as rice paddies (Duan et al., 2021) (Figure 2). The subtropical northern boundary of the monsoon humid climatic type describes the area's climate.

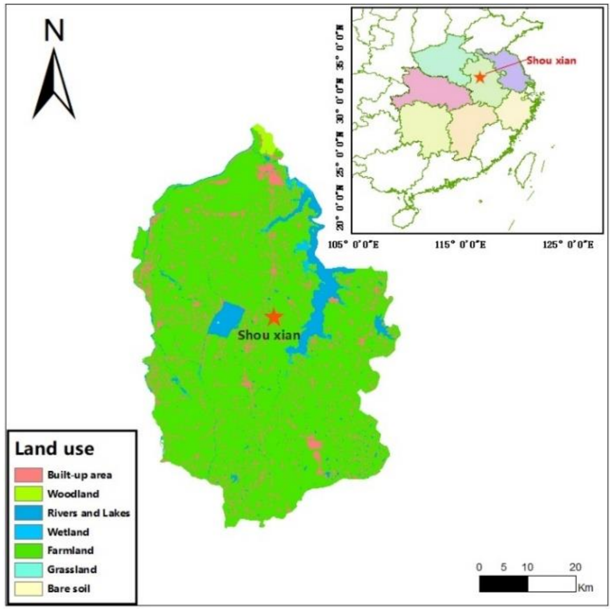


Figure 1. Geographical location and land-cover map of Shouxian County.

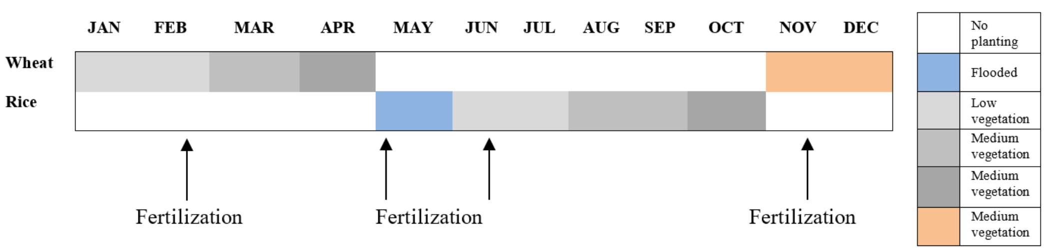


Figure 2. Crop calendars for the rice and wheat in the North Yangtze River Delta region.

2.2 Data

Over the site described above, EC sensors (EC 150, Campbell Scientific Inc., Logan, UT, USA) were installed at 2.5 meters above the ground, including a three-dimensional sonic anemometer (CSAT3, Campbell Scientific Inc., Logan, UT, USA) and a CO_2/H_2O open-path infrared gas analyzer. The sensible and latent heat fluxes were computed half-hourly using EddyPro software, with time lag compensation, double coordinate rotation, spectrum correction, and Webb-Pearman-Leuning density correction (Wutzler et al., 2018; Anapalli et al., 2019). Poor-quality fluxes (Eddypro quality check flag value=2) were discarded. And a quality check based on the relationship between the measured flux and friction velocity was carried out to

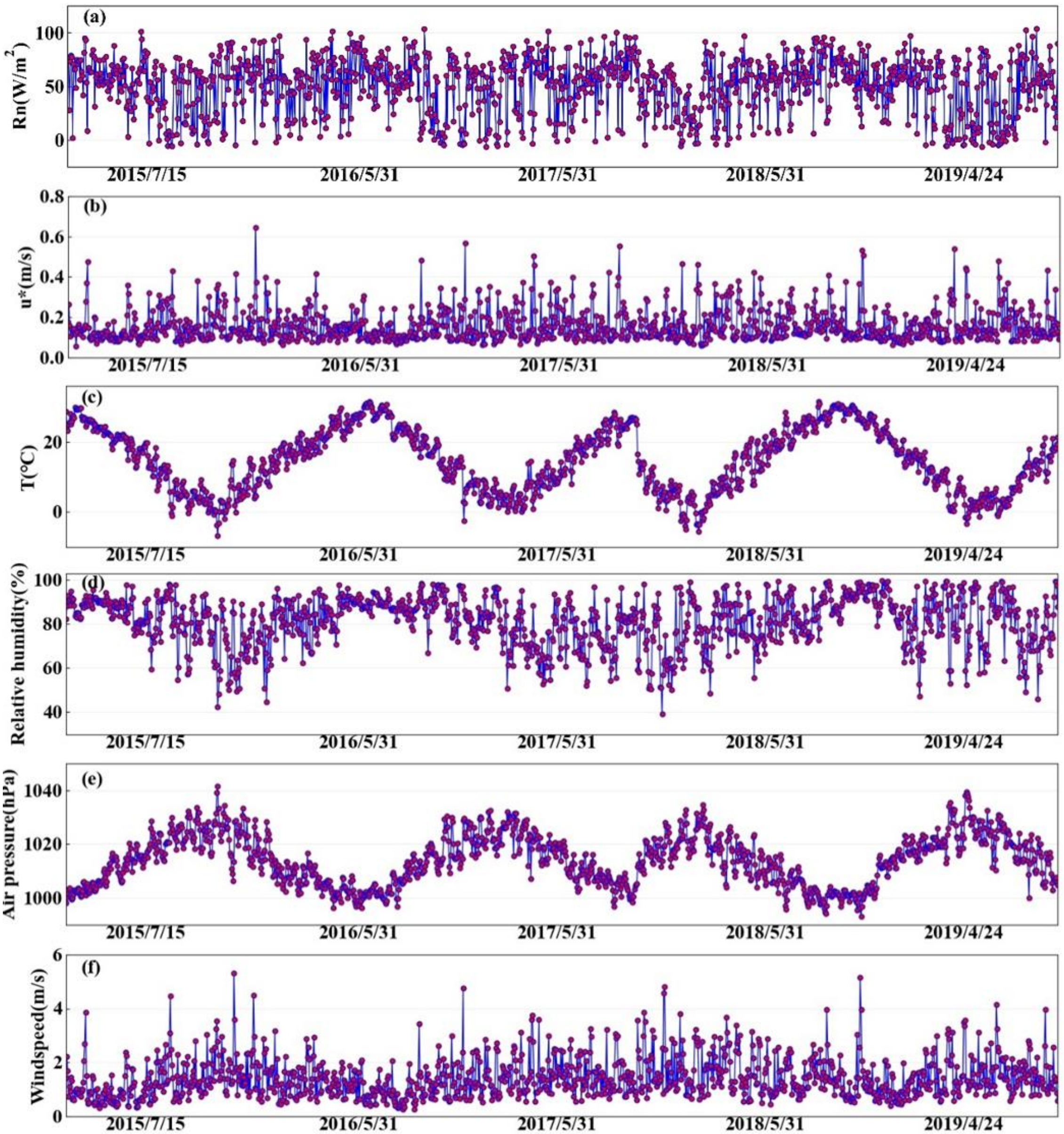


Figure 3. Daily averaged a) Rn: net radiation(Wm^{-2}) , b) u*: friction velocity(m/s), c) T: air temperature(℃), d) RH: relative humidity(%), e) P: air pressure(hPa), and f) WS:wind speed(m s^{-1}).

remove the biased data (Papale et al., 2006). Then, using the marginal distribution sampling technique, the flow data were gap-filled (Reichstein et al., 2005). The time series of air temperature, relative humidity, wind speed, air pressure, friction velocity, and net radiation were also subjected to quality control. The missing data which need gapfilling are H and LE, with 7205 and 16013 missing, accounting for 12.09% and 26.87% respectively. According to the criteria of X(h) < (X − 4σ) or X(h) > (X + 4σ), where X(h) indicates the time series of the component, X is the mean across the averaging interval, and σ is the standard deviation, noisy data were eliminated (Gao et al., 2003). Data observed from 15 July 2015 to 24 April 2019 are used in this study, and Figure 3 shows the daily average data of Rn: net radiation(W m^{-2}), u*: friction velocity(m/s), T: air temperature(℃), RH: relative humidity(%), P: air pressure(hPa), and WS: wind speed(m s^{-1}).

2.3 The RF Model

RF is a machine learning method that is quick, adaptable, and frequently used to analyze classification and regression jobs (Breiman, 2001). This model can successfully evaluate highly dimensional and multicollinear data and is resistant to overfitting (Belgiu et al., 2016). The RF model provides a feature-selection tool to assist in determining the importance of the predictor. The contribution of each variable to the model, with important variables having a higher effect on the results of the model evaluation, is the definition of feature significance (Liu et al., 2021). 90% of the data collected at the Shouxian observation site throughout the study period were used to train the RF model, while the remaining 10% was used to independently validate the model (hereafter, validation dataset). To lessen the overfitting in this case, a 10-fold cross-validation (CV) procedure was used (Cai et al., 2020). All training data used here was randomly divided into ten subsamples of equal size for the 10-fold CV tests. And nine out of the ten subsamples were used as training data (hereafter, training dataset), while the remaining subsample was used as testing data (hereafter, testing dataset). All ten of the subsamples were utilized as testing data exactly once for each of the 10 iterations of the CV procedure. One estimate was created by averaging the 10 findings from the folds. We modified the four RF model hyperparameters based on Bayesian optimization to get the optimal model (Baareh et al., 2021; Frazier, P.I., 2018): the maximum number of features considered to split a node (Max features), the maximum number of trees to build (n estimators), the minimum sample number placed in a node prior to the node being split (min split), and the maximum number of levels for each decision tree (Max depth). Bayesian optimizer is used to tune parameters, you can quickly find an acceptable hyperparameter value, compared with grid search, the advantage is that the number of iterations is less (time saving), the granularity can be very small. For example, if we want to adjust the regularized hyperparameters of linear regression, we set the black box function to linear regression, the independent variable is a hyperparameter, the dependent variable is linear regression in the training set accuracy, set an acceptable black box function dependent variable value, such as 0.95, the obtained hyperparameter result is a hyperparameter that can make the linear regression accuracy exceed 0.95. The simulated performance of the 10-fold CV outcomes was evaluated using four statistical metrics: the correlation coefficient (*r*), mean absolute error (MAE), root mean square error (RMSE), and standard deviation(σ_n). As a result, the final RF model's parameters were adjusted to n estimators = 246, min split = 2, Max features = 10, and Max depth = 35, to have the best statistical metrics.

The four statistical metrics are calculated by:

$$r = \frac{\sum_{i=1}^{N}(S_i-\bar{S})(O_i-\bar{O})}{\sqrt{\sum_{i=1}^{N}(S_i-\bar{S})^2}\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2}}, \quad (1)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|S_i-O_i|, \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(S_i-O_i)^2}{N}}, \quad (3)$$

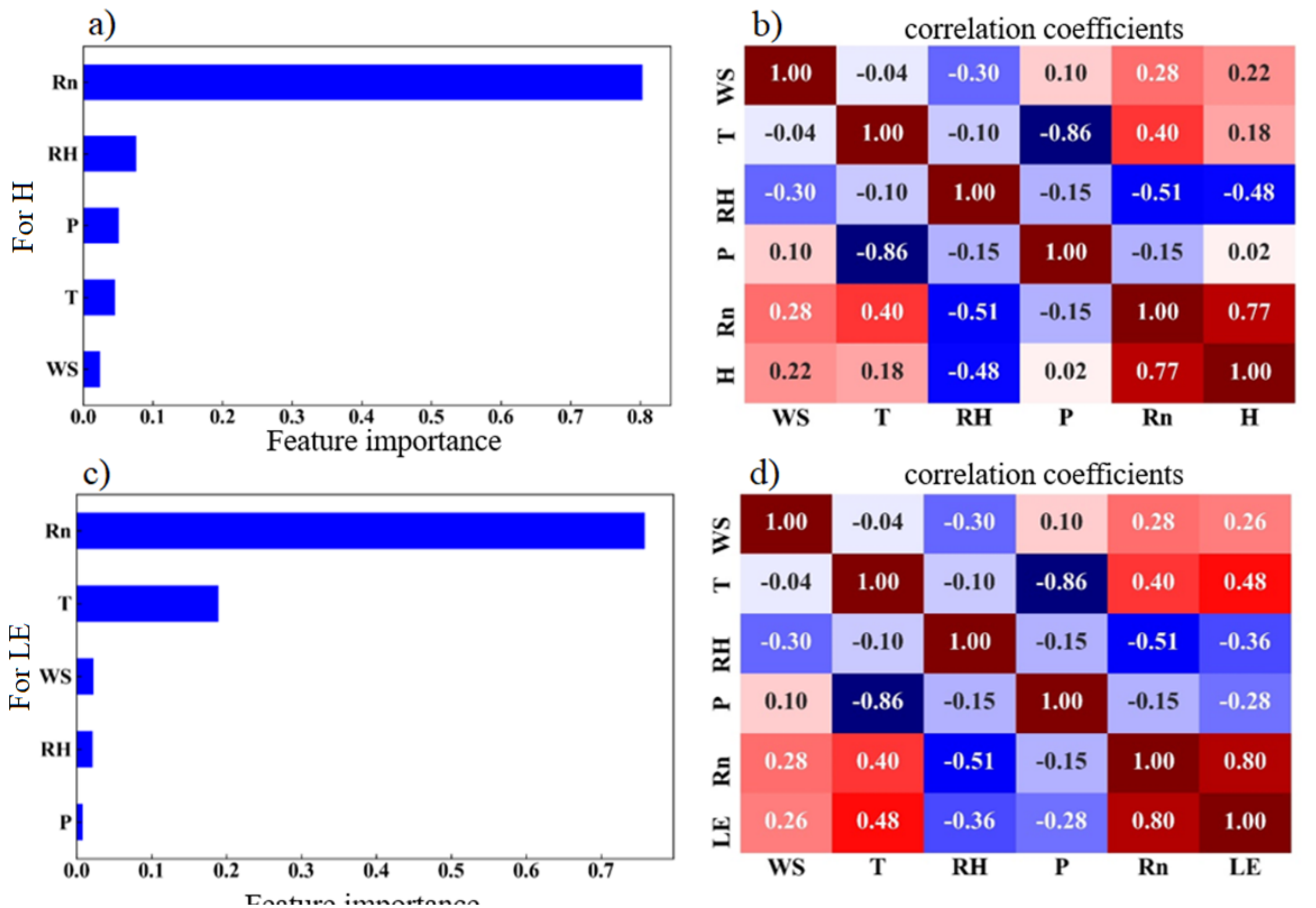
$$\sigma_n = \frac{\sqrt{\sum_{i=1}^{N}(S_i-O_i)^2}}{N}. \quad (4)$$

where S stands for the modeled value, O is the observation, $\bar{O}$ is the mean observed value, and $\bar{S}$ is the mean modeled observation, σ_n indicates the standard deviation. The subscript i represents the serial number of samples, and N represents the total number of samples.

3 Results and discussion

3.1 Driving Factors of H and LE on a Seasonal Scale

The possible driving factors of H and LE were investigated to determine their respective contributions by the RF model as shown in Figure 4. Rn, which accounted for 78% and 76% of the total variable significance of H and LE, respectively, was the most crucial variable in regulating the heat fluxes (Figures 4a and 4c). Consistent with the high variable significance values, H and LE also had the highest r of 0.79 and 0.75 with H and LE, respectively, as shown in Figures 4b and 4d. The other four factors contributed much smaller than Rn, and WS, T, RH, and P had importance values of 2%, 4%, 7%, and 5% (2.2%, 19%, 2%, and 0.6%) for H (LE), respectively. All these elements such as Rn, T, WS, RH are normalized before the model starts training. When these elements are normalized, it ensures uniformity and comparability. In general, all of these predictors played a role in the H and LE calculation, and for H, the sequence of importance was Rn, RH, P, T, and WS; while for LE, it was Rn, T, WS, RH, and P. The most significant impact on the change of H and LE came from Rn, which was the most important energy source of the surface and modulated the surface temperature directly. RH and T had a minor impact on the H and LE changes in terms of climatic parameters, which carried the information of the light-dependent reactions of H and LE fluxes. Particularly, WS and P had the minimal impacts on the H and LE fluxes. The WS, T, and RH also affected H and LE according to the Monin-Obukhov similarity theory (Monin and Obukhov, 1954), while P represented the contributions from the background weather systems.


a)
For H
Rn
RH
P
T
WS
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
Feature importance
b)
correlation coefficients
WS 1.00 -0.04 -0.30 0.10 0.28 0.22
T -0.04 1.00 -0.10 -0.86 0.40 0.18
RH -0.30 -0.10 1.00 -0.15 -0.51 -0.48
P 0.10 -0.86 -0.15 1.00 -0.15 0.02
Rn 0.28 0.40 -0.51 -0.15 1.00 0.77
H 0.22 0.18 -0.48 0.02 0.77 1.00
WS T RH P Rn H
c)
For LE
Rn
T
WS
RH
P
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
d)
correlation coefficients
WS 1.00 -0.04 -0.30 0.10 0.28 0.26
T -0.04 1.00 -0.10 -0.86 0.40 0.48
RH -0.30 -0.10 1.00 -0.15 -0.51 -0.36
P 0.10 -0.86 -0.15 1.00 -0.15 -0.28
Rn 0.28 0.40 -0.51 -0.15 1.00 0.80
LE 0.26 0.48 -0.36 -0.28 0.80 1.00
WS T RH P Rn LE

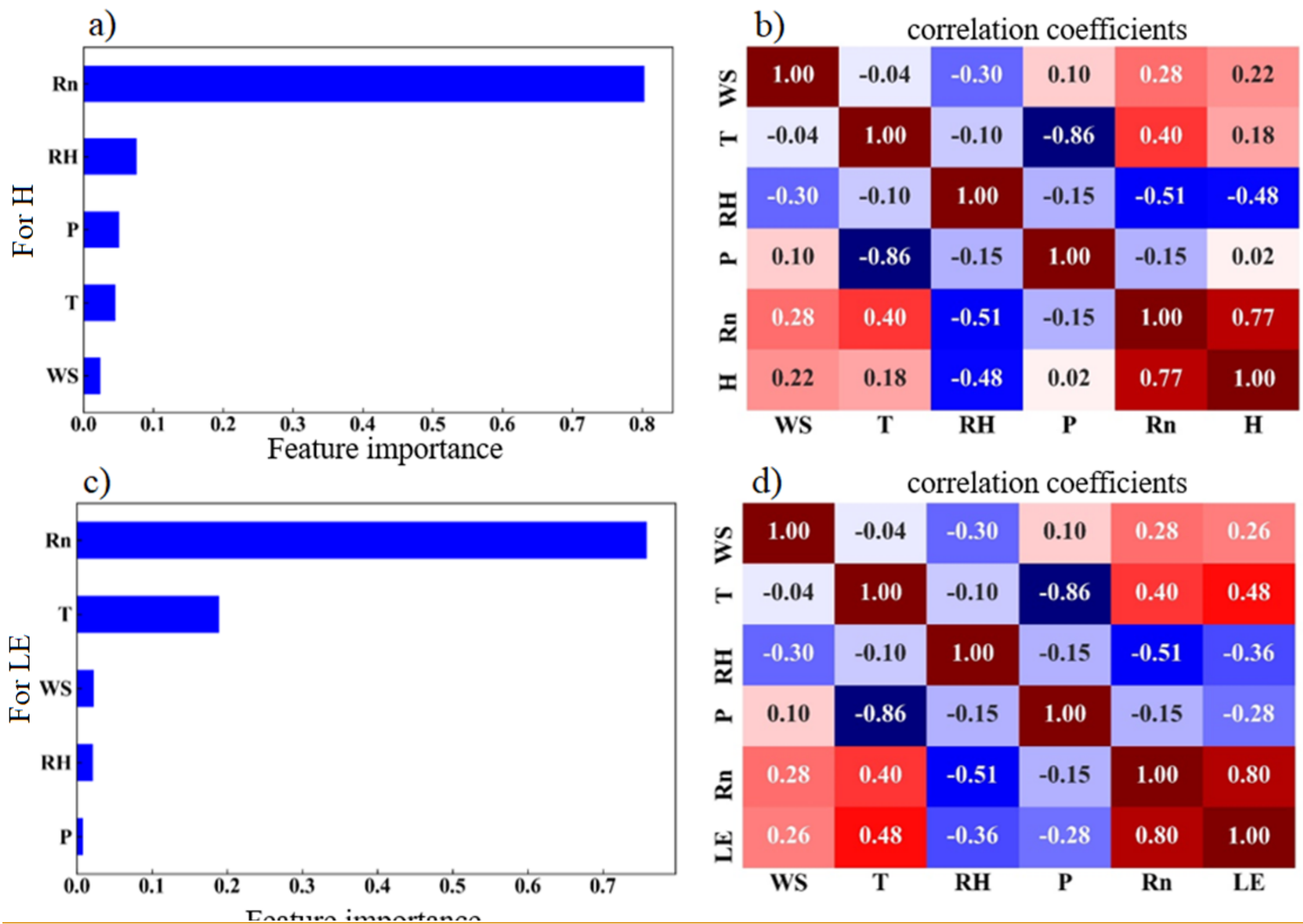
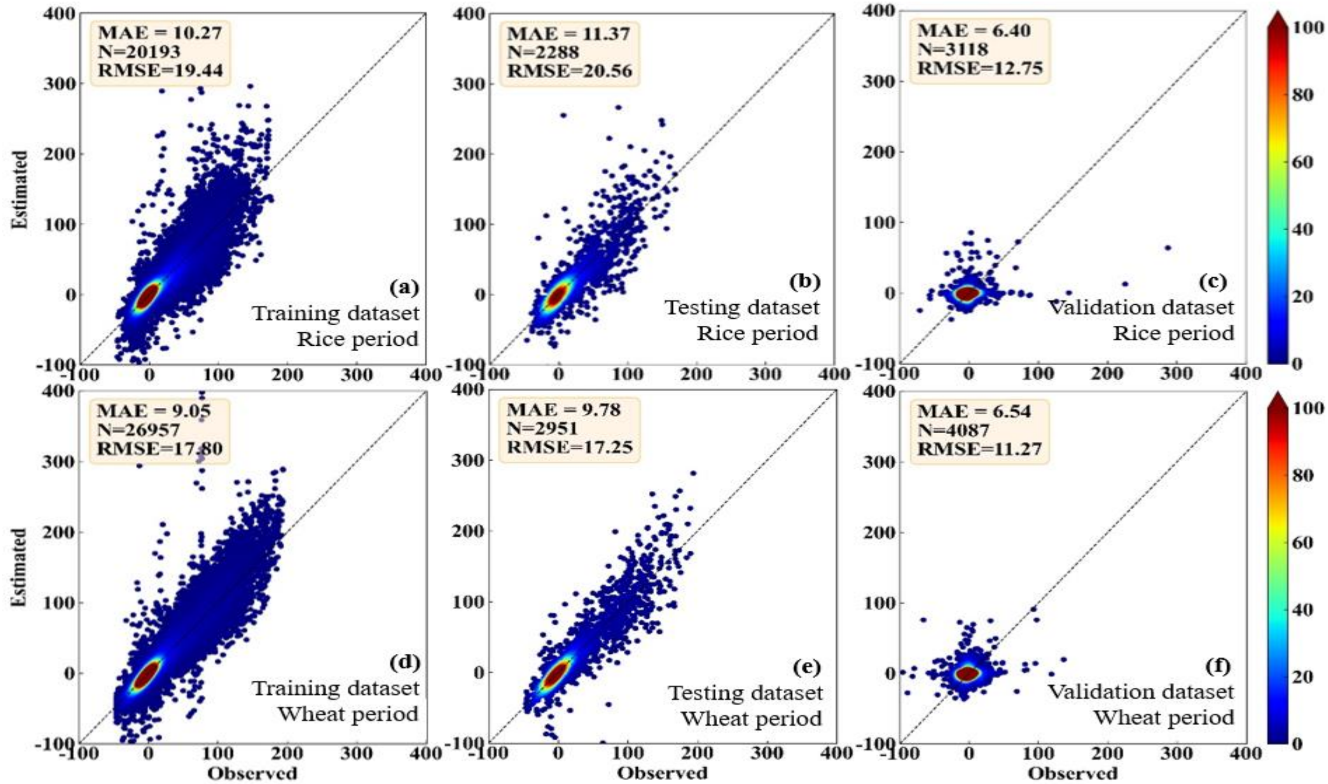


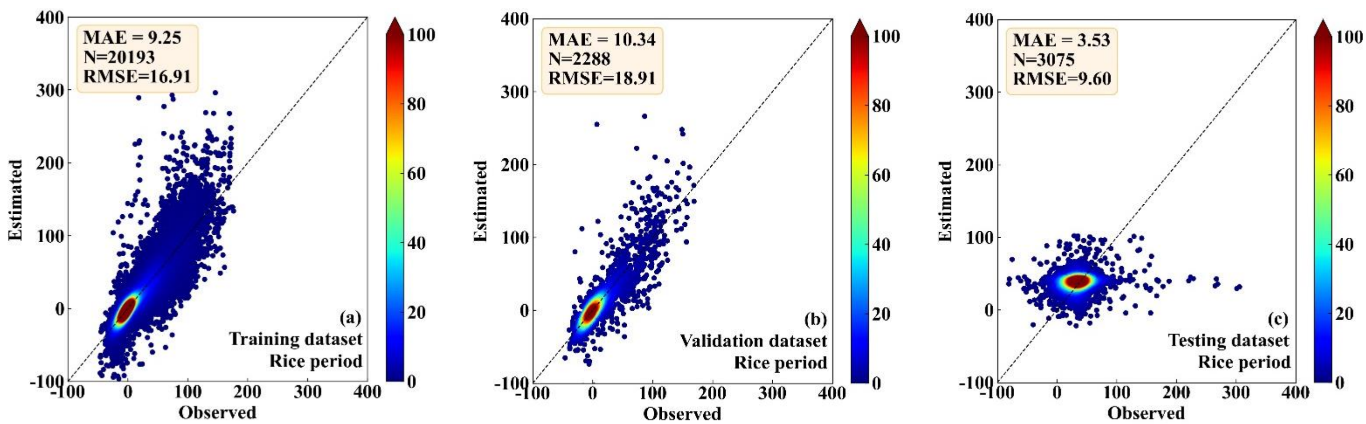
Figure 4. The feature importance of the variables for a) H and c) LE, and the correlation coefficient between each of the input variables for b) H and d) LE.

3.2 RF Model Evaluation

Figures 5-6 show the comparison between the observed and the RF-estimated H and LE, respectively. In the period of rice, the RF model showed good performance for both the training dataset (MAE =8.51 and 17.89 Wm^{-2}; RMSE =14.11 and 29.82 Wm^{-2}, for H and LE, respectively) and the testing dataset (MAE =9.61 and 10.34 Wm^{-2}, RMSE = 15.63 and 17.21Wm^{-2}, for H and LE, respectively) (Figures 5a, 5b, 6a, and 6b). RF model also showed high consistency with direct measurements for the validation dataset (MAE=5.88 and 20.97 Wm^{-2}, RMSE = 10.67 and 29.46 Wm^{-2}, for H and LE, respectively), (Figures 5c and 6c). In the period of wheat, the performance of the RF model for the training, testing, and validation datasets of H and LE was similar to that in the period of rice. For the training, testing, and validation datasets, respectively, the MAEs are 7.18, 8.01, and 6.01 Wm^{-2} for H, and 13.58, 8.82, and 19.93 Wm^{-2} for LE; and the RMSEs are12.27, 13.61, and 9.86 Wm^{-2} for H, and

24.92, 15.17, and 28.74 Wm^{-2} for LE (Figure 5d,e,f, Figure 6 d,e,f). These results demonstrate that the RF model is capable of effectively calculating the H and LE with input variables of Rn, WS, T, RH, and P.





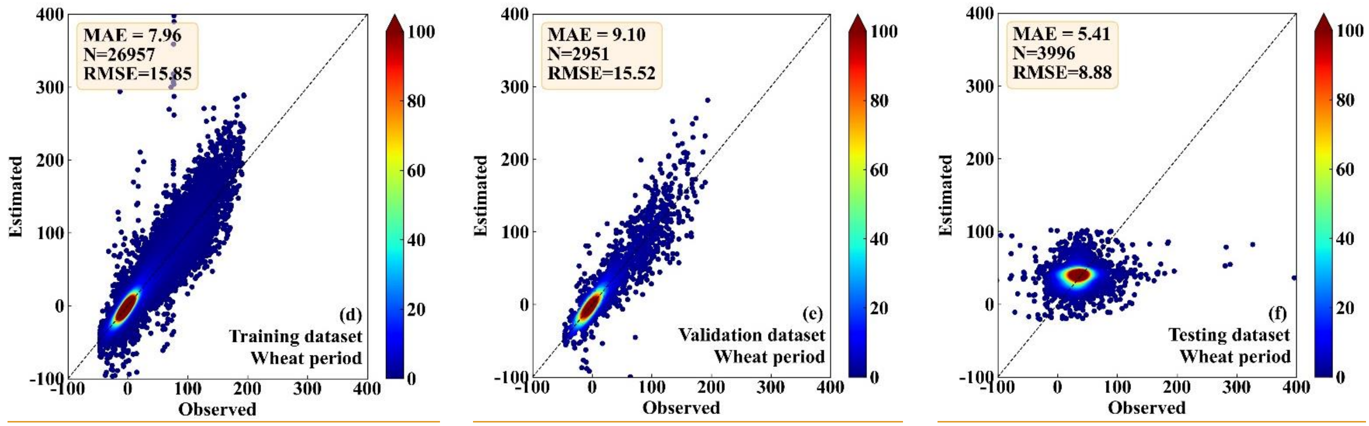
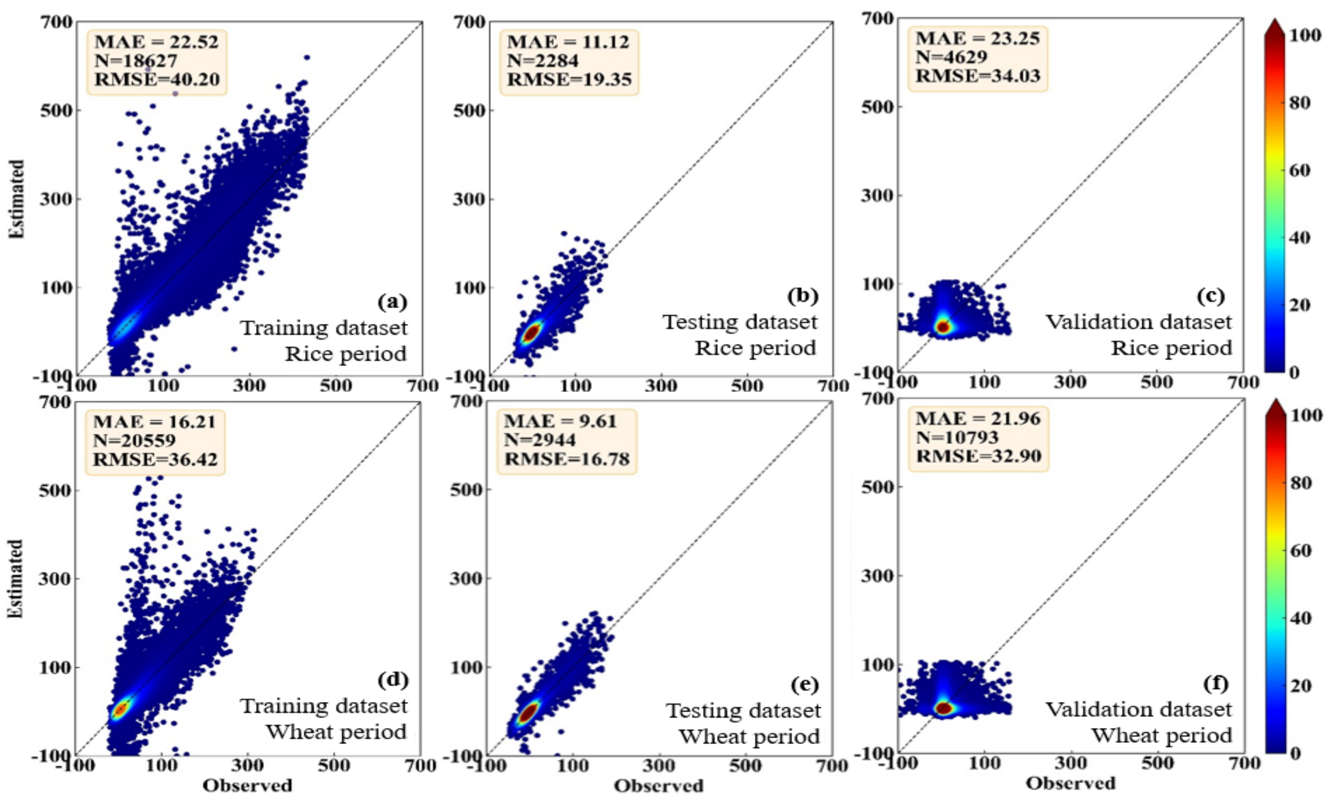


Figure 5. Scatter density plots of the observed and the RF-estimated H values, a) and d) for the training dataset, b) and e) for the ~~testing~~validation dataset, and c) and f) for the ~~validation~~testing dataset. And a), b) and c) are in the period of rice, while d), e) and f) are in the period of wheat.



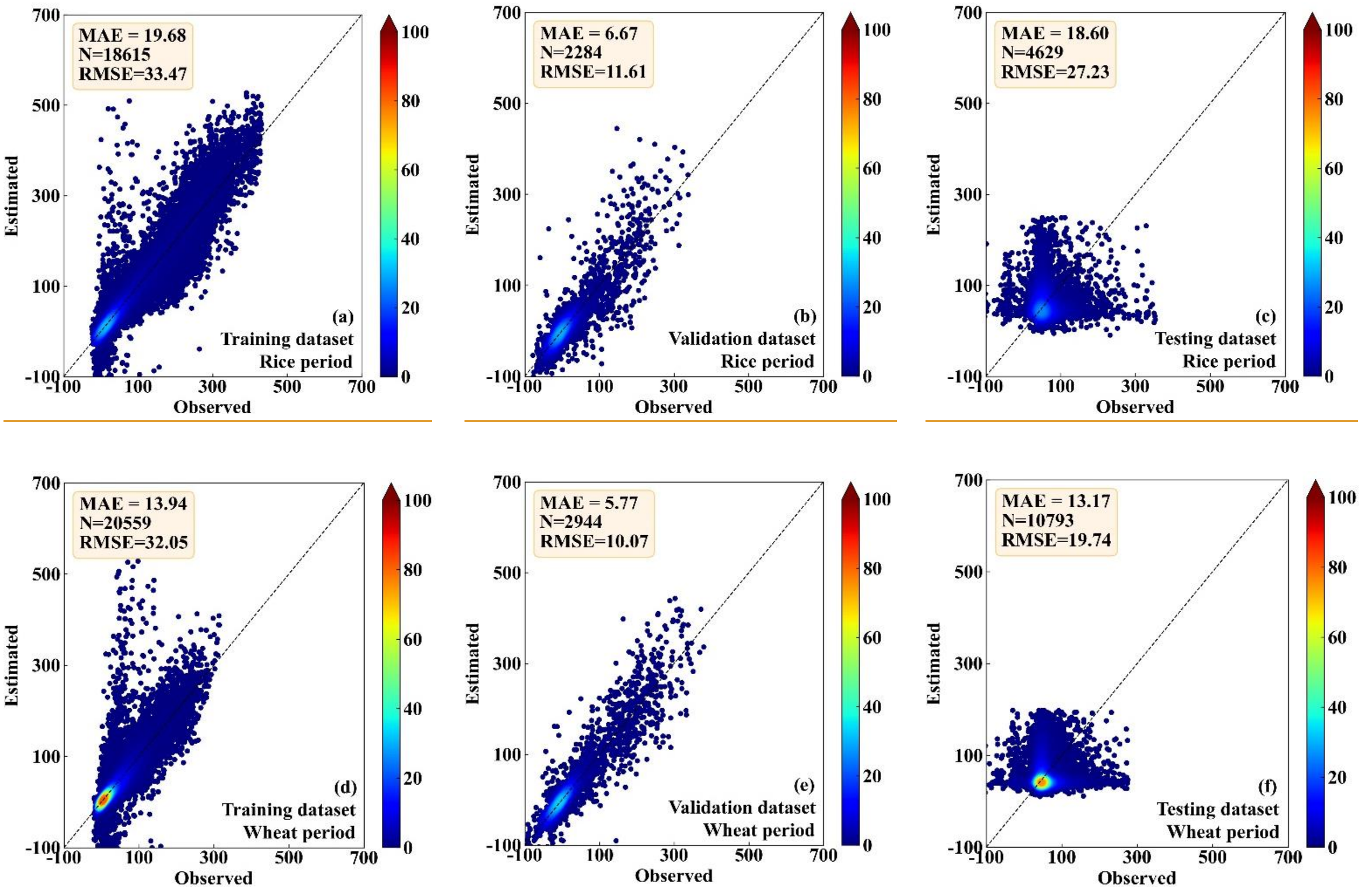


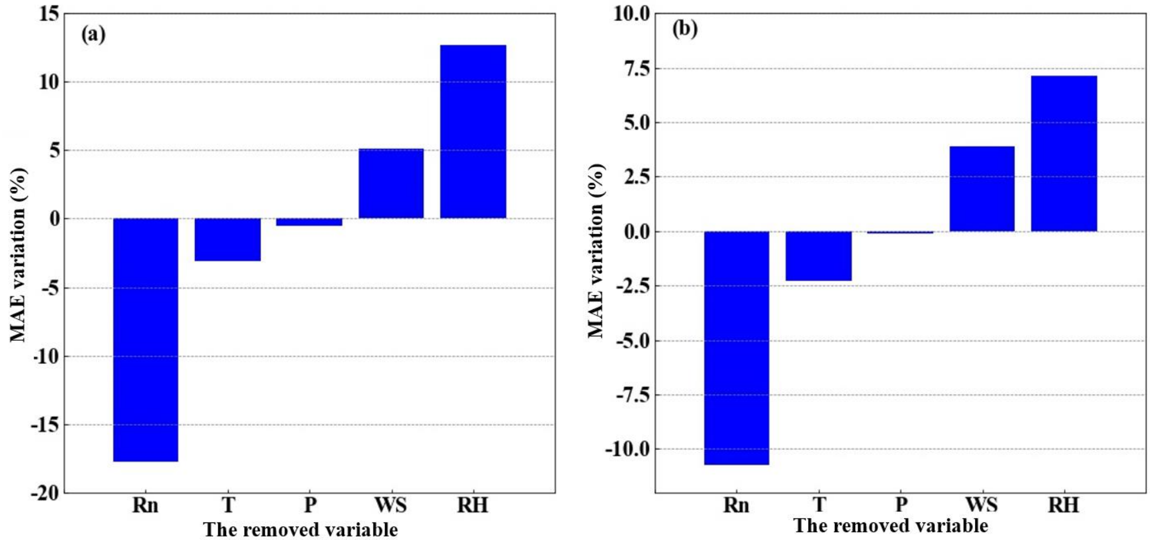
Figure 6. Same as Figure 5, but for LE.

3.3 Examination of Input Combinations

Meteorological elements may occasionally be unavailable due to the failure of sensors so the 5-variable input combination derived in Section 3.2 is not always applicable. Therefore, examination of other alternative input combinations is important to have substitute choices for data gap-filling when the 5-variable input combination is unavailable. In this subsection, we investigated the RF model's performance under the situation of lacking one element in the 5-variable input combination, i.e., we tested the 4-variable input combinations of (WS, T, RH, P), (Rn, T, RH, P), (Rn, WS, RH, P), (Rn, WS, T, P), and (Rn, WS, T, RH), by removing Rn, WS, T, RH, and P from the 5-variable input combination, respectively. The MAEs and RMSEs for these combinations are shown in Table 1, and it demonstrates that the RF model's accuracy may either increase or decrease as a result of the removal of a meteorological element during the training phase. For instance, it was found that the model's performance greatly improved once RH was eliminated from the input combination, with the MAE and RMSE of H decreasing from 6.48 and 11.94 Wm^{-2} to 5.66 and 11.06 Wm^{-2}, respectively, and LE from 19.1 and 39.39 Wm^{-2} to 17.74 and 35.27 Wm^{-2}. This outcome is logical given that RH and H do not have a strong correlation, as a result, performance will be enhanced if RH is not included in the gap-filling processing pipeline. According to our findings, the RF model's performance may be greatly enhanced by excluding irrelevant meteorological elements from the study and choosing only those that have a significant impact on the variable. Our findings imply that in order to attain the best gap-filling accuracy, it is necessary to take into account both the advantages and disadvantages of ML-based models as well as the ideal input components. The results suggested that RH at a single level was not well correlated to the fluxes as shown in Section 3.1, because the one-level RH was strongly affected by the irrigation activity which was an external factor of the weather system. As a result, RF model performance was enhanced when the irrelevant variable (i.e., RH) was removed from the input list. The same condition also happened to the removal of WS, as could be seen from Section 3.1, WS showed small correlations with the fluxes. WS over this site was rather small, and frequently below 2 m s^{-1}, and under this light wind condition, the fluxes were mostly driven by the buoyancy rather than the wind shear. Figure 7 presents the MAE variation percentage of the 4-variable input combinations from the 5-variable input combination. After RH was removed from the input list, the RF model showed favorable performance for both H and LE, as shown in Figure 7, with MAE values improvements of 12.65 and 7.12%, respectively. Notably, the removal of Rn from the input combination resulted in a considerable decline in the RF model's performances, with MAE degradation percentage values reaching 16.20% and 10.73%, respectively. This outcome makes sense since Rn is highly associated with H and LE; hence, performance will be declined if Rn is left out of the input training dataset. As a consequence, our findings demonstrated that choosing strongly associated components could greatly increase the gap-filling accuracy. According to our findings, the best input combination is (Rn, WS, T, P).

Table 1.The MAEs and RMSEs of the RF-estimated heat fluxes for the 4-variable input combinations, and the corresponding changes from the 5-variable input combination.

Factors Included	Factors Eliminated		MAE (change)	RMSE (change)
WS, T, RH, P	Rn	H	7.63 (+1.15)	10.72 (−1.22)
		LE	21.15 (+2.05)	39.38 (−4.62)
Rn, T, RH, P	WS	H	6.15 (−0.33)	11.42 (−0.52)
		LE	18.36 (−0.74)	36.13 (−2.34)
Rn, WS, RH, P	T	H	6.68 (+0.20)	11.48 (−0.46)
		LE	19.54 (+0.44)	38.54 (−1.46)
Rn, WS, T, P	RH	H	5.66 (−0.82)	11.06 (−0.88)
		LE	17.74 (−1.36)	35.27 (−4.12)
Rn, WS, T, RH	P	H	6.49 (+0.03)	11.77 (−0.17)
		LE	19.12 (+0.02)	38.13 (−1.07)



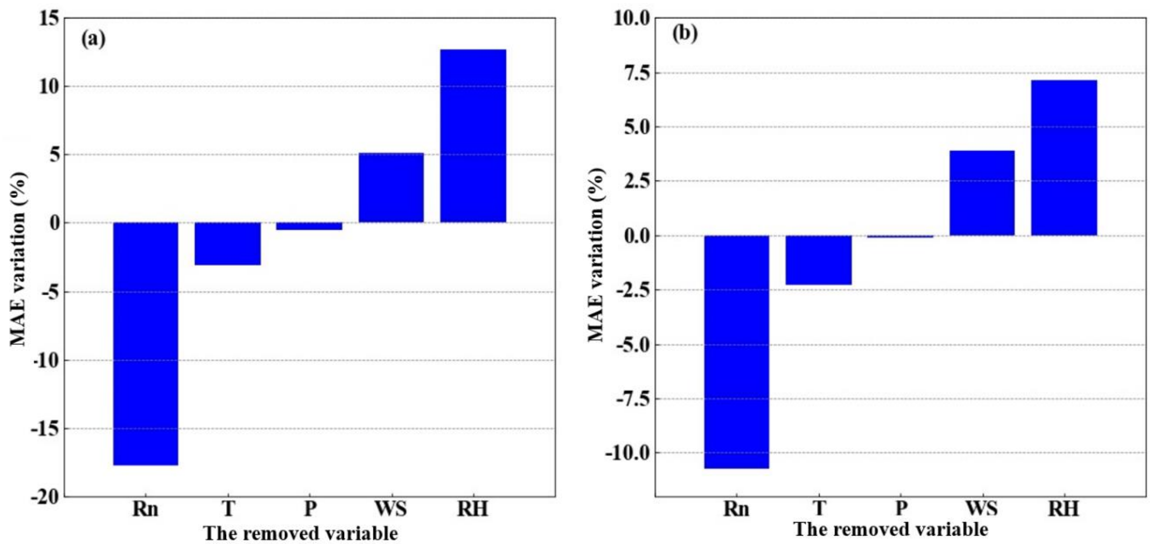


Figure 7. The MAE percentage variation after changing the 5-variable input combinations to the 4-variable input combinations, a) for H, and b) for LE, respectively. The x-axis labels indicate the removed variables.

It should be noted that other variables that might have an impact on the H and LE were not investigated here. For example, given that our research site was over farmland and plants were growing, knowledge of the variations of the leaf area index (LAI) and inclusion of it to the training dataset should also be useful to increase the accuracy of the RF model in H and LE gap-filling. The monsoonal climate here also incurred considerable precipitation variations, which might as well potentially contribute to the RF model accuracy improvement. However, due to the lack of LAI and precipitation observations, the inclusion of the two variables into the RF model training dataset was not applicable in this study. Additionally, as shown above, more variables would bring a higher observation demand, and lead to more complexity and potentially decreased results, such as the adding variable of RH.

3.4 Comparison with other four ML methods

3.4.1 Comparison in H estimation

To further investigate the reliability of the RF model, we used a Taylor diagram to compare its performance in H estimation with other four ML models: ~~linear regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), and gradient boosting decision tree (GBDT).~~ support vector machine (SVM), k-nearest neighbor (KNN), gradient boosting decision tree (GBDT), and linear regression (LR). SVM is a data-oriented classification algorithm, and the basic model is to find the best separation hyperplane on the feature space so that the positive and negative sample intervals on the training set are maximum. Its advantages are that the kernel function can be used to map to a high-dimensional space; the use of the kernel function can solve the nonlinear classification ; the classification idea is very simple, that is, to maximize the interval between the sample and the decision-making surface ;the classification effect is better ;and the nonlinear relationship between data and features is easy to obtain when the small and medium-sized sample size is large.KNN is particularly suitable for multi-classification problems. Its advantage is that it is simple in thought, easy to understand, easy to implement; No estimation parameters, no training; High accuracy, insensitive to outliers.GBDT can flexibly handle various types of data, including continuous and discrete values. With relatively few parameter adjustment times, the prediction preparation rate can also be relatively high. If the data dimension is high, the computational complexity of the algorithm will increase. Using some robust loss functions, the robustness to outliers is very strong.LR is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables that depend on each other.The results have good interpretability, can intuitively express the importance of each attribute in the prediction, and the calculation of entropy is not complicated.

All the models were optimized with the same technique described above for the RF model. The results are shown in Figure 8. The EC measurements were used as the benchmark. It can be seen that the RF model generally outperforms the other four models, with the standard deviations (σ_n) and correlation values of 1.05 and 0.98 during the period of rice planting, and 0.96 and 0.95 during the period of wheat planting, respectively. The SVM model is the second most accurate model, with the σ_n and correlation of 0.92 and 0.98 during the period of rice planting, and 0.91 and 0.93 during the period of wheat planting, respectively. The LR model performs the worst, with the σ_n and correlation of 0.60 and 0.76 during the period of rice planting, and 0.80 and 0.72 during the period of wheat planting, respectively. The accuracy of KNN and the GBDT models is in between the above-discussed models, and the $\sigma_n$and correlation during the rice and wheat period for KNN are 0.68 and 0.73, and 0.77 and 0.82; and for GBDT are 0.79 and 0.80, and 0.81 and 0.9, respectively.

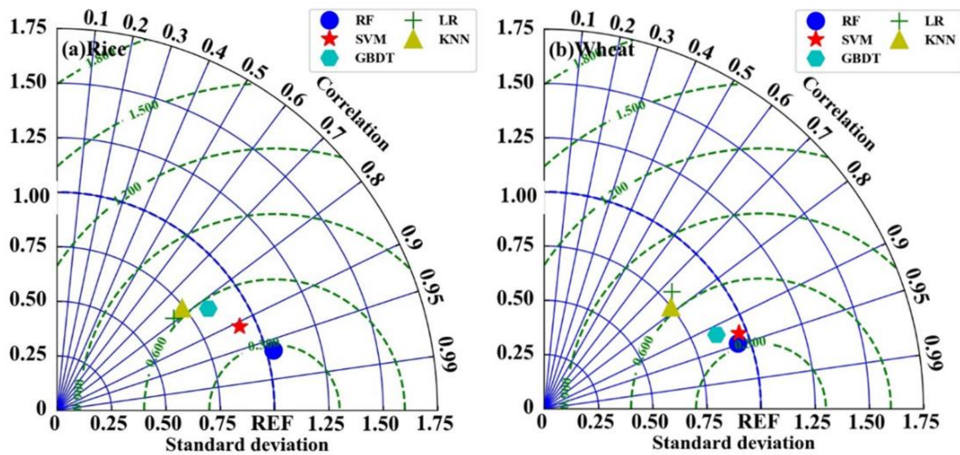


Figure 8. The performances of the five models for estimating H in the period of a) rice and b) wheat.

3.4.2 Comparison in LE estimation

Figure 9 illustrates a comparison of the estimated LE by all five models during the period of rice and wheat planting. The results are similar to those in the H estimation, and the RF model is found to perform better than the other four models, with the σ_n and correlation values of 0.95 and 0.97 during the period of rice planting, and 0.97 and 0.96 during the period of wheat planting, respectively. Nonetheless, the KNN model performs the worse for LE estimating and has the σ_n and correlation values of 0.68 and 0.82 during the period of rice planting, and 0.62 and 0.79 during the period of wheat planting, respectively. Overall, as shown by the Taylor diagram of Figures 8 and 9, in this study the RF model has the best accuracy in either H or LE estimation for data gap-filling.

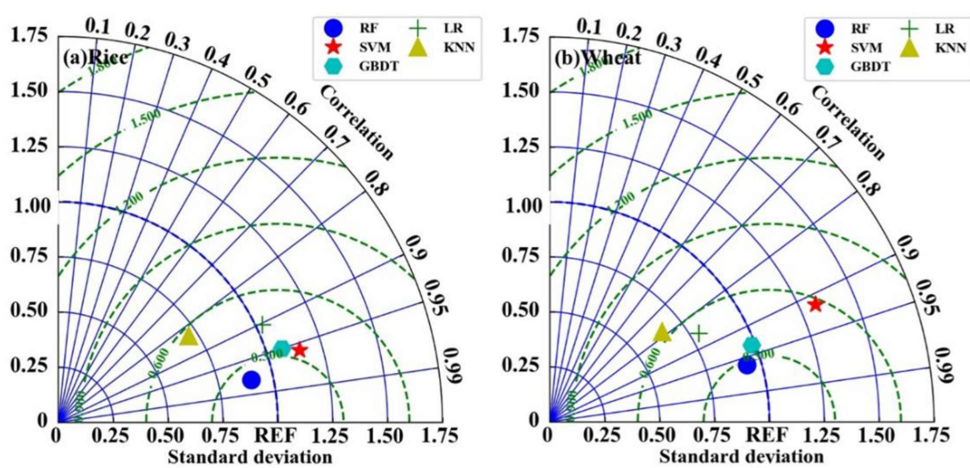


Figure 9. Same as Figure 8, but for LE

4 Summary and Conclusions

To assess the RF model's capacity for gap-filling the sensible and latent heat flux measurements over rice-wheat rotation croplands, 90% of the total observation data gathered at Shouxian were utilized for training and testing, and the remaining 10% for independent validation. Our findings demonstrate that Rn is the most important variable in regulating H and LE, and it accounts for 78% and 76% of the total variable significance in the RF model construction for H and LE calculation, respectively. The least important variables are WS and P, and their total variable significances are 2% and 0.6%, respectively. During the periods of rice and wheat planting, the RF model with a 5-variable input combination shows reliable performance, with MAE values of 5.88 Wm^{-2} and 20.97 Wm^{-2}, and RMSE values of 10.67 Wm^{-2} and 29.46 Wm^{-2}, respectively. However, further analysis of the RF model with 4-variable input combinations indicates that the performance of the model is improved when RH is removed from the input list, and the MAE values decrease by 12.65% and 7.12% for H and LE, respectively. Nonetheless, the 4- variable input combination without Rn causes an increase in the MAE values of the model, by 16.20% and 10.73% for H and LE, respectively. Therefore, the best input combination found in this study for heat fluxes gap-filling is (Rn, WS, T, P). Statistical comparison of RF and other four typical ML models (LR, KNN, SVN, and GBDT) by Tylar diagram further shows that RF is the most accurate, with the standard deviations and correlation values of 0.95 and 0.97 during the period of rice planting, and 0.97 and 0.96 during the period of wheat planting, respectively. While the LR and KNN models perform the worst for H and LE gap-filling, respectively, according to the statistical metrics of the Tylor diagram.

This study is based on only the data collected over rice–wheat-rotation croplands, but the method presented above to find a reliable gap-filling ML model can also be used over other types of the underlying surface and in other climate zones. It should be noted that over different types of the underlying surface and climates, the variable significances can vary and a careful check of the input combinations is needed. For example, over polar oceans with strong winds, Rn probably is not the most important driving factor, while the winds which cause mostly the turbulence may take the first place. On the other hand, over areas without human irrigation activity, RH will possibly be strongly related to the latent heat flux, and hence the inclusion of it into the input list may increase the ML model performance. Besides the examination of the input combinations, the choice of an ML model and the method to optimize its parameters are also important.

Overall, this study shows the potential to use the RF model to produce trustworthy gap-filling data of H and LE over rice–wheat-rotation croplands, and the ML methods are suggested to be used to derive the fluxes' estimations when direct EC observations are not available.

References

Alavi, N., Warland, J.S., Berg, A.A.: Filling gaps in evapotranspiration measurements for water budget studies: Evaluation of a Kalman filtering approach, J. Agric. For. Meteorol., 141 (1), 57–66, https://doi.org/10.1016/j.agrformet.2006.09.011, 2006.

Anapalli, S.S.; Fisher, D.K.; Reddy, K.N.; Krutz, J.L.; Pinnamaneni, S.R.; Sui, R.: Quantifying water and CO_2 fluxes and water use efficiencies across irrigated C3 and C4 crops in a humid climate, J. Sci. Total Environ., 663, 338–350, https://doi.org/10.1016/j.scitotenv.2018.12.471, 2018.

Baareh, A.K.; Elsayad, A.; Al-Dhaifallah, M.: Recognition of splice-junction genetic sequences using random forest and Bayesian optimization, J. Multimed. Tools Appl., 80, 30505–30522, https://doi.org/10.1007/s11042-021-10944-7, 2021.

Belgiu, M.; Dragut, L.: Random forest in remote sensing: A review of applications and future directions, Isprs J. Photogramm. J.Remote Sensing., 114, 24–31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Beringer, J., McHugh, I., Hutley, L. B., Isaac, P., and Kljun, N.: Technical note: Dynamic INtegrated Gap-filling and partitioning for OzFlux (DINGO), J. Biogeosciences., 14, 1457–1460, https://doi.org/10.5194/bg-14-1457-2017, 2017.

Best, M. J., M. Pryor, D. B. Clark, G. G. Rooney, et al.: The Joint UK Land Environment Simulator (JULES), model description - Part 1: Energy and water fluxes, J. Geosci. Model Dev., 4, 677-699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.-A.: Machine learning in acoustics: Theory and applications, J. Acoust. Soc. Am., 146, 3590–3628, https://doi.org/10.1121/1.5133944,2019.

Breiman, L.: Random Forests, J. Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cai, J.C.; Xu, K.; Zhu, Y.H.; Hu, F.; Li, L.H.: Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest, J. Appl. Energy., 262, 114566, https://doi.org/10.1016/j.apenergy.2020.114566, 2020.

Duan, Z.; Grimmond, C.; Gao, C.Y.; Sun, T.; Liu, C.; Wang, L.; Li, Y.; Gao, Z.: Seasonal and interannual variations in the surface energy fluxes of a rice–wheat rotation in Eastern China, J. Appl. Meteorol. Climatol., 60, 877–891, https://doi.org/10.1175/JAMC-D-20-0233.1, 2021.

Duan, Z.; Yang, Y.; Wang, L.; Liu, C.; Fan, S.; Chen, C.; Tong, Y.; Lin, X.; Gao, Z.: Temporal characteristics of carbon dioxide and ozone over a rural-cropland area in the Yangtze River Delta of eastern China, J. Sci. Total Environ., 757, e143750, https://doi.org/10.1016/j.scitotenv.2020.143750, 2021.

Falge, E.; Baldocchi, D.; Olson, R.; Anthoni, P.; Aubinet, M.; Bernhofer, C.; Burba, G.; Ceulemans, R.; Clement, R.; Dolman, H.: Gap filling strategies for defensible annual sums of net ecosystem exchange, J. Agric. For. Meteorol., 107, 43–69, https://doi.org/10.1016/S0168-1923(00)00225-2, 2001.

Foltýnová, L.; Fischer, M.; McGloin, R.P.: Recommendations for gap-filling eddy covariance latent heat flux measurements using marginal distribution sampling, J. Theor. Appl. Climatol., 139, 677–688, https://doi.org/10.1007/s00704-019-02975-w, 2020.

Frazier, P.I.: A Tutorial on Bayesian Optimization, arXiv 2018, https://doi.org/10.48550/arXiv.1807.02811, 2018.

Gao, Z. Q., L. G. Bian, and X. J. Zhou.: Measurements of turbulent transfer in the near-surface layer over a rice paddy in China, J. Geophys. Res., 108(D13), 4387–4387, https://doi.org/10.1029/2002JD002779, 2003.

Garratt, J. R.: The atmospheric boundary layer. Cambridge Atmospheric and Space Science Series, Cambridge University Press, 316, https://doi.org/10.1017/CBO9781316117422, 2015.

Hui, D.; Wan, S.; Su, B.; Katul, G.; Monson, R.; Luo, Y.: Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations, J. Agric. For. Meteorol., 121, 93–111, https://doi.org/10.1016/S0168-1923(03)00158-8, 2004.

Isaac, P., Cleverly, J., McHugh, I., Van Gorsel, E., Ewenz, C., and Beringer, J.: OzFlux data: Network integration from collection to curation, J. Biogeosciences, 14, 2903–2928, https://doi.org/10.5194/bg-14-2903-2017, 2017.

Jiang, L.; Zhang, B.; Han, S.; Chen, H.; Wei, Z.: Upscaling evapotranspiration from the instantaneous to the daily time scale: Assessing six methods including an optimized coefficient based on worldwide eddy covariance flux network, J. Hydrol., 596, 126135, https://doi.org/10.1016/j.jhydrol.2021.126135, 2021.

Kepert, J.: Choosing a boundary layer parameterization for tropical cyclone modelling, J. Mon. Wea. Rev. [serial online]., 140(5), 1427-1445, DOI: https://doi.org/10.1175/MWR-D-11-00217.1, 2012.

Khan, M.S.; Jeon, S.B.; Jeong, M.H.: Gap-Filling Eddy Covariance Latent Heat Flux: Inter-Comparison of Four Machine Learning Model Predictions and Uncertainties in Forest Ecosystem, J. Remote Sens., 13, 4976. https://doi.org/10.3390/rs13244976, 2021.

Khan, M.S.; Liaqat, U.W.; Baik, J.; Choi, M.: Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach, J. Agric. For. Meteorol., 252, 256–268, https://doi.org/10.1016/j.agrformet.2018.01.022, 2018.

Kim, Y.; Johnson, M.S.; Knox, S.H.; Black, T.A.; Dalmagro, H.J.; Kang, M.; Kim, J.; Baldocchi, D.: Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis, J. Glob. Chang. Biol., 26, 1499–1518, https://doi.org/10.1111/gcb.14845, 2020.

Kunwor, S., Starr, G., Loescher, H. W., and Staudhammer, C. L.: Preserving the variance in imputed eddycovariance measurements: Alternative methods for defensible gap filling, J. Agr. Forest Meteorol., 232, 635–649, https://doi.org/10.1016/j.agrformet.2016.10.018, 2017.

Li, X., Z. Gao, Y. Li, and B. Tong.: Comparison of sensible heat fluxes measured by a large aperture scintillometer and eddy covariance system over a heterogeneous farmland in East China, J. Atmosphere., 8, 101, https://doi.org/10.3390/ atmos8060101, 2017.

Liu, J.; Zuo, Y.; Wang, N.; Yuan, F.; Zhu, X.; Zhang, L.; Zhang, J.; Sun, Y.; Guo, Z.; Guo, Y.; et al.: Comparative Analysis of Two Machine Learning Algorithms in Predicting Site-Level Net Ecosystem Exchange in Major Biomes, J. Remote Sens., 13, 2242, https://doi.org/10.3390/rs13122242, 2021.

McCandless, T., Gagne, D. J., Kosović, B., Haupt, S. E., Yang, B., Becker, C., & Schreck, J. (2022). Machine Learning for Improving Surface-Layer-Flux Estimates. Boundary-Layer Meteorology, 185(2), 199-228.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkin G., Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D., Jarvis, A. J., Kattge, J., Noormets, A., and Stauch, V. J.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, J. Agr. Forest Meteorol., 147, 209–232, https://doi.org/10.1016/j.agrformet.2007.08.011, 2007.

Moncrieff, J.; Clement, R.; Finnigan, J.; Meyers, T.: Averaging, Detrending, and Filtering of Eddy Covariance Time Series. In Handbook of Micrometeorology: A Guide for Surface Flux Measurement and Analysis, The Netherlands, pp., 7–31, https://doi.org/10.1007/1-4020-2265-4_2, 2006.

Monin, A. S., & Obukhov, A. M. (1954). Basic laws of turbulent mixing in the surface layer of the atmosphere. Contrib. Geophys. Inst. Acad. Sci. USSR, 151(163), e187.

Nisa, Z.; Khan, M.S.; Govind, A.; Marchetti, M.; Lasserre, B.; Magliulo, E.; Manco, A.: Evaluation of SEBS, METRIC-EEFlux, and QWaterModel Actual Evapotranspiration for a Mediterranean Cropping System in Southern Italy, J. Remote Sens., 13, 4976 18 of 19, https://doi.org/10.3390/agronomy11020345, 2021.

Papale, D.; Reichstein, M.; Aubinet, M.; Canfora, E.; Bernhofer, C.; Kutsch, W.; Longdoz, B.; Rambal, S.; Valentini, R.; Vesala, T.; et al.: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation, J. Biogeosciences., 3, 571–583, https://doi.org/10.5194/bg-3-571-2006, 2006.

Reichstein, M.; Falge, E.; Baldocchi, D.; Papale, D.; Aubinet, M.; Berbigier, P.; Bernhofer, C.; Buchmann, N.; Gilmanov, T.; Granier, A.; et al.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, J. Glob. Change Biol., 11, 1424–1439, https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Richard, A.; Fine, L.; Rozenstein, O.; Tanny, J.; Geist, M.; Pradalier, C.: Filling Gaps in Micro-Meteorological Data, Switzerland, https://doi.org/10.1007/978-3-030-67670-4_7, 2020.

Stauch, V.J.; Jarvis, A.J.: A semi-parametric gap-filling model for eddy covariance CO2 flux time series data, J. Glob. Chang. Biol., 12, 1707–1716, https://doi.org/10.1111/j.1365-2486.2006.01227.x, 2006.

Vitale, D., Bilancia, M., Papale, D.: A multiple imputation strategy for eddy covariance data, J. Environ. Inform., 34, 68–87, https://doi.org/10.3808/jei.201800391, 2018.

Wang, L.; Wu, B.; Elnashar, A.; Zeng, H.; Zhu, W.; Yan, N.: Synthesizing a Regional Territorial Evapotranspiration Dataset for Northern China, J. Remote Sens., 13, 1076, https://doi.org/10.3390/rs13061076, 2021.

Webb, E.K.; Pearman, G.I.; Leuning, R. Correction of flux measurements for density effects due to heat and water vapor Transfer, Q. J. R. Meteorol. Soc., 106, 85–100, https://doi.org/10.1002/qj.49710644707, 1980.

Wilson, K.B.; Hanson, P.J.; Mulholland, P.J.; Baldocchi, D.D.; Wullschleger, S.D.: A comparison of methods for determining forest evapotranspiration and its components: Sap-flow, soil water budget, eddy covariance and catchment water balance, J. Agric. For. Meteorol., 106, 153–168, https://doi.org/10.1016/S0168-1923(00)00199-4, 2001.

Wutzler, T., Lucas-Moffat, A., Migliavacca, M., Knauer, J., Sickel, K., Sigut, L., Reichstein, M.: Basic and extensible post-processing of eddy covariance flux data with REddyProc, J. Biogeosciences., 15 (16): 5015–5030, https://doi.org/10.5194/bg-15-5015-2018, 2018.

Yu, T.C.; Fang, S.Y.; Chiu, H.S.; Hu, K.S.; Tai, P.H.Y.; Shen, C.C.F.; Sheng, H.: Pin accessibility prediction and optimization with deep learning-based pin pattern recognition, J. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 40, 2345–2356, https://doi.org/10.1145/3316781.3317882, 2019.