Response to the reviewers' comments

Response to Referee #1

We greatly appreciate Dr. René Orth providing valuable and constructive comments on our manuscript HESS-2020-590. We seriously considered each comment and revised the manuscript accordingly. The individual comments are replied below. The comments are shown in black font and our responses are shown in blue font.

Comment (1): This study derives land evapotranspiration with a machine learning approach applied to daily meteorological station data from across the world. More decreasing than increasing trends are found across the Earth's land areas. The controls for these trends are determined and distinguished by jointly considering trends in evapotranspiration and precipitation minus evapotranspiration as a proxy for runoff.

Recommendation: I think the paper requires major revisions.

The methodology and research question addressed in this paper are novel and relevant, making it a potentially good fit for HESS. Also the joint consideration of evapotranspiration and (proxy) runoff to interpret the reported trends and associate them with potential causes is an important contribution to the land surface science community. However, before the paper is suitable for publication in HESS, some critical shortcomings need to be addressed.

Reply: Thanks for your encouraging and constructive comments. All your comments are addressed in the revised manuscript, and we hope you will find the latest version suitable for publication.

Comment (2): I like that the authors validated their simulation results against observed streamflow and a reference ET product. However, the diagnosed agreement with these products is actually not very convincing, particularly in terms of the trends, as shown in Figure S8, and in the comparison between Figures 4c and S11. I think it is critical to understand these differences between the products, as otherwise I am missing convincing evidence that the data simulated here can be used to assess global ET and runoff trends.

Reply: We agree with the point that the differences between the products should be better understood. The difference between our ET and the ET estimated by model tree ensemble (MTE) can be caused by the different models, driven data, and time scales. Meanwhile, we should not fully expect the P-ET to be completely consistent with observed streamflow, because (1) there are inherent differences between atmospheric scale and hydrologic scale, (2) the conversion of P-ET to streamflow largely depends on the underlying surface, and (3) the observed streamflow is strongly affected by human activities especially on long time scales. We also note that because our work is based on a boundary layer perspective where physics does not change in time, whereas direct estimates of ET are affected by additional things like nutrients or increased CO_2 whose trends cannot be captured by the Fluxnet data.

According to the comment, we added more discussions and expanded the cross-validation in terms of trends as suggested in the comment (3).

Comment (3): Adding to (1), it would be insightful to expand the cross-validation analysis from Figure 2 to validate the derived data also in terms of the trends observed at the independent cross-validation stations, as the final conclusions of this study are build on the trends rather than the short-term variability of the data.

Reply: Thanks for the suggestion. We expanded the cross-validation analysis in terms of trends in the revised manuscript. It was suggested that the trends predicted by the ANN model were highly correlated with the observed trends, and even in most cases, the estimation of the trends were more reliable than the estimation of the values (please see the Attached Figures at the end).

Comment (4): The comparison of ET and runoff trends between the CMIP5 scenario simulations and the machine learning-derived, historical simulations, does not really make sense as totally different time periods are considered to compute the respective trends, if I get this right?

Reply: Yes, the trends vary from different time periods. We acknowledge that a consistent time period is better, but it is difficult to achieve due to paucity of data. Moreover, the main purpose here is to identify the change directions of ET or runoff rather than their magnitudes. Thus, we can determine whether the trends of decrease (increase) in ET or runoff is a long-term existing phenomenon by comparing the change directions over different time periods in history and in simulated future scenarios.

After considering the comment, we will add an explanation in section 3.3 in the revised manuscript.

Comment (5): The description of the employed machine learning algorithm is not clear. The choice of artificial neural networks over other machine learning methods is not sufficiently motivated. Why is it more suitable than for example random forests in this context?

Reply: Different machine learning algorithms have their own advantages. The artificial neural networks tend to work well, as it has strong nonlinear function approximation capability and fault tolerance. As our selected neural networks demonstrated good performance, to save space, we did not compare different machine learning algorithms in this study.

We still fully considered your comments and added a description on the motivation of using neural networks in section 2.4.

Comment (6): Why not simply use FLUXCOM instead of deriving yet another estimate?

Reply: The data in the FLUXCOM or the results estimated by the model tree ensemble (MET) are driven by monthly remote-sensing and meteorological reanalysis data, and thus they rely on the satellite era and cannot be used for long-term trends. In addition to the too short record required for correct trend estimates, evapotranspiration can be modified by subtle changes in CO_2 or nutrients, which can modify transpiration for instance or there can be even a decoupling between surface and deeper soil moisture (Berg et al., 2016), none of which are captured by FLUXCOM. This is why we use an opposite view – we use in essence a boundary layer budget based on Salvucci and Gentine (2015) and Gentine et al. (2016) except that we lump the non-linearity for the boundary layer budget in a neural network. The physics of the boundary layer is not changing in time and weather stations have been available for many decades, allowing the development of the first long-term record of hydrologic trends. Finally, FLUXNET-MTE and FLUXCOM are based on data that are highly localized, especially in the northern hemisphere. The network of weather station is both much longer but also extend to much more remote places such as in the tropics, providing much more constrain in those places, where other retrievals typically display very large uncertainties. The aim of our strategy in this study is therefore to infer a longer surface fluxes (as well as better generalization to the tropics and other remote regions). Thus, the length of the surface heat fluxes in existing products does not match with our research purpose. After considering this comment, we will add an explanation in introduction section.

Comment (7): Further, the setup of the ANN model is unclear, i.e. how the hyper-parameters are chosen (why exactly 2 hidden layers? why 500 epochs?). Why different performance metrics (RMSE and MSE) are chosen, how the training is done, and how overfitting is prevented. I acknowledge that some of these choices are necessarily arbitrary, but this would be good to mention, including some tests on the relevance of these choices for the conclusions of the study.

Reply: The more hidden layers and neurons in the ANN, the stronger the nonlinear ability of the model, but the complexity and training time are also increasing. In theory, a neural network with 2 hidden layers can realize any complex mapping, as the nonlinear ability can be enhanced by adding neurons. The ANN model with 2 hidden layers and 15 neurons shows good performance and appropriate time consumption (see the Attached Fig. S2). As for the optimal number of neurons, we initially determined it according to an empirical formula, i.e.,

$$h = \sqrt{(n+m)} + a .$$

Where *n* is the number of input neurons, and *m* is the number of output neurons, and *a* is a constant ranging from 0 to 10.

MSE was an indicator used to evaluate the performance of neural works in the training process of adjusting weight, and RMSE is an indicator used to analyze the bias between the ANN predicted values and the observed values in validation set.

To avoid over-fitting, the early stopping method was used to avoid overfitting this study, that is, we recorded the best validation accuracy during the training process, and the training was stopped when the MSE was no longer reduced after going through the entire dataset.

We apologize for the omission of some descriptions on the setup of training the ANN, and we have added more information on the process in the section 2.4 accordingly.

Comment (8): There are many small language errors (such as missing articles or wrong grammar) throughout the manuscript. The authors need to take special care of these when revising the manuscript.

Reply: The language has been polished by a language editing agency, and we will take special care for each sentence and add missing articles in the references.

Comment (9): lines 33 & 36: Please explain what you mean with "offline".

Reply: The calculation of potential evaporation in drought index using meteorological variables from climate model outputs is offline. We have modified this sentence to "Using potential evaporation rather than actual ET or calculating offline ET using meteorological variables from climate model outputs in the traditional drought indices, the calculation implicitly assumes that soil can always supply moisture to meet the atmospheric evaporation demand".

Comment (10): lines 57, 72, 83, 123/124: You mention in these places different sets of variables which are (not?) used by the ANN algorithm, please clarify.

Reply: Top-of-atmosphere shortwave radiation, relative humidity, temperatures (mean, maximum, and minimum temperatures), and surface wind speed are the variables used by the ANN algorithm (in line 123/124).

The expressions of temperatures, humidity, and solar radiation (in line 57) are a broad concept and does not refer to a specific variable. We have rephrased this sentence to "This approach utilizes daily observations of meteorological variables such as temperatures, humidity, and solar radiation."

The expressions of top-of-atmosphere shortwave radiation, vapor pressure deficit (VPD), mean temperature, and surface wind speed (in line 72) are referring to the data collected from the integrated daily product of FLUXNET2015. VPD is used to calculate relative humidity, and daily maximum and minimum temperatures are obtained from half-hourly/hourly flux tower measurements. These data were used to train the ANN model.

The expressions of precipitation, temperatures (mean, maximum, and minimum temperatures), dew point temperature, and surface wind speed (in line 83) are referring to the data collected from weather stations. Dew point temperature was used to calculate relative humidity at weather stations, and the meteorological data collected from global weather stations were used to drive the trained ANN models.

According to the comment, we have modified the expressions to ensure that the information is clear in the revised manuscript.

Comment (11): line 69: How is this gap filling done?

Reply: The gap-filling data were provided by the FLUXNET.

Comment (12): line 71: Why are you targeting daily resolution? To infer trends, monthly resolution might be sufficient?

Reply: We agree with that monthly scale might be sufficient for inferring trend, but daily scale is a true time scale in reality and thus it can capture the daily-cycle signal of water and heat fluxes. There is no contradiction between retrieving surface fluxes on daily scale and on monthly scale, as monthly fluxes can be converted from daily results.

Comment (13): lines 86-91: I do not fully understand this paragraph. Do the original and target station groups differ by the number of stations in some grid cells? If yes, do you average across multiple stations in one grid cell, or remove stations to only keep one (and which one?)?

Reply: Thanks for the comment, and we have improved the writing of this paragraph. The target stations were obtained according to the following three steps, i.e., (1) The stations with a time series spanning less than ten years were excluded; (2) if the stations had the same geographic coordinates, we used the stations with longer observation to replace the stations with shorter observation; (3) if there were multiple stations showing different coordinates in a 0.1-degree grid, we removed the stations with a shorter observation length.

Comment (14): line 101: Why do you employ top-of-atmosphere radiation instead of surface radiation which is what the vegetation is actually exposed to?

Reply: Because there do not exist reliable long-term surface observational solar radiation data, top-of-atmosphere shortwave radiation is a good replacement. After considering this comment, we have added an explanation in section 2.3.

Comment (15): lines 156/157: Where is the influence of the resistances shown, or how do you arrive to this conclusion?

Reply: We demonstrate the influence in an indirect way by inferring changes from a big leaf model (which is not used to derive the ANN). According to this comment and the Eq. (14), we have modified the sentences as following:

"Therefore, a decline in EF is linked with surface resistance (r_s) and there is a negative relationship between r_s and EF. Annual EF ranges from 0 to 1 and r_a is a function of wind speed. Thus changes in r_a are relatively small while changes in r_s can be strong."

Comment (16): line 160: How was the cross validation done? How/which stations or times have been chosen as independent validation data?

Reply: We randomized the samples of five flux towers (moderate number) from 212 sites as the validation set, and then used the remaining samples to train the ANN model. The ANN predicted daily λE (H) of the validation set were compared with their observed values. As for validation under different land covers, we randomly select samples from five flux towers under this land type as the validation set.

According to this comment, we have added a description of the random selection of cross-validation samples. Meanwhile, we have redrawn the Fig. 2 by randomly selecting 10 flux towers from different plant function types as the validation set (see Fig. 2 in the Attached Figures).

Comment (17): line 168: I would think that also SAV is water-limited?

Reply: Agree and done.

Comment (18): lines 165, 169: Why are there different correlations given for OSH?

Reply: Apology for the typo. We have corrected it in the revised manuscript.

Comment (19): line 184/185: What does 0.78~0.79 and 0.77~0.78 mean?

Reply: The range of correlation coefficients. We have corrected these sentences.

Comment (20): line 195: The reference for FLUXCOM would be Jung et al. 2019 which is also in the reference list, or did you actually use the MTE product from Jung et al. 2010?

Reply: Thanks for the reminder. We used the ensemble data of latent heat flux on land from the Department of Biogeochemical Integration (BGI) of the Max Planck Institute (https://www.bgc-jena.mpg.de/geodb/projects/Data.php). The data were retrieved using the model tree ensemble (MTE) approach for upscaling FLUXNET measurements (Jung et al., 2011). We have corrected the reference and relevant expressions in the revised manuscript.

Comment (21): line 220: I cannot see a cooling trend in northern Europe in Figure 3.

Reply: We have corrected it in the revised manuscript.

Comment (22): lines 226-228: Phrasing could be improved, the "when" is not needed here.

Reply: We have rephrased this sentence.

Comment (23): line 235: Not sure if I would see a "persistent long-term trend" anywhere here as the spatial patterns of trends vary quite a lot across time periods in Figure 5.

Reply: We have revised the sentence to avoid inaccurate expression.

Comment (24): Which models did you consider in particular?

Reply: The simulation is an ensemble from Phase 5 of the Coupled Model Intercomparison Project (CMIP5) under the RCP8.5 scenario. We have modified the relevant expression.

Comment (25): line 261: There is no "RCP8.5 climate model".

Reply: Apology for the wrong expression, and we have corrected it accordingly.

Comment (26): section 3.4: Very nice approach and analysis to infer potential causes of the observed trends.

Reply: Thanks for the favorable evaluation.

Comment (27): Figure 1: What are the black empty bars which are superimposed on the colored bars?

Reply: Black empty bars represent the towers in the FLUXNET2015 Dataset and the solid bars represent all towers registered in FLUXNET.

We have redrawn the Figure 1 and its related Figure S1a to make the illustration clear (see Fig. 1 and Fig. S1 in the Attached Figures).

Comment (28): Figure 3: How is the Sahara desert defined? Why are other deserts not excluded, too? What is the time period over which the trends are computed? How is the spatial interpolation between the station locations done?

Reply: There is no strict definition of the scope of the Sahara desert. We prefer excluding the Sahara because it has the largest desert area and scarce meteorological observation data. We referred to previous drought studies such as Vicente-Serrano et al., 2015, which also did not consider the Sahara region.

The trends are computed over the period of 1950-2017. If the observational data is not that long, the trends are converted to a uniform length of 68 years. The spatial interpolation in this study uses the Kriging interpolation method based on ArcGIS platform.

Comment (29): Figure S10: This is cool, but where is the information coming from?

Reply: Data are collected from the Food and Agriculture Organization of the United Nations (http://www.fao.org/nr/water/aquastat/irrigationmap/index10.stm). We have added the information of data source in the revised manuscript.

Comment (30): Tables S1 and S2: Some of the variable names here need more explanation. And why was only the u-component of the wind speed used?

Reply: We have provided more explanations about the variable names in Tables S1 and S2 (see Attached Tables), and have improved the writing of this section.

The wind speed used in this study is mean surface wind speed of the day, not the u-component of the wind speed. We have modified the abbreviation of wind speed.

Some papers are added in the References, i.e.,

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Pagpale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res.: Biogeo., 116, G00J07, http://dx.doi.org/10.1029/2010JG001566, 2011.

Orth, R., and Destouni, G: Drought reduces blue-water fluxes more strongly than green-water fluxes in Europe. Nat. Commun., 9, 3602, http://dx.doi.org/10.1038/s41467-018-06013-7, 2018.

Vicente-Serrano, S. M., Van Gerard, V. D. S., Beguería, S., Azorin-Molina, C., and Lopez-Moreno, J. I.: Contribution of precipitation and reference evapotranspiration to drought indices under different climates. J. Hydrol., 526, 42–54. http://dx.doi.org/10.1016/j.jhydrol.2014.11.025, 2015.

Attached Figures

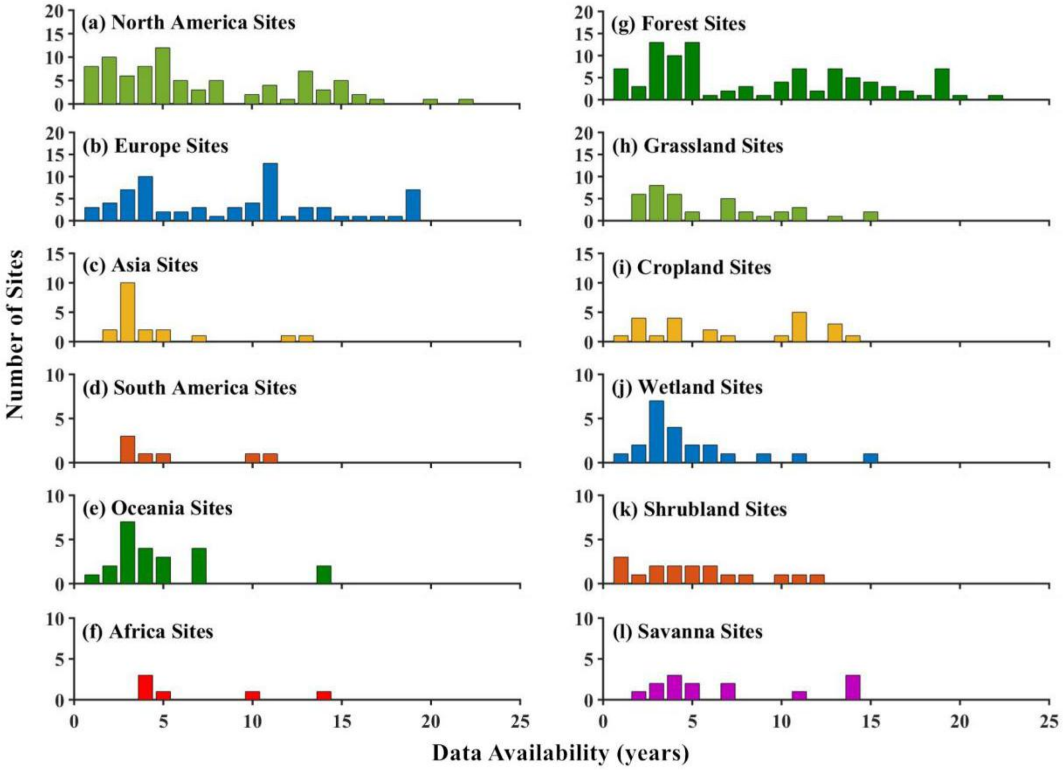
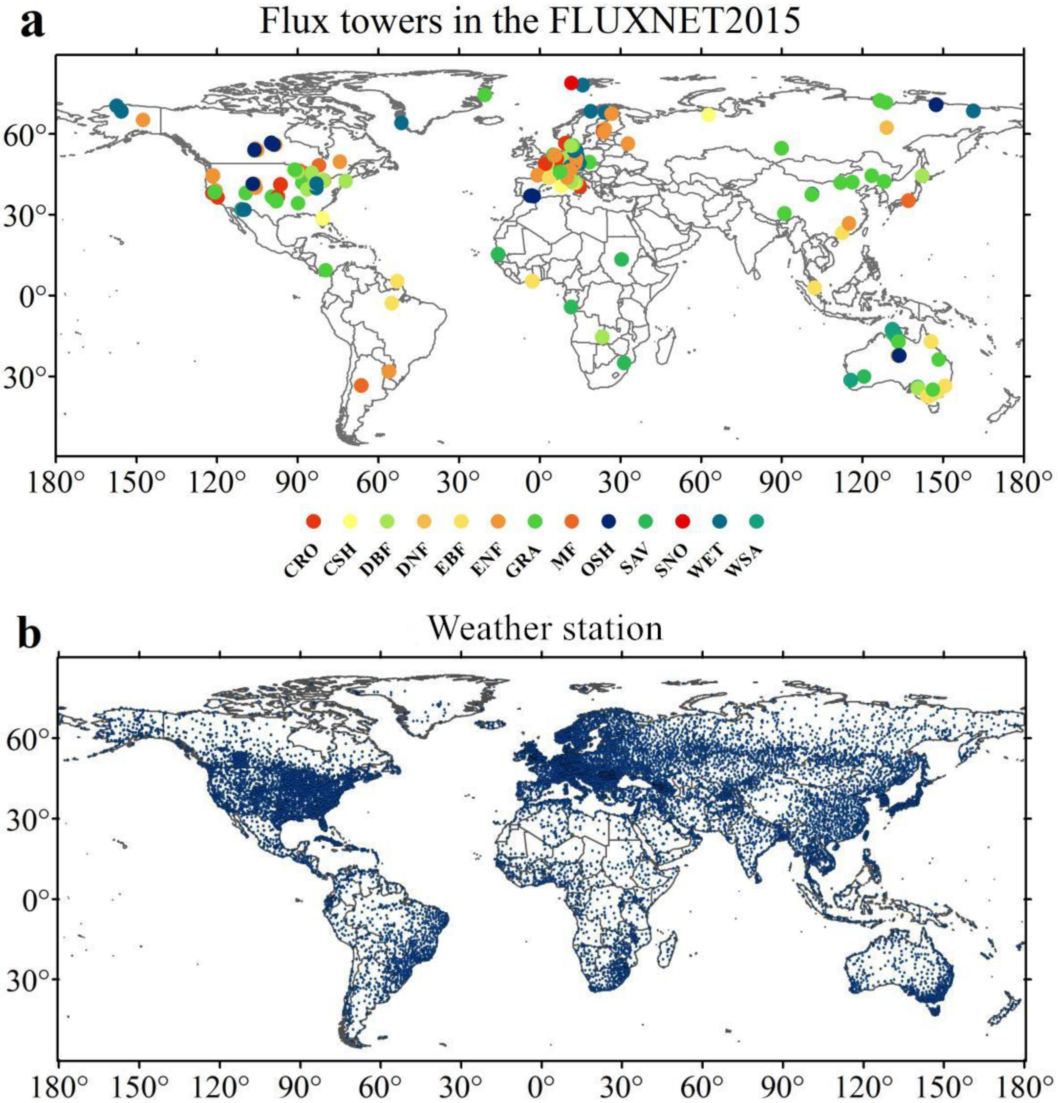
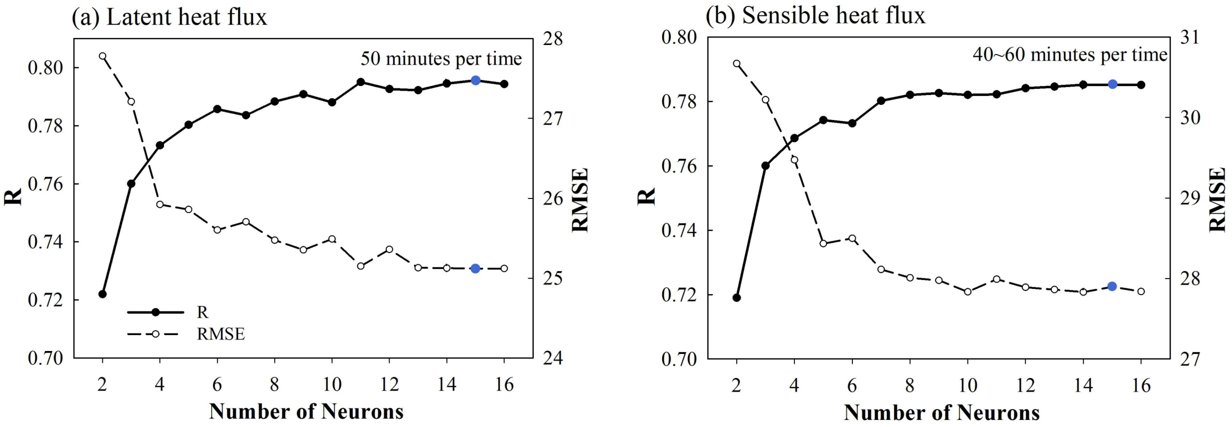


Figure 1. Data summary of the flux towers used in this study.



Supplementary Figure S1. Spatial distribution of (a) the flux towers in the FLUXNET2015 and (b) the weather stations used in this study. The plant function types of the flux towers include Croplands (CRO), Deciduous Needleleaf Forests (DNF), Evergreen Needleleaf Forest (ENF), Evergreen Broadleaf Forest (EBF), Deciduous Broadleaf Forest (DBF), Mixed Forest (MF), Grasslands (GRA), Savannas (SAV), Woody Savannas (WSA), Closed Shrublands (CSH), Open Shrublands (OSH), Wetlands (WET), and Snow and Ice (SNO).



Supplementary Figure S2. The performance of the ANN model using different number of neurons.

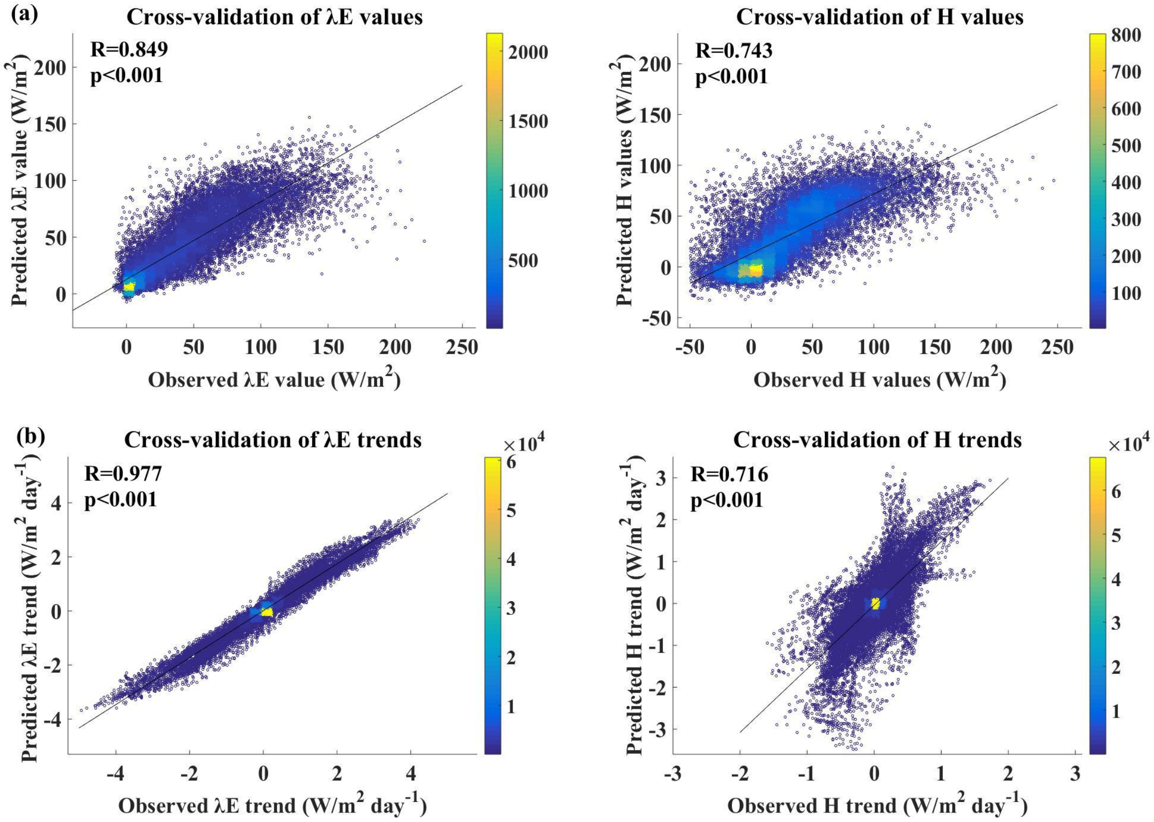
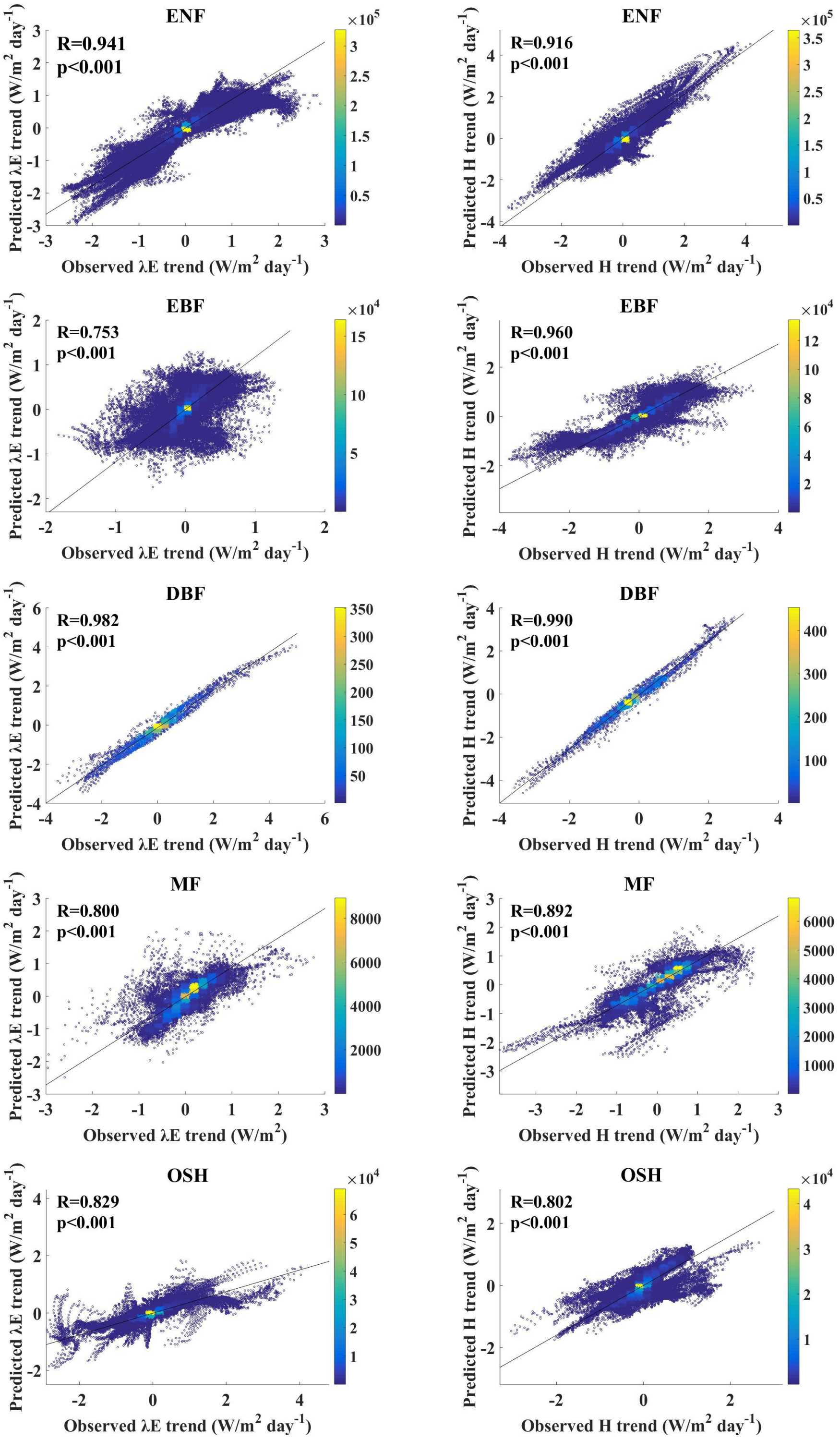
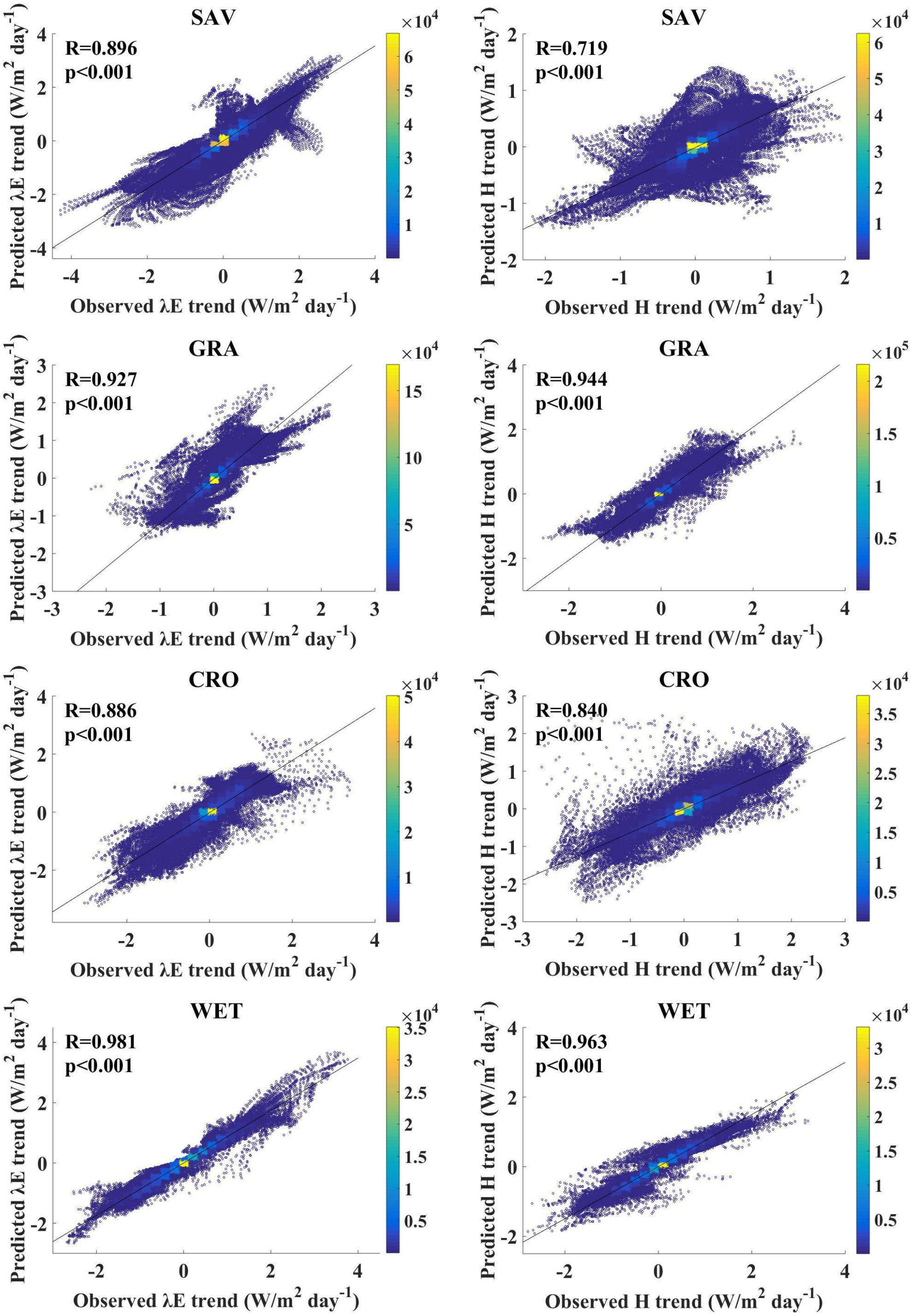


Figure 2. Density scatter plot for (a) the cross-validation in terms of values and (b) the cross-validation in terms of trends. The validation set of values cross-validation is randomly composed of 10 flux towers from different plant function types, and the validation set of trends cross-validation is composed of the trends calculated from all time periods.



Supplementary Figure S5. Density scatter plot for the cross-validation in terms of trends for different samples from ENF, EBF, DBF, MF, and OSH, respectively. The validation set is randomly composed of one flux tower from one plant function type, and the trends are estimated for all time periods.



Supplementary Figure S6. Density scatter plot for the cross-validation in terms of trends for different samples from SAV, GRA, CRO, and WET, respectively. The validation set is randomly composed of one flux tower randomly selected from one plant function type, and the trends are estimated for all time periods.

Attached Tables

Table S1. Test results of model training using different variable combinations*

Combination of different variables	λE		H	
	R	RMSE ($W\ m^{-2}$)	R	RMSE ($W\ m^{-2}$)
{Tmax; Tmin}	0.60	36.22	0.52	43.75
{RH; Tmax; Tmin}	0.66	32.11	0.60	40.97
{RH; Tmean; Tmax; Tmin}	0.67	32.00	0.61	40.06
{RH; Tmax; Tmin; DTR}	0.67	30.89	0.62	39.48
{SW_IN_POT; Tmax; Tmin}	0.70	30.75	0.78	31.81
{SW_IN_POT; RH}	0.70	30.72	0.69	37.01
{SW_IN_POT; RH; Tmean; Tmax; Tmin}	0.74	28.80	0.72	35.8
{SW_IN_POT; RH; Tmean; Tmax; Tmin; WS}	0.75	28.65	0.74	34.72
{SW_IN_POT; RH; Tmax; Tmin; WS}	0.74	28.78	0.73	35.06
{SW_IN_POT; RH; Tmax; Tmin; WS; P}	0.75	28.90	0.75	33.01
{SW_IN_POT; RH; Tmax; WS; Tmin; P}	0.76	28.11	0.75	34.80
{SW_IN_POT; RH; Tmean; Tmax; Tmin; DTR; WS; P}	0.77	27.12	0.74	34.34

*Tmax, Tmin, and Tmean are maximum, minimum, and mean temperature, respectively. RH, DTR, and SW_IN_POT are relative humidity, daily temperature range, and top-of-atmosphere shortwave, respectively. WS and P are mean wind speed and total precipitation of the day.

Table S2. Variables and data sources for training ANN model*

Variables	Units	Data sources	Usage
SW_IN_POT	W/m^2	The daily integrated dataset	Input variable
Tmean	°C	The daily integrated dataset	Input variable
Tmax	°C	Half-hourly or hourly data	Input variable
Tmin	°C	Half-hourly or hourly data	Input variable
VPD	hPa	The daily integrated dataset	VPD was used to calculate RH
WS	m/s	The daily integrated dataset	Input variable
λE	W/m^2	The daily integrated dataset	Output variable
H	W/m^2	The daily integrated dataset	Output variable

*Vapor pressure deficit (VPD) was used to calculate relative humidity. λE and H are the latent heat flux and sensible heat flux, respectively.