HYDROLOGICAL PROCESSES
Hydrol. Process. **30**, 2861–2879 (2016)
Published online 19 April 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.10832

Multi-model ensemble prediction of terrestrial evapotranspiration across north China using Bayesian model averaging

Gaofeng Zhu,[1]* Xin Li,[2] Kun Zhang,[1] Zhenyu Ding,[3] Tuo Han,[1] Jinzhu Ma,[1] Chunlin Huang,[2] Jianhua He[1] and Ting Ma[1]

[1] *Key Laboratory of Western China's Environmental Systems (Ministry of Education), Lanzhou University, Lanzhou 730000, China*
[2] *Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Science, Lanzhou 730000, China*
[3] *Chinese Academy For Environmental Planning, Beijing 100012, China*

Abstract:

Using high-quality dataset from 12 flux towers in north China, the performance of four evapotranspiration (ET) models and the multi-model ensemble approaches including the simple averaging (SA) and Bayesian model average (BMA) were systematically evaluated in this study. The four models were the single-layer Penman–Monteith (P–M) model, the two-layer Shuttleworthe–Wallace (S–W) model, the advection–aridity (A–A) model, and a modified Priestley–Taylor (PT-JPL). Based on the mean value of Taylor skill (*S*) and the regression slope between measured and simulated ET values across all sites, the order of overall performance of the individual models from the best to the worst were: S–W (0.88, 0.87), PT-JPL (0.80, 1.17), P–M (0.63, 1.73) and A–A (0.60, 1.68) [statistics stated as (Taylor skill, regression slope)]. Here, all models used the same values of parameters, LAI and fractional vegetation cover as well as the forcing meteorological data. Thus, the differences in model performance were mainly attributed to errors in model structure. To the ensemble approach, the BMA method has the advantage of generating more skillful and reliable predictions than the SA scheme. However, successful implementation of BMA requires accurate estimates of its parameters, and some degradation in performance were observed when the BMA parameters generated from the training period were used for the validation period. Thus, it is necessary to explore the seasonal variations of the BMA parameters according the different growth stages. Finally, the optimal conditional density function of half-hourly ET approximated well by the double-exponential distribution. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS evapotranspiration; multi-model intercomparison; water-limited ecosystems; Bayesian model averaging; Penman–Monteith; advection–aridity; Priestley–Taylor

Received 17 November 2015; Accepted 25 February 2016

INTRODUCTION

Terrestrial evapotranspiration (ET) is a phase transition of water from liquid (or ice) to gas (Wang and Dickinson, 2012). This process serves as one of the main components of the hydrological cycle, accounting for ~60% of terrestrial precipitation (Shiklomanov, 1998). The latent heat (λET) accompanying ET is particularly effective in cooling the land surface (Katul *et al.*, 2012), thus critical in atmospheric processes and biogeochemical cycles (Teuling *et al.*, 2010; Jung *et al.*, 2010; Mu *et al.*, 2011; Sheffield *et al.*, 2012). Therefore, accurately estimating and measuring ET (or λET) are of the long-term interest to hydrologists (Xu and Singh, 2005), climate modellers (Seneviratne *et al.*, 2010), ecologists (Fisher *et al.*, 2011), and agriculturalists (Zhu *et al.*, 2014a).

The development of instrumentation for measuring scalar fluxes and vertical wind in the 1970s led to the development of the eddy covariance (EC) technique (Baldocchi *et al.*, 2001; Wilson *et al.*, 2002). To date, the technique is the key measurement tool used by several large observational projects, such as the FLUXNET (Baldocchi *et al.*, 2001), the Integrated Carbon Observation System in Europe (www.icos-infrastructure.eu), and the Coordinated Enhanced Observation Project (CEOP) in the arid and semi-arid regions of northern China (Yao *et al.*, 2013). These projects provide a high-quality and valuable dataset of surface hydrological and meteorological variables across a wide range of biomes and climate conditions (Wang and Dickinson, 2012). On the other hand, a number of models have been developed for estimating ET since the 1950s. These models vary in degrees of complexity from the empirical equations using

*Correspondence to: Zhu Gaofeng, Tianshui Road 222, Lanzhou, Gnasu Province 730000, China.
E-mail: zhugf@lzu.edu.cn

a single climatic variable to formulations based on the conservation of either energy or mass, or both (Brutsaert, 2005). Among them, the renowned Penman–Monteith (P–M) model (Monteith, 1965) is physically sound and rigorous, and has been widely used in estimating ET. However, the P–M model treats the land surface as one homogeneous layer and cannot distinguish between the plant transpiration and soil evaporation (Monteith, 1965); the latter may be the main component of ET in arid region because of low vegetation cover fraction. Shuttleworth and Wallace (1985) extend the P–M model to the sparse canopies, and developed a two-layer model (S–W) to separately account for plant transpiration and soil evaporation. However, these two ET models require numerous parameters and proper determination of these parameters is usually difficult (Zhu *et al.*, 2013, 2014b), thus limiting their applications (Wang and Dickinson, 2012). The complementary relationship (CR) hypothesized by Bouchet (1963) provides an avenues for estimating ET from only routine meteorological variables without detailed knowledge of the surface states. Brutsaert and Stricker (1979) first put Bouchet's hypothesis into practice in their advection–aridity (A–A) model, which has been applied at hourly (Parlange and Katul, 1992), daily, and monthly time steps (Brutsaert and Stricker, 1979). Finally, Priestley and Taylor (1972) proposed a radiation-based model for equilibrium ET under conditions of unlimited soil moisture supply. To scale-down the equilibrium ET to actual ET, Fisher *et al.* (2008) presented a novel mode (PT-JPL) based on bio-physiological constraints and soil evaporation partitioning. The PT-JPL is attractive for its simplicity and potential to obtain regional or global ET using satellite data (Mu *et al.*, 2007).

Nevertheless, there are still some insufficiencies and limitations in the applying these models. First, most previous studies have focused either on intercomparison of different models over single (or a few) locations (Stannard, 1993; Fisher *et al.*, 2005, 2009; Zhang *et al.*, 2008; Gharsallah *et al.*, 2013; Zhu *et al.*, 2013, 2014b), or on evaluating a specific model over different land surfaces (Fisher *et al.*, 2008; García *et al.*, 2013). As far as we know, systemic intercomparison and evaluation of different models over a wide range of biomes and climatic conditions are relatively few with the exception of Ershadi *et al.* (2014), who evaluated the performance of multiple ET models based on dataset from 20 FLUXNET sites. However, only three of the 20 sites included semiarid and arid vegetation with annual precipitation below 400 mm (Ershadi *et al.*, 2014). Hence, the performances of the ET models over semiarid and arid climatic conditions remain uncertain. Second, the traditional approach to ET prediction is to postulate a model structure and assume the mismatches between observed and simulated values are solely attributed to parameter uncertainties (Vrugt and Robinson, 2007). To date, numerous studies have focused on obtaining good matches between observed and simulated ET by locally calibrating the model parameters (i. e. Ortega-Farias *et al.*, 2004, 2006; Shi *et al.*, 2008; Hu *et al.*, 2009; Doody *et al.*, 2011; Zhu *et al.*, 2013, 2014b). However, this procedure does not take the structural error of the model into consideration, and also limits the utility of the model to those specific locations (Ershadi *et al.*, 2014). Thus, much attention should be paid in identifying and diagnosing the model structure errors by intercomparing the performances of different models with a set of fixed parameters for similar biomes. Third, aiming to extract as much information as possible from the existing models, the multi-model ensemble approaches have become popular in reliable prediction and uncertainty analysis (Raftery *et al.*, 2005; Ajami *et al.*, 2007; Duan *et al.*, 2007; Vrugt and Robinson, 2007). Recently, Ershadi *et al.* (2014) showed that even the simple averaging (SA) method performed better than any individual model in estimating ET across the 20 selected FLUXNET sites. Also, the Bayesian model averaging (BMA) approach has been used to merge a range of satellite-based models for regional/global ET estimations (Vinukollu *et al.*, 2011; Mueller *et al.*, 2011; Yao *et al.*, 2014; Chen *et al.*, 2015). The results indicated that the BMA method can generally outperforms the best individual model, and provides a useful tool for generating a long-term regional/global terrestrial ET product (Yao *et al.*, 2014; Chen *et al.*, 2015). However, the success of the BMA method depends on the skill and performance of the individual members of the ensemble (Vrugt and Robinson, 2007). In some instances, the forecast error of the BMA approach was of similar magnitude as the forecast error of the best model in the ensemble (Georgekakos *et al.*, 2004; Vrugt and Robinson, 2007). Thus, systematic evaluation of the performance of the BMA method for ET models varying in structural complexity is urgently needed over various biomes and climates. In addition, the normal conditional density function was commonly used in practice when merging the multiple ET products (Yao *et al.*, 2014; Chen *et al.*, 2015). As ET is a complicated variable coupling budgets of energy, hydrology, and carbon (Yao *et al.*, 2014), another avenue of research is to develop appropriate structures for errors in ET.

Using a collection of high-quality tower based data from the CEOP experiments across north China over the main growing seasons (from July to September) of two years, the objectives of the present study were to: (1) evaluate the performance of the widely-used models in estimating actual ET over different biomes and climatic conditions; (2) identify and diagnose the model structure errors of the selected models by using a set of fixed

parameters for similar biomes; and (3) evaluate the BMA method by comparison with SA method and the individual ET models to generate possible improvements of the BMA method for estimating ET.

DATA AND METHODOLOGY

Observations from eddy covariance flux towers

To evaluate the performance of ET models and the multi-model ensemble approaches, we used the ground-observed data from 12 flux towers (Figure 1), which were set up under the CEOP in the arid and semi-arid regions of northern China (http://observation.tea.ac.cn/). The climate at the flux tower locations varies from semi-humid to arid with associated variations in vegetation (i.e. grassland and cropland, which are the major land-surface biomes in north China) (Table I). These datasets include rainfall (TE525MM, Campbell Scientific Instruments Inc.), air temperature, relative humidity (HMP45C, Vaisala Inc., Helsinki, Finland), wind speed/direction (034B, Met One Instruments Inc., USA), downward and upward solar and longwave radiation (PSP, The EPPLEY Laboratory Inc., USA), soil temperature (Campbell-107, Campbell Scientific Instruments Inc.) and moisture (CS616, Campbell Scientific Instruments Inc.) profiles at depths of 0.02, 0.04, 0.1, 0.2, 0.4, 0.8, 1.2, and 1.6 m, and surface soil heat flux (HFT3, Campbell Scientific Instruments Inc.). All turbulent flux observations were measured by the EC method. Data gaps because of instrument malfunction, power failure, and bad weather conditions were filled using artificial neural network (ANN) and mean diurnal variations (MDV) methods (Falge *et al.*, 2001). The energy closure which may be affected by many factors is still a key indicator to assess the quality of flux data (Drexler *et al.*, 2004). Liu *et al.* (2011) examined the energy closure at different sites of the CEOP sites. About 85% of the energy balance closure was found in EC data, indicating the measurements are reasonable.

Remote sensing based measurements

The 16-day MODIS NDVI (MOD13Q1) product (Solano *et al.*, 2010) with 250-m spatial resolution was extracted at each tower location from the Simple Object Access Protocol (SOAP) web service site (http://daac.ornl.gov/MODIS/). The 16-day gaps between successive NDVI records were temporally interpolated using linear interpolation. The leaf area index (LAI) and fractional vegetation cover were calculated from the NDVI data using the method described by Jiménez-Muñoz *et al.* (2009) and Ershadi *et al.* (2014), respectively. All ET

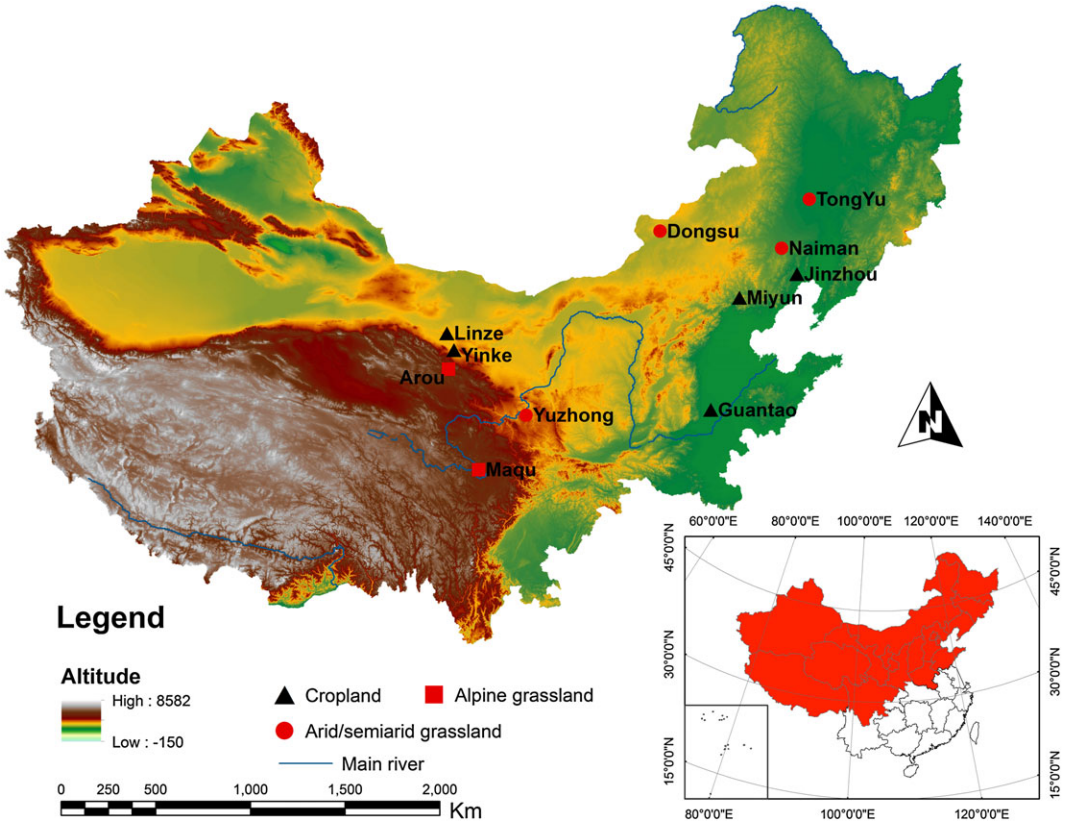

Figure 1. Location of the eddy covariance towers used to provide forcing and validation data in this study across north China

Table I. Main characteristics of the 12 flux sites in the study region

Biome	Site name	Site No.	Location (lat./long.)	Elevation (m)	EC above canopy (m)	Annual precipitation (mm)	Climate zone	Vegetation type	Year	References
Arid/semiarid grassland	Tongyu	1	44.88°/122.88°	151	2.0	388	Semi-arid	Degraded grassland	2008,2009	Wang *et al.* (2014)
	Naiman	2	42.93°/120.70°	361	2.0	366	Semi-arid	Degraded grassland	2008	Wang *et al.* (2014)
	Dongsu	3	44.09°/113.57°	990	2.0	287	Arid	Desert steppe	2008,2009	Wang *et al.* (2014)
	Yuzhong	4	35.95°/104.13°	1965	2	382	Semi-arid	Typical steppe	2008,2009	Wang *et al.* (2014)
Alpine grassland	Arou	5	38.04°/100.46°	3033	2	396	Semi-arid	Alpine meadow	2008,2009	Zhu *et al.* (2013, 2014b)
	Maqu	6	33.89°/102.14°	3423	2	599	Semi-humid	Alpine wetland	2008	Wang *et al.* (2014)
Cropland	Tongyu	7	44.88°/122.88°	151	2	388	Semi-arid	Cropland (maize)	2008, 009	Wang *et al.* (2014)
	Jinzhou	8	41.15°/121.20°	22	4.0	600	Semi-arid	Cropland (maize)	2008,2009	Xu *et al.* (2011)
	Miyun	9	40.63°/117.32°	352	25	584	Semi-arid	Cropland (maize)	2008,2009	Xu *et al.* (2011)
	Guantao	10	36.52°/115.11°	42	15	536	Semi-arid	Cropland (wheat)	2008	Xu *et al.* (2011)
	Yingke	11	38.86°/100.58°	1556	4.5	125	Arid	Cropland (maize)	2008,2009	Zhu *et al.* (2014a)
	Linze	12	39.25°/100.13°	1384	4.0	120	Arid	Cropland (maize)	2008,2009	Zhu *et al.* (2014a)

models use the same values of LAI and fractional vegetation cover for their parameterization.

Model descriptions

P–M model. The P–M model can be formulated as (Monteith, 1965):

$$\lambda \mathrm{ET} = \frac{\Delta(R_\mathrm{n} - G) + \rho C_\mathrm{p} D / r_\mathrm{a}}{\Delta + \gamma(1 + r_\mathrm{s}/r_\mathrm{a})} \tag{1}$$

where λ is the latent heat of evaporation ($\mathrm{J\,kg^{-1}}$); Δ is the slope of the saturation water vapour pressure curve ($\mathrm{Pa\,K^{-1}}$); ρ is the density of the air ($\mathrm{kg\,m^{-3}}$); C_p is the specific heat capacity of dry air at constant pressure ($\mathrm{J\,kg^{-1}\,K^{-1}}$); D is the water vapour pressure deficit (kPa); γ is the psychrometric constant ($\mathrm{kPa\ K^{-1}}$); r_a is the aerodynamic resistance ($\mathrm{s\ m^{-1}}$); and r_c is the surface canopy resistance ($\mathrm{s\,m^{-1}}$).

The aerodynamic resistance r_a is usually computed with the following equation, assuming neutral stability conditions (Brutsaert, 1982):

$$r_a = \frac{1}{k^2 u_z} \ln\left(\frac{z-d}{h_c - d}\right) \ln\left(\frac{z-d}{z_0}\right) \tag{2}$$

where h_c is the mean vegetation height (m), z is the height of wind speed measurements (m), d is the zero plane displacement (m) estimated as $d = 0.67h_c$, z_0, the roughness length for momentum transfer (m), is estimated by $z_0 = 0.123h_c$, k is the von Karman's constant ($k = 0.41$), and u_z is wind speed at the reference height ($\mathrm{m\,s^{-1}}$).

The canopy resistance r_s can be calculated using the Jarvis-type model (Jarvis, 1976):

$$r_\mathrm{s} = \frac{r_\mathrm{smin}}{\mathrm{LAI} \prod_i F_i(X_i)} \tag{3}$$

where r_smin represents the minimal stomatal resistance of individual leaves under optimal conditions. The value of r_smin was acquired based on the vegetation lookup tables used in the Simple Biosphere model (i.e. r_smin is equal to $30\,\mathrm{s\,m^{-1}}$ for crops and $60\,\mathrm{s\,m^{-1}}$ for grasslands; Dorman and Sellers, 1989). $F_i(X_i)$ is the stress function of a specific environmental variable X_i, with $0 \le F_i(X_i) \le 1$. Following Chen and Dudhia (2001), the stress functions were expressed as:

$$F_1(R_n) = \frac{r_\mathrm{smin}/r_\mathrm{smax} + f}{1 + f} \quad \text{with} \tag{4}$$

$$f = 0.55 \frac{R_g}{R_{gl}} \frac{2}{\mathrm{LAI}}$$

$$F_2(T_\mathrm{a}) = 1 - 0.0016(298 - T_a)^2 \tag{5}$$

$$F_3(D) = 1 - gD \tag{6}$$

$$F_4(\theta) = \begin{cases} 1 & \theta > \theta_{cr} \\ \dfrac{(\theta - \theta_{wp})}{(\theta_{cr} - \theta_{wp})} & \theta_{wp} \leq \theta \leq \theta_{cr} \\ 0 & \theta < \theta_{wp} \end{cases} \tag{7}$$

where r_{smax} is the maximum canopy resistance set equal to 5000 s m^{-1} (Chen and Dudhia, 2001); R_{gl} is the species-dependent threshold value of solar radiation for transpiration (W m^{-2}), which is equal to 100 W m^{-2} for crops and grasslands; R_g is the incident solar radiation (W m^{-2}); T_a is the air temperature (K) at the reference height; g is a parameter associated with the water vapour deficit D (kPa), set equal to 0.0025 kPa^{-1} (Noilhan and Planton, 1989); θ is the actual volumetric soil water content in the root-zone (m^3 m^{-3}); θ_{wp} is water content at the wilting point (m^3 m^{-3}); and θ_{cr} is the critical water content (m^3 m^{-3}) at which plant stress starts set, taken as $0.75\theta_{sat}$. θ_{sat} is the saturated soil water content (m^3 m^{-3}), which was estimated empirically through the near-surface soil texture.

Two-layers Shuttleworthe–Wallace (S–W) model. The S–W model (Shuttleworth and Wallace, 1985) combined two P–M type equations for plant transpiration and soil evaporation. The S–W model is expressed as follows:

$$\lambda ET = \lambda E + \lambda T = C_s ET_s + C_c ET_c \tag{8}$$

$$ET_s = \frac{\Delta A + \left[\rho C_p D - \Delta r_a^s (A - A_s)\right] / \left(r_a^a + r_a^s\right)}{\Delta + \gamma \left[1 + r_s^s / \left(r_a^a + r_a^s\right)\right]} \tag{9}$$

$$ET_c = \frac{\Delta A + \left[\rho C_p D - \Delta r_a^c A_s\right] / \left(r_a^a + r_a^c\right)}{\Delta + \gamma \left[1 + r_s^c / \left(r_a^a + r_a^c\right)\right]} \tag{10}$$

$$C_s = \frac{1}{1 + [R_s R_a / R_c (R_s + R_a)]} \tag{11}$$

$$C_c = \frac{1}{1 + [R_c R_a / R_s (R_c + R_a)]} \tag{12}$$

$$R_a = (\Delta + \gamma) r_a^a \tag{13}$$

$$R_c = (\Delta + \gamma) r_a^c + \gamma r_s^c \tag{14}$$

$$R_s = (\Delta + \gamma) r_a^s + \gamma r_s^s \tag{15}$$

where ET_s and ET_c are terms to describe evaporation from soil and transpiration from the plant (W m^{-2}), respectively; C_s and C_c are soil surface resistance and canopy resistance coefficients (dimensionless), respectively; r_s^c and r_s^s are the surface resistance for plant canopy and soil surface (s m^{-1}), respectively; r_a^c and r_a^s are aerodynamic resistances from the leaf to canopy height and soil surface to canopy height (s m^{-1}), and r_a^a is aerodynamic resistances from canopy height to reference height (s m^{-1}). A and A_s (W m^{-2}) are the available energy input above the canopy and above the soil surface, respectively, and are calculated as:

$$A = R_n - G \tag{16}$$

$$A_s = R_{ns} - G \tag{17}$$

where R_n and R_{ns} are net radiation fluxes into the canopy and the substrate (W m^{-2}), respectively; G is the soil heat flux (W m^{-2}). R_{ns} was calculated using a Beer's law relationship of the form:

$$R_{ns} = R_n \exp(-K_A LAI) \tag{18}$$

in which K_A is the extinction coefficient of light attenuation, and set 0.60 for fully grown plant (Sene, 1994).

The soil surface resistance r_s^s is interpreted as the resistance from the water vapour to diffuse through the top layer of the soil. It can be expressed as a function of the top layer (0- to 5-cm) of soil water content (Sellers *et al.*, 1992; Zhu *et al.*, 2013):

$$r_s^s = \exp\left(8.206 - 4.255 \frac{\theta_s}{\theta_{sat}}\right) \tag{19}$$

in which θ_s is soil water content in the top layer. The canopy resistance r_s^c was calculated following Equation 3, while the three aerodynamic resistance (i.e. r_a^a, r_a^c, and r_a^s) were computed as reported by Shuttleworth and Wallace (1985) and Shuttleworth and Gurney (1990).

Modified Priestley–Taylor (PT-JPL) model. The Priestley–Taylor (Priestley and Taylor, 1972) model is a simplified but surprising successful form of the P–M model. This model was introduced to estimate potential ET from an extensive wet surface in conditions of minimal advection (Stannard, 1993; Sumner and Jacobs, 2005), and is expressed as:

$$\lambda ET = \alpha_{PT} \frac{\Delta}{\Delta + \gamma} (R_n - G) \tag{20}$$

where α_{PT} is a unitless coefficient. For well-water surface, the value of α_{PT} has theoretical significance and was estimated to be 1.26 (Priestley and Taylor, 1972). Scaling of the Priestley–Taylor potential ET to actual ET has been performed by modification of α_{PT} as a function of the environmental variables (Flint and Childs, 1991). Here, we used the modified form of the Priestley–Taylor model developed by Fisher *et al.* (2008) (hereafter PT-JPL model). In this model, total ET is partitioned into soil evaporation (λE_s), wet canopy evaporation (λE_{wc}), and canopy transpiration (λT_c) and are defined as:

$$\lambda E_s = k_s \times \alpha_{\mathrm{PT}} \frac{\Delta}{\Delta+\gamma}(R_{ns} - G) \quad (21)$$

$$\lambda E_{wc} = k_{wc} \times \alpha_{\mathrm{PT}} \frac{\Delta}{\Delta+\gamma} R_{nc} \quad (22)$$

$$\lambda T_c = k_c \times \alpha_{\mathrm{PT}} \frac{\Delta}{\Delta+\gamma} R_{nc} \quad (23)$$

where R_{ns} is the net radiation for soil ($\mathrm{W\,m^{-2}}$) given by Equation 18; R_{nc} is the net radiation for canopy ($\mathrm{W\,m^{-2}}$), $R_{nc} = R_n - R_{ns}$; k_s, k_{wc}, and k_c are reduction functions for scaling of potential ET in each of soil, wet canopy surface, and canopy to their actual values, and defined as:

$$k_s = f_{\mathrm{wet}} + f_{\mathrm{SM}}(1 - f_{\mathrm{wet}}) \quad (24)$$

$$k_{\mathrm{wc}} = f_{\mathrm{wet}} \quad (25)$$

$$k_c = (1 - f_{\mathrm{wet}}) f_g f_T f_M \quad (26)$$

where f_{wet} is relative surface wetness (RH^4); f_{SM} is a soil moisture constraint ($\mathrm{RH}^{D/\beta}$); f_g is green canopy fraction ($f_{\mathrm{APAR}}/f_{\mathrm{IPAR}}$); f_T is a plant temperature constraint (exp $[-((T_a - T_{\mathrm{opt}})/T_{\mathrm{opt}})^2]$); f_M is a pant moisture constraint ($f_{\mathrm{APAR}}/f_{\mathrm{APARmax}}$). RH represents relative humidity (%), β is soil moisture constraint parameter, and $\beta = 1$ kPa was used in the original model (Fisher *et al.*, 2008), T_{opt} is the optimum plant growth temperature (298 K; García *et al.*, 2013), and f_{APAR} and f_{IPAR} are fraction of photosynthetically active radiation (PAR) that is absorbed and intercepted by vegetation cover, respectively, and are calculated as reported by Ershadi *et al.* (2014).

A–A model. The A–A model was first proposed by Brutsaert and Stricker (1979) and further improved by Parlange and Katul (1992). They are based on Bouchet's (1963) CR hypothesis that actual (λET) and potential ET ($\lambda\mathrm{ET_p}$; $\mathrm{W\ m^{-2}}$) should converge to wet surface ET ($\lambda\mathrm{ET_w}$; $\mathrm{W\ m^{-2}}$) at wet surface conditions. As an initially wet surface dries, λET and $\lambda\mathrm{ET_p}$ derive from $\lambda\mathrm{ET_w}$ with opposite changes in flux. Its general form is:

$$\lambda\mathrm{ET} = \left(\frac{b+1}{b}\right)\lambda\mathrm{ET_w} - \frac{\lambda\mathrm{ET_p}}{b} \quad (27)$$

$$\lambda\mathrm{ET_w} = \alpha_{\mathrm{PT}} \frac{\Delta}{\Delta+\gamma}(R_{\mathrm{n}} - G) \quad (28)$$

$$\lambda\mathrm{ET_p} = \frac{\Delta}{\Delta+\gamma}(R_{\mathrm{n}} - G) + \frac{\gamma}{\Delta+\gamma}\frac{\rho(q^* - q)}{r_{\mathrm{a}}} \quad (29)$$

where b is the proportionality constant and is equal to 1 in the A–A model; α_{PT} is the Priestley–Taylor coefficient, considered here as 1.26 (Priestley and Taylor, 1972); q^* and q are the saturation-specific humidity at air temperature and the specific humidity of the atmosphere ($\mathrm{kg\,kg^{-1}}$), respectively; r_{a} is the aerodynamic resistance ($\mathrm{s\,m^{-1}}$) and its formula in similar to that used for the P–M model.

BMA

BMA was proposed by Raftery *et al.* (2005) as a statistical probabilistic scheme for model combination. To explicate the BMA method, let y denotes the quantity to be forested, $\mathbf{D} = \left[y_1^{\mathrm{obs}}, y_2^{\mathrm{obs}}, \cdots, y_T^{\mathrm{obs}}\right]$ to be the training data with length T, and $\boldsymbol{f} = [f_1, f_2, \cdots, f_k]$ the ensemble of predictions obtained from k different models (i.e. $k = 4$ in this study). The $p_i(y|f_i, \mathbf{D})$ $(i = 1, 2, \cdots, k)$ is the posterior distribution of y given model predication f_i and observational data set **D**. According to the law of total probability, the posterior distribution of the BMA predication of y can be expressed as (Raftery *et al.*, 1997, 2005):

$$p(y|\mathbf{D}) = \sum_{i=1}^{k} p(f_i|\mathbf{D}) p_i(y|f_i, \mathbf{D}) \quad (30)$$

where $p(f_i|\mathbf{D})$ is the posterior probability of forecast f_i being the best one given the observational data **D**. If we denote $w_i = p(f_i|\mathbf{D})$, we should obtain $\sum_{i=1}^{k} w_i = 1$. One premise of the BMA scheme is that the weights (w_i) should reflect relative model performance as they are the probabilistic likelihood measures of a model being correct given the observational data **D** (Duan *et al.*, 2007). Supposing that $p_i(y|f_i, \mathbf{D})$ to be Gaussian distribution centred at a linear function of the original forecast $a_i + b_i f_i$ (where a_i and b_i are bias correction terms that are derived by simple linear regression of y on f for each of the individual ensemble members from the training data) with standard deviation σ_i, the posterior mean and variance of the BMA predication for variable y are:

$$E[y|\mathbf{D}] = \sum_{i=1}^{k} p(f_i|\mathbf{D}) \cdot E[p_i(y|f_i, \mathbf{D})] = \sum_{i=1}^{k} w_i(a_i + b_i f_i) \quad (31)$$

$$\mathrm{Var}[y|\mathbf{D}] = \sum_{i=1}^{k} w_i(a_i + b_i f_i - \sum_{i=1}^{k} w_i(a_i + b_i f_i))^2 + \sum_{i=1}^{k} w_i \sigma_i^2 \quad (32)$$

In essence, BMA prediction [Equation 31] is the average of individual predictions, and it receives higher weighting from better performing models; Variance of BMA prediction [Equation 32] consists of two terms, the

first representing the between-model-variance and the second representing the within-model-variance.

Successful implementation of the BMA method requires specification of w_i and σ_i^2 $(i=1,2,\cdots,k)$ on the basis of training observational data **D**. Let $\theta=\{w_i,\sigma_i^2,i=1,2,\cdots,k\}$, the log-likelihood function of Equation 30 can be approximated as:

$$\ell(\theta)=\sum_{t=1}^{T}\log(\sum_{i=1}^{k}w_i{\cdot}p_i(y_t|f_{i,t},\mathbf{D})). \quad (33)$$

Because of its high dimensionality of this problem, it is hard to obtain analytical solution of θ. In this study, the Expectation–Maximization (EM) algorithm was used to search the optimal value of θ (Raftery *et al.*, 2003). In brief, the EM algorithm casts the maximum likelihood problem as a 'missing data' problem. The missing data $Z_{i,t}$ has value 1 if the i th model ensemble is the best prediction at time t and value 0 otherwise. Hence, at any time t, only one of $\{Z_{1,t},Z_{2,t},\cdots Z_{k,t}\}$ is equal to 1, and the rest are zeros. The EM algorithm starts with an initial guess for θ, and then alternates between the expectation step, which estimates $Z_{i,t}$ on the current value of θ, and the maximization step, where new value of θ is estimated given the current value of $Z_{i,t}$. The expectation and maximization steps are repeated continually until certain convergence criteria are satisfied. The detailed description of the EM algorithm is given in Supporting Information A. In this study, the half-hourly EC-measured λET data from each of the 17 flux towers in 2008 was used for BMA training, whereas data set in 2009 was used to evaluate the performance of the BMA method during an independent validation period.

In addition, a SA method to merge the four ET model estimates (with equal weights) was included to develop an overall evaluation of performance, which can be expressed as:

$$\lambda\mathrm{ET}_{\mathrm{SA}}=\frac{1}{K}\sum_{i=1}^{K}\lambda\mathrm{ET}_i \quad (34)$$

where $\lambda\mathrm{ET}_{\mathrm{SA}}$ and $\lambda\mathrm{ET}_i$ are half-hourly actual ET predicted using the SA method and each individual ET model ($\mathrm{W\,m^{-2}}$), respectively.

Analysis of model-data mismatch

The model performance was quantified using statistical analysis based on half-hourly λET for each model-data pair. Model-data mismatch was evaluated using the coefficient of determination (R^2), slope, y-intercept, bias, root-mean-square error (RMSE), relative error (RE), and the Nash–Sutcliffe efficiency coefficient (NSE) (Legates and McCabe, 1999). The skills were calculated as:

$$\mathrm{Bias}=\frac{\sum_{t=1}^{n}(O(t)-M(t))}{n} \quad (35)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}[O(t)-M(t)]^2} \quad (36)$$

$$\mathrm{RE}=\frac{\mathrm{RMSE}}{\overline{O}} \quad (37)$$

$$\mathrm{NSE}=1-\frac{\sum_{t=1}^{n}[O(t)-M(t)]^2}{\sum_{t=1}^{n}\left(O(t)-\overline{O}\right)^2} \quad (38)$$

where n is the total number of observations; $O(t)$ is the observed values at time t, $\bar{O}$ is the mean of the observed data, and $M(t)$ is the simulated λET value at time t. NSE indicates how well the scatter plot observed versus simulated data fits the 1:1 line and ranges between $-\infty$ to 1, with a NSE = 1 being the optimal value.

A final characterization of model performance uses the Taylor diagram (Taylor, 2001), in which a single point indicate the coefficient of determination (R), the ratio of the standard deviations between the simulation and the observation ($\sigma_{\mathrm{norm}}=\sigma_{\mathrm{m}}/\sigma_{\mathrm{o}}$, where σ_{m} and σ_{o} are the standard deviations of simulation and observation, respectively) and the root-mean-square difference of the two patterns on a two-dimensional plot. More generally, each point of the Taylor diagram for any arbitrary data group can be scored as:

$$S=\frac{2(1+R)}{(\sigma_{\mathrm{norm}}+1/\sigma_{\mathrm{norm}})^2} \quad (39)$$

where S is the model skill metric bound by zero and unity where unity indicates perfect agreement with observations.

RESULTS

Performance of four ET models over the entire data period

Comparisons between observed and simulated half-hourly λET in Table II and Figure 2 provided an overview of performances among the four ET models over the entirety of the available period of data collected for each tower. Generally, the four models had a similar R range (i.e. mostly ranging between 0.70 and 0.95; Figure 2), but S–W and PT-JPL performed better than P–M and A–A with σ_{norm} closer to 1 (Figure 2), lower Bias, RMSE and RE, and greater NSE (Table II). The diagram of Taylor

Table II. Summary of statistical performance of the four ET models over different biomes over the entire data period (2008–2009). The bold number represents the best performance in each site

Site No.	R^2				Slope				Bias				RMSE				RE				NSE			
	P–M	S–W	PT-JPL	A–A	P–M	S–W	PT-JPL	A–A	P–M	S–W	PT-JPL	A–A	P–M	S–W	PT-JPL	A–A	P–M	S–W	PT-JPL	A–A	P–M	S–W	PT-JPL	A–A
1	0.65	**0.70**	0.65	0.64	2.00	**1.09**	2.00	1.99	-64.5	3.67	**−2.2**	-27.1	139	**48.8**	71.9	125	2.51	**0.99**	1.36	2.47	-3.06	**0.48**	-0.06	-2.32
2	**0.73**	**0.73**	0.65	0.63	2.84	**0.49**	1.73	2.75	-66.9	15.5	**−14.2**	-29.5	113	**25.7**	53.1	99.8	3.69	**0.98**	2.01	3.74	-8.55	**0.45**	-1.38	-7.13
3	**0.6**	0.53	0.49	0.43	2.20	**0.72**	1.60	2.81	-28.1	11.9	**−6.55**	-37.8	79.7	**28.3**	61.1	139	2.89	**1.02**	2.2	4.78	-4.16	**0.35**	-2.02	-14.0
4	0.59	**0.70**	0.63	0.57	1.82	**0.95**	1.20	1.89	-56.8	7.42	**−0.74**	-28.4	134	**42.5**	65.8	132	2.32	**0.87**	1.32	2.48	-2.89	**0.60**	0.06	-2.62
5	**0.84**	**0.84**	**0.84**	0.82	1.68	1.17	**0.96**	1.52	-58.8	**2.65**	22.61	-5.55	129	**59.2**	52.1	97.2	1.35	**0.75**	0.63	1.15	-0.26	0.71	**0.79**	0.23
6	0.68	**0.81**	0.80	0.81	1.36	**0.96**	0.89	1.39	-35.3	11.5	16.7	**−1.91**	96.2	**47.6**	49.8	77.7	1.24	**0.61**	0.59	0.97	0.04	0.77	**0.78**	0.41
7	**0.76**	0.75	0.69	0.58	1.93	0.68	**1.20**	1.27	-58.4	23.7	**−2.04**	21.3	121	**44.3**	60.4	94.5	2.15	**0.79**	1.07	1.34	-1.70	**0.64**	0.33	-0.33
8	0.62	**0.63**	0.60	0.61	1.55	0.89	**1.02**	1.29	-68.3	**−0.67**	-10.6	-42.9	158	**76.1**	89.6	135	1.95	**0.96**	1.13	1.16	-1.19	**0.52**	0.29	-0.27
9	**0.68**	0.67	0.67	0.63	1.36	0.80	**0.95**	1.23	-52.1	**1.50**	-3.67	-22.9	120	**64.8**	71.7	108	1.65	**0.89**	0.99	1.49	-0.23	**0.64**	0.56	0.01
10	0.75	**0.77**	**0.77**	0.74	1.44	0.89	**1.04**	1.37	-38.9	**5.70**	-4.45	-16.3	100	**49.2**	54.1	85.7	1.39	**0.68**	0.76	1.15	-0.06	**0.75**	0.69	0.19
11	**0.82**	**0.82**	0.81	0.80	1.29	0.80	**0.83**	1.18	-36.6	30.7	33.7	**2.83**	119	**75.6**	79.9	105	0.89	**0.62**	0.66	0.84	0.49	**0.79**	0.76	0.61
12	0.72	**0.75**	0.65	0.76	1.32	**0.95**	0.66	1.44	-38.8	**11.3**	39.1	-29.3	124	**75.4**	87.7	128	1.11	**0.69**	0.81	1.17	0.15	**0.68**	0.56	0.09

skill provided further statistical details on the model performances (Figure 3). The values of Taylor skill (*S*) for the S–W model over all towers showed a very narrow range between 0.69 and 0.95 with a mean of 0.88 (Figure 3a), which indicated that S–W model has a good overall performance. The second good model is PT-JPL model, and the values of *S* varied between 0.52 and 0.96 with a mean of 0.80 (Figure 3a). Among the models, the performance of the P–M model is ranked third (i.e. *S* ranging from 0.35 to 0.84 with a mean of 0.63; Figure 3 a), while the A–A model exhibited a relative large variation in *S* values ranging from 0.16 to 0.88 with a mean of 0.61 (Figure 3a). Further statistical details on the performance of the models are provided as scatter plots and summary tables in the Supporting Information B.

As expected, the model performances showed a high degree of site-specific variation. Generally, the model performances were satisfactory over croplands with the mean *S* values ranging from 0.72 to 0.89 (Figure 3b). However, except for the two alpine grasslands on Qinghai–Tibetan plateau (i.e. Maqu and Arou), the models have lower performances over the other four arid/semiarid grasslands (Figure 3b), which was mainly attributed to the significant overestimations of λET by the P–M and A–A models with $\sigma_{\text{norm}} > 1.5$ and slopes > 2 (Figure 2 and Supporting Information B). The relationship between soil moisture and ET fraction ($\text{EF} = \lambda\text{ET}/\lambda\text{ET}_{\text{potential}}$; where $\lambda\text{ET}_{\text{potential}}$ was potential daily ET which was calculated by setting $r_s = 0$ s m^{-1} in Equation 1; W m^{-2}) over the grasslands was presented to distinguish the soil moisture-limited and energy-limited ET regimes (Figure 4). Soil moisture in two alpine grasslands on Qinghai–Tibetan plateau was maintained at a relatively high level (>25%), and EF was independent of the soil moisture content. In contrast, soil moisture in arid/semiarid grasslands mainly lied in a transitional regime (i.e. 4–25%), and EF was a linear function of soil moisture content (Figure 4), confirming that ET over these ecosystems are mainly constrained by soil moisture supply (Seneviratne *et al.*, 2010). Thus, it seems to be challenges for the P–M and A–A models to accurately estimate λET over soil moisture-limited ecosystems in arid/semi-arid regions, where errors may attribute to the heterogeneous land surface conditions (i.e. low LAI; data not show) or the soil moisture stress on ET processes (discussed below).

Testing the BMA scheme using the data from training period

The BMA scheme was applied to obtain a set of BMA weights for each site using the measured half-hourly λET data from the training period (2008 year). The weights for the four ET models across all sites are shown in Figure 5, from which we can visually notice that the BMA weights

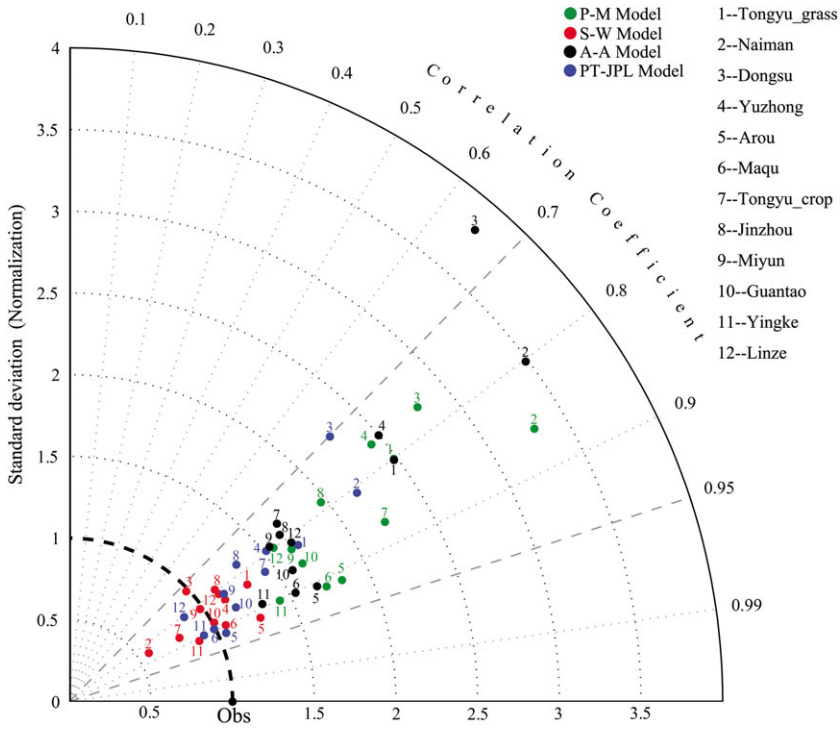

Figure 2. Performance of the four individual ET models for the 12 selected tower sites (number 1–12). Statistics in the Taylor diagram are derived from simulated and observed half-hourly λET fluxes during two year periods. An ideal model would have a standard deviation ratio (σ_{norm}) of 1.0 and a correlation coefficient of 1.0 (Obs, the reference point)

roughly reflects the individual model performance at each site. For examples, S–W ranked first in RE and NSE performances for arid/semiarid grasslands in the training period (Table S1 in Supporting Information B), and gained the highest weight ranging from 0.38 to 0.42; PT-LPJ performed best for two alpine grasslands on Qinghai–Tibetan plateau and has the highest weight (i.e. 0.32 and 0.31 for Arou and Maqu, respectively). In addition, the relative contributions of the four ET models varied for different sites. For examples, the A–A model weights for arid/semiarid grasslands (i.e. 0.11–0.15) were lower than that for crops (i.e. 0.18–0.21) and that for two alpine grasslands on Qinghai–Tibetan plateau (i.e. 0.17 and 0.19 for Arou and Maqu, respectively). On the contrary, the S–W model weights decrease from 0.38 to 0.42 for arid/semiarid grasslands to 0.29–0.37 for crops and to 0.27 and 0.28 for Arou and Maqu, respectively. Thus, the BMA weights did indeed reflect relative model performance over different sites.

During the training period, estimates of half-hourly λET calculated using the BMA method were compared with those for the SA method and the individual ET models for each site (Figure 6a and Supporting Information B). The most prominent merit of the BMA method is its robustness over all sites as indicated by a narrow variations in regression slope (0.99–1.01) and σ_{norm} (1.08–1.38) values. If the mean values of model skill (S) for all towers were considered as measures of model performances, the BMA method presented the best overall performance (Figure 6). Notably, the predictive capabilities of the BMA method for some sites (i.e. Dongsu, Miyun, Guantao, and Yingke) were similar or not better than that of the best performing model (S–W) in the ensemble members, as indicated by relative larger σ_{norm}, RMSE and RE, and lower NSE values (Figure 6a and Table S1 in Supporting Information B). However, the SA method tended to overestimate λET over the arid/semiarid grasslands with $\sigma_{norm} > 2$, and is ranked in the forth (after BMA, S–W and PT-JPL) in the overall model performances. This suggested that SA merging multiple model estimations does not necessarily yield rational λET estimates, especially over the moisture-limited land surface conditions.

Validation of BMA predictions using data from independent periods

The previous sections show that the BMA scheme is a promising tool for yielding proper λET estimations during the training period. A natural question to ask is how the BMA predictions perform when they are evaluated using data from an independent validation period (2009). In this section, we used the weights (w_i) and bias correction coefficients (a_i and b_i, $i = 1, \cdots, 4$) of each model obtained from the training periods to compute BMA predictions for

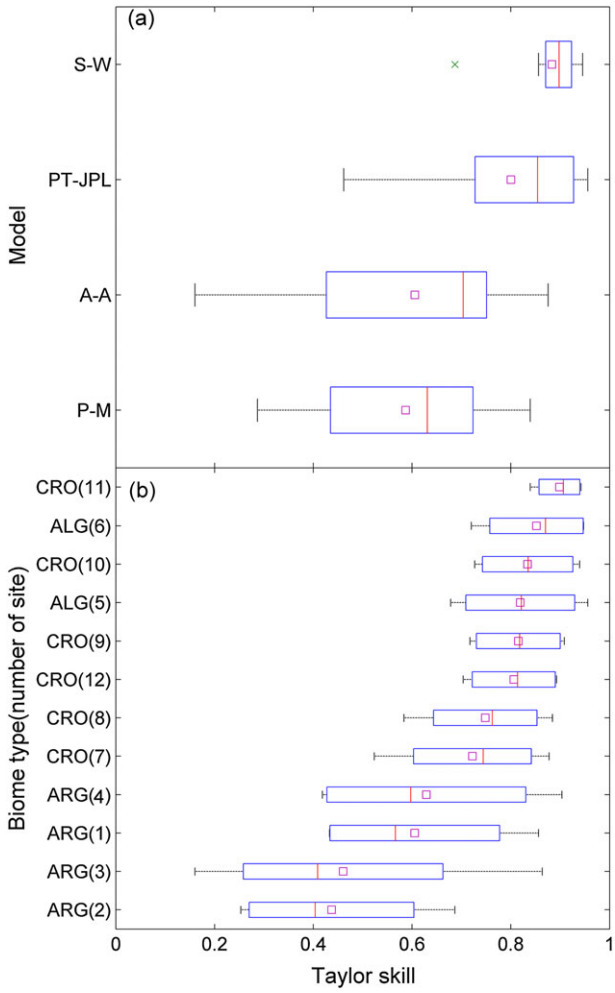

Figure 3. Boxplots of Taylor skill (*S*) for half-hourly λET by (**a**) every model over the selected 12 sites, and (**b**) each site represented by the number in the bracket of the four models. Panels show interquartile range (box), mean (square), median (solid line), range (whiskers), and outliers (cross). CRO = cropland, ALG = alpine grassland, ARG = arid/semiarid grassland. The site number is given in Table I

the validation period. In term of performance statistics (i.e. R^2, RMSE, RE and NSE), the BMA predications in the validation period was similar to that in the training period (Supporting Information B), suggesting that the BMA predictions were consistent over different periods. Notably, the Taylor skill for BMA method (i.e. 0.83–0.95) showed a relatively wider variation over different sites than that for S–W (i.e. 0.87–0.95), and the mean value of Taylor skill for BMA method (0.89) was slightly lower than that for S–W (0.91) (Figure 6b). The results indicated that the overall performance of the BMA predictions in the validation period were not necessary better than the best performing model (S–W) in the ensemble. However, the advantage of using BMA method was still obvious compared to the SA method and other individual models (Figure 6b and Table S2 in Supporting Information B).

One feature of the BMA method is that it can derive probabilistic ensemble predictions from competing individual deterministic predictions (see details in Supporting Information C). In this study, we generated 1000 BMA ensemble predictions to get a reasonable empirical probability density function (PDF) at each time step. Ideally, the spread of the ensemble predictions should be as small as possible, but consistent with observations, so that the predictive PDF is as sharp as possible (Vrugt and Robinson, 2007). For each biome type, two sites (where the mean Taylor skill value of models was highest and lowest, respectively; Figure 2) were selected to evaluate the precision of the BMA probabilistic predications. Figure 7 presented the excepted BMA predictions given by Equation 30 along with the 95% confidence interval of the BMA ensemble for a representative day in each month at the selected sites. To put Figure 7 in a proper perspective, the corresponding predictions derived from the SA method and the best individual model were also shown. Over sites where the mean Taylor skill value was highest for the grass and crop ecosystems (i.e. Arou and Yingke), the uncertainty bounds defined as the 95% confidence intervals are sharp and consistent with the observations (Figures 7a and 7c). Also, the excepted BMA predications at each site were comparable with the best performing model in the ensemble (i.e. PT-JPL and S–W at Arou and Yingke, respectively). On the contrary, the uncertainty ranges were clearly inconsistent with the observations (i.e. significant overestimations) over sites where the mean Taylor skill value was lowest for the grasslands and croplands (i.e. Dongsu and Tongyu; Figures 7b and 7d), and there is some degradation in performance of the excepted BMA predications comparing to the best model of the ensemble (S–W model). Thus, the skill and performance of the individual members of the ensemble ultimately determine the success of BMA-derived predications. There is no guarantee that the BMA deterministic predications were certainly superior to that of the best model in the ensemble at particular sites.

DISCUSSION

Evaluation of the individual model performance

The A–A model requires only widely available meteorological measurements to estimate actual ET without detailed information about the surface biophysical and hydrological states (Szilagyi and Jozsa, 2008). For this reason, there has been a renewed interest among hydrologists in evaluating and testing the model and its underlying assumptions (see details in Crago and Qualls, 2013). In this study, we found that the A–A model showed an overall low performance as indicated by

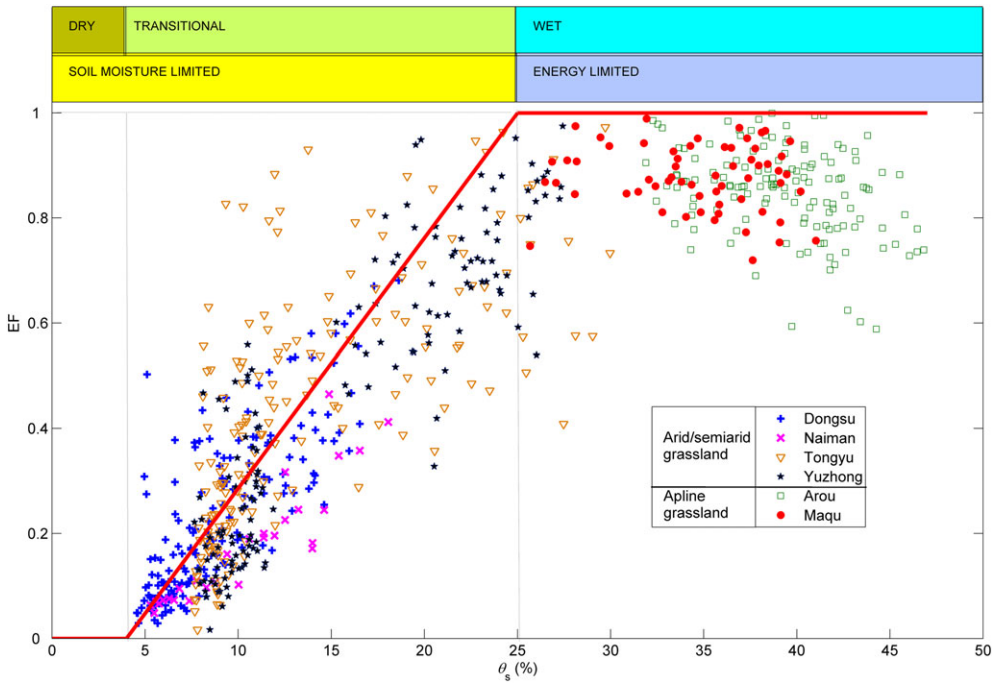

Figure 4. The relationship between the soil moisture regimes and corresponding evapotranspiration regimes of the grasslands in north China. EF denotes the evapotranspiration fraction

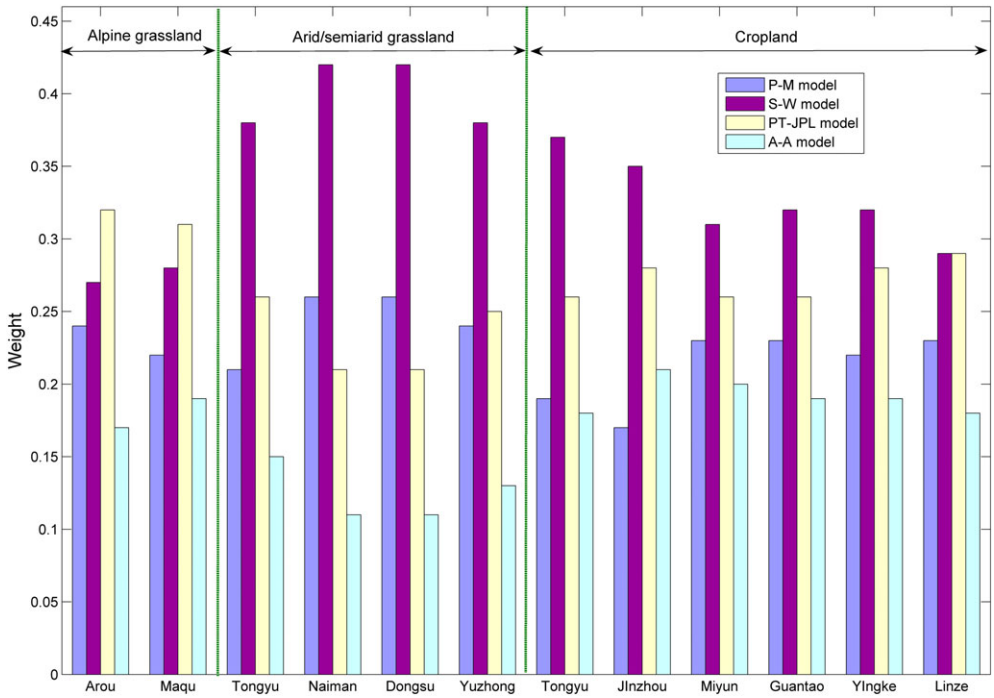

Figure 5. The BMA weight computed over the entire training period (2008) across all sites in north China. The models that perform better receive higher weights than that perform worse

relatively large overestimation of half-hourly λET across all tower sites (i.e. regression slope ≥1.09; Tables in Supporting Information B). Similar model performance has been reported by Ershadi *et al.* (2014), with slope ≥1.05 for an average 5-year period using half-hourly or hourly data in 20 FLUXNET sites over a wide range of biomes. To get better match between observed and estimated ET, some authors stated that the Priestley–

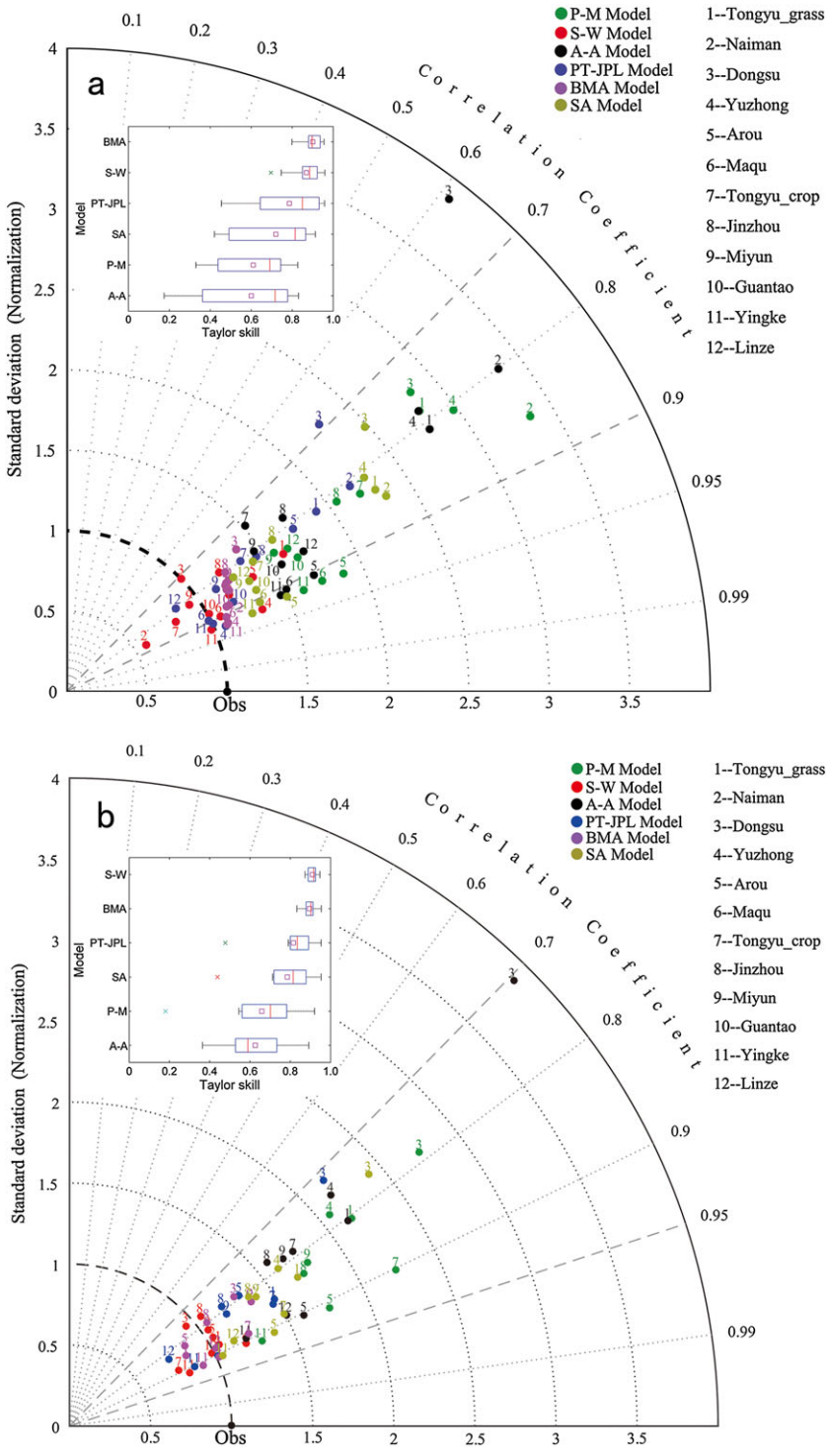

Figure 6. Performance of the four individual ET models for the 12 selected tower sites (number 1–12). (**a**) training period; (**b**) validation period. The inserted plot was the Taylor skill (*S*) of model over the selected 12 sites for corresponding period. Data for Naiman (site no. 2), Maqu (no. 6) and Guantao (no.10) was not available in 2009

Taylor coefficient α_{PT} (Hobbins *et al.*, 2001; Xu and Singh, 2005; Szilagyi and Jozsa, 2008; Gao *et al.*, 2011) or the proportionality constant b (Kahler and Brutsaert, 2006; Szilagyi, 2007; Szilagyi *et al.*, 2009; Han *et al.*, 2012) should be treated as parameters and need to be calibrated over varying land states (e.g. soil moisture; Garcia *et al.*, 2009) and climate conditions (e.g. seasonality; Yang *et al.*, 2013). However, the calibration-free merit of the A–A

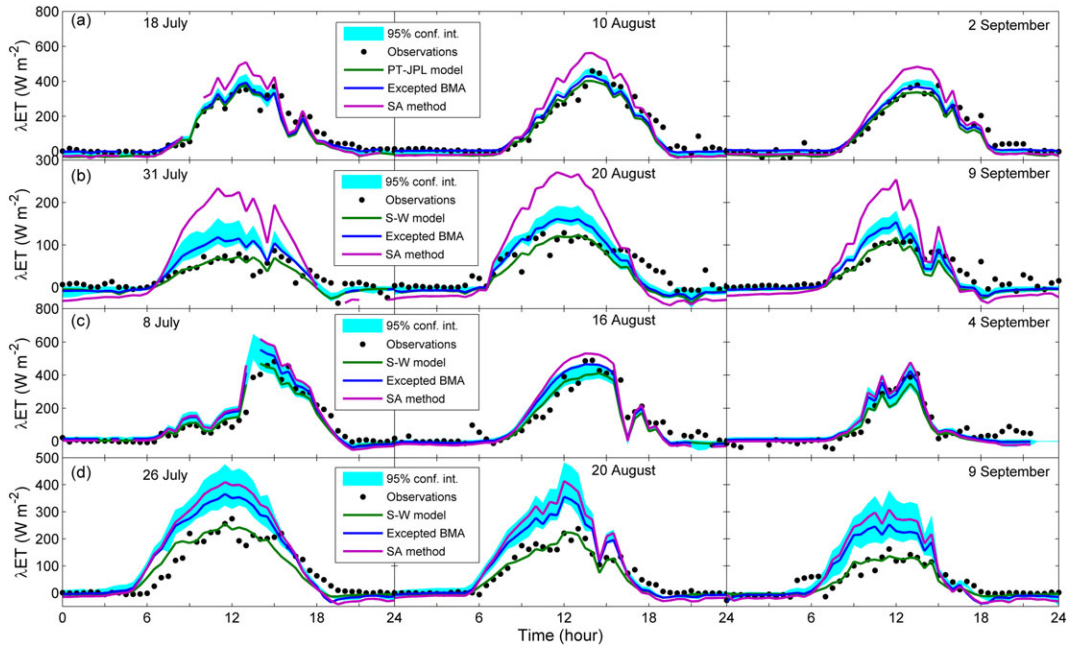

Figure 7. Expected BMA predictions and 95% confidence interval compared to observations at selected sites. (**a**) Arou, (**b**) Dongsu, (**c**) Yingke, and (**d**) Tongyu

model will ultimately be lost by doing so. For different locations, we found that the prediction precision of the A–A model over croplands (which is often associated with high soil water availability because of irrigation) and two alpine grasslands on Qinghai–Tibetan plateau were better than that over the four soil moisture-limited grasslands in north China (Figure 3 and Supporting Information B). The results confirmed the conclusion of previous studies, which showed that the predictive power of the A–A model increases in moving toward regions of increased energy control of ET rates (i.e. humid regions) and decreases in moving toward regions of increased soil moisture control (i.e. arid regions) (Lemeur and Zhang, 1990; Hobbins *et al.*, 2001; Xu and Li, 2003; Xu and Singh, 2005). Thus, it may still be challenging for the A–A model to properly descript the soil–moisture constraints on ET processes over arid environments.

Despite of the common theoretical basis (i.e. the Penman model; Penman, 1948), the P–M and S–W models performed significantly different in our study. Generally, the P–M model overestimated half-hourly λET across all sites, especially over the four soil moisture-limited grasslands (i.e. slope ≥ 1.73; Tables in Supporting Information B); the S–W model showed a satisfactory accuracy in the estimation of half-hourly λET over all sites. These results are in general agreement with some previous studies. For example, Zhang *et al.* (2008) compared the performance of the P–M and S–W models with measured half-hourly λET over a vineyard located in arid region of northwest China, and found the P–M model overestimated λET significantly, while the estimated λET from the S–W model was approximately equal to the measured λET. However, some studies reported that the P–M and S–W models yielded similar results (Fisher *et al.*, 2005). Such difference in model performance may be strongly related to the effects of soil moisture stress and variations in LAI (Burba and Verma, 2005). In the study of Fisher *et al.* (2005), the dense canopy ($LAI > 2.9\,m^2\,m^{-2}$) and high vegetation cover (more than 70%) of the conifer forest in Northern California resulted in insignificant contribution of soil evaporation to the total ET. Under such conditions, the S–W model reduced back to the P–M model and gave similar results. But in sparse canopies such as the soil moisture-limited grasslands in this study, soil surface resistance (about $1000\,s\,m^{-1}$ for dry soil; Dorman and Sellers, 1989) was generally higher than the canopy resistance (about $200–400\,s\,m^{-1}$; Dorman and Sellers, 1989; Zhang *et al.*, 2008). Thus, the P–M model overestimates λET because the canopy resistance in the model is lower than the actual surface resistance, which is an integration of canopy and soil surface resistances (Stannard, 1993). Another possible explanation for the overestimation of the P–M model might be related to model parameters (i.e. r_{smin}). The value of r_{smin} used in this study was relative smaller than some locally calibrated values of previous studies. For example, Zhang *et al.* (2008) reported the optimal value of r_{smin} for vineyard in northwest China was $146\,s\,m^{-1}$; r_{smin} was set to be $252\,s\,m^{-1}$ for maize in North Italy (Gharsallah *et al.*, 2013). Although constant parameter values were used, the S–W still performed best among the models, which indeed

highlighted the import influence of model structure on modelling ET (Ershadi *et al.*, 2015). It can be expected that the performance of the S–W model would further be improved using locally calibrated parameters. This has been confirmed by numerous studies (Stannard, 1993; Teh *et al.*, 2001; Zhang *et al.*, 2008; Zhu *et al.*, 2013, 2014b). Thus, the models with ET partitioning structure (canopy transpiration and soil evaporation) were highly recommended for ET simulations especially in semiarid and arid areas.

The PT-LPJ mode relies on atmospheric and ecophysiological constraints to scale down the Priestley–Taylor model (Priestley and Taylor, 1972). Despite its simplicity (i.e. not requiring specification of aerodynamic and surface resistances), the PT-JPL model performed well in croplands and two alpine grasslands on Qinghai–Tibetan plateau (Figure 2). Similar model performance has been reported in previous studies (Fisher *et al.*, 2008, 2009; Vinukollu *et al.*, 2011; Ershadi *et al.*, 2014). In this study, we found that the PT-JPL model exhibited reduced performance over the four sparse and soil moisture-limited grasslands in north China (i.e. overestimation with slope ranging from 1.03 to 1.76; Tables in Supporting Information B). This may be mainly attributed to errors in the parameterization of soil moisture constraint f_{SM} (= $RH^{D/\beta}$). Recently, García *et al.* (2013) reported that the model using the original parameterization of f_{SM} ($\beta = 1$ kPa) did not provide meaningful λET estimation over the Mediterranean grasslands (i.e. $R^2 \sim 0.16$ and negative bias $\sim -16.5\,W\,m^{-2}$ which indicated an overestimation). However, the model performance was significantly improved (i.e. $R^2 = 0.53$–0.64 and bias = 15–17 $W\,m^{-2}$) by setting $\beta = 0.1$ kPa. Similar to works of García *et al.* (2013), we also compared the model performance with two different values of β across all sites during the whole data available period (Table III). The PT-JPL model performed better (i.e. slope closer to 1, smaller RMSE and larger NSE) using $\beta = 0.1$ kPa over the four soil moisture-limited grasslands, while better results corresponding to $\beta = 1$ kPa were found over the alpine grasslands on Qinghai–Tibetan plateau. Over the croplands, the differences in model performance by using the two β values were not significant. Thus, parameterization using f_{SM} should be tuned according to the conditions for successful results (García *et al.*, 2013). Until now, much attention has been paid to properly parameterize f_{SM}, such as using Apparent Thermal Inertia (ATI) (García *et al.*, 2013; Yao *et al.*, 2013) and *in site* measured volumetric soil water content (García *et al.*, 2013). In the future, more works on the performance intercomparisons of different f_{SM} parameterizations across a wide range of biomes in different climatic regions may be needed. In addition, the uncertainty in remote sensing data (i.e. NDVI) may also cause errors in λET estimation (Ershadi *et al.*, 2014).

Evaluation of the BMA method performance

Because of intrinsic uncertainty in model structure, predictions from a single model often lead to over-confidence and significant bias (Hoetting *et al.*, 1999; Raftery *et al.*, 2003; Parrish *et al.*, 2012). Multimodel ensemble approaches have therefore become increasing popular in climate change projections (Tebaldi *et al.*, 2006; Yang *et al.*, 2012), land surface component simulation (i.e. soil moisture, Guo *et al.*, 2007; surface longwave radiation, Wu *et al.*, 2012), groundwater assessment (Neuman, 2003), hydrologic streamflow predictions (Duan *et al.*, 2007; Vrugt and Robinson, 2007; Zhang *et al.*, 2009), and terrestrial ET estimation (Ershadi *et al.*, 2014; Yao *et al.*, 2014). With the

Table III. Evaluation of PT-JPL half-hourly λET with EC data for $\beta = 1$kPa and $\beta = 0.1$ kPa. The bold number represents the best performance in each site

Site No.	$\beta = 1$						$\beta = 0.1$					
	R^2	Slope	Bias	RMSE	RE	NSE	R^2	Slope	Bias	RMSE	RE	NSE
1	0.65	2.00	-2.2	71.9	1.36	-0.06	**0.67**	**1.32**	3.13	**56.9**	1.25	**0.04**
2	**0.65**	1.73	-14.2	53.1	2.01	-1.38	**0.65**	**1.63**	**−7.05**	**47.0**	**1.80**	**−0.84**
3	0.49	1.60	-6.55	61.1	2.2	-2.02	**0.51**	**1.16**	**1.76**	**40.5**	**1.43**	**−0.30**
4	0.63	1.20	-0.74	65.8	1.32	0.06	**0.65**	**0.99**	9.68	**44.5**	0.99	**0.46**
5	0.84	**0.96**	**22.6**	52.1	**0.63**	**0.79**	**0.86**	0.88	29.9	51.1	0.66	0.78
6	**0.80**	**0.89**	**16.7**	**49.8**	**0.59**	**0.78**	**0.80**	0.79	24.6	51.6	0.64	0.74
7	**0.69**	1.20	-2.04	60.4	1.07	0.33	0.63	**0.86**	12.3	**50.7**	0.9	**0.52**
8	0.60	**1.02**	-10.6	89.6	1.13	0.29	**0.67**	1.14	-16.3	**78.1**	1.20	0.31
9	0.67	**0.95**	-3.67	71.7	0.99	0.56	**0.69**	0.91	**1.66**	**71.2**	**0.90**	**0.61**
10	**0.77**	1.04	**−4.45**	**54.1**	**0.76**	**0.69**	0.75	**1.01**	-5.25	57.5	0.86	0.66
11	0.81	0.83	33.7	79.9	0.66	0.76	**0.85**	**0.88**	**28.0**	**63.0**	**0.60**	**0.81**
12	0.65	**0.66**	**39.1**	87.7	**0.81**	0.56	**0.70**	0.62	44.5	**84.7**	**0.81**	**0.57**

ensemble techniques, the SA approach focused on point predictions, while the BMA approach was mainly concerned with producing bias-corrected probabilistic forecasts (Diks and Vrugt, 2010). Various case studies have shown that the BMA approach can produce more accurate and reliable predictions than the SA method (Barnston *et al.*, 2003; Vrugt *et al.*, 2008). Here, we also found that the BMA approach in performances was superior to the SA approach both in the training and validation periods. For example, the average RMSE over all sites is 55.6 $W m^{-2}$ and NSE = 0.62 for the BMA approach, while the average RMSE is 79.7 $W m^{-2}$ and NSE = −0.19 for the SA approach (Tables in Supporting Information B). This may because the BMA considers a weighted average of ensemble distribution that is centred on the bias-corrected individual forecasts with data over a training period, whereas the SA method directly averages individual forecasts without using any auxiliary data (Raftery *et al.*, 2003). Recently, Yao *et al.* (2014) showed that the BMA approach by merging five satellite-based ET algorithms yielded better ET results compared to the SA approach. Noticeably, the SA approach in this study overestimated λET in most cases and ranked fourth (after BMA, S–W and PT-LPJ) in overall model performances. This was different from the results of Ershadi *et al.* (2014), who reported that SA of four ET models produced the best λET estimations over 20 FLUXNET sites. Thus, the performance of the SA approach seems to vary over different biomes and climatic conditions.

Successful implementation of BMA needs proper estimates of its related parameters (i.e. w_i, σ_i, a_i, and b_i) of the individual models in the ensemble (Vrugt *et al.*, 2008; Sloughter *et al.*, 2010). In our implementation, these parameters were fitted from a given training period (2008) and assumed to be static over calibration period (2009). It was found that the BMA method overall performed best during the training period (Figure 6a). Recently, Chen *et al.* (2015) also illustrated that the BMA method can outperform the best of eight satellite-based ET models in China. However, some degradation in BMA performance was observed during the calibration period, and its predication skill at some sites (i.e. Dongsu, Tongyu, Yuzhong) was equal to or not better than the best performing model (S–W) in the ensemble (Figure 6b). Thus, the essential avenue in obtaining reliable large scale ET estimation still greatly depends on the developments of physically-based accurate and applicable ET models (Wang and Dickinson, 2012). This finding is consistent with results from previous studies (Pavan and Doblas-Reyes 2000; Palmer *et al.* 2000; Peng *et al.* 2002; Barnston *et al.*, 2003; Georgekakos *et al.*, 2004; Vrugt *et al.*, 2008; Zhang *et al.*, 2009). The degradation in BMA performance may be attributed to the following reasons. First, the static assumption of the BMA parameters may not be true in real world. In this study, we found that the regression relationship between observed and simulated λET of each model was generally different in the two years (Tables in Supporting Information B). Thus, the static coefficients (a_i and b_i) may produce biassed

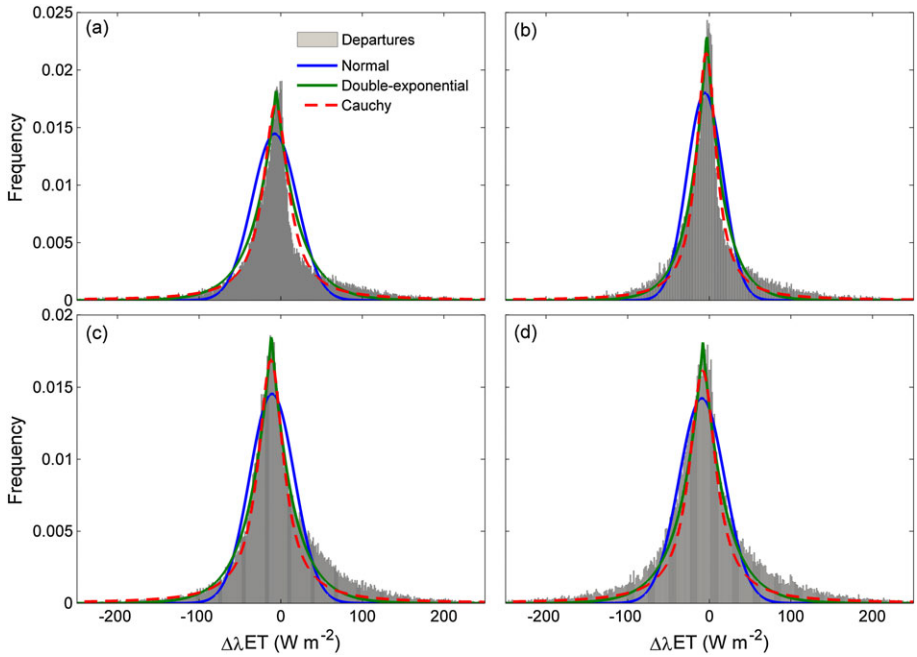

Figure 8. Histograms depicting the frequency distribution of the departures between observed and bias-corrected simulations for the four ET models. (**a**) P–M, (**b**) S–W, (**c**) PT-JPL, and (**d**) A–A

predictions in the validation period. To date, various adaptations of BMA have been proposed to solve this problem. For example, Raftery *et al.* (2005) used the sliding window technique to obtain unbiassed forecasts. That is, the training period is limited to a shorter sliding window surrounding the forecast, and each forecast is based on the dynamically updated parameters. Hsu *et al.* (2009) proposed the use of sequential Bayesian approach for recursively adjusting the BMA parameters. These approaches may provide better forecasts for state-space models. However, the majority of ET algorithms do not belong to these type models, which makes these approaches unsuitable for λET estimation. Thus, a feasible solution to this problem is to split the training period into different intervals according to growing seasons (i.e. LAI and land cover fraction), and compute separate BMA parameters for each interval. This is the scope of our future work.

Second, the conditional density function $p_i(y|f_i, \mathbf{D})$ is widely recognized to have significant influences on the performance of the BMA approach (Raftery *et al.*, 2005; Vrugt *et al.*, 2008; Duan and Phillips, 2010; Yao *et al.*, 2014). Generally, normal density works well for weather quantities such as temperature, sea-level pressure (Raftery *et al.*, 2005; Vrugt *et al.*, 2008; Miao *et al.*, 2013), shortwave and longwave radiation (Wu *et al.*, 2012), and Gamma distribution provides good fits to precipitation products (Vrugt *et al.*, 2008; Sloughter *et al.*, 2010; Yang *et al.*, 2012). As λET is a variable that links energy, water, and plant productivity, a priori assessment of the conditional density for the BMA method may be difficult (Zhu *et al.*, 2014b; Yao *et al.*, 2014). Figure 8 showed the distribution of the departures of bias-corrected simulations and observations of half-hourly λET, which were approximated better by the double-exponential and Cauchy PDFs. These results were in agreement with previous studies (Richardson *et al.*, 2006; Zhu *et al.*, 2014b). However, the Cauchy distribution may not be appropriate for model-data fusion, because its first four moments are undefined (Richardson *et al.*, 2008).

Third, the performances of individual models have significant impacts on the accuracy of the BMA approach (Figure 7). In this study, fixed parameters were used for specified land cover types without considering the parameter variations over different land surface conditions and plant growing seasons. Bonan *et al.* (2012) and Chen *et al.* (2013) pointed out that uncertainties in model estimates were because of parameter errors are equivalent to those from different model structures. Our previous study also showed that the seasonal variations in parameters have significant impacts on long-term ET simulations (Zhu *et al.*, 2013). Unfortunately, a proper data set of model parameters for different land cover types and climate spaces is not available at present. To improve the accuracy of ET estimates at large scales, it is therefore urgently needed to calibrate the ET models using the data from FLUXNET sites over a wide range of biomes and climatic conditions. In addition, the accuracy of EC observations may have influences on the BMA performances because EC observations were assumed as true values in calculating the BMA parameters of individual ET models. Recently, Wang *et al.* (2014) systemically studied the flux uncertainties of EC systems equipped in the some CEOP sites, and reported that the uncertainties for λET were about 13%. However, measurement uncertainty was not explicitly accounted in present work. Thus, an integrated BMA framework that accounts for the parameter and measurement uncertainties is needed to improve the accuracy of long-term global terrestrial ET estimates (Ajami *et al.*, 2007).

CONCLUSIONS

In this study, four ET models and the multi-model ensemble approaches (i.e. SA and BMA) were evaluated at half-hourly time steps over 12 flux tower sites in north China. Although the focus of this paper was on evaluating the performances of the selected ET models at the tower scale, the results would be helpful in identifying proper models for generating robust regional or global ET products. When using tower-based forcing data, the S–W model, followed by the PT-JPL model, outperformed the single-source P–M model and the A–A model. Thus, the models with ET partitioning (i.e. soil evaporation and plant transpiration) structure were recommended for large scale or global ET simulations. However, some crucial variables (i.e. soil water content, vegetation cover and plant root depth) in controlling the ET process in water-limited areas are still unavailable at the regional or global scale. In the future, efforts by integrating *in-situ* measurements, satellite observations and data assimilation technique should been done to improve the estimations of these variables in semi-arid and arid areas. As far as the multi-model ensemble approaches were concerned, the BMA method yielded better performance than the SA method. During the validation period, there was some degradation of the BMA approach, which may be attributed to the improper assumption of static BMA parameters. Thus, it is still necessary to explore the seasonal variations of the BMA parameters according the different growth stages. Finally, the double-exponential probability distribution may be appropriate in the half-hourly λET context.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Tanya M Doody (Editor) for her continued help during the revisions of the

article. We also thank the anonymous reviewers for their critical reviews and helpful comments. This research was supported by the Chinese Academy of Sciences Action Plan for West Development Program Project (KZCX2-XB3-15), National Natural Science Foundation of China (nos. 31370467 and 41571016), the Fundamental Research Funds for the Central Universities (no. 861944), and the CAS Interdisciplinary Innovation Team of the Chinese Academy of Science.

REFERENCES

Ajami NK, Duan Q, Sorooshian S. 2007. An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.* **43**: W01403. DOI:10.1029/2005WR004745

Baldocchi D, Falge E, Gu L, Olson R, Hollinger D, Running S. 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities. *Bull. Am. Meteorol. Soc.* **82**(11): 2415–2434.

Barnston AG, Mason SJ, Goddard L, DeWitt DF, Zebiak SE. 2003. Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Am. Meteorol. Soc.* **84**: 1783–1796.

Bonan GB, Oleson KW, Fisher RA, Lasslop G, Reichstein M. 2012. Reconciling leaf physiological traits and canopy flux data: use of the TRY and FLUXNET databases in the Community Land Model version 4. *J. Geophys. Res.* **117**: G02026. DOI:10.1029/2011JG001913

Bouchet RJ. 1963. Evapotranspiration réelle, evapotranspiration potentielle, et production agricole. *Ann. Agron.* **14**: 743–824.

Brutsaert W. 1982. Evaporation into the atmosphere: theory history and applications. Reidel Publishing, Dordrecht etc., pp. 299.

Brutsaert W. 2005. *Hydrology: an introduction*. Cambridge Univ. Press: Cambridge; 605.

Brutsaert W, Stricker H. 1979. An advection–aridity approach to estimate actual regional evapotranspiration. *Water Resour. Res.* **15**(2): 443–450.

Burba GG, Verma SB. 2005. Seasonal and interannual variability in evapotranspiration of native tallgrass prairie and cultivated wheat ecosystems. *Agric. Forest Meteorol.* **135**(1–4): 190–201.

Chen F, Dudhia J. 2001. Coupling and Advanced land surface-hydrology model with the penn state-NCAR MM5 Modelling System. Part I: Model implementation and sensitivity. *Mon. Weather Rev.* **129**: 569–585.

Chen J, Chen B, Black TA, Innes JL, Wang G, Kiely G, Hirano T, Wohlfahrt G. 2013. Comparison of terrestrial evapotranspiration estimates using the mass transfer and Penman–Monteith equations in land surface models. *J. Geophys. Res. Biog.* **118**: 1715–1731.

Chen Y, Yuan W, Xia J, Fisher JB, Dong W, Zhang X, Liang S, Ye A, Cai W, Feng J. 2015. Using Bayesian model averaging to estimate terrestrial evapotranspiration in China. *J. Hydrol.* **528**: 537–549.

Crago RD, Qualls RJ. 2013. The value of intuitive concepts in evaporation research. *Water Resour. Res.* **49**: 6100–6104.

Diks CGH, Vrugt JA. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stoch. Environ. Res. Risk Assess.* **24**(6): 809–820.

Doody TM, Nagler PL, Glenn EP, Moore GW, Morino K, Hultine KR, Benyon RG. 2011. Potential for water salvage by removal of non-native woody vegetation from dryland river systems. *Hydrol. Process.* **25**: 4117–4131.

Dorman JL, Sellers PJ. 1989. A global climatology of albedo, roughness length and stomatal resistance for atmospheric general circulation models as represented by the Simple Biosphere model (SiB). *Journal of Applied Meteorology.* **28**: 833–855.

Drexler J, Snyder R, Spano D, Paw UKT. 2004. A review of models and micrometeorological methods used to estimate wetland evapotranspiration. *Hydrol. Process.* **18**: 2071–2101.

Duan Q, Ajami NK, Gao X, Sorooshian S. 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**(5): 1371–1386.

Duan Q, Phillips TJ. 2010. Bayesian estimation of local signal and noise in multimodel simulations of climate change. *J. Geophys. Res.* **115**: D18123. DOI:10.1029/2009JD013654

Ershadi A, McCabe MF, Evans JP, Chaney NW, Wood EF. 2014. Multi-site evaluation of terrestrial evaporation models using FLUXNET data. *Agricultural and Forest Meteorology.* **187**: 46–61.

Ershadi A, McCabe MF, Evans JP, Wood EF. 2015. Impact of model structure and parameterization on Penman–Monteith type evaporation models. *J. Hydrol.* **525**: 521–535.

Falge E, Baldocchi D, Olson R, Anthoni P, Aubinet M, Bernhofer C, Burba G, Ceulemans R, Clement R, Dolman H, Granier A, Gross P, Grünwald T, Hollinger D, Jensen N-O, Katul G, Keronen P, Kowalski A, Lai CT, Lawc BE, Meyers T, Moncrieff J, Moors E, Munger JW, Pilegaard K, Rannik Ü, Rebmann C, Suyker A, Tenhunen J, Tu K, Verma S, Vesala T, Wilson K, Wofsy K. 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agr. Forest Meteorol.* **107**: 43–69.

Fisher JB, DeBiase TA, Qi Y, Xu M, Goldstein AH. 2005. Evapotranspiration models compared on a Sierra Nevada forest ecosystem. *Environmental Modelling & Software.* **20**: 783–796.

Fisher JB, Whittaker RJ, Malhi Y. 2011. ET come home: potential evapotranspiration in geographical ecology. *Glob. Ecol. Biogeogr.* **20**: 1–18.

Fisher JB, Malhi Y, Bonal D, Da Rocha HR, De Arajuo AC, Gamo M. 2009. The land–atmosphere water flux in the tropics. *Global Change Biol.* **15**: 2694–2714.

Fisher JB, Tu KP, Baldocchi DD. 2008. Global estimates of the land atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Rem. Sens. Environ.* **112**: 901–919.

Flint AL, Childs SW. 1991. Use of the Priestley–Taylor evaporation equation for soil water limited conditions in a small forest clearcut. *Agric. Forest Meteorol.* **56**(3–4): 247–260.

Gao G, Xu C, Chen D, Singh VP. 2011. Spatial and temporal characteristics of actual evapotranspiration over Haihe River basin in China. *Stochastic Environmental Research and Risk Assessment.* **26**: 65–669.

Garcia CA, Andraski BJ, Stonestrom DA, Cooper CA, Johnson MJ, Michel RL, Wheatcraft SW. 2009. Transport of tritium contamination to the atmosphere in an arid environment. *Vadose Zone J.* **8**: 450–461.

García M, Sandholt I, Ceccato P, Ridler M, Mougin E, Kergoat L, Morillas L, Timouk F, Fensholt R, Domingo F. 2013. Actual evapotranspiration in drylands derived from in-situ and satellite data: assessing biophysical constraints. *Remote Sens., Environ.* **131**(0): 103–118.

Georgekakos KP, Seo DJ, Gupta H, Schaake J, Butts MB. 2004. Characterizing streamflow simulation uncertainty through multi-model ensembles. *J. Hydrol.* **298**(1–4): 222–241.

Gharsallah O, Facchi A, Gandolfi C. 2013. Comparison of six evapotranspiration models for a surface irrigated maize agro-ecosystem in Northern Italy. *Agricultural Water Management.* **130**: 119–130.

Guo Z, Dirmeyer PA, Gao X, Zhao M. 2007. Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Q. J. R. Meteorol. Soc.* **133**(624): 731–747.

Han S, Hu H, Tian F. 2012. A nonlinear function approach for the normalized complementary relationship evaporation model. *Hydrol. Process.* **26**: 3973–3981.

Hobbins MT, Ramírez JA, Brown TC. 2001. The complementary relationship in estimation of regional evapotranspiration: an enhanced advection–aridity model. *Water Resour. Res.* **37**: 1389–1403.

Hoetting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian modeling averaging: a tutorial. *Stat. Sci.* **14**(4): 382–417.

Hsu KL, Moradkhani H, Sorooshian S. 2009. A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resour. Res.* **45**: W00B12. doi: 10.1029/2008WR006824.

Hu ZM, Yu GR, Zhou YL, Sun XM, Li YN, Shi PL, Wang YF, Song X, Zheng ZM, Zhang L, Li SG. 2009. Partitioning of evapotranspiration and its controls in fourgrassland ecosystems: application of a two-source model. *Agric. Forest Meteorol.* **149**: 1410–1420.

Jarvis PG. 1976. The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philos. T. Roy. Soc. B.* **273**: 563–610.

Jiménez-Muñoz J, Sobrino J, Plaza A, Guanter L, Moreno J, Martinez P. 2009. Comparison between fractional vegetation cover retrievals from vegetation indices and spectral mixture analysis: case study of PROBA/CHRIS data over an agricultural area. *Sensors.* **9**(2): 768–793.

Jung M, Reichstein M, Ciais P, Seneviratne SI, Sheffield J, Goulden ML, Bonan G, Cescatti A, Chen J, de Jeu R, Dolman AJ, Eugster W, Gerten D, Gianelle D, Gobron N, Heinke J, Kimball J, Law BE, Montagnani L, Mu Q, Mueller B, Oleson K, Papale D, Richardson AD, Roupsard O, Running S, Tomelleri E, Viovy N, Weber U, Williams C, Wood E, Zaehle S, Zhang K. 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* **467**: 951–954.

Kahler DM, Brutsaert W. 2006. Complementary relationship between daily evaporation in the environment and pan evaporation. *Water Resour. Res.* **42**: W05413. DOI:10.1029/2005WR004541

Katul, GG, Oren, R, Manzoni, S, Higgins, C, Parlange, MB, 2012. Evapotranspiration: a process driving mass transport and energy exchange in the soil–plant–atmosphere–climate system. *Rev. Geophys.* **50**: RG3002. doi:10.1029/2011RG000366.

Legates DR, McCabe GJ. 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35**: 233–241.

Lemeur R, Zhang L. 1990. Evaluation of three evapotranspiration models in terms of their applicability for an arid region. *J. Hydrol.* **114**: 395–411.

Liu SM, Xu ZW, Wang WZ, Jia ZZ, Zhu MJ, Bai J, Wang JM. 2011. A comparison of eddy-covariance and large aperture scintillometer measurements with respect to the energy balance closure problem. *Hydrol. Earth Syst. Sci.* **15**: 1291–1306.

Miao C, Duan Q, Sun Q, Li J. 2013. Evaluation and application of Bayesian multi-model estimation in temperature simulations. *Prog. Phys. Geogr.* **37**(6): 727–744.

Monteith JL. 1965. Evaporation and environment. *Symp. Soc. Exp. Biol.* **19**: 205–234.

Mu QZ, Heinsch FA, Zhao M, Running SW. 2007. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote Sensing of Environment.* **111**: 519–536.

Mu Q, Zhao M, Running SW. 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment.* **115**(8): 1781–1800.

Mueller B, Seneviratne SI, Jimenez C, Corti T, Hirschi M, Balsamo G, Ciais P, Dirmeyer P, Fisher JB, Guo Z, Jung M, Maignan F, McCabe MF, Reichle R, Reichstein M, Rodell M, Sheffield J, Teuling AJ, Wang K, Wood EF, Zhang Y. 2011. Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. *Geophys. Res. Lett.* **38**(6): L06402.

Neuman SP. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stoch. Environ. Res. Risk Assess.* **17**(5): 291–305.

Noilhan J, Planton S. 1989. A simple parameterisation of land and surface process for meteorological models. *Am. Meteorol. Soc.* **117**: 536–549.

Ortega-Farias S, Olioso A, Antonioletti R, Brisson N. 2004. Evaluation of the Penman–Monteith model for estimating soybean evapotranspiration. *Irrig. Sci.* **23**: 1–9.

Ortega-Farias S, Olioso A, Fuentes S, Valdes H. 2006. Latent heat flux over a furrow-irrigated tomato crop using Penman–Monteith equation with a variable surface canopy resistance. *Agricultural Water Management.* **82**: 421–432.

Palmer TN, Brankovic C, Richardson DS. 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.* **126**: 2013–2033.

Parlange MB, Katul GG. 1992. An advection–aridity evaporation model. *Water Resour. Res.* **28**: 127–132.

Parrish MA, Moradkhani H, DeChant CM. 2012. Toward reduction of model uncertainty: integration of Bayesian model averaging and data assimilation. *Water Resour. Res.* **48**: W03519. DOI:10.1029/2011WR011116

Pavan V, Doblas-Reyes J. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. *Climate Dyn.* **16**: 611–625.

Peng P, Kumar A, van den Dool HM, Barnston AG. 2002. An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.* **107**: 4710. DOI:10.1029/2002JD002712

Penman HL. 1948. Natural evaporation from open water, bare soil and grass. Royal Society of London. Series App., 120–146.

Priestley CHB, Taylor RJ. 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. *Mon. Weather Rev.* **100** (2): 81–92.

Raftery AE, Balabdaoui F, Gneiting T, Polakowski M. 2003. Using Bayesian Model Averaging to calibrate forecast ensembles. Technical Report No.440. Department of Statistics. University of Washington.

Raftery AE, Madigan D, Hoeting JA. 1997. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**: 179–191.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**: 1155–1174.

Richardson AD, Hollinger DY, Burba GG, Davis KJ, Flanagan LB, Katul GG, Williammunger J, Ricciuto DM, Stoy PC, Suyker AE, Verma SB, Wofsy SC. 2006. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agr. Forest Meteorol.* **136**: 1–18.

Richardson AD, Mahecha MD, Falge E, Kattge J, Moffat AM, Papale D, Reichstein M, Stauch VJ, Braswell BH, Churkina G, Kruijt B, Hollinger DY. 2008. Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. *Agr. Forest Meteor.* **148**: 38–50.

Sellers PJ, Heiser MD, Hall FG. 1992. Relations between surface conductance and spectral vegetation indices at intermediate ($100\,m^2$ to $15\,km^2$) length scales. *J. Geophys. Res.* **97**(D17): 19033–19059.

Sene KJ. 1994. Parameterisations for energy transfers from a spare vine crop. *Agric. For. Meteor.* **71**: 1–18.

Seneviratne SI, Corti T, Davin EL, Hirschi M, Jaeger EB, Lehner I, Orlowsky B, Teuling AJ. 2010. Investigating soil moisture-climate interactions in a changing climate: a review. *Earth-Science Reviews.* **99**: 125–161.

Sheffield J, Wood EF, Roderick ML. 2012. Little change in global drought over the past 60 years. *Nature* **491**: 435–438.

Shi T, Guan D, Wang A, Wu J, Jin C, Han S. 2008. Comparison of three models to estimate evapotranspiration for a temperate mixed forest. *Hydrol. Process.* **22**(17): 3431–3443.

Shiklomanov AI. 1998. World water resources: a new appraisal and assessment for the twenty-first century: a summary of the monograph "World Water Resources" Report, 37 pp., UNESCO, Paris.

Shuttleworth WJ, Gurney RJ. 1990. The theoretical relationship between foliage temperature and canopy resistance in sparse crops. *Q. J. Roy. Meteorol. Soc.* **116**: 497–519.

Shuttleworth WJ, Wallace JS. 1985. Evaporation from sparse crops—an energy combination theory. *Q. J. Roy. Meteor. Soc.* **111**: 839–855.

Sloughter JM, Gneiting T, Raftery AE. 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association.* **105**(489): 25–35.

Solano R, Didan K, Jacobson A, Huete A. 2010. *MODIS vegetation index user's guide vegetation index and phenology lab*. The University of Arizona.

Stannard DI. 1993. Comparison of Penman–Monteith, Shuttleworth–Wallace, and modified Priestley–Taylor evapotranspiration models for wildland vegetation in semiarid rangeland. *Water Resour. Res.* **29**(5): 1379–1392.

Sumner DM, Jacobs JM. 2005. Utility of Penman–Monteith Priestley–Taylor reference evapotranspiration and pan evaporation methods to estimate pasture evapotranspiration. *J. Hydrol.* **308**(1–4): 81–104.

Szilagyi J. 2007. On the inherent asymmetric nature of the complementary relationship of evaporation. *Geophysical Research Letters.* **34**: L02405. DOI:10.1029/2006GL028708

Szilagyi J, Hobbins MT, Jozsa J. 2009. Modified advection–aridity model of evapotranspiration. *J. Hydrol. Eng.* **14**(6): 569–574.

Szilagyi J, Jozsa J. 2008. New findings about the complementary relationship based evaporation estimation methods. *J. Hydrol.* **354**: 171–186.

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **106**(D7): 7183–7192.

Tebaldi C, Hayhoe K, Arblaster JM, Meehl GA. 2006. Going to the extremes, an intercomparison of model-simulated historical and future changes in extreme events. *Climatic Change.* **79**: 185–211.

Teh CBS, Simmonds LP, Wheeler TR. 2001. Modelling the partitioning of solar radiation capture and evapotranspiration intercropping systems. In: Proceedings of the 2nd International Conference on Tropical Climatology. Meteorology and Hydrology TCMH-2001. Brussels. Belgium.

Teuling AJ, Seneviratne SI, Stöckli R, Reichstein M, Moors E, Ciais P, Luyssaert S, van den Hurk B, Ammann C, Bernhofer C, Dellwik E, Gianelle D, Gielen B, Grünwald T, Klumpp K, Montagnani L, Moureaux C, Sottocornola M, Wohlfahrt G. 2010. Contrasting response of European forest and grassland energy exchange to heatwaves. *Nature Geoscience* **3**: 722–727.

Vinukollu RK, Wood EF, Ferguson CR, Fisher JB. 2011. Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: evaluation of three process-based approaches. *Remote Sens. Environ.* **115**(3): 801–823.

Vrugt JA, Diks CGH, Clark MP. 2008. Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environ. Fluid Dyn.* **8**: 579–595.

Vrugt JA, Robinson BA. 2007. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. *Water Resour. Res.* **43**: W01411. DOI:10.1029/2005WR004838

Wang JM, Zhuang JX, Wang WZ, Liu SM, Xu ZW. 2014. Assessment of uncertainties in eddy covariance flux measurement based on intensive flux matrix of HiWATER-MUSOEXE. *IEEE Geosci. Remote Sens.* **12** (2): 259–263.

Wang K, Dickinson RE. 2012. A review of global terrestrial evapotranspiration: observation, modeling, climatology and climatic variability. *Rev. Geophys.* **50**: RG2005. DOI:10.1029/2011RG000373

Wilson K, Goldstein A, Falge E, Aubinet M, Baldocchi P, Bernhofer C, Ceulemans G, Dolmab H, Field C, Grelle A, Ibrom A. 2002. Energy balance closure at FLUXNET sites. *Agric. For. Meteorol.* **113**(1–4): 223–243.

Wu H, Zhang X, Liang S, Yang H, Zhou G. 2012. Estimation of clear-sky land surface longwave radiation from MODIS data products by merging multiple models. *J. Geophys. Res.* **117**: D22107. DOI:10.1029/2012JD017567

Xu CY, Singh VP. 2005. Evaluation of three complementary relationship evapotranspiration models by water balance approach to estimate actual regional evapotranspiration in different climatic regions. *J. Hydrol.* **308**: 105–121.

Xu ZX, Li JY. 2003. A distributed approach for estimating catchment evapotranspiration: comparison of the combination equation and the complementary relationship approaches. *Hydrol. Process.* **17**: 1509–1523.

Yang H, Yang D, Lei Z. 2013. Seasonal variability of the complementary relationship in the Asian monsoon region. *Hydrol. Process.* **27**(19): 2736–2741.

Yang T, Hao X, Shao Q, Xu CY, Zhao C, Chen X, Wang W. 2012. Multi-model ensemble projections in temperature and precipitation extremes of the Tibetan Plateau in the 21st century. *Global and Planetary Change* **80–81**: 1–13.

Yao Y, Liang S, Cheng J, Liu S, Fisher JB, Zhang X, Jia K, Zhao X, Qin Q, Zhao B, Han S, Zhou G, Zhou G, Li Y, Zhao S. 2013. MODIS-driven estimation of terrestrial latent heat flux in China based on a modified Priestley–Taylor algorithm. *Agric. For. Meteor.* **171–172**: 187–202.

Yao Y, Liang SL, Li XL, Hong Y, Fisher JB, Zhang NN, Chen JQ, Cheng J, Zhao SH, Zhang XT, Jiang B, Sun L, Jia K, Wang KC, Chen Y, Mu QZ, Feng F. 2014. Bayesian multimodel estimation of global terrestrial latent heat flux from eddy covariance, meteorological, and satellite observations. *J. Geophys. Res. Atmos.* **119**: 4521–4545.

Zhang B, Kang S, Li F, Zhang L. 2008. Comparison of three evapotranspiration models to Bowen ratio-energy balance method for a vineyard in an arid desert region of northwest China. *Agr. Forest Meteorol.* **148**: 1629–1640.

Zhang X, Srinivasan R, Bosch D. 2009. Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging. *J. Hydrol.* **347**: 307–317.

Zhu GF, Su YH, Li X, Zhang K, Li CB. 2013. Estimating actual evapotranspiration from an alpine grassland on Qinghai–Tibetan plateau using a two-source model and parameter uncertainty analysis by Bayesian approach. *J. Hydrol.* **476**: 42–51.

Zhu GF, Li X, Su YH, Zhang K, Bai Y, Ma JZ, Li CB, Hu XL, He JH. 2014a. Simultaneous parameterization of the two-source evapotranspiration model by Bayesian approach: application to spring maize in an arid region of northwest China. *Geosci. Model Dev.* **7**: 1467–1482.

Zhu GF, Lu L, Su YH, Wang XF, Cui X, Ma JZ, He JH, Zhang K, Li CB. 2014b. Energy flux partitioning and evapotranspiration in a sub-alpine spruce forest ecosystem. *Hydrol. Process.* **28**(19): 5093–5104.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.