Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2018JD028422

Special Section:
Water-soil-air-plant-human nexus: Modeling and observing complex land-surface systems at river basin scale

Intercomparison of Six Upscaling Evapotranspiration Methods: From Site to the Satellite Pixel

Xiang Li[1], Shaomin Liu[1], Huaixiang Li[1], Yanfei Ma[2], Jianghao Wang[3], Yuan Zhang[1], Ziwei Xu[1], Tongren Xu[1], Lisheng Song[4], Xiaofan Yang[1], Zheng Lu[1], Zeyu Wang[1], and Zhixia Guo[1]

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China, [2]Department of Geography, Handan College, Handan, China, [3]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, [4]Chongqing Engineering Research Center for Remote Sensing Big Data Application, School of Geographical Sciences, Southwest University, Chongqing, China

Abstract Land surface evapotranspiration (ET) is an important component of the surface energy budget and water cycle. To solve the problem of the spatial-scale mismatch between in situ observations and remotely sensed ET, it is necessary to find the most appropriate upscaling approach for acquiring ground truth ET data at the satellite pixel scale. Based on a data set from two flux observation matrices in the middle stream and downstream of the Heihe River Basin, six upscaling methods were intercompared via direct validation and cross validation. The results showed that the area-weighted method performed better than the other five upscaling methods introducing auxiliary variables (the integrated Priestley-Taylor equation, weighted area-to-area regression kriging [WATARK], artificial neural network, random forest [RF], and deep belief network methods) over homogeneous underlying surfaces. Over moderately heterogeneous underlying surfaces, the WATARK method performed better. However, the RF method performed better over highly heterogeneous underlying surfaces. A combined method (using the area-weighted and WATARK methods for homogeneous and moderately heterogeneous underlying surfaces, respectively, and using the RF method for highly heterogeneous underlying surfaces) was proposed to acquire the daily ground truth ET data at the satellite pixel scale, and the errors in the ground truth ET data were evaluated. The Dual Temperature Difference (DTD) and ETMonitor were validated using ground truth ET data, which solve the problem of the spatial-scale mismatch and quantify uncertainties in the validation process.

Key Points:
- Six upscaling evapotranspiration methods were intercompared based on direct validation and cross validation
- Daily "ground truth" evapotranspiration data were acquired at the satellite pixel scale over the Heihe River Basin
- The remotely sensed evapotranspiration products DTD and ETMonitor were validated using daily ground truth ET data

Correspondence to:
S. Liu,
smliu@bnu.edu.cn

Citation:
Li, X., Liu, S., Li, H., Ma, Y., Wang, J., Zhang, Y., et al. (2018). Intercomparison of six upscaling evapotranspiration methods: From site to the satellite pixel. *Journal of Geophysical Research: Atmospheres*, *123*. https://doi.org/10.1029/2018JD028422

Received 5 FEB 2018
Accepted 7 JUN 2018
Accepted article online 25 JUN 2018

1. Introduction

Land surface evapotranspiration (ET) includes evaporation from soil and water bodies, vegetation interception evaporation, and transpiration. As an important link in the surface water cycle and energy transfer, accurate estimations of ET benefit global climate change research and have great practical value for water resource management at the regional to global scale (Anderson et al., 2012; Jung et al., 2010).

Since the 1970s, with the development of remote sensing technology, the estimation of remotely sensed evapotranspiration (RS_ET) has become an effective method to obtain regional and global ET (Kalma et al., 2008; Li et al., 2009; Wang & Dickinson, 2012). Currently, a variety of medium- to low-resolution (≥1-km) RS_ET products have been released, including MOD16 (MODIS global ET; 8 day, 1 km; Mu et al., 2011), Global Land Evaporation Amsterdam Model ET (daily, 25 km; Miralles et al., 2011), ET-ITC (monthly, 10 km; Chen et al., 2014), Breathing Earth System Simulator ET (8 day, 1 km; Jiang & Ryu, 2016), Global Land Surface Satellite ET (8 day, 5 km before 2000/1 km after 2000; Yao et al., 2013), ETWatch (daily/monthly, 1 km; Wu et al., 2012), ETMonitor (8 day, 1 km; Hu & Jia, 2015), and DTD (Dual Temperature Difference; daily, 1 km; L. S. Song et al., 2016) products. However, uncertainties with respect to model mechanisms, parameterization schemes, model inputs, and scaling issues (Jia et al., 2012) can limit the performance of various RS_ET estimation models. The heterogeneity of underlying surfaces and the complexity of near-surface meteorological conditions both reduce the accuracy of RS_ET products (Liou & Kar, 2014; Liu et al., 2016). The mean absolute percentage errors (MAPEs) of the current RS_ET products at the instantaneous, daily, monthly, and annual scales are approximately 15–30%, 14–44%, 9–35%, and 5–21%, respectively (Ershadi et al., 2014; Kalma et al., 2008; Miralles et al., 2016; Velpuri et al., 2013; Wang et al., 2008; Yao et al., 2013). Therefore, RS_ET products must be validated using ground measurements before application.

In recent years, most validations for RS_ET products have been conducted over homogeneous underlying surfaces (Jia et al., 2012). When examining heterogeneous underlying surfaces, however, the validation of RS_ET products remains a challenge. A major problem is the spatial-scale mismatch between in situ measurements and RS_ET products (L. S. Song et al., 2016). Therefore, it is important to select appropriate validation pixels. When a pixel located in the instrument or multiple pixels around the instrument are selected as the validation pixels, the error is small over homogeneous underlying surfaces; larger errors are generated over heterogeneous underlying surfaces. When the pixels located in the flux source areas are selected, the footprint value within the source areas is normalized as the relative weight. The footprint-weighted average of the remote sensing estimates within these pixels is compared with in situ measurements. Although the problem of spatial-scale mismatch can be partially addressed, the heterogeneity of the subpixel will have a greater impact on the results of the validation (Bai et al., 2015; Jia et al., 2012). Therefore, to reduce the uncertainty caused by the spatial-scale mismatch that arises during validation, it is necessary to acquire "ground truth" ET data at the satellite pixel scale over heterogeneous underlying surfaces. At present, methods for the acquisition of ground truth ET data at the satellite pixel scale over heterogeneous underlying surfaces mainly include average method, simulation method, geostatistical method, and machine learning method.

The average methods include arithmetic average (Liu et al., 2016; Peng et al., 2008), area-weighted (AW) average (Beyrich et al., 2006; Ezzahar et al., 2009; Gottschalk et al., 1999; Liu et al., 2016), and footprint-weighted (Evans et al., 2012; Liu et al., 2016) methods. Based on the measurements from the Heihe Watershed Allied Telemetry Experimental Research-Multi-Scale Observation Experiment on Evapotranspiration (HiWATER-MUSOEXE) and a high-resolution land cover map derived from aircraft remote sensing, F. N. Xu et al. (2017) applied the AW method over a heterogeneous landscape with footprint analysis and multivariate regression. The simulation method for upscaling land surface water and heat fluxes can be divided into empirical and statistical models (Fang et al., 2016; Sun et al., 2011; Zhang et al., 2001), semiempirical and semitheoretical models (Liu et al., 2016), and theoretical models (Gottschalk et al., 1999; Heinemann & Kerschgens, 2005). Based on mulitsite measurement data from HiWATER-MUSOEXE, Liu et al. (2016) proposed an upscaling method integrated Priestley-Taylor equation (P-T) over heterogeneous surface to overcome the limitation of AW method that is mostly appropriate for homogeneous surface. Geostatistical methods mainly involve these upscaling methods based on kriging (Ge et al., 2015; Hu et al., 2015; Stein, 2012) and the Bayesian theory (Bernardo & Smith, 2001; Gao et al., 2014; Qin et al., 2013; Shi et al., 2015). Based on the data from the HiWATER-MUSOEXE, Ge et al. (2015) used the normalized difference vegetation index (NDVI), land surface temperature (LST), fraction of vegetation cover (FVC), and wind speed as auxiliary information along with the area-to-area regression kriging) method to upscale the *H* from the eddy covariance (EC) scale to large aperture scintillometer (LAS) scale. The upscaled results were generally close to the LAS observations. Upscaling methods based on machine learning include artificial neural network (ANN) (Chen et al., 2013; Landeras et al., 2008; Wang et al., 2016), regression tree (Ghahreman & Sameti, 2014; Jung et al., 2011; Metzger et al., 2013; Rahimikhoob, 2014; K. Xu et al., 2017), support vector machine (Yang et al., 2006), and random forest (RF; Bodesheim et al., 2018), which have been used in surface flux upscaling. Alternative machine learning methods have been utilized to upscale other parameters; for example, the deep belief network (DBN; X. D. Song et al., 2016) is used in soil moisture upscaling. At present, machine learning methods generally conduct upscaling from site to regional scale. Jung et al. (2011) trained model tree ensembles and then upscaled the monthly LE from the site scale to the global scale and produced global products at 0.5° spatial resolution for monthly average value. In the most recent study of Bodesheim et al. (2018), they performed the upscaling using the trained random decision forest and produced continuous half-hourly products at 0.5° spatial resolution in the global scale.

Although there are many upscaling methods, most upscaling methods currently developed for ET are in the exploratory stage. Due to the heterogeneity of the underlying surfaces and the complexity of near-surface meteorological conditions, the applicability of different upscaling methods varies, and each method has cons and pros. Therefore, carrying out a comparative study of various upscaling methods is necessary in order to determine the most appropriate upscaling method over heterogeneous surfaces to obtain ground truth ET data at the satellite pixel scale and to reduce uncertainty due to spatial-scale mismatch in the validation process.

In summary, this study was based on two matrices data from the HiWATER experiment: the MUSOEXE conducted in the middle stream of the Heihe River Basin (HRB) from May to September in 2012 and the Hydrometeorological Observation Network conducted in the downstream of the HRB from 2013 to 2015.

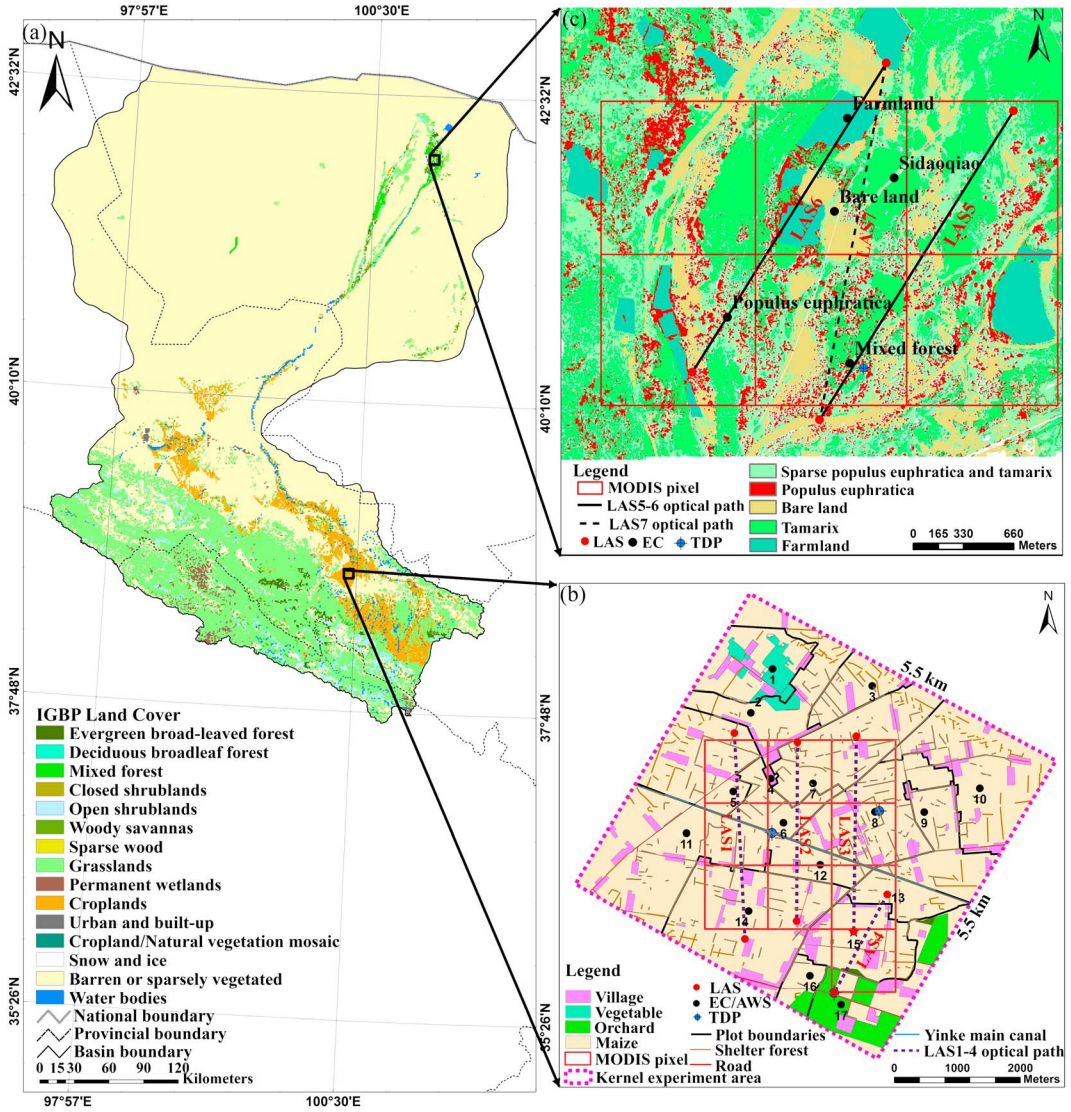

Figure 1. Spatial locations of (a) HRB and stations in the (b) middle stream and (c) downstream areas.

Six upscaling methods—the AW method, P-T method, weighted area-to-area regression kriging (WATARK) method, ANN method, RF method, and DBN method—were intercompared via direct validation taking the LAS measurements as the satellite pixel reference and cross validation using three-cornered hat method (TCH; Galindo & Palacio, 1999). Finally, daily ground truth ET data were acquired using a combined method at the satellite pixel scale to validate RS_ET products over heterogeneous surfaces.

2. Study Area and Data

2.1. Study Area and Experiment

The HiWATER experiment was performed in the HRB (37.7°–42.7°N, 97.1°–102.0°E), which is located in a semi-arid region of northwestern China and covers approximately 1,432,000 km^2 (Figure 1a). The underlying surfaces of the HRB are primarily characterized by glaciers, frozen soil, alpine meadow, and forest in the upstream area; by desert and crops such as maize, wheat, and vegetable in the middle stream area; and by riparian ecosystems and widespread desert in the downstream area (Li et al., 2013).

The HiWATER-MUSOEXE, the thematic experiment in HiWATER, was conducted in the middle stream of the HRB between May and September 2012 and involved a flux observation matrix composed of two nested matrices: one large experimental area (30 km × 30 km) in an oasis-desert area and one kernel

experimental area (5.5 km × 5.5 km) within the Zhangye Oasis (Figure 1b). According to the distribution of crops, shelterbelts, residential areas, roads, and canals, as well as the soil moisture and irrigation status, the kernel experimental area was divided into 17 elementary sampling plots (including site 1 = vegetable, site 4 = village, site 17 = orchard, and other sites = maize). One EC system (two at site 15—Daman superstation) and one automatic weather station (AWS) were included per plot for the synchronous observation of water and energy fluxes and meteorological elements. In addition, three poplar trees near sites 6, 8, and 17 were each installed with three thermal dissipation probes (TDPs) installed at a height of 1.3 m to observe transpiration of the shelterbelt. Three groups of optical LASs were installed in three 3 × 1 MODIS pixels (from west to east: LAS1, LAS2, and LAS3). In addition, one group of LASs was installed in one 2 × 1 MODIS pixel (LAS4). More details regarding the HiWATER-MUSOEXE field campaign can be found in Li et al. (2017). The sites located in the three 3 × 1 and one 2 × 1 MODIS pixels were selected to study the ET upscaling methods based on the 11 EC and AWS sets, 4 LAS groups, and 3 TDP groups at sites 4–8, 11–15, and 17 (Table 1).

The Hydrometeorological Observation Network of the HRB has been operational since July 2013 during the HiWATER experiment. There are five sites installed in Ejina Oasis in the downstream of the HRB, constituting a flux observation matrix (3 km × 2 km; Figure 1c), including the Sidaoqiao superstation (*Tamarix*), Farmland station (farmland, mainly planting melon), Bare land station (bare land), Mixed forest station (sparse *Populus euphratica* and *Tamarix*), and *Populus euphratica* station (sparse *Populus euphratica*). Each site was installed with one EC and one AWS to simultaneously observe *H*, LE, and meteorological elements. In addition, according to the different heights and diameters at breast height of *Populus euphratica*, three *Populus euphratica* trees were outfitted with three TDPs at a height of 1.3 m to observe the transpiration of *Populus euphratica* near the Mixed forest station. From July 2013 to May 2015, two groups of LASs (LAS5 and LAS6) were installed in two 2 × 2 MODIS pixels. Since May 2015, two groups of LASs have been adjusted to one group of LAS (LAS7), which was installed in one 2 × 1 MODIS pixel. Detailed information regarding the sites and instruments is provided in Table 1 and Figure 1c.

To obtain reliable observation data, the consistency analysis of observation instruments is particularly important. Before the HiWATER experiment, two comparison experiments for surface energy flux measurement systems were conducted, including one in the Bajitan Gobi desert (a nearly flat and open surface) to the west of Zhangye City in the middle stream from 14 to 24 May 2012 and the other in the shrubs (relatively homogeneous underlying surfaces) around Ejin Banner in the downstream from 27 June to 3 July 2013 (Xu et al., 2013). The comparison experiment in the middle stream involves 7 LASs, 20 ECs, and 18 radiometers; the comparison experiment in the downstream involves 2 LASs, 6 ECs, and 6 radiometers. The results of two comparison experiments showed that sensible heat fluxes between the LAS1–LAS7 and the corresponding ECs were consistent in the middle stream and downstream, with regression slope values 3%, 1%, 1%, 5%, 8%, and 4% and *R* values 0.97, 0.98, 0.98, 0.97, 0.96, and 0.95, respectively. It is noted that LAS5 and LAS7 have the same consistency because their corresponding ECs are the same. Therefore, EC and LAS measurements were consistent with each other for homogeneous underlying surface, leading to the conclusion that the LAS measurements and related ECs were also comparable. More details can be found in Xu et al. (2013) and Liu et al. (2016).

2.2. Data Processing

2.2.1. Flux and Meteorological Data

The raw EC data were stored at a sampling frequency of 10 Hz and processed using the EdiRe (http://www.geos.ed.ac.uk/homes/jbm/micromet/EdiRe/) software packages, including spike detection, coordinate rotation (2-D rotation), sonic virtual temperature correction, corrections for density fluctuation (Webb-Pearman-Leuning correction), and frequency response correction. The EC data were finally converted into half-hour-averaged flux data. A quality assessment was performed using the quality flags (0, 1, and 2) according to both stationary test and integral turbulence characteristics test. The flux data for flag 2 were discarded. The 30-min flux data were screened as follows: data from periods of sensor malfunction; data within 1 hr before or after precipitation; incomplete 30-min data when the missing data constituted more than 3% of the 30-min raw record; were rejected (Liu et al., 2011, 2013). The Bowen ratio closure method was used to force energy balance closure in our study (Twine et al., 2000; Z. W. Xu et al., 2017). To obtain daily ET, the nonlinear regression method was used to establish the relationship between LE and R_n and to fill the gaps between the 30-min flux data. Daily ET was calculated using the continuous 30-min data. In addition, days

Table 1
Details Regarding Meteorological and Flux Sites in Two Flux Observation Matrices for Middle Stream and Downstream of the HRB

Observation items	Type, manufactures	Height/depth (m)	Site	Duration
Eddy covariance system sensible and latent heat flux	CSAT3/Li7500A, Cambell/Li-cor, USA	4.2 (6.2 after 19 August)	4	site 4: 2012.5.31–9.17
		4.6	6	site 6: 2012.5.28–9.21
		3.8	7	site 7: 2012.5.29–9.18
		5	13	site 13: 2012.5.27–9.20
		4.5, 34	15	site 15: 2012.5.25-
		3.5 (2013.7.14–2015.10.29)	F	site F: 2013.7.14–2015.10.29
		8	S	site S: 2013.7.6-
	CSAT3/Li7500, Cambell/Li-cor, USA	3	5	site 5: 2012.6.03–9.18
		3.2	8	site 8: 2012.5.28–9.21
		3.5	11	site 11: 2012.5.29–9.18
		3.5	12	site 12: 2012.5.28–9.21
		4.6	14	site 14: 2012.5.30–9.21
		22	P	site P: 2013.7.12–2016.4.21
		22	M	site M: 2013.7.12-
		3.5	B	site B: 2013.7.10–2016.3.14
	CSAT3/EC150, Campbell, USA	7	17	site 17: 2012.5.31–9.17
		3.5 (2013.7.14–2014.4.15)	F	site F: 2013.7.14–2014.4.15
Large aperture scintillometer sensible heat flux	BLS900, Scintec, Germany zzlas, RR9340, Rainroot, China	33.45 (path length: 3256)	LAS1	2012.6.07–9.19 2012.6.16–9.19
	BLS900, Scintec, Germany BLS450, Scintec, Germany	33.45 (path length: 2841)	LAS2	2012.6.07–9.19 2012.6.18–9.19
	BLS900, Scintec, Germany LAS, Kipp&zonen, Netherland	33.45 (path length: 3111)	LAS3	2012.6.06–9.20 2012.6.19–9.20
	BLS450, Scintec, Germany zzlas, RR9340, Rainroot, China	22.45(path length: 1854)	LAS4	2012.6.02- 2012.6.21–9.20
	BLS900, Scintec, Germany	25.5 (path length: 2390)	LAS5	2013.7.11–2015.4.24
	BLS900, Scintec, Germany	25.5 (path length: 2380)	LAS6	2013.9.16–2015.4.24
	BLS900, Scintec, Germany	25.5 (path length: 2350)	LAS7	2015.4.26–2015.10.30
Air pressure	PTB110, Vaisala, Finland	-	17	AWS4: 2012.5.10–9.17
	CS100, Campbell, USA	-	4, 5, 6, 7, 8, 11, 12, 13, 14, 15, S	AWS5: 2012.6.04–9.18
	AV-410BP, Avalon, USA	-	M	AWS6: 2012.5.09–9.21
Precipitation	TE525MM, Texas Electric, USA	-	4, 5, 6, 7, 8, 11, 12, 13, 14, 15, S	AWS7: 2012.5.28–9.18 AWS8: 2012.5.14–9.21
	52203, RM Young, USA	-	17, M	AWS11: 2012.6.02–9.18
Wind speed/direction	010C/020C, Met One, USA	5, 10	6	AWS12: 2012.5.10–9.21
		10	4, 5, 7, 8	AWS13: 2012.5.06–9.20
		5, 7, 10, 15, 20, 28	S	AWS14: 2012.5.06–9.21
	3001, RM Young, USA	10	11	AWS15: 2012.5.10-
		28	M, P	AWS17: 2012.5.12–9.17

Table 1 *(continued)*

Observation items	Type, manufactures	Height/depth (m)	Site	Duration
	034B, Met One, USA	10	12, 13, 14, 17	AWS P: 2013.7.11–2016.4.21
	Windsonic, Gill, UK	3, 5, 10, 15, 20, 30, 40	15	AWS M: 2013.7.13-
Air temperature/humidity	HMP45D, Vaisala, Finland	5	12, 13, 14	AWS S: 2013.7.12-
	HMP45C, Vaisala, Finland	5	4, 5, 17	AWS F: 2013.7.10–2015.10.29
		28	M, P	AWS B: 2013.7.10–2016.4.21
	HMP45AC, Vaisala, Finland	5	7, 8, 11	
		5, 10	6	
	AV-14TH, Avalon, USA	3, 5, 10, 15, 20, 30, 40	15	
	HC2S3, Campbell, USA	5, 7, 10, 15, 20, 28	S	
Four-component radiation	CNR4, Kipp&Zonen, Netherland	6	6, 8, 13, 14	
		4	7, 12	
		10	S	
		6, 24	P	
		6	B, F	
		24	M	
	CNR1, Kipp&Zonen, Netherland	6	4, 17	
		4	5, 11	
	PSP&PIR, Eppley, USA	12	15	
	CM21, Kipp&Zonen, Netherland(2013.8–2014.4)	6	F	
Soil moisture	ECH$_2$O-5, Decagon Devices, USA	−0.02, −0.04, −0.1, −0.2, −0.4, −0.6, −1	12, 13, 14	
	CS616, Campbell, USA		4, 5, 6, 7, 8, 11, 17	
	CS616, Campbell, USA	−0.02, −0.04, −0.1, −02, −0.4, −0.8, −1.2, −1.6	15	
	ML2X, Delta-Tdevices, UK	−0.02, −0.04, −0.1, −0.2, −0.4, −0.8, −1.2, −1.6	S	
		−0.02, −0.04, −0.1, −0.2, −0.4, −0.6, −1.0	M	
		−0.02, −0.04	P, B, F	
Soil heat flux	HFP01, Hukseflux, Netherland	−0.06	4, 5, 6, 7, 8, 11, 17	
	HFT3, Campbell, USA		12, 13, 14, M, P, B, F	
	HFP01SC, Hukseflux, Netherland		15, S	
Soil temperature	109ss-L, Campbell, USA	0, −0.02, −0.04, −0.1, −0.2, −0.4, −0.6, −1	4, 6, 8, M	
	AV-10 T, Avalon, USA		5, 12, 13, 14	
	109, Campbell, USA		7, 11, 17	
	AV-10 T, Avalon, USA	0, −0.02, −0.04, −0.1, −0.2, −0.4, −0.8, −1.2, −1.6	15	
		0, −0.02, −0.04	P, B, F	
Infrared radiation temperature	SI-111, Apogee, USA	4	5, 7, 11	
		6	4, 6, 8, 17, S, M, P, F, B	
	IRTC3, Avalon, USA	4	12, 13, 14, 15	

Note. P, M, S, F, and B represent the *Populus euphratica* station, Mixed forest station, Sidaoqiao superstation, Farmland station, and Bare land station, respectively.

with a percentage of missing 30-min data greater than 50% were considered missing days and were not used in the upscaling analysis. Detailed information about the EC raw data processing steps can be found in Xu et al. (2013) and Liu et al. (2013).

The LAS system provided a measurement of the structure parameter for the refractive index of air (C_n^2). The raw LAS data were first averaged to 30 min. Then, the path average sensible heat fluxes were iteratively calculated combining meteorological data (e.g., wind speed, air temperature, and pressure) based on the Moninin-Obukhov Similarity Theory. To perform quality control for the raw LAS data, the criterion $C_n^2 < 0.193\ L^{8/3}\lambda^{1/3}D^{5/3}$ (L is the path length, D is the diameter of optical aperture, and λ is wavelength) was applied to remove data with values exceeding the saturated condition (Ochs & Wilson,1993). Data were rejected under the following circumstances: if the demodulation signal was small, if collected within 1 hr

before or after precipitation, or if collected at night when weak turbulence occurred (friction wind speed [u_*] less than 0.1 m/s). The nonlinear regression method was used to fill the gaps between the 30-min data. The optical LAS can only measure an integrated H over its path length. To estimate the LAS-LE, the energy balance equation $LE = R_n - G_0 - H$ was used. The area-averaged R_n and surface soil heat flux (G_0) were obtained using the area-weighted method of related R_n and G_0 values from several plots within each LAS source area. However, during advection conditions ($H < 10$ W/m^2 in the daytime), the LAS-LE was obtained directly from the linear interpolation using the relationship of LAS-LE and the area-averaged R_n under nonadvection conditions ($H > 10$ W/m^2; Liu et al., 2016). Finally, daily ET was calculated by summing the half-hourly gap-filled ET to a 24-hr value. In addition, days with a percentage of missing 30-min data greater than 50% were considered missing days and were not used in the upscaling analysis.

The AWS data include wind speed and direction, precipitation, air temperature and humidity, pressure, downward and upward shortwave and longwave radiation, radiometric surface temperature, soil heat flux, soil temperature, and moisture profile. The processing and quality control steps are important, including the following steps: (1) AWS data were processed to a 30-min average period; (2) data out of the range of physical possibility were rejected, and the linear interpolation was used to fill the gaps; and (3) soil heat flux plates were buried at depths of 0.06 m, with three replicates located underground at each site (two plates were buried under bare soil, while another plate was buried under plants). The surface soil heat flux G_0 was calculated using the "PlateCal" approach (Liebethal et al., 2005) based on the combination of weighted vegetation fraction, soil temperature, and moisture measured above the heat plates. The LST data were calculated with downward and upward longwave radiation measured by AWSs and with surface emissivity measured by a FT-IR spectrometer (102F; Mu et al., 2012).

The transpiration of the shelterbelts in the middle stream and of *Populus euphratica* in the downstream was observed using the TDP. The raw TDP data were temperature differences with a sampling frequency of 30 s, averaged over 10 min, and processed to a 30-min average period. The sap flux density and total sap flow were calculated using the temperature difference between the probes and the cross-sectional area of the sapwood. Daily transpiration of the shelterbelt in the middle stream and of *Populus euphratica* in the downstream was calculated using total sap flow and the forest belt area. Finally, the TDP data were postprocessed to ensure the quality of the data, including removal of data that were significantly beyond physical meaning or instrument range and elimination of suspicious data due to probe failures (Qiao et al., 2015).

2.2.2. Remote Sensing and Atmospheric Forcing Data

The remote sensing data used in this study are based on the Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model data fusion algorithm. For the middle stream, 6 ETM+ images from May to September 2012 and 12 ASTER images were fused with the MODIS data corresponding to the time of the MODIS overpass. For the downstream, 25 ETM+ images from 2013 to 2015 were fused with the MODIS data corresponding to the time of MODIS overpass. The reflectivity data and land surface temperature in the overpass time of the satellite at 100-m resolution were obtained from May to September 2012 in the middle stream and from May to September 2013–2015 in the downstream, respectively. Finally, the land surface parameters (e.g., R_n, LST, NDVI, and FVC) were calculated using the fused data (Ma, 2015).

Land use/land cover data for the Zhangye region at 30-m resolution in the 2005 data set (Yan, 2011) were used in the middle stream. Based on Landsat ETM data obtained in 2012 and the results of field verification, the 2005 land cover data set was updated using the "Editor" in ArcGIS. Land use/land cover data (Xu et al., 2015) at 30-m resolution were used in the downstream.

Based on the DTD model derived from the two-source energy balance model scheme, L. S. Song et al. (2016) estimated the daily ET in the HRB for the years 2012–2015 at a spatial resolution of 1 km. Based on the Shuttleworth-Wallace dual-source model, using a variety of biophysical parameters derived from microwave and optical remote sensing observations as forcing data, Hu and Jia (2015) evaluated the daily ET (ETMonitor) at the local scale and global scale for the years since 2009 at a spatial resolution of 1 km. The DTD and ETMonitor products were used in the present study for validation, as discussed in section 4.3.

In this study, hourly atmospheric forcing data with a spatial resolution of 5 km, produced by the Weather Research and Forecasting model, were collected from the data set produced by the Environmental and Ecological Science Data Center for West China, National Natural Science Foundation of China (Pan et al.,

2012; http://card.westgis.ac.cn/). Atmospheric forcing data with a spatial resolution of 100 m, such as atmosphere pressure, air temperature, relative humidity, and precipitation, were produced using the regression kriging method combined with ground measurements from 21 flux towers in the middle stream and 5 flux towers in the downstream with Weather Research and Forecasting data (Ma, 2015). Finally, these data were resampled to 30-m resolution.

3. Methodology

3.1. Upscaling Methods

According to the general classification of upscaling methods (as described in section 1), in this study, the AW, P-T, and WATARK methods were selected from the average, simulation, and geostatistical methods, respectively. Among the machine learning methods, the ANN, RF, and DBN algorithms were selected for ET upscaling. Results from the selected upscaling methods were intercompared by taking the LAS measurements as the satellite pixel reference. More details of the AW, P-T, WATARK, ANN, RF, and DBN methods to obtain ET at the satellite pixel scale can be found in Appendix A.

3.2. Footprint Model

The footprint of flux measurement is the transfer function between the measured value and the set of forcings on the surface-atmosphere interface, while the source area can be interpreted as the integral of the footprint function over a specified domain (Schmid, 1994, 2002). The flux contributing source area of the EC and LAS measurements can be estimated using the footprint model. To estimate the flux footprint of the EC measurements, a method proposed by Kormann and Meixner (2001), namely, a Eulerian analytic flux footprint model, was implemented. The LAS source area can be calculated by combining the path-weighting function of the LAS with the footprint model for point fluxes (Meijninger et al., 2002). The resolution of the source area was 30 m for both EC and LAS measurements, and the flux contribution of the chosen total source area was set to 90%.

4. Results and Discussion

4.1. Analysis of Surface Heterogeneity

The spatial heterogeneity of ET was evaluated using the coefficient of variation CV (see Appendix B) as the assessment index based on the LE measured by ECs in the middle stream and downstream matrices (taking the growing periods in 2012 and 2014 as two examples). The results of CV are shown in Figure 2.

As shown in Figure 2a, the spatial heterogeneity of ET in the middle stream matrix has a U shape. Before the crop at full cover (before 10 July) and at the end of the crop growing period (after 3 September), the spatial heterogeneity of ET (0.23–0.35) is large. However, in the middle of the crop growing period, the heterogeneity of ET (0.10–0.22) is relatively small. In addition, the spatial heterogeneity of ET decreases significantly after rainfall and increases during the period of rotational irrigation.

According to the variation trend of spatial heterogeneity of ET over the underlying surfaces, combined with the maize phenophase in the study area, in this study, days before the crop at full cover (before 10 July), at the end of the crop growing period (after 3 September) and the period of rotational irrigation in the middle of the crop growing period, were defined as the moderately heterogeneous stages of ET. The period of nonirrigation in the middle of the crop growing period was defined as the homogeneous stage of ET.

The underlying surfaces in the downstream matrix are very fragmentary, with less precipitation and no irrigation during the growing period. Using the growing period in 2014 as an example, the variation trend of spatial heterogeneity of ET over time was obtained using the CV as the evaluation index (Figure 2b). The heterogeneity of ET (0.35–0.58) is obviously greater than that in the middle stream (0.10–0.35) and with a smaller variation range. Thus, the entire growing period was defined as the highly heterogeneous stage of ET.

4.2. Comparison of Upscaling Methods

4.2.1. Direct Validation: Comparison With LAS Measurements

Based on the footprint model, the averaged footprints in the daytime (10:30–18:00 BST) during the growing period were calculated with four groups of LASs in 2012, two groups of LASs in 2014, and one group of LAS in 2015 (Figure 3). As shown in Figure 3, the proportion of daytime-averaged LAS source area in the

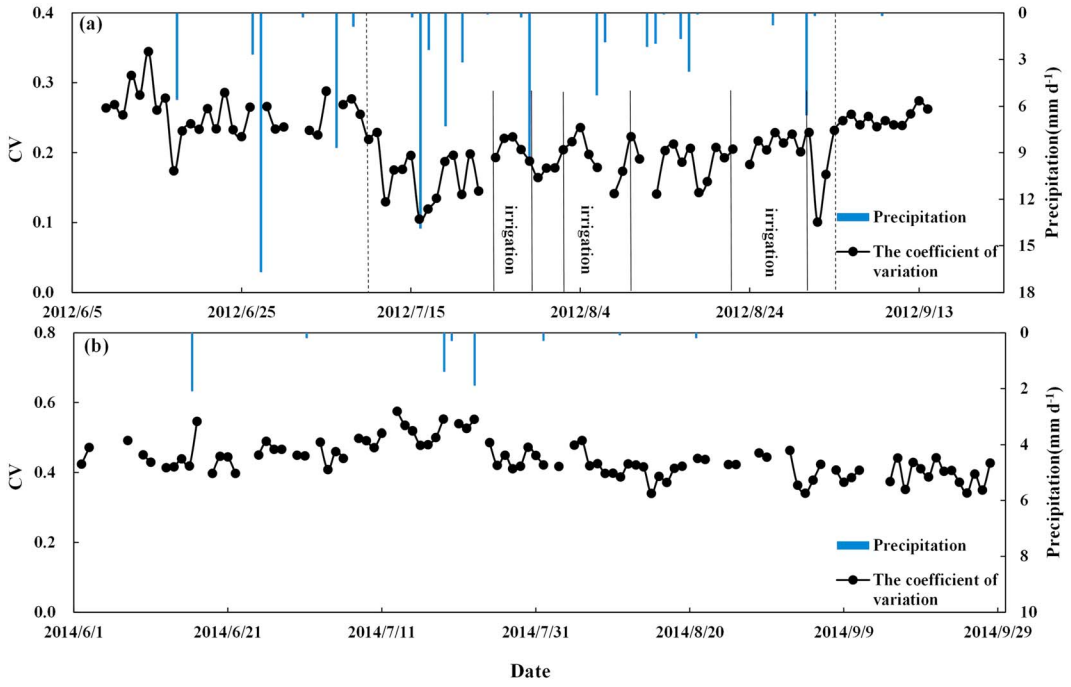

Figure 2. Variation trend of spatial heterogeneity of ET over time in (a) the middle stream matrix in 2012 and (b) downstream matrix in 2014. Dotted lines in (a) represent the time before the crop at full cover and the end of the crop growing period.

corresponding MODIS pixels was approximately 80% in 2012 and 2014 and more than 95% in 2015, which basically covered the corresponding MODIS pixels and were located in the central area. It is assumed that the four groups of LAS measurements in the middle stream matrix can represent the measurements in the 3 × 1 or 2 × 1 MODIS pixel scale. The two groups of LASs in the downstream matrix in 2014 can represent the measurements in the 2 × 2 MODIS pixel scale. One group of LAS in 2015 can represent the measurements in the 2 × 1 MODIS pixel scale in the downstream matrix. Therefore, the LAS measurements can be used as a reference to evaluate the upscaled results in the 3 × 1, 2 × 2, or 2 × 1 MODIS pixels.

Based on the data in the middle stream matrix during the growing period (8 June to 14 September 2012) and the downstream matrix during the growing period (1 June to 30 September 2014 and 2015), the upscaled ET results of six upscaling methods were obtained and compared with the LAS measurements.

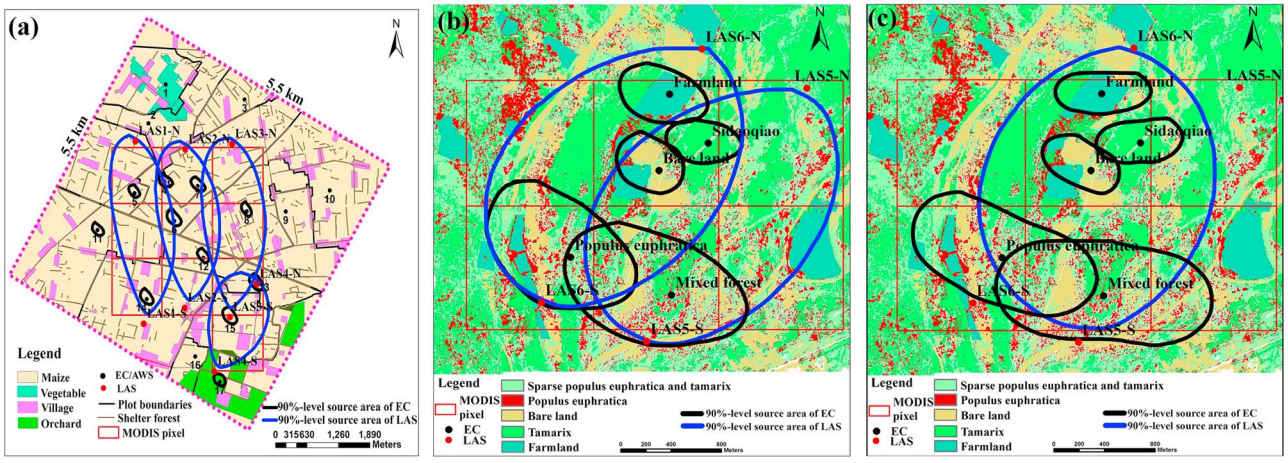

Figure 3. Daytime averaged LAS and EC source areas (90% flux contribution) in (a) the middle stream matrix in 2012, (b) downstream matrix in 2014, and (c) downstream matrix in 2015.

Table 2
Results of Global Sensitivity Analysis of ANN, RF, and DBN Models in Middle Stream and Downstream Matrices

Input parameters	The middle stream matrix			The downstream matrix		
	ANN	RF	DBN	ANN	RF	DBN
R_n	0.25	0.43	0.52	0.25	0.43	0.50
VPD	0.08	0.04	0.16	0.01	0.03	0.07
LST	0.35	0.21	0.36	0.51	0.34	0.53
NDVI	0.26	0.18	0.30	0.16	0.13	0.38
FVC	0.16	0.15	0.22	0.21	0.24	0.27

The upscaling of ET results over heterogeneous underlying surfaces at the satellite pixel scale can be significantly improved by using these upscaling methods introducing auxiliary variables that can characterize the heterogeneity of the surface water and heat conditions (Liu et al., 2016). In this study, certain auxiliary variables (i.e., R_n, VPD, LST, NDVI, and FVC) retrieved from remote sensing data were introduced to the P-T, WATARK, ANN, RF, and DBN methods. Their accuracy will directly affect the accuracy of the upscaled ET results. To reduce the error of the upscaled ET results caused by the uncertainty of the auxiliary variables, we must analyze the sensitivity of these parameters to determine the contribution of each parameter to the upscaled ET results. Table 2 shows the results of the global sensitivity analysis of ANN, RF, and DBN (P-T and WATARK methods cannot perform global sensitivity analysis because they are not model based) in the middle stream and downstream matrices using Extend Fourier Amplitude Sensitivity Test (Saltelli et al., 1999). This method not only estimates the independent impact of the input parameters of a model on the estimated results but also considers the combined effect of the interaction among different model parameters on the results. A larger value indicates a greater contribution rate of the parameters to the estimated results.

From the results of the global sensitivity analysis in Table 2, generally, it can be seen that the ANN, RF, and DBN models are more sensitive to R_n, LST, and NDVI than VPD and FVC. Thus, R_n, LST, and NDVI were selected as auxiliary variables introducing to the three upscaling methods in this study. Using LAS1 in the middle stream as an example, Figure 4a shows the relative error trend for the five upscaling methods introducing auxiliary variables over time. Figures 4b and 4c show the trend of the retrieval error of LST and R_n at the high resolution (30 m) over time, respectively. In Figures 4a–4c, it can be seen that the retrieval accuracy of LST and R_n significantly affects the accuracy of the upscaled ET results. In most occasions, for example, on 22 June, 4 July, and 13 September, the retrieval errors of LST and R_n are large, and thus, the corresponding errors of the upscaled ET results are also larger. However, on certain days, for example, on 10 June and 19 July, the relative errors of upscaled results are large but the retrieval errors of LST and R_n are small. This is because other than LST and R_n, other sensitive parameters and the errors derived from observation data (e.g., LAS, EC, LST, and R_n) will also affect the relative errors of upscaled results. Besides, the algorithms of upscaling methods also have errors, which would affect the upscaled results.

The current retrieval error of LST is approximately 2–3 K (Guillevic et al., 2014; Kabsch et al., 2008; Zhou et al., 2015). The retrieval error of R_n is approximately 10% (De Oliveira & Moraes, 2013; Huang et al., 2016; Wang et al., 2015). Therefore, 2 K and 15 W/m^2 and 3 K and 15 W/m^2, respectively, were taken as the error threshold for remotely sensed LST and R_n in middle stream and downstream matrices (because the underlying surfaces in the downstream are fragmentary, the error threshold for remotely sensed LST is relatively large). In the following comparison, data retrieval errors for LST or R_n that are greater than the threshold will be removed. Using direct validation (i.e., taking the LAS measurements as the satellite pixel reference), a comparison of the upscaled ET results (for the assessment index, see Appendix B) is presented in Figure 5, with statistical results in Table 3.

Figure 5 and Table 3 show that for the middle stream matrix, the AW method has good accuracy over homogeneous underlying surfaces (RMSE is 0.19 mm/d, and MAPE is 3.15%) and is slightly superior to the other five upscaling methods introducing auxiliary variables. This finding is consistent with the conclusion of Liu et al. (2016). Over moderately heterogeneous underlying surfaces, the WATARK method performs slightly better in the middle stream matrix, with the slightly smaller error (RMSE is 0.29 mm/d, and MAPE is 4.80%). The RF method follows, with an RMSE of 0.29 mm/d and MAPE of 5.44%, and performs slightly better among the three machine learning methods. In addition to the DBN method, the performance of upscaling methods introducing auxiliary variables is slightly better than the AW method (RMSE is 0.36 mm/d, and MAPE is 6.43%). Over highly heterogeneous underlying surfaces, the RF method introducing auxiliary variables performs slightly better, with an RMSE of 0.55 mm/d and MAPE of 13.65%. The AW method provides an RMSE of 0.63 mm/d and MAPE of 14.19%. The performances of other four upscaling methods introducing auxiliary variables are relatively poor.

In the above analysis, for moderately heterogeneous underlying surfaces, the WATARK method performs better in the middle stream matrix, followed by the RF method. However, for highly heterogeneous underlying

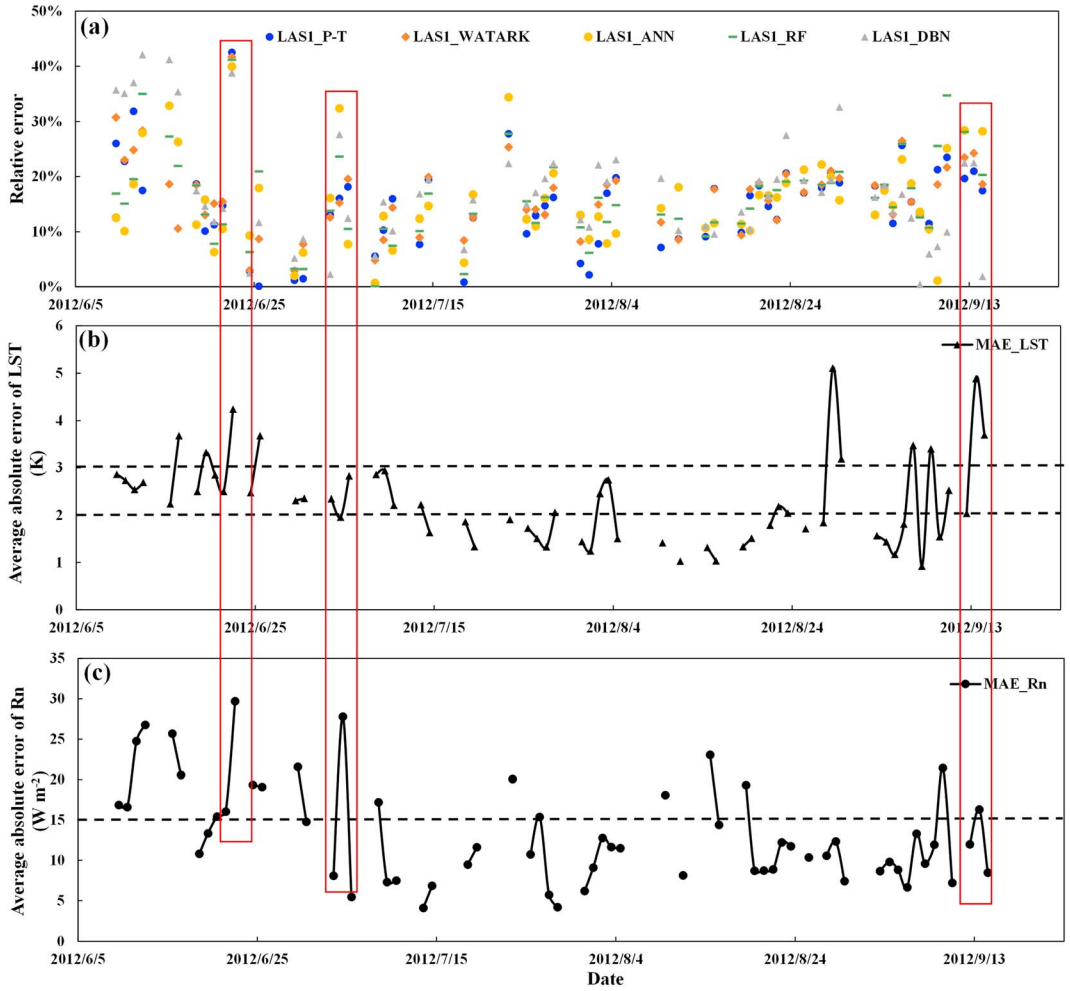

Figure 4. (a) Trend of the error of upscaling methods introducing auxiliary variables and trend of the retrieval error of (b) LST and (c) R_n over time. MAE represents the mean absolute error calculated by the average of auxiliary variables at sites.

surfaces, the RF method performs better in the downstream matrix. To analyze the cause of the differences between the two methods in middle stream and downstream matrices, here using the middle stream matrix as an example, the accuracy of the upscaled ET results from the WATARK and RF methods is compared for different numbers of EC sites (Figure 6). As shown in Figure 6, with the decreasing in site number, the

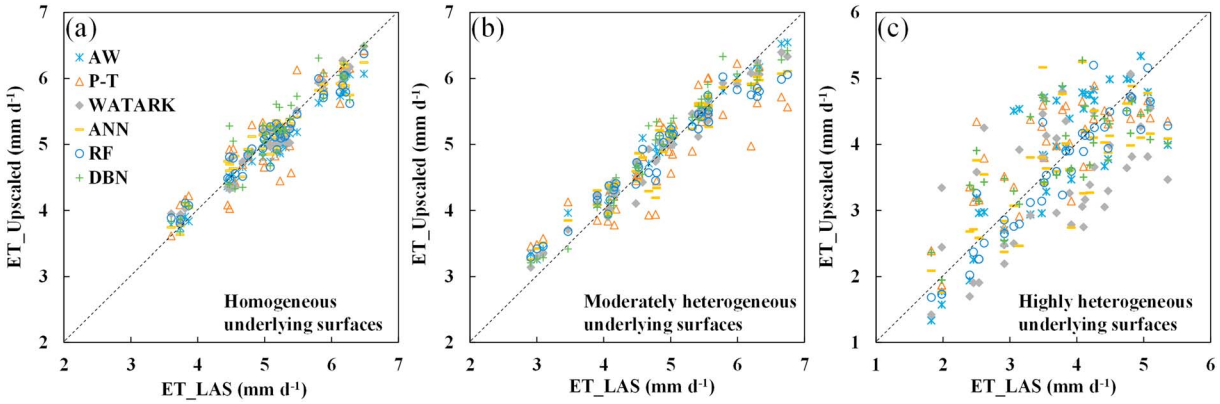

Figure 5. Comparison of upscaled ET results obtained from six upscaling methods with LAS measurements when (a/b) MAE_LST $<$ 2 K and MAE_R_n $<$ 15 W/m^2 in the middle stream matrix and (c) MAE_LST $<$ 3 K and MAE_R_n $<$ 15 W/m^2 in the downstream matrix.

Table 3
Statistics for Comparison Between Upscaled ET Results Obtained From Six Upscaling Methods and LAS Measurements

	Homogeneous underlying surfaces (N = 40; middle stream matrix)				Moderately heterogeneous underlying surfaces (N = 42; middle stream matrix)				Highly heterogeneous underlying surfaces (N = 43; downstream matrix)			
Method	Slope	R	RMSE (mm/d)	MAPE (%)	Slope	R	RMSE (mm/d)	MAPE (%)	Slope	R	RMSE (mm/d)	MAPE (%)
AW	0.99	0.97	0.19	3.15	1.03	0.92	0.36	6.43	0.88	0.82	0.63	14.19
P-T	1.01	0.95	0.29	4.36	0.99	0.93	0.30	5.56	1.04	0.77	0.63	14.55
WATARK	1.01	0.96	0.27	4.22	1.02	0.97	0.29	4.80	0.91	0.58	0.88	20.37
ANN	1.02	0.93	0.32	5.27	1.05	0.93	0.34	6.25	0.97	0.75	0.66	15.19
RF	1.01	0.93	0.27	4.06	0.99	0.96	0.29	5.44	0.84	0.95	0.55	13.65
DBN	1.03	0.90	0.38	5.96	0.98	0.83	0.46	7.98	1.01	0.70	0.77	17.84

overall trends show that MAPE of the two methods is getting larger and R^2 is getting lower. The MAPE (R^2) of the WATARK method is smaller (larger) than the RF method when the number of sites is relative larger; for example, the number of sites is larger than about 7. When the number of sites is relative smaller, the MAPE (R^2) of the RF method is smaller (larger) than the WATARK method; for example, the number of sites is smaller than about 7. For the WATARK method, the upscaled results are calculated by adding the spatial trend of ET obtained by a regression equation consisting of auxiliary variables and the residual obtained by area-to-area kriging. However, the magnitude of the residuals is relatively small, and the absolute value is generally less than 0.3 mm/d (ET is generally 5 mm/d). Therefore, the accuracy of the WATARK method is mainly affected by the accuracy of the spatial trend of ET, which depends on the number and representativeness of sites; the RF method is not affected in this way. This difference is also the main reason that the upscaling effect of WATARK method outperforms the RF method in the middle stream matrix with more sites, whereas the RF method outperforms the WATARK method in the downstream matrix with relatively few sites.

To compare the accuracy of the three machine learning methods, the EC observation data of the middle stream and downstream matrices were used as the training samples. In a tenfold cross validation, the training sample was randomly partitioned into 10 subsamples, a single subsample was retained as the validation data for testing the model, and the remaining nine subsamples were used as training data. The cross-validation process was repeated 10 times, and the results were averaged as the validation results. Tenfold cross-validation was performed for the ANN, RF, and DBN methods in the middle and downstream matrices (Figure 7). Figure 7 shows that the predicted results of the ANN, RF, and DBN models in the middle and downstream matrices are in good agreement with the measured values. In the middle stream matrix, the R (RMSE) values were 0.86, 0.92, and 0.88 (0.72, 0.59, and 0.69 mm/d) for the ANN, RF, and DBN models, respectively; in the downstream matrix, the R (RMSE) values were 0.81, 0.84, and 0.73 (0.91, 0.84, and 0.99 mm/d), respectively. Among the three machine learning methods, in the middle stream and downstream matrices, the performance of RF method is best, consistent with the results in Figure 5 and

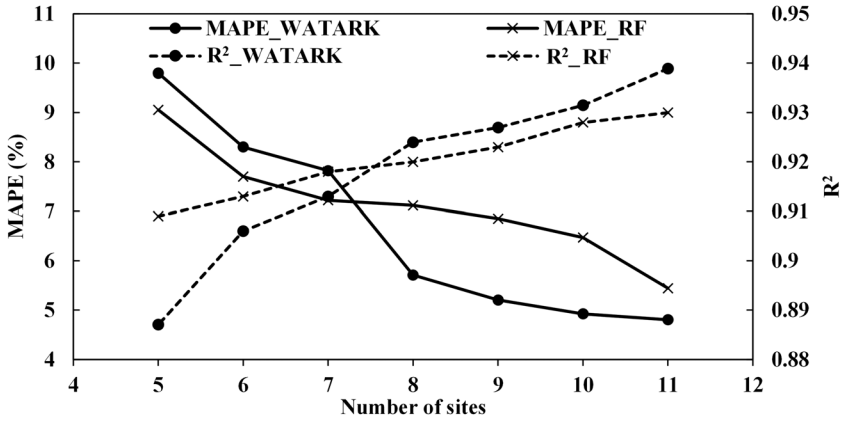

Figure 6. Comparison of the accuracy of the WATARK and RF methods as the number of sites changes.

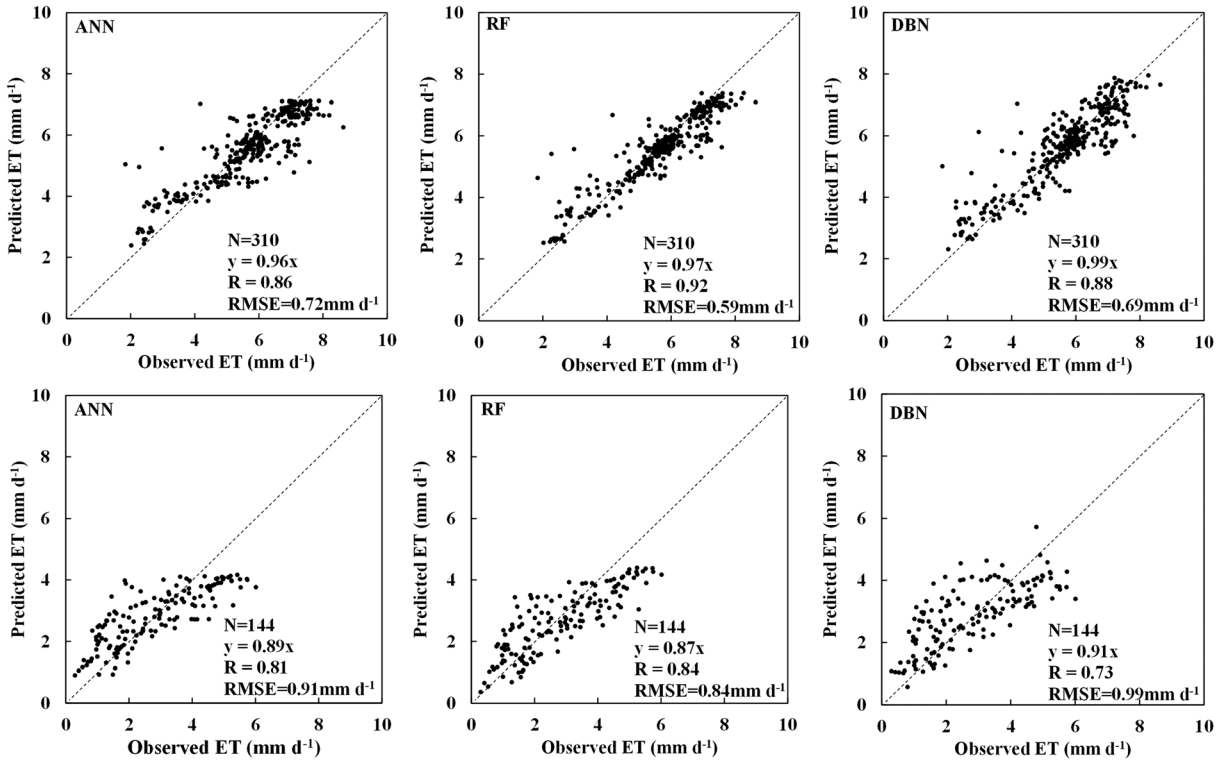

Figure 7. Tenfold cross-validation results in the middle stream matrix (upper) and downstream matrix (lower).

Table 3. This finding further shows that the RF method is superior to the other two machine learning methods in the middle stream and downstream matrices.

Relative error analysis was conducted to quantify the impacts of spatial heterogeneity, LST, and R_n on the accuracy of the upscaled results (Figure 8). The overall trend of the relative error of the upscaled results increases with increasing CV. When the CV is small ($<$0.22, i.e., homogeneous underlying surfaces), the average relative error of the upscaled results is small. The average relative error of the AW method is smaller than in other upscaling methods introducing auxiliary variables, consistent with previous analyses. With increasing CV (0.23–0.35, i.e., moderately heterogeneous underlying surfaces), the average relative error of the upscaled results is increased. The average relative error of the WATARK method is smaller than that of the other upscaling methods, which is also consistent with the previous analysis. As CV continues to increase ($>$0.35, i.e., highly heterogeneous underlying surfaces), the relative error of the upscaled results reaches its peak. The average relative error of the RF method is smaller than that of other upscaling methods, consistent with the previous analysis. In addition, when CV $>$ 0.35, the error of the upscaled results shows an obvious linear correlation with CV and increases with the retrieval error of auxiliary variables. When the retrieval accuracy of LST and R_n is low (i.e., the retrieval errors of LST and R_n are large), the relative error of the upscaled results is high. In contrast, with the increase in retrieval accuracy (i.e., the retrieval errors of LST and R_n decrease), the relative error of the upscaled results is correspondingly reduced. Therefore, for the upscaling methods introducing auxiliary variables, the accuracy of the upscaled results is affected not only by the spatial heterogeneity of the underlying surfaces but also by the retrieval accuracy of the auxiliary variables.

4.2.2. Cross Validation: Intercomparison of Six Upscaling Methods

To further validate various upscaling methods, a cross-validation method, the TCH (see Appendix C), was performed for intercomparison and quantification of uncertainties of the various upscaling methods (Figure 9). The magnitude of the relative uncertainty varies among different upscaling methods, which depends on the heterogeneity of the underlying surfaces. For homogeneous underlying surfaces in the middle stream matrix, the relative uncertainty of the AW method is slightly smaller. For moderately heterogeneous underlying surfaces in the middle stream matrix, the relative uncertainty of the WATARK method is slightly smaller than other methods, while the relative uncertainty of the DBN method is slightly larger. For highly

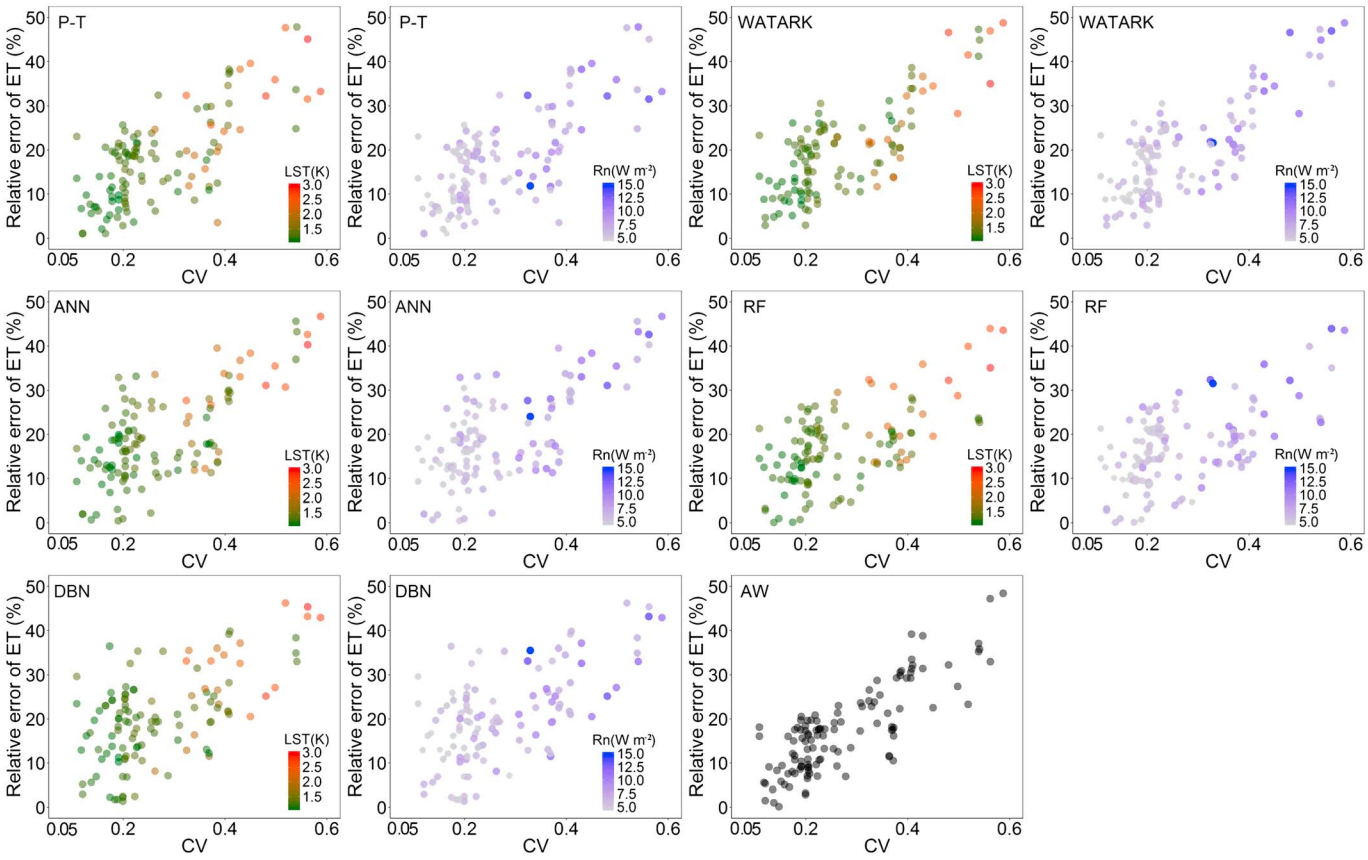

Figure 8. Impacts of CV, LST, and R_n on the relative error of ET from upscaling methods. Color bars represent the retrieval error values of LST (K) and R_n (W/m^2).

heterogeneous underlying surfaces in the downstream matrix, the relative uncertainty of the RF method is slightly smaller, while the WATARK method has slightly larger relative uncertainty. The results of six upscaling methods by the cross-validation method (TCH) are basically consistent with the direct validation results taking the LAS observation as a reference (Table 3). On the whole, the sequence of relative uncertainty from high to low over different underlying surfaces is homogeneous underlying surfaces, followed by moderately heterogeneous underlying surfaces and highly heterogeneous underlying surfaces. For most upscaling methods, if the error from direct validation is larger, the relative uncertainty from cross validation is larger.

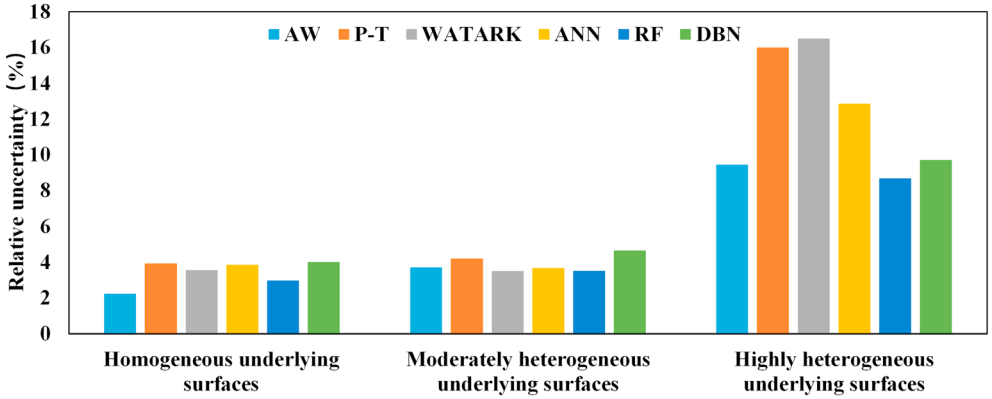

Figure 9. Intercomparison of six upscaling methods based on the TCH cross validation method.

4.2.3. Spatial Distribution Pattern of Upscaled ET

According to the differences in vegetation phenology, weather conditions, and soil moisture, 24 June (before the crop is at full cover), 2 August (the crop is at full cover), and 3 September 2012 (at the end of the crop growing period) in the middle stream matrix (3 × 3 km) and 24 July and 9 September 2014 and 16 July and 23 August 2015 in the downstream matrix (2 × 2 km) were selected to compare the spatial distribution patterns of upscaled results among the five upscaling methods.

As shown in Figure 10, the spatial distribution patterns of the upscaled ET from five upscaling methods in the middle stream matrix (Figure 10a) and the downstream matrix (Figure 10b) are similar. For the middle stream matrix, higher values were found in the maize underlying surface, while lower values were observed in the village underlying surface. For the time points of before the crop at full cover (24 June), crop at full cover (2 August), and the end of the crop growing period (3 September), daily ET (the average of the five upscaling methods) over the maize underlying surface showed a tendency to first increase and then decrease, with values of approximately 5.7, 6.7, and 4.6 mm/d, respectively. Daily ET over the village underlying surface during the same period was approximately 2.6, 3.3, and 1.8 mm/d, respectively. For the downstream matrix, higher values were found in the vegetation cover underlying surface (e.g., *Populus euphratica* and *Tamarix*) with a daily ET of approximately 4.8 mm/d, while lower values were observed in the bare land, with a daily ET of approximately 2.4 mm/d.

Taking the WATARK method with the higher accuracy in the middle stream matrix and the RF method with the higher accuracy in the downstream matrix as references, the Pearson correlation coefficients (see Appendix B) and the difference in upscaled ET between WATARK or RF methods and the other four methods were calculated (Figure 11). In the middle stream matrix, the smallest Pearson correlation coefficient and the greatest difference of spatial distribution pattern between the upscaled ET of WATARK and other methods were found in village and road areas. However, for the downstream matrix, the smallest Pearson correlation coefficient and the greatest difference in spatial distribution pattern between the upscaled ET of RF and other methods were found in areas of bare land. For the P-T method, the correlation coefficient between the α and T_s-T_a in site 4 or Bare land station, which represent the village and bare land underlying surfaces, was smaller than that of the other sites. This condition results in a huge difference between the upscaled result of the P-T method and other methods in these two underlying surfaces. In addition, when training ANN, RF, and DBN models, the number of training samples in the village or bare land underlying surfaces was much smaller than that in the vegetation cover underlying surface, which may also be a reason for the large difference in the village and bare land underlying surfaces.

4.2.4. Upscaled ET Over Typical Underlying Surfaces

According to the land use classification in the study area, the ET of the three main underlying surfaces of maize, orchard, and village was extracted in the middle stream. The five main underlying surfaces of *Tamarix*, *Populus euphratica*, sparse *Populus euphratica* and *Tamarix*, bare land, and farmland were extracted in the downstream. EC measurements were used as a reference for comparison to the upscaled ET of the different upscaling methods (Figure 12).

In Figure 12, the difference in the upscaled ET over the vegetation underlying surface according to the upscaling methods in the middle stream matrix and downstream matrix is smaller, similar to EC measurements. However, over the village underlying surface in the middle stream matrix and bare land underlying surface in the downstream matrix, the upscaled ET using the P-T upscaling method was smaller than those of the other four methods because the relationship between α and T_s-T_a in site 4 and Bare land site, which represent the village and bare land underlying surfaces, is poorer than that of the other sites.

4.3. Acquisition of Daily Ground Truth ET Data at the Satellite Pixel Scale

According to the above analysis, the upscaled accuracy of the AW method is mainly affected by the heterogeneity of the underlying surfaces. Because the representativeness of EC measurements decreases with the increase in the heterogeneity, the accuracy of the AW method decreases with the increase in the heterogeneity of the underlying surfaces. The upscaled accuracy of the other five upscaling methods used high-resolution remote sensing data is affected not only by the heterogeneity of the underlying surfaces but also by the retrieval accuracy of the auxiliary variables. In addition, for the WATARK method, the upscaled accuracy is also significantly influenced by the number of EC sites. Three machine learning methods use the measurement data as the training sample to train the model at sites and then apply the trained model to achieve

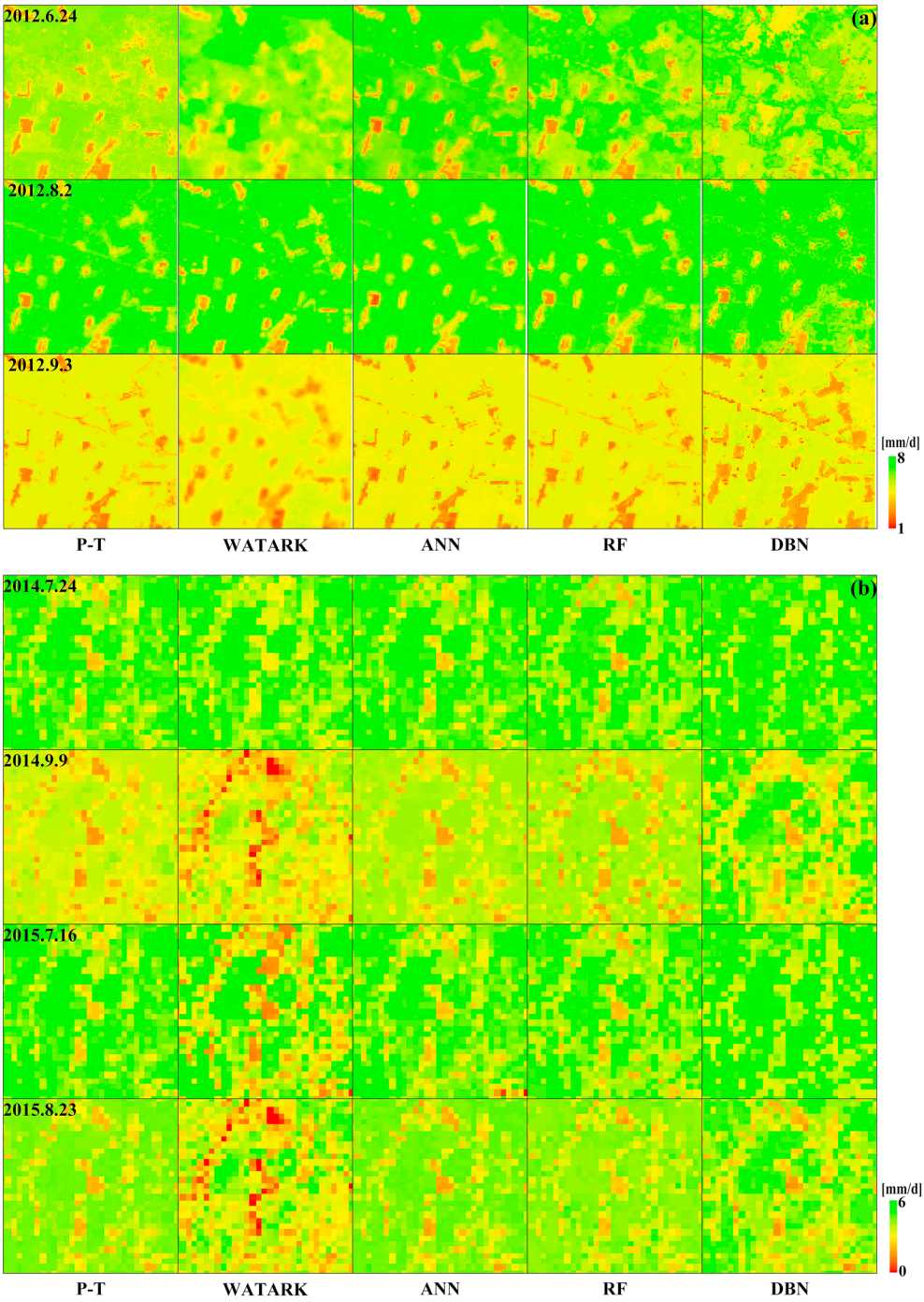

Figure 10. Comparison of spatial distribution patterns of upscaled ET from the five upscaling methods (from left to right: P-T, WATARK, ANN, RF, and DBN methods) in (a) the middle stream and (b) downstream matrices.

the upscaled ET at the satellite pixel scale. Therefore, the accuracy of the upscaled ET is also affected by the amount of the training samples and the representativeness of the sites. In addition, the upscaled results of the three machine learning methods are different. The accuracy of the RF method is highest because the ANN belongs to the machine learning algorithm in the same way, the RF is an improved regression tree algorithm, and it integrates the characteristics of bagging and random selects feature to split, and also have advantages of antinoise and hard to fall into overfitting (Breiman, 2001; Fang et al., 2011). The DBN is essentially a multilayer neural network learning algorithm, and its performance is better for complex

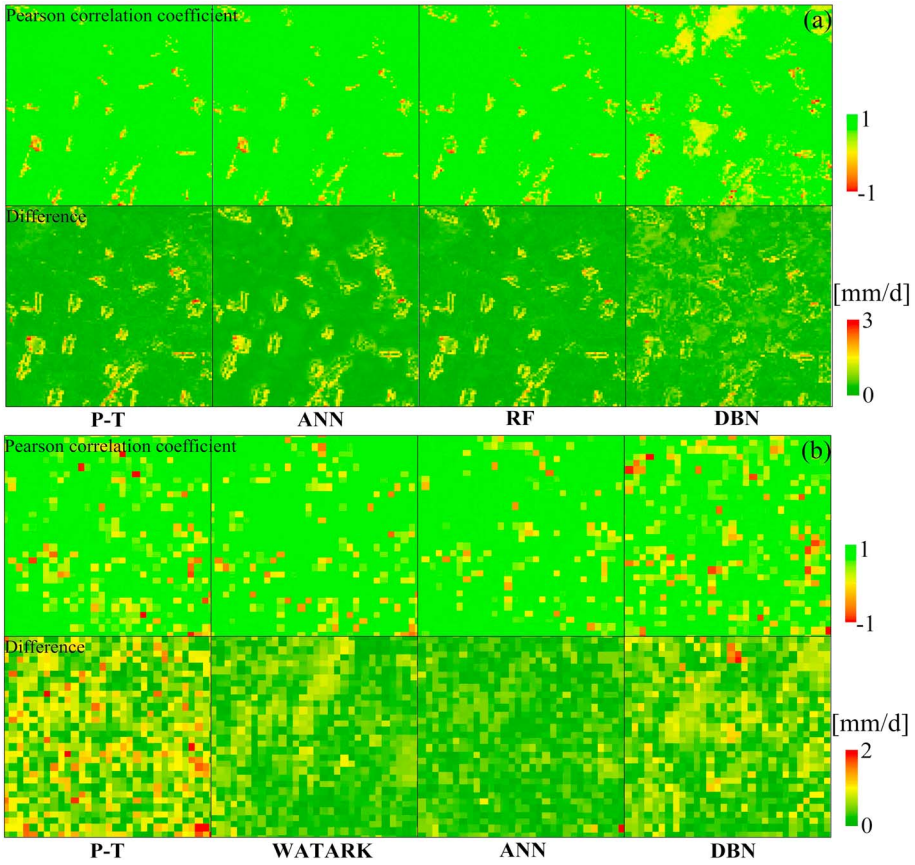

Figure 11. Comparison of the spatial correlation coefficient and difference in the upscaled ET in (a) the middle stream matrix (from left to right: P-T, ANN, RF, and DBN methods) and (b) downstream matrix (from left to right: P-T, WATARK, ANN, and DBN methods) during the growing period.

functions. It is suitable for modeling based on mass data samples. With sparse data, it is impossible to estimate the rule of the data without bias. Thus, the accuracy of the upscaled ET may not be as good as the results generated using simpler algorithms (Hinton et al., 2006).

Based on the analysis in section 4.2 and the study by Liu et al. (2016), the daily ground truth ET data at the satellite pixel scale are acquired by using the AW method when the underlying surface is homogeneous during the crop at full cover in the middle stream matrix. When the underlying surface is moderately

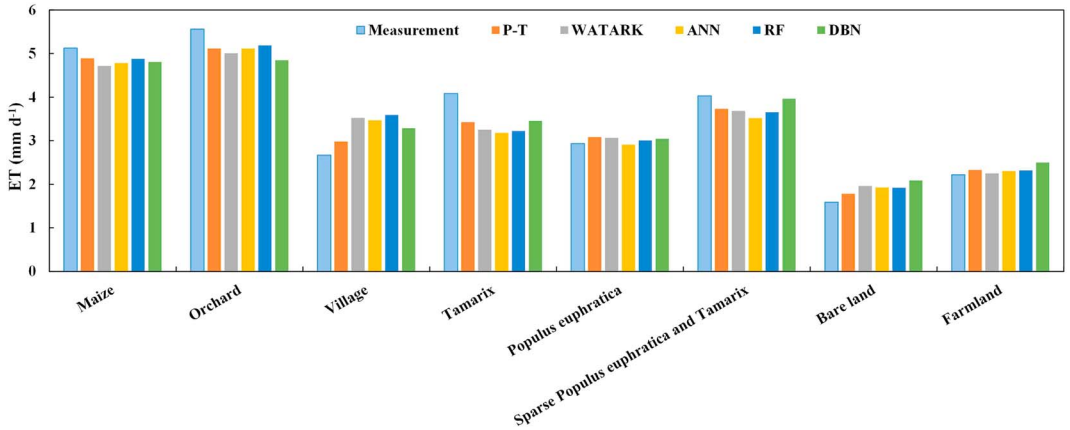

Figure 12. Comparison of upscaled ET over the typical underlying surfaces from upscaling methods.

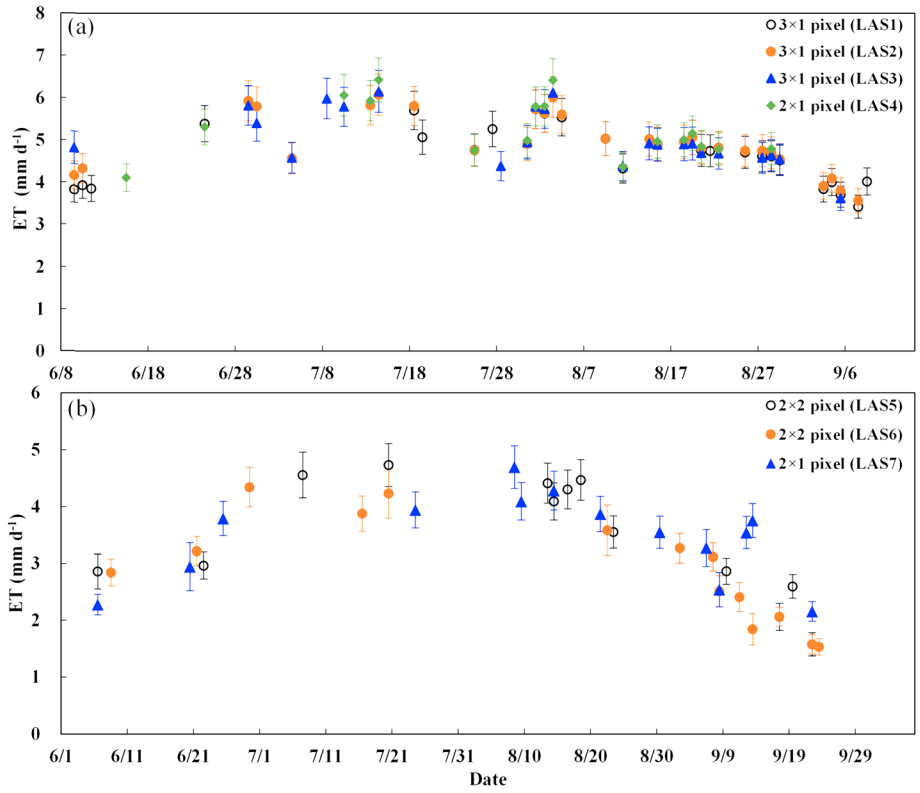

Figure 13. Daily ground truth ET data at the satellite pixel scale in (a) the middle stream matrix in 2012 and (b) downstream matrix in 2014 and 2015. The error bar indicates the uncertainty in the ground truth ET at the satellite pixel scale.

heterogeneous, namely, before the crop at full cover, at the end of the crop growing period and irrigation time during the crop at full cover, the daily ground truth ET data at the satellite pixel scale are acquired by the WATARK method in the middle stream matrix. When the underlying surface is highly heterogeneous, the daily ground truth ET data at the satellite pixel scale are acquired by the RF method in the downstream matrix. As shown in Figure 13, in the middle stream matrix (Figure 13a), the MAPE values of the daily ground truth ET data in the LAS1, LAS2, LAS3, and LAS4 pixels were 3.5%, 3.7%, 4.8%, and 3.9%, respectively. The accuracy of the upscaled ET was very high in the LAS1, LAS2, and LAS4 pixels and was relatively poor in the LAS3 pixels compared with LAS observations. In the downstream matrix (Figure 13b), the MAPE values of the daily ground truth ET data in the LAS5, LAS6, and LAS7 pixels were 11.7%, 15.1%, and 14.2%, respectively, compared with LAS observations. From the above analysis, the accuracy of daily ground truth ET data at the satellite pixel scale was good. The ground truth ET data can therefore meet the requirements for validating RS_ET products.

The errors in daily ground truth ET data at the satellite pixel scale were primarily derived from the EC and LAS measurements and from the upscaling process (Liu et al., 2016). The errors in LAS observation mainly include the error of LAS observation of sensible heat flux, the error associated with latent heat flux when using the residual method and spatial representation in LAS measurements (i.e., the proportion of daytime-averaged LAS source area in the corresponding MODIS pixels will change with the different atmospheric conditions, such as wind speed/direction and atmospheric stability). In addition, the accuracy of the upscaling methods itself also differs with the choice of algorithm.

The ground truth ET data at the satellite pixel scale were selected to validate the DTD and ETMonitor products (2012, 2014, and 2015; Figure 14). As shown in Figure 14a, the daily ET obtained by DTD in the middle stream and downstream matrices is distributed in the vicinity of the 1:1 line, and the RMSE and MAPE are 0.94 mm/d and 21.6%, respectively. As shown in Figure 14b, the accuracy of the ET obtained by ETMonitor in the downstream matrix is not as good as that in the middle stream matrix, with an RMSE and MAPE of 1.25 mm/d and 26.5%, respectively. The uncertainties of DTD and ETMonitor in the validation process are 8.91% and 8.95%,

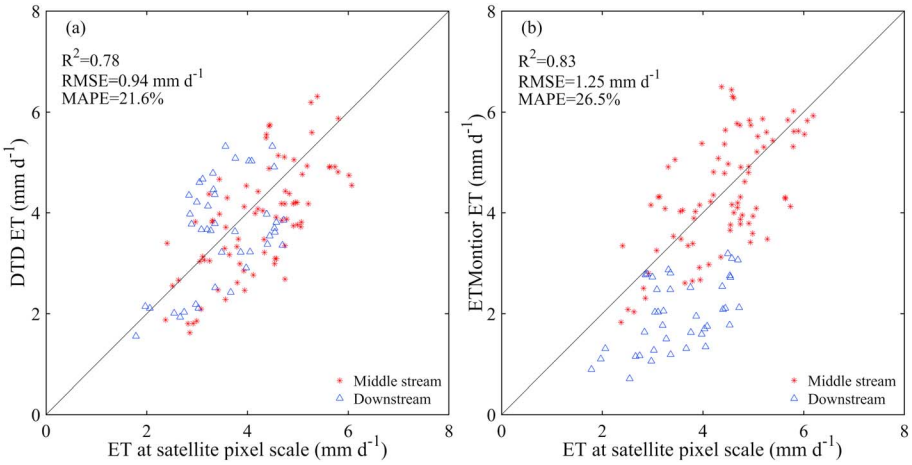

Figure 14. Validation of daily RS_ET from (a) DTD and (b) ETMonitor with ground truth ET data at the satellite pixel scale.

respectively. The acquisition of ground truth data at the satellite pixel scale not only solves the problem of the spatial scale mismatch between remotely sensed estimates and the in situ observations but also quantifies uncertainties in the validation process.

5. Conclusions

Based on multisite measurements in the middle stream during the HiWATER-MUSOEXE, and the Hydrometeorological Observation Network in the downstream of the HRB, six upscaling methods were intercompared and a combined method was developed to acquire ground truth ET data at the satellite pixel scale.

For direct validation, over homogeneous underlying surfaces in the middle stream matrix, the results of six upscaling methods had good accuracy, and the AW method was slightly superior to the other five upscaling methods introducing auxiliary variables, with RMSE and MAPE values of 0.19 mm/d and 3.15%, respectively. Over moderately heterogeneous underlying surfaces in the middle stream matrix, when the retrieval accuracy of the auxiliary variables was high and the number of observation sites was high, the WATARK method performed slightly better, with RMSE and MAPE values of 0.29 mm/d and 4.80%, respectively, followed by the RF method. In addition to the DBN method, the methods introducing auxiliary variables were slightly superior to the AW method, and the RF methods have the relatively better performance in three machine learning methods. Over highly heterogeneous underlying surfaces in the downstream matrix, when the retrieval accuracy of LST and R_n was high, the RF method performed slightly better, with RMSE and MAPE values of 0.55 mm/d and 13.65%, respectively, and the other four upscaling methods introducing auxiliary variables performed relatively poor. For cross validation, the relative uncertainties of the upscaling methods over homogeneous underlying surfaces were slightly smaller than that over moderately heterogeneous underlying surfaces, and the relative uncertainties of upscaling methods over highly heterogeneous underlying surfaces were largest. For homogeneous underlying surfaces, the relative uncertainty of the AW method was slightly smaller than other methods, and for moderately heterogeneous underlying surfaces, the relative uncertainty of the WATARK method was slightly smaller. The relative uncertainty of the RF method was slightly smaller than other methods over highly heterogeneous underlying surfaces. The spatial distribution patterns of the upscaled ET from the upscaling methods introducing auxiliary variables were similar. The difference in upscaled ET obtained by various methods was small over the vegetation underlying surfaces, while the differences in the village and the bare land underlying surfaces were large. For the AW method, the degree of heterogeneity of the underlying surfaces was the main affecting factor. For the other five upscaling methods introducing auxiliary variables, in addition to the heterogeneity of the underlying surfaces, the retrieval accuracy of LST and R_n was a predominant influencing factor. In addition, the WATARK method requires more observation sites. Machine learning upscaling methods (i.e., ANN, RF, and DBN) must consider the amount and representativeness of samples for training models.

Thus, a combined method (using the AW and WATARK methods for homogeneous and moderately heterogeneous underlying surfaces, respectively, and using the RF method for highly heterogeneous underlying surfaces) was used to acquire daily ground truth ET data at the satellite pixel scale. Taking LAS measurements as the satellite pixel reference, the error of the ground truth ET data was evaluated. The MAPE values of the daily ground truth ET data for the LAS1, LAS2, LAS3, and LAS4 pixels in the middle stream matrix were 3.5%, 3.7%, 4.8%, and 3.9%, respectively. The MAPE values of the daily ground truth ET data for the LAS5, LAS6, and LAS7 in the downstream matrix were 11.7%, 15.1%, and 14.2%, respectively. The DTD and ETMonitor were validated by the daily ground truth ET data, with RMSE and MAPE values of 0.94 mm/d and 21.6% and 1.25 mm/d and 26.5%, respectively. The uncertainties of the DTD and ETMonitor in the validation process were 8.91% and 8.95%, respectively.

In this study, LAS measurements were taken as the "reference values" at the satellite pixel scale to evaluate the upscaled results. However, the LE of the LAS was calculated using the energy balance residual method or the relationship between the LAS-LE and area-averaged R_n. Therefore, in future studies, the uncertainty (e.g., LAS observation error, the error of the energy balance residual method, and the error from the LAS spatial representativeness) derived from the LAS measurements as the reference values at the satellite pixel scale and the influence of advection and secondary circulation will require further analysis. In addition, the upscaled ET at the satellite pixel scale and the direct ET observations derived from the combination of LAS and microwave scintillometer measurements should be compared in future work.

Appendix A

Appendix A shows the details of six upscaling methods including AW, P-T, WATARK, ANN, RF, and DBN used in this study.

A1. The Area-Weighted (AW) Method

First, the study area was divided into different plots according to the heterogeneity of the underlying surfaces and the land use/land cover. The fraction of each plot (weighted value) was calculated after combining the land use/land cover maps. Finally, ET at the satellite pixel scale was computed using the following equation:

$$ET_{AW} = \sum_{i=1}^{N} W_i \theta_i \tag{A1}$$

where ET_{AW} is the ET calculated by the AW method, N is the number of plots, W_i is the area proportion of the ith plot relative to the satellite pixel and θ_i is the EC measurement in the ith plot.

The weighted values of the EC measurements in different MODIS pixels in the middle stream and downstream matrices are listed in Table A1.

A2. The Integrated Priestley-Taylor Equation (P-T) Method

The Priestley-Taylor equation (Priestley & Taylor, 1972) can be used to calculate the potential ET from wet surfaces under conditions of minimal advection as follows:

$$ET_P = \alpha \frac{\Delta}{\Delta + \gamma} (R_n - G_0) \tag{A2}$$

where ET_P is the ET for saturated surfaces; α is the Priestley-Taylor parameter, usually taken as 1.26; R_n and G_0 are the net radiation and surface soil heat flux, respectively; Δ is the slope of the saturation vapor pressure to air temperature; and γ is the psychrometer constant.

It has been demonstrated that relationships between the Priestley-Taylor parameter α and soil moisture, land surface temperature, and vegetation parameters can be used in the Priestley-Taylor equation. Based on such relationships, the Priestley-Taylor equation can estimate the actual ET over unsaturated surfaces (Davies & Allen, 1973; Fisher et al., 2008; Jiang & Islam, 2001).

In this study, first, the Priestley-Taylor parameter α for each EC site was calculated using the Priestley-Taylor equation by combining site observations and the relationship between α and the difference of the land surface temperature and air temperature (T_s-T_a) was established in each plot. Second, ET in each plot was

Table A1
Weighted Value of Each EC Measurement in Different MODIS Pixels in Middle Stream and Downstream Matrices

	3 × 1 MODIS pixels (LAS1)	3 × 1 MODIS pixels (LAS2)	3 × 1 MODIS pixels (LAS3)	2 × 1 MODIS pixels (LAS4)	2 × 2 MODIS pixels (LAS5)	2 × 2 MODIS pixels (LAS6)	2 × 1 MODIS pixels (LAS7)
Related EC	Area weighted%	Area weighted%	Area weighted%	Area weighted%	Area weighted%	Area weighted%	Area weighted%
EC4 (village)	15	15	13	7			
EC5 (maize)	20	-	-				
EC6 (maize)	13	17	-				
EC7 (maize)	-	18	12				
EC8 (maize)	-	9	36				
EC11 (maize)	19	-	-				
EC12 (maize)	-	14	10	10			
EC13 (maize)	-	-	9	21			
EC14 (maize)	22	20	-				
EC15 (maize)	-	-	11	50			
EC17 (orchard)	-	-	-	6			
TDP-shelterbelt	11	7	9	6			
EC-P (sparse *Populus euphratica*)					-	35	-
EC-M (sparse *Populus euphratica* and *tamarix*)					37	-	32
EC-S (*Tamarix*)					29	31	32
EC-F (melon)					9	7	9
EC-B (bare land)					14	13	14
TDP-*Populus euphratica*					11	14	13

Note. P, M, S, F, and B represent the *Populus euphratica* station, Mixed forest station, Sidaoqiao superstation, Farmland station, and Bare land station, respectively.

estimated using the Priestley-Taylor equation by combining the T_s-T_a retrieved from remote sensing images and atmospheric forcing data. Finally, ET at the satellite pixel scale was acquired using the equation (A3).

$$ET_{P\_T} = \sum_{i=1}^{N} W_i ET_i + W_s ET_s \quad (A3)$$

where ET_{P_T} is the ET in the three 3 × 1, two 2 × 2, or two 2 × 1 MODIS pixels; N is the number of plots or land cover types in the pixel; ET_i is the ET in the ith plot or land cover type calculated by the Priestley-Taylor equation; ET_s is the ET measured by TDP in the shelterbelt in the middle stream matrix or in the *Populus euphratica* in the downstream matrix; W_i is the fraction of the ith plot or land cover type relative to the corresponding MODIS pixels; and W_s is the fraction of the shelterbelt or *Populus euphratica* relative to the corresponding MODIS pixels.

According to Liu et al. (2016), the relationship between α and T_s-T_a was established in this study. Figure A1 shows the variation in α and T_s-T_a during the growing period (using the Daman superstation maize in 2012 and Sidaoqiao superstation *Tamarix* in 2014 as examples of the middle stream and downstream, respectively). As shown in Figure A1, in the middle stream, there was a clear negative correlation between α and T_s-T_a before the crop at full cover (before 29 June) and at the end of the growing period (after 30 August). In contrast, α was approximately equal to 1.45 (blue dotted line, T_s-T_a = 0 K) during the middle of the crop growing period. For the downstream, there was a clear negative correlation between α and T_s-T_a before 29 June and after 4 September in 2014. The α was approximately equal to 0.51 in 2014 (blue dotted line, T_s-T_a = 8 K) during the middle of the vegetation growing period. The LSTs retrieved from remote sensing images were used in the following upscaling process. Therefore, the relationship between α and T_s-T_a was estimated for the satellite passing times (11:30–12:00 BST and 12:00–12:30 BST). The relationship between α and T_s-T_a was established using a linear equation in each plot or land cover type before the crop at full cover and at the end of the growing period (Table A2 and Table A3). Good correlation coefficients were achieved, in general.

A3. The Weighted Area-to-Area Regression Kriging (WATARK) Method

Based on the correlation with ET and the availability of the remote sensing data for implementation of the WATARK method, we selected R_n, LST, and NDVI as auxiliary variables (the same as these machine learning

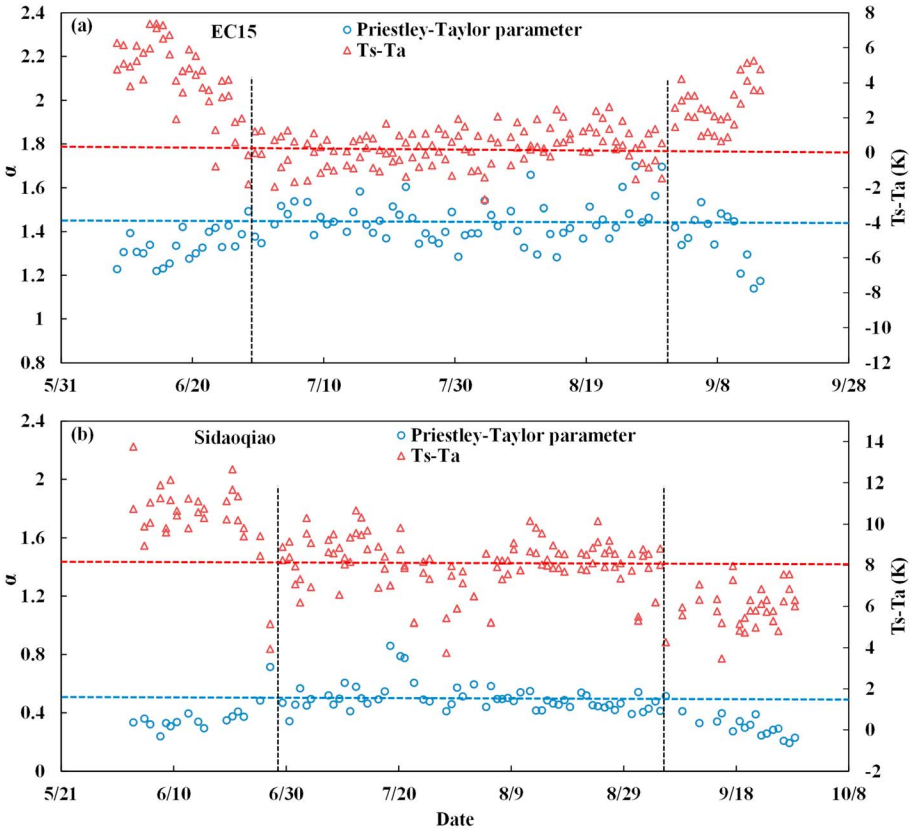

Figure A1. Variation in the Priestley-Taylor parameter α and the difference between the surface and air temperature (T_s-T_a) in (a) the middle stream and (b) downstream in 2014.The black dotted lines indicate 29 June and 30 August at EC15 in 2012 (29 June and 4 September at Sidaoqiao in 2014), the blue dotted line indicates α = 1.45 at EC15 (0.51 at Sidaoqiao), and the red dotted line indicates T_s-T_a = 0 K at EC15 (8 K at Sidaoqiao).

methods in section 4.2.1). The processing steps were as follows: First, the selected auxiliary variables were extracted with footprint models, that is, the weighted average. The ET observed by EC was used as the dependent variable, and the auxiliary variables extracted by the footprint model were used as independent variables to establish a regression model for the ET spatial trend and residual as

$$\overline{\mathrm{ET}}_i = \beta_0 + \beta_1 \cdot \overline{Rn}_i + \beta_2 \cdot \overline{\mathrm{LST}}_i + \beta_3 \cdot \overline{\mathrm{NDVI}}_i + \bar{r}_i \quad \text{(A4)}$$

where $\overline{\mathrm{ET}}_i$ is the ET of the ith EC; $\overline{Rn}_i$, $\overline{\mathrm{LST}}_i$, and $\overline{\mathrm{NDVI}}_i$ are remotely sensed auxiliary variables extracted by the ith EC footprint; β_0, …, β_3 are regression coefficients; and $\bar{r}_i$ is the residual of the ith EC.

Second, the regression coefficients were obtained by the stepwise regression method, and then the spatial trend and residual of the ET at the footprint scale were achieved by the regression model. Third, the residual at the satellite pixel was calculated using the weighted area-to-area kriging equations. Finally, the ET was estimated by adding the spatial trend (regression estimations) to the weighted area-to-area kriging results at the satellite pixels corresponding to the LAS.

Table A2
Linear Relationship (y = ax + b) Between the Priestley-Taylor Parameter α (y Axis) and the Difference of the Land Surface and Air Temperatures, T_s-T_a (x Axis) in the Middle Stream

	8 to 28 June 2012				31 August to 14 September 2012			
Site	*a*	*b*	*R*	*n*	*a*	*b*	*R*	*n*
EC4	−0.07	1.86	0.56	35	−0.10	1.56	0.58	30
EC5	−0.03	1.46	0.82	39	−0.09	1.59	0.72	30
EC6	−0.02	1.39	0.92	42	−0.09	1.57	0.73	30
EC7	−0.02	1.09	0.69	42	−0.03	1.34	0.71	30
EC8	−0.02	1.31	0.60	41	−0.09	1.49	0.77	30
EC11	−0.02	1.40	0.91	42	−0.12	1.54	0.75	30
EC12	−0.02	1.38	0.69	41	−0.08	1.55	0.73	30
EC13	−0.01	1.45	0.57	40	−0.07	1.76	0.77	30
EC14	−0.03	1.37	0.84	41	−0.08	1.50	0.78	30
EC15	−0.02	1.44	0.74	42	−0.06	1.54	0.71	28
EC17	−0.06	1.54	0.82	42	−0.09	1.64	0.72	30

A4. Artificial Neural Network (ANN) Method

Among numerous ANN algorithms, the multilayer networks using the error back propagation algorithm (Haykin, 1998) are the most popular. Sufficient training data are input into the input layer, which consists of auxiliary variables (i.e., R_n, VPD, LST, NDVI, and FVC), to train the network. The

Table A3
Linear Relationship (y = ax + b) Between the Priestley-Taylor Parameter α (y Axis) and the Difference of the Land Surface and Air Temperatures, T_s-T_a (x Axis) in the Downstream

	1 to 28 June 2014				5 to 30 September 2014			
Site	*a*	*b*	*R*	*n*	*a*	*b*	*R*	*n*
Populus euphratica	−0.02	0.63	0.77	46	−0.03	0.55	0.65	45
Mixed forest	−0.03	0.67	0.80	36	−0.05	0.57	0.67	44
Sidaoqiao	−0.05	0.85	0.82	32	−0.05	0.63	0.66	33
Farmland	−0.05	0.67	0.75	30	−0.03	0.40	0.66	48
Bare land	−0.01	0.46	0.73	44	−0.02	0.37	0.64	45

weights, which represent the linkages between the variables and ET, are obtained in the hidden layer, and the final processing results are output at the output layer. There are two training phases for an ANN. In the first phase (also called the forward pass), the input variables are propagated from the input to the output layer. The differences between the predicted and measured ET are calculated by propagating selected variables from the input layer to the output layer. In the second phase (also called the reverse pass), the errors are propagated backward from the output layer to the hidden layer to adjust the weights. Back propagation uses a steepest descent method on the error surfaces to adjust the weights to minimize the difference between the predicted and expected output. To prevent an ANN model from being trapped in a local minimum error during training, a momentum term (Rumelhart et al., 1986) to the weight change was added in obtaining ET of the ANN.

A5. Random Forest (RF) Method

The RF regression algorithm is an ensemble-learning algorithm that combines a large set of regression trees, which was used in this study. A regression tree represents a set of restrictions or conditions that are hierarchically organized and successively applied from a root to a terminal node or leaf in a tree (Breiman et al., 1984; Quinlan, 1993). At each node in the tree, a random subset of predictor variables is used to identify the most efficient split, which is defined by identifying the predictor variable and the split point that results in the largest reduction in the residual sum of squares between the sample observations and the node mean. All trees are grown to the maximum extent, as controlled by the node size set by the user. The result is an ensemble of low-bias, high-variance regression trees, where the final predictions are derived by averaging the predictions of the individual trees (Breiman, 2001). In this study, three parameters were determined in RF: the parameter *ntree* (i.e., the number of regression trees grown based on a bootstrap sample of the observations) was set to 1,000; for the parameter *mtry* (i.e., the number of different predictors tested at each node), the default value of the square root of the total number of the variables was used; and the parameter *nodesize* (i.e., the minimal size of the terminal nodes of the trees) was set to the default value (one).

A6. Deep Belief Network (DBN) Method

First, the DBN is decomposed into a series of RBMs composed of two adjacent layers. As a special type of Markov random field, RBM is a bipartite undirected graphical model that consists of a layer of visible units (random variables) $v \in \{0, 1\}^V$ (v and V are the visible units and visible layer, respectively) and a layer of binary hidden units (random variables) $h \in \{0, 1\}^H$ (h and H are the hidden units and hidden layer, respectively). The visible units represent the observed data, and the hidden units learn to capture the higher-order features based on the input. The visible and hidden layers are connected by a symmetrical weight matrix $W \in M^{V \times H}$ (W and M are the weight and matrix, respectively), but there are no connections within a layer. Second, training parameters are used layer by layer to initialize the DBN. Feature activations can be implemented through RBMs and hence adopted to effectively learn within the DBN architecture. The hidden layer will be visible to the next RBM if the previous RBM is trained. Finally, the back-propagation algorithm is used to better tune and optimize the entire DBN using the gradient descent method (X. D. Song et al., 2016).

Appendix B

Appendix B shows the assessment methods used in this study. The coefficient of variation (CV) is defined as the ratio of the standard deviation and the average. The CV was calculated as

$$CV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\overline{x}} \tag{B1}$$

where x_i, $\overline{x}$, and n are the ith sample value, sample average value, and number of samples, respectively. A greater CV is clearly associated with greater spatial heterogeneity of the physical quantity. In section 4.1, CV is used to evaluate the spatialheterogeneity of ET before comparison.

Taking LAS measurements as the satellite pixel reference, the indices of mean absolute percentage error (MAPE), relative error (RE), root-mean-square error (RMSE), and Pearson correlation coefficient (R) were selected to assess the accuracy and spatial distribution patterns of the various upscaled ET results, and the respective equations are as follows:

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{|P_i - O_i|}{\overline{O}} \tag{B2}$$

$$\mathrm{RE} = \frac{P_i - O_i}{\overline{O}} \tag{B3}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(P_i - O_i)^2/n} \tag{B4}$$

$$R = \sum_{i=1}^{n}\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)/\left[\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2\right]^{1/2} \tag{B5}$$

where P_i is the estimated value, O_i is the measured value, $\overline{P}$ is the mean estimated value, $\overline{O}$ is the mean measured value, and n is the number of samples. In sections 4.2.1 and 4.3, P indicates the upscaling values, and O indicates the LAS measurements. In section 4.2.3, P_i indicates the ith pixel value of the compared upscaled ET, and O_i indicates the ith pixel value of the referenced upscaled ET. $\overline{P}$ indicates the mean value of the relative pixel from the compared upscaled ET, and O indicates the mean value of the relative pixel from the referenced upscaled ET.

Appendix C

The TCH can be used to estimate the relative uncertainties among various data sets under the condition of no measurements when at least three different series of data (e.g., the upscaled ET derived from several upscaling methods in this study) are available. In section 4.2.2, we used the TCH method for cross validation of six upscaling methods without requiring that the data sources be entirely independent. The theory of the TCH is described below.

The observational errors in the TCH method are commonly assumed to be normally distributed. Suppose that N is the number of different time series, denoted by $\{X_i\}_{i=1,2,\ldots,N}$, and ? corresponds to different time series. Thus, each time series can be expressed as

$$X_i = X_t + \varepsilon_i, \quad \forall_i = 1, 2, \ldots, N \tag{C1}$$

where X_t is the true signal and ε_i is the error of the ith time series. The symbol $\forall$ is called the universal quantifier. Since no true estimate of X_t is available, any single time series is chosen arbitrarily as the reference. The series of differences matrix can be obtained by calculating the difference between each time series and the reference. The corresponding covariance matrix S of the series of difference matrices can thus be obtained. Notably, the valuation of uncertainties is not related to the selection of reference series.

The unknown $N \times N$covariance matrix of the individual noise R is introduced, and R is related to S as follows:

$$\mathrm{S} = J{\cdot}R{\cdot}J^T \tag{C2}$$

where J is described as follows:

$$J_{N-1,N} = \begin{bmatrix} 1 & 0 & \ldots & 0 & -1 \\ 0 & 1 & \ldots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 \end{bmatrix}$$

The equation (C2) cannot be solved because the number of equations is less than the number of unknowns (when N is greater than 3). Therefore, the remaining free parameters require a reasonable method to obtain the unique solution. Galindo and Palacio (1999) proposed a constraint function to meet $|R| > 0$ as follows:

$$H_{(r_{1N},\ldots,r_{NN})} = -\frac{|R|}{|S|\cdot K} < 0 \tag{C3}$$

where $r_{1N}, \ldots, r_{NN}$ are the elements of corresponding R and $K = \sqrt[N-1]{|S|}$ is introduced to better obtain a numerical solution.

This constraint function constrains the free parameters within the solution domain but is not sufficient to determine a unique solution for the free parameters. The following objective function is also used to give the optimal pick criterion to obtain the unique parameter solution as follows:

$$F_{(r_{1N},\ldots,r_{NN})} = \frac{1}{K^2}\cdot\sum_{i<j}^{N} r_{ij}^2 \tag{C4}$$

The R is calculated using the objective function under the constraint condition by combining with the equation (C2). The uncertainty of the time series $\{X_i\}_{i = 1, 2, \ldots, N}$, called $\{\sigma_i\}_{i = 1, 2, \ldots, N}$, can be obtained by calculating the square root of the diagonal values in R (i.e., $\{r_{ii}\}_{i = 1, 2, \ldots, N}$). The relative uncertainty is the ratio of σ_i to the mean of X_i.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (41531174) and National Basic Research Program of China (2015CB953702). We would like to thank all the scientists, engineers, and students who participated in HiWATER field campaigns. We appreciate the anonymous reviewers for their constructive comments. All data in this study were provided by the data center of the "Integrated research on the eco-hydrological process of the Heihe River Basin" (http://www.heihedata.org). ET products from DTD and ETMonitor were provided by Prof. Lisheng Song and Li Jia, respectively.

References

Anderson, M. C., Allen, R. G., Morse, A., & Kustas, W. P. (2012). Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote Sensing of Environment*, *122*, 50–65. https://doi.org/10.1016/j.rse.2011.08.025

Bai, J., Jia, L., Liu, S. M., Xu, Z. W., Hu, G. H., Zhu, M. J., & Song, L. S. (2015). Characterizing the footprint of eddy covariance system and large aperture scintillometer measurements to validate satellite-based surface fluxes. *IEEE Geoscience and Remote Sensing Letters*, *12*(5), 943–947. https://doi.org/10.1109/LGRS.2014.2368580

Bernardo, J. M., & Smith, A. M. (2001). Bayesian theory. *Measurement Science and Technology*, *12*(2), 221–222.

Beyrich, F., Leps, J.-P., Mauder, M., Bange, J., Foken, T., Huneke, S., et al. (2006). Area-averaged surface fluxes over the Litfass region based on eddy-covariance measurements. *Boundary-Layer Meteorology*, *121*(1), 33–65. https://doi.org/10.1007/s10546-006-9052-x

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., & Reichstein, M. (2018). Upscaled diurnal cycles of land-atmosphere fluxes: A new global half-hourly data product. *Earth System Science Data*, 1–47. https://doi.org/10.5194/essd-2017-130

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Chapman and Hall. EE. UU.: Wadsworth International Group

Chen, X. L., Su, Z. B., Ma, Y. M., Liu, S. M., Yu, Q., & Xu, Z. W. (2014). Development of a 10-year (2001-2010) 0.1° data set of land-surface energy balance for mainland China. *Atmospheric Chemistry and Physics*, *14*(23), 13,097–13,117. https://doi.org/10.5194/acp-14-13097-2014

Chen, Z. Q., Shi, R. H., & Zhang, S. P. (2013). An artificial neural network approach to estimate evapotranspiration from remote sensing and AmeriFlux data. *Frontiers of Earth Science*, *7*(1), 103–111. https://doi.org/10.1007/s11707-012-0346-7

Davies, J. A., & Allen, C. D. (1973). Equilibrium, potential and actual evaporation from cropped surfaces in southern Ontario. *Journal of Applied Meteorology*. https://doi.org/10.1175/1520-0450(1973)012<0649:EPAAEF>2.0.CO;2

De Oliveira, G., & Moraes, E. C. (2013). Validação do balanço de radiação obtido a partir de dados MODIS/TERRA na Amazônia com medidas de superfície do LBA. *Acta Amazonica*, *43*(3), 353–363. https://doi.org/10.1590/S0044-59672013000300011

Ershadi, A., McCabe, M. F., Evans, J. P., Chaney, N. W., & Wood, E. F. (2014). Multi-site evaluation of terrestrial evaporation models using FLUXNET data. *Agricultural and Forest Meteorology*, *187*, 46–61. https://doi.org/10.1016/j.agrformet.2013.11.008

Evans, J. G., McNeil, D. D., Finch, J. W., Murray, T., Harding, R. J., Ward, H. C., & Verhoef, A. (2012). Determination of turbulent heat fluxes using a large aperture scintillometer over undulating mixed agricultural terrain. *Agricultural and Forest Meteorology*, *166–167*, 221–233. https://doi.org/10.1016/j.agrformet.2012.07.010

Ezzahar, J., Chehbouni, A., Hoedjes, J., Ramier, D., Boulain, N., Boubkraoui, S., et al. (2009). Combining scintillometer measurements and an aggregation scheme to estimate area-averaged latent heat flux during the AMMA experiment. *Journal of Hydrology*, *375*(1–2), 217–226. https://doi.org/10.1016/j.jhydrol.2009.01.010

Fang, K., Wu, J., Zhu, J., & Xie, B. (2011). A review of technologies on random forests (in Chinese). *Statistics & Information Forum*, *26*(3), 32–37.

Fang, Y., Sun, G., Caldwell, P., Mcnulty, S. G., Noormets, A., Domec, J. C., et al. (2016). Monthly land cover-specific evapotranspiration models derived from global eddy flux measurements and remote sensing data. *Ecohydrology*, *9*(2), 248–266. https://doi.org/10.1002/eco.1629

Fisher, J. B., Tu, K. P., & Baldocchi, D. D. (2008). Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of Environment*, *112*(3), 901–919. https://doi.org/10.1016/j.rse.2007.06.025

Galindo, F. J., & Palacio, J. (1999). Estimating the instabilities of N correlated clocks. Proceedings of the 31st Annual Precise Time and Time Interval (PTTI) Meeting, (April), 285–296.

Gao, S. G., Zhu, Z. L., Liu, S. M., Jin, R., Yang, G. C., & Tan, L. (2014). Estimating the spatial distribution of soil moisture based on Bayesian maximum entropy method with auxiliary data from remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *32*(1), 54–66. https://doi.org/10.1016/j.jag.2014.03.003

Ge, Y., Liang, Y. Z., Wang, J. H., Zhao, Q. Y., & Liu, S. M. (2015). Upscaling sensible heat fluxes with area-to-area regression kriging. *IEEE Geoscience and Remote Sensing Letters*, *12*(3), 656–660. https://doi.org/10.1109/LGRS.2014.2355871

Ghahreman, N., & Sameti, M. (2014). Comparison of M5 model tree and artificial neural network for estimating potential evapotranspiration in semi-arid climates. *Desert*, *19*(1), 75–81.

Gottschalk, L., Batchvarova, E., Gryning, S., Lindroth, A., Melas, D., Motovilov, Y., et al. (1999). Scale aggregation—Comparison of flux estimates from NOPEX. *Agricultural and Forest Meteorology*, *98–99*, 103–119. https://doi.org/10.1016/S0168-1923(99)00142-2

Guillevic, P. C., Biard, J. C., Hulley, G. C., Privette, J. L., Hook, S. J., Olioso, A., et al. (2014). Validation of land surface temperature products derived from the Visible Infrared Imaging Radiometer Suite (VIIRS) using ground-based and heritage satellite measurements. *Remote Sensing of Environment*, *154*, 19–37. https://doi.org/10.1016/j.rse.2014.08.013

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
Heinemann, G., & Kerschgens, M. (2005). Comparison of methods for area-averaging surface energy fluxes over heterogeneous land surfaces using high-resolution non-hydrostatic simulations. *International Journal of Climatology*, *25*(3), 379–403. https://doi.org/10.1002/joc.1123
Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527
Hu, G. C., & Jia, L. (2015). Monitoring of evapotranspiration in a semi-arid inland river basin by combining microwave and optical remote sensing observations. *Remote Sensing*, *7*(3), 3056–3087. https://doi.org/10.3390/rs70303056
Hu, M. G., Wang, J. H., Ge, Y., Liu, M. X., Liu, S. M., Xu, Z. W., & Xu, T. R. (2015). Scaling flux tower observations of sensible heat flux using weighted area-to-area regression kriging. *Atmosphere*, *6*(8), 1032–1044. https://doi.org/10.3390/atmos6081032
Huang, G. H., Li, X., Ma, M. G., Li, H. Y., & Huang, C. L. (2016). High resolution surface radiation products for studies of regional energy, hydrologic and ecological processes over Heihe River Basin, northwest China. *Agricultural and Forest Meteorology*, *230–231*, 67–78. https://doi.org/10.1016/j.agrformet.2016.04.007
Jia, Z. Z., Liu, S. M., Xu, Z. W., Chen, Y. J., & Zhu, M. J. (2012). Validation of remotely sensed evapotranspiration over the Hai River Basin, China. *Journal of Geophysical Research*, *117*, D13113. https://doi.org/10.1029/2011JD017037
Jiang, C., & Ryu, Y. (2016). Multi-scale evaluation of global gross primary productivity and evapotranspiration products derived from Breathing Earth System Simulator (BESS). *Remote Sensing of Environment*, *186*, 528–547. https://doi.org/10.1016/j.rse.2016.08.030
Jiang, L., & Islam, S. (2001). Estimation of surface evaporation map over Southern Great Plains using remote sensing data. *Water Resources Research*, *37*(2), 329–340. https://doi.org/10.1029/2000WR900255
Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396
Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research*, *116*, G00J07. https://doi.org/10.1029/2010JG001566
Kabsch, E., Olesen, F. S., & Prata, F. (2008). Initial results of the land surface temperature (LST) validation with the Evora, Portugal ground-truth station measurements. *International Journal of Remote Sensing*, *29*(17–18), 5329–5345. https://doi.org/10.1080/01431160802036326
Kalma, J. D., McVicar, T. R., & McCabe, M. F. (2008). Estimating land surface evaporation: A review of methods using remotely sensed surface temperature data. *Surveys in Geophysics*, *29*(4–5), 421–469. https://doi.org/10.1007/s10712-008-9037-z
Kormann, R., & Meixner, F. X. (2001). An analytical footprint model for non-neutral stratification. *Boundary-Layer Meteorology*, *99*(2), 207–224. https://doi.org/10.1023/A:1018991015119
Landeras, G., Ortiz-Barredo, A., & López, J. J. (2008). Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). *Agricultural Water Management*, *95*(5), 553–565. https://doi.org/10.1016/j.agwat.2007.12.011
Li, X., Cheng, G. D., Liu, S. M., Xiao, Q., Ma, M. G., Jin, R., et al. (2013). Heihe watershed allied telemetry experimental research (HiWater) scientific objectives and experimental design. *Bulletin of the American Meteorological Society*, *94*(8), 1145–1160. https://doi.org/10.1175/BAMS-D-12-00154.1
Li, X., Liu, S. M., Xiao, Q., Ma, M. G., Jin, R., Che, T., et al. (2017). A multiscale dataset for understanding complex eco-hydrological processes in a heterogeneous oasis system. *Scientific Data*, *4*, 1–11. https://doi.org/10.1038/sdata.2017.83
Li, Z. L., Tang, R. L., Wan, Z. M., Bi, Y. Y., Zhou, C. H., Tang, B. H., et al. (2009). A review of current methodologies for regional evapotranspiration estimation from remotely sensed data. *Sensors*, *9*(5), 3801–3853. https://doi.org/10.3390/s90503801
Liebethal, C., Huwe, B., & Foken, T. (2005). Sensitivity analysis for two ground heat flux calculation approaches. *Agricultural and Forest Meteorology*, *132*(3–4), 253–262. https://doi.org/10.1016/j.agrformet.2005.08.001
Liou, Y. A., & Kar, S. K. (2014). Evapotranspiration estimation with remote sensing and various surface energy balance algorithms—A review. *Energies*, *7*(5), 2821–2849. https://doi.org/10.3390/en7052821
Liu, S. M., Xu, Z. W., Song, L. S., Zhao, Q. Y., Ge, Y., Xu, T. R., et al. (2016). Upscaling evapotranspiration measurements from multi-site to the satellite pixel scale over heterogeneous land surfaces. *Agricultural and Forest Meteorology*, *230*, 97–113. https://doi.org/10.1016/j.agrformet.2016.04.008
Liu, S. M., Xu, Z. W., Wang, W. Z., Jia, Z. Z., Zhu, M. J., Bai, J., & Wang, J. M. (2011). A comparison of eddy-covariance and large aperture scintillometer measurements with respect to the energy balance closure problem. *Hydrology and Earth System Sciences*, *15*(4), 1291–1306. https://doi.org/10.5194/hess-15-1291-2011
Liu, S. M., Xu, Z. W., Zhu, Z. L., Jia, Z. Z., & Zhu, M. J. (2013). Measurements of evapotranspiration from eddy-covariance systems and large aperture scintillometers in the Hai River Basin, China. *Journal of Hydrology*, *487*, 24–38. https://doi.org/10.1016/j.jhydrol.2013.02.025
Ma, Y. (2015). Estimating evapotranspiration with multi-source remote sensing data—A case study of Zhangye desert-oasis in the middle stream of the Heihe River watershed (in Chinese), (Doctoral dissertation). Beijing, Beijing Normal University.
Meijninger, W. M. L., Hartogensis, O. K., Kohsiek, W., Hoedjes, J. C. B., Zuurbier, R. M., & De Bruin, H. A. R. (2002). Determination of area-averaged sensible heat fluxes with a large aperture scintillometer over a heterogeneous surface–Flevoland field experiment. *Boundary-Layer Meteorology*, *105*(1), 37–62. https://doi.org/10.1023/A:1019647732027
Metzger, S., Junkermann, W., Mauder, M., Butterbach-Bahl, K., Trancón, Y., Widemann, B., et al. (2013). Spatially explicit regionalization of airborne flux measurements using environmental response functions. *Biogeosciences*, *10*(4), 2193–2217. https://doi.org/10.5194/bg-10-2193-2013
Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., & Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, *15*(2), 453–469. https://doi.org/10.5194/hess-15-453-2011
Miralles, D. G., Jiménez, C., Jung, M., Michel, D., Ershadi, A., Mccabe, M. F., et al. (2016). The WACMOS-ET project—Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences*, *20*(2), 823–842. https://doi.org/10.5194/hess-20-823-2016
Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019
Mu, X., Hu, R., Huang, S., & Chen, Y. (2012). HiWATER: Dataset of emissivity in the middle reaches of the Heihe River Basin in 2012. Beijing normal university; cold and arid regions environmental and engineering research institute. *Chinese Academy of Sciences*. https://doi.org/10.3972/hiwater.042.2013.db
Ochs, G. R., & Wilson, J. J. (1993). A second-generation large aperture scintillometer, NOAA Tech. Memor. ERL ETL-232. NOAA Environmental Research Laboratories, Boulder, CO, USA, 24.
Pan, X. D., Li, X., Shi, X. K., Han, X. J., Luo, L. H., & Wang, L. X. (2012). Dynamic downscaling of near-surface air temperature at the basin scale using WRF-a case study in the Heihe River Basin, China. *Frontiers of Earth Science*, *6*(3), 314–323. https://doi.org/10.1007/s11707-012-0306-2

Peng, G. L., Liu, S. M., Cai, X. H., Lu, L., & Xu, Z. W. (2008). Footprint analysis of turbulent flux measurement over heterogeneous surface (in Chinese). *Chinese Journal of Atmospheric Sciences*, *32*(5), 1064–1070.
Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2
Qiao, C., Sun, R., Xu, Z. W., Zhang, L., Liu, L. Y., Hao, L. Y., & Jiang, G. Q. (2015). A study of shelterbelt transpiration and cropland evapotranspiration in an irrigated area in the middle reaches of the Heihe river in northwestern China. *IEEE Geoscience and Remote Sensing Letters*, *12*(2), 369–373. https://doi.org/10.1109/LGRS.2014.2342219
Qin, J., Yang, K., Lu, N., Chen, Y. Y., Zhao, L., & Han, M. L. (2013). Spatial upscaling of in-situ soil moisture measurements based on MODIS-derived apparent thermal inertia. *Remote Sensing of Environment*, *138*, 1–9. https://doi.org/10.1016/j.rse.2013.07.003
Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaurmann.
Rahimikhoob, A. (2014). Comparison between M5 model tree and neural networks for estimating reference evapotranspiration in an arid environment. *Water Resources Management*, *28*(3), 657–669. https://doi.org/10.1007/s11269-013-0506-x
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). In D. E. Rumelhart & J. L. McClelland (Eds.), *Learning internal representation by error propagation parallel distributed processing foundations* (Vol. 1). Cambridge, MA: MIT press.
Saltelli, A., Tarantola, S., & Chan, K. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, *41*(1), 39–56. https://doi.org/10.1080/00401706.1999.10485594
Schmid, H. P. (1994). Source areas for scalars and scalar fluxes. *Boundary-Layer Meteorology*, *67*(3), 293–318. https://doi.org/10.1007/BF00713146
Schmid, H. P. (2002). Footprint modeling for vegetation atmosphere exchange studies: A review and perspective. *Agricultural and Forest Meteorology*, *113*(1–4), 159–183. https://doi.org/10.1016/S0168-1923(02)00107-7
Shi, Y. C., Wang, J. D., Qin, J., & Qu, Y. H. (2015). An upscaling algorithm to obtain the representative ground truth of LAI time series in heterogeneous land surface. *Remote Sensing*, *7*(10), 12,887–12,908. https://doi.org/10.3390/rs71012887
Song, L. S., Liu, S. M., Kustas, W. P., Zhou, J., Xu, Z. W., Xia, T., & Li, M. S. (2016). Application of remote sensing-based two-source energy balance model for mapping field surface fluxes with composite and component surface temperatures. *Agricultural and Forest Meteorology*, *230*, 8–19. https://doi.org/10.1016/j.agrformet.2016.01.005
Song, X. D., Zhang, G. L., Liu, F., Li, D. C., Zhao, Y. G., & Yang, J. L. (2016). Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model. *Journal of Arid Land*, *8*(5), 734–748. https://doi.org/10.1007/s40333-016-0049-0
Stein, M. L. (2012). *Interpolation of spatial data: Some theory for kriging*. New York: Springer Science & Business Media.
Sun, G., Alstad, K., Chen, J. Q., Chen, S. P., Ford, C. R., Lin, G. H., et al. (2011). A general predictive model for estimating monthly ecosystem evapotranspiration. *Ecohydrology*, *4*(2), 245–255. https://doi.org/10.1002/eco.194
Twine, T. E., Kustas, W. P., Norman, J. M., Cook, D. R., Houser, P. R., Meyers, T. P., et al. (2000). Correcting eddy-covariance flux underestimates over a grassland. *Agricultural and Forest Meteorology*, *103*(3), 279–300. https://doi.org/10.1016/S0168-1923(00)00123-4
Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35–49. https://doi.org/10.1016/j.rse.2013.07.013
Wang, D., Liang, S., He, T., & Shi, Q. (2015). Estimating clear-sky all-wave net radiation from combined visible and shortwave infrared (VSWIR) and thermal infrared (TIR) remote sensing data. *Remote Sensing of Environment*, *167*, 31–39. https://doi.org/10.1016/j.rse.2015.03.022
Wang, K. C., & Dickinson, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Reviews of Geophysics*, *50*, RG2005. https://doi.org/10.1029/2011RG000373
Wang, S., Fu, Z. Y., Chen, H. S., Nie, Y. P., & Wang, K. L. (2016). Modeling daily reference ET in the karst area of northwest Guangxi (China) using gene expression programming (GEP) and artificial neural network (ANN). *Theoretical and Applied Climatology*, *126*(3–4), 493–504. https://doi.org/10.1007/s00704-015-1602-z
Wang, W. H., Liang, S. L., & Meyers, T. (2008). Validating MODIS land surface temperature products using long-term nighttime ground measurements. *Remote Sensing of Environment*, *112*(3), 623–635. https://doi.org/10.1016/j.rse.2007.05.024
Wu, B. F., Yan, N. N., Xiong, J., Bastiaanssen, W. G. M., Zhu, W. W., & Stein, A. (2012). Validation of ETWatch using field measurements at diverse landscapes: A case study in Hai Basin of China. *Journal of Hydrology*, *436–437*, 67–80. https://doi.org/10.1016/j.jhydrol.2012.02.043
Xu, F. N., Qi, Y., Wang, J. H., & Zhang, J. L. (2015). Riparian forest vegetation coverage information classification based on object-oriented method in Heihe River (in Chinese). *Remote Sensing Technology and Application*, *30*(5), 996–1005.
Xu, F. N., Wang, W. Z., Wang, J. M., Xu, Z. W., Qi, Y., & Wu, Y. R. (2017). Area-averaged evapotranspiration over a heterogeneous land surface: Aggregation of multi-point EC flux measurements with high-resolution land-cover map and footprint analysis. *Hydrology and Earth System Sciences*, *21*(8), 4037–4051. https://doi.org/10.5194/hess-21-4037-2017
Xu, K., Metzger, S., & Desai, A. R. (2017). Upscaling tower-observed turbulent exchange at fine spatio-temporal resolution using environmental response functions. *Agricultural and Forest Meteorology*, *232*, 10–22. https://doi.org/10.1016/j.agrformet.2016.07.019
Xu, Z. W., Liu, S. M., Li, X., Shi, S. J., Wang, J. M., Zhu, Z. L., et al. (2013). Intercomparison of surface energy flux measurement systems used during the HiWATER-MUSOEXE. *Journal of Geophysical Research: Atmospheres*, *118*, 13,140–13,157. https://doi.org/10.1002/2013JD020260
Xu, Z. W., Ma, Y. F., Liu, S. M., Shi, W. J., & Wang, J. M. (2017). Assessment of the energy balance closure under advective conditions and its impact using remote sensing data. *Journal of Applied Meteorology and Climatology*, *56*(1), 127–140. https://doi.org/10.1175/JAMC-D-16-0096.1
Yan, C. (2011). Land use/Land cover data of Zhangye city in 2005. *Heihe Plan Science Data Center*. https://doi.org/10.3972/heihe.011.2013.db
Yang, F. H., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., et al. (2006). Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(11), 3452–3461. https://doi.org/10.1109/TGRS.2006.876297
Yao, Y. J., Liang, S. L., Cheng, J., Liu, S. M., Fisher, J. B., Zhang, X. D., et al. (2013). MODIS-driven estimation of terrestrial latent heat flux in China based on a modified Priestley-Taylor algorithm. *Agricultural and Forest Meteorology*, *171–172*, 187–202. https://doi.org/10.1016/j.agrformet.2012.11.016
Zhang, L., Dawes, W. R., & Walker, G. R. (2001). Response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resources Research*, *37*(3), 701–708. https://doi.org/10.1029/2000WR900325
Zhou, J., Li, M. S., Liu, S. M., Jia, Z. Z., & Ma, Y. F. (2015). Validation and performance evaluations of methods for estimating land surface temperatures from ASTER data in the middle reach of the Heihe River Basin, Northwest China. *Remote Sensing*, *7*(6), 7126–7156. https://doi.org/10.3390/rs70607126