Agricultural and Forest Meteorology 169 (2013) 12–24

Contents lists available at SciVerse ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

What controls the error structure in evapotranspiration models?

Aaron Polhamus[a], Joshua B. Fisher[a,*], Kevin P. Tu[b]

[a] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, United States
[b] Pioneer Hi-Bred Int'l, 18369 County Road 96, Woodland, CA 95695, United States

ARTICLE INFO

Article history:
Received 25 January 2012
Received in revised form 16 July 2012
Accepted 1 October 2012

Keywords:
Evapotranspiration
Decoupling
Stomatal resistance
Machine learning
Error
Uncertainty

ABSTRACT

Evapotranspiration models allow climate modelers to describe surface–atmosphere interactions, ecologists to understand the impact that global temperature change and increased radiation budgets will have on ecosystems, and farmers to decide how much irrigation to give their crops. Physically based algorithms for estimating evapotranspiration must manage a trade-off between physical realism and the difficulty of parameterizing key inputs, namely resistance factors associated with water vapor transport through the canopy and turbulent transport of water vapor from the canopy to ambient air. In this study we calculate predicted evapotranspiration at 42 AmeriFlux sites using two types of dedicated evapotranspiration models—one using physical resistances from the Penman–Monteith equation (Monteith, 1965) (Mu et al., 2007, 2011) and another based on the Priestley–Taylor (1972) equation, substituting functional constraints for resistances (Fisher et al., 2008). We analyze the structure of the residual series with respect to various meteorological and biophysical inputs, specifically Jarvis and McNaughton's (1986) decoupling coefficient, Ω, which is designed to represent the degree of control that plant stomata versus atmospheric demand and net radiation exercise over transpiration. We find that vegetation indices, magnitude of daytime fluxes, and bulk canopy resistance (r_c)—which largely drives Ω—are strong predictors of patterns in model bias for all flux products. Though our analysis suggests a consistently negative relationship between Ω and mean predicted error for all evapotranspiration models, we found that vegetation indices and flux magnitudes were the most significant drivers of model error. Before addressing error associated with canopy resistance and Ω, refinements to existing models should focus on correcting biases with respect to flux magnitudes and canopy indices. We suggest a dual-model approach for backsolving r_c (rather than estimating it from lookup tables and canopy indices), and increased attention to water availability, which largely drives stomatal opening and closure.

1. Introduction

Evapotranspiration (ET) is a central component of the global hydrological cycle (Fisher et al., 2009; Hasler and Avissar, 2006; Jimenez et al., 2011; Jung et al., 2010), and therefore of interest to hydrologists, climate modelers, ecologists, and agriculturalists. Accurate description of ET allows climate modelers to predict how atmospheric concentrations of water vapor are likely to change with increasing global temperatures (Arnell et al., 1996; Takahashi, 2008), and how these trends are likely to relate to other phases of the hydrological cycle, such as river runoff volumes (Arnell et al., 1996). Ecologists may be interested to know how forest canopies will respond to changes in the radiation budget and precipitation (Malhi et al., 2009). Finally multiple authors (Fisher et al., 2008; Mu et al., 2007; Sheffield et al., 2010) have sought to develop ET products for applications in water resource management. The Food and Agricultural Organization of the United Nations publishes irrigation guidelines that estimate crop water demand in part through estimating the ET rate (Allen et al., 1998).

In spite of the relevance of ET prediction to multiple fields of research and industry, the ET products currently in use vary widely with respect to predicted ET (Jimenez et al., 2011; Jung et al., 2010). Our results will show that the models considered here from Mu et al. (2007, 2011) and Fisher et al. (2008) have moderate success in predicting observed data, but with room for substantial improvement. Multiple authors, particularly those who implement a version of the Penman–Monteith equation (Monteith, 1965) have discussed the difficulty of parameterizing resistance inputs, particularly stomatal or canopy resistance, associated with the biophysical controls over ET (Fisher et al., 2008; Mu et al., 2007, 2011; Sheffield et al., 2010). All models, regardless of whether they directly or indirectly parameterize stomatal resistance, must ultimately account for the influence of stomatal control over transpiration. To determine if the models from Mu et al. and Fisher et al. account for this control, we examined prediction errors in relation to Jarvis and McNaughton's (1986) decoupling coefficient,

* Corresponding author. Tel.: +1 323 540 4569; fax: +1 818 354 9476.
E-mail addresses: jbfisher@jpl.nasa.gov, joshbfisher@gmail.com (J.B. Fisher).

0168-1923/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.agrformet.2012.10.002

Ω, an index of stomatal control which explains the degree to which transpiration is driven by radiation versus controlled by stomatal resistance. We tested the null hypothesis that Ω is not correlated with ET model performance. If there are grounds for rejecting this hypothesis, it may suggest that more accurate means of parameterizing stomatal resistance inputs to ET are necessary to improve current models. Our goal is to recommend a strategy for improving evapotranspiration estimation via a statistical analysis of the structure of model prediction errors for two widely used ET products.

The paper proceeds with: (1) a short overview of the primary ET products currently implemented in the literature and discussion of the decoupling coefficient; (2) discussion of the methodology used to parameterize resistances and describe the structure of ET model error with respect to the independent variables; (3) presentation and summary of the data; (4) results and discussion and (5) and conclusion.

2. Theoretical framework

2.1. Modeling evapotranspiration

Penman (1948) noted that there are two theoretical approaches to describe the evaporation of water from saturated surfaces, "the first being on an aerodynamic basis in which evaporation is regarded as due to turbulent transport of vapor by a process of eddy diffusion, and the second being on an energy basis in which evaporation is regarded as one of the ways of degrading incoming radiation." He concluded that the energy balance model was a more realistic method of describing the process of evaporation than aerodynamic transport. The premise of an energy balance model is that when solar energy encounters the earth's surface some of it is reflected, and some of it is converted to sensible and latent heat. Sensible heat refers to energy that can be directly sensed as heat through the transfer of energy from the canopy or soil to the air and the air's subsequent change in temperature, while latent heat refers to energy that is used to evaporate water as it changes phase from a liquid to a vapor. Energy degraded through increasing soil temperature is referred to as ground heat flux (G). Penman's equation for evapotranspiration is[1]:

$$\lambda E = \frac{\Delta(R_n - G) + \rho c_P[e_s(T_a) - e_a]/r_a}{\Delta + \gamma} \quad (1)$$

where λ is the latent flux of heat, taken to be 2.257 (MJ kg^{-1}); E is the ET rate (kg m^{-2} s^{-1}); Δ is the slope of the saturation-to-vapor pressure curve (Pa K^{-1}); R_n is net radiation (W m^{-2}); ρ is the density of dry air at approximately 12 °C, taken to be 1.234 kg m^{-3}; c_P is the specific heat capacity of air, taken to be 1005 J kg^{-1} K^{-1}; $e_s(T_a)$ is the saturation vapor pressure and e_a is the actual vapor pressure (Pa); r_a is the aerodynamic resistance to transfer of water vapor from the surface to ambient air (s m^{-1}); and γ is the psychrometric constant, taken to be 0.066 kPa K^{-1}. The difference between e_s and e_a is commonly referred to as vapor pressure deficit (VPD).

Penman's formula describes potential ET (ET_p): the amount of evaporation that would occur if the surface were sufficiently well-watered. Monteith (1965) modified Penman's formulation to include a measure of surface resistance (r_s). Typically considered one of the most theoretically grounded means of estimating ET (Cleugh et al., 2007; Langensiepen et al., 2009), this formula is referred to as the "Penman–Monteith" (PM) formula:

$$\lambda E = \frac{\Delta(R_n - G) + \rho c_P[e_s(T_a) - e_a]/r_a}{\Delta + \gamma(1 + (r_s/r_\alpha))} \quad (2)$$

where all terms are as in Eq. (1) and r_s describes the surface resistance to water vapor transport (s m^{-1}). If we assume that an extensive canopy acts as a uniform leaf over a given area of land r_s becomes equal to bulk canopy resistance, r_c. Though we rely on this assumption for our formulation of r_c, a more accurate conception of surface resistance involves those arising from soil resistance at ground level in addition to stomatal resistance to transpiration (Leuning et al., 2008). Note that when the ratio of r_c to r_a is low, this formula converges on Penman's original. As bulk canopy resistance grows in relation to aerodynamic resistance, however, r_c begins to play an increasingly active role in limiting ET.

Mu et al. (2007) expand upon Cleugh et al.'s (2007) method for parameterizing the PM equation, partitioning incoming radiation into components intercepted by the soil and by the canopy, and from these summing soil- and canopy-based evapotranspiration. They estimate soil resistance and r_c separately, estimate r_a as a function of temperature, and constrain potential evaporation from the soil via a multiplier that is a function of VPD and relative humidity. Recently the Mu et al. (2007) approach was updated by "(1) simplifying the calculation of vegetation cover fraction; (2) calculating ET as the sum of daytime and nighttime components; (3) adding soil heat flux calculation; (4) improving estimates of stomatal conductance, aerodynamic resistance and boundary layer resistance; (5) separating dry canopy surface from the wet; and (6) dividing soil surface into saturated wet surface and moist surface" (Mu et al., 2011). We denoted these models as MOD16_2007 and MOD16_2011 in reference to their versions of implementation as the MODIS MOD16 evapotranspiration product.

Despite the conceptual power of PM, reliable estimation of canopy and aerodynamic resistances can be difficult, and the necessarily broad assumptions on resistance parameterization in Mu et al. (2007, 2011) may be too simplistic. Priestley and Taylor (1972) suggested a simplification of PM for ET_o that replaces resistance parameters with a coefficient α, such that:

$$\lambda E = \alpha \frac{\Delta}{\Delta + \gamma}(R_n - G) \quad (3)$$

where all terms are as in Eq. (1) and G represents the ground heat flux (W m^{-2}). Here α is a constant that is tuned to reflect the overall magnitude of latent heat relative to sensible heat flux, and is related to the Bowen ratio $\beta = H/LE$, the ratio of sensible to latent heat. Priestley and Taylor derive a value of $\alpha = 1.26$ for well-watered surfaces, though multiple studies adjust α to reflect different surface conditions (Fisher et al., 2005, 2009; Garcia and Andre, 2000; Pereira and Villa Nova, 1992).

Fisher et al. (2008) note that Priestley and Taylor's formulation, though less sophisticated than PM, is easier to parameterize and has in many studies been shown to have similar if not greater explanatory power than versions of PM for well-watered surfaces (Jin et al., 2005; Sumner and Jacobs, 2005). They design an ET model that incorporates water availability constraints and disaggregates sources of ET by parsing out the total ET budget among canopy transpiration, soil evaporation, and interception evaporation. They then calculate potential ET from each component according to PT, and constrain these estimates with canopy and soil moisture constraint indices designed to be sensitive to water limitation and the density of the vegetative cover. The goal of the Fisher et al. (2008) model is to establish a technique for estimating evapotranspiration whose inputs can be reliably estimated from data–sparse environments (e.g. via remote sensing), while at the same time describing the physiological and non-physiological processes driving ET. The model is driven by five inputs: net radiation (R_n), normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), maximum air temperature (T_{max}), and water vapor pressure (e_a), all of which can be approximated via remote

[1] We unitize these terms as in Fisher et al. (2009).

sensing (Fisher et al., 2008). We denote the Fisher et al. (2008) model as "PT-JPL". When used to predict ET from 16 eddy covariance flux tower sites, the PT-JPL model performed well, achieving $R^2 = 0.90$ (Fisher et al., 2008). 1 contains information about every variable used in the calculations of the PT-JPL and ET_MU models.

2.2. The decoupling coefficient

Jarvis and McNaughton (1986) identify a conflict in the ET literature between those who believe that plant stomata control transpiration, and those who believe that atmospheric variables such as vapor pressure and net radiation are the primary drivers. They create a unitless index, the decoupling coefficient, Ω, which describes the extent to which the VPD of a leaf (or canopy) boundary layer is coupled to the ambient air:

$$\Omega = \frac{1 + \varepsilon}{1 + \varepsilon + (r_c/r_a)} \tag{4}$$

where $\varepsilon = \Delta/\gamma$, and r_c and r_a are defined as in Eq. (3). As r_c increases relative to r_a the degree of decoupling approaches 0, i.e. the VPD of the leaf/canopy boundary layer becomes closely matched to that of the ambient air. Conversely, when r_a is large relative to r_c, the decoupling coefficient approaches 1, i.e. the VPD of the leaf/canopy boundary layer differs substantially from the ambient air. McNaughton and Jarvis begin by considering Ω in the context of a single, amphistomatous leaf, before scaling up to the canopy level. They conclude that at large spatial scales (e.g. 10^5 m across) overall transpiration will be relatively insensitive to changes in bulk stomatal resistance, citing De Bruin and Holtslag's (1982) study which found that a 3-fold change in canopy conductance from 0.4 to 1.2 mol m^{-2} s^{-1} would be expected to cause only a 20 percent change in transpiration. This present study is concerned with a spatial scale of approximately 1 km^2, leaving open the possibility that changes in bulk stomatal resistance could have a significant impact on the accuracy of ET models.

McNaughton and Jarvis also derive an expression for the sensitivity of a change in transpiration (at the leaf scale) to a change in the stomatal conductance of that leaf ($g_{stom}{}^{-1} = r_{stom}$). We can extrapolate from the leaf to canopy level, substituting g_{stom} for the bulk stomatal resistance of the canopy, g_c, while E represents the overall transpiration from the surface:

$$\frac{(\delta E/E)}{(\delta g_c/g_c)} = 1 - \Omega \tag{5}$$

The intuition here is that when Ω is close to 0, a fractional change in canopy conductance results in an equal fractional change in transpiration (e.g. in the fully closed scenario above, a small degree of opening corresponds to a large percentage increase in transpiration). As canopies couple with the ambient air (Ω approaches 1), radiation becomes the dominant driver of transpiration, and fractional changes in stomatal conductance result in much smaller fractional changes in transpiration. When stomatal control of transpiration is high and atmospheric inputs like net radiation and VPD are large there will be a disconnect between the level of atmospheric demand for ET and the 'willingness' of the canopy to provide it. The difficulty of accurately describing these resistances, however, may result in ET models being prone to bias in scenario.

3. Data

The flux tower data for this study comes from the level 3 products of the AmeriFlux network, and was downloaded from http://cdiac.ornl.gov/ftp/ameriflux/data/Level3/. AmeriFlux is part of a global network of over 500 towers spanning 5 continents (FLUXNET), providing half-hourly to hourly measurements of carbon dioxide, water vapor, and energy exchanges between the land and atmosphere across a diverse range of ecosystems and climates using the eddy covariance method (Baldocchi, 2008). In addition to meteorological variables such as temperature, water vapor saturation, and incoming net radiation, each tower measures fluxes of latent heat, sensible heat, and soil heat flux. Fluxes are reported as averages over a half-hour interval. Daily and monthly values expressed in W m^{-2} likewise reflect the average rate of flux over that interval. It is essential that the averaging interval of the modeling inputs matches that of the predicted flux, and we have taken to assure that this is the case here.

Though the original data are reported as half-hourly averages, we use a two-step averaging process to construct a data set of monthly averages. This approach averages much of the random variation in the data, resulting in better model fits than we obtain at larger time scales. We compared the results of the daily and monthly averages and found that the loss of resolution from the monthly averaging did result in an increase in fit as measured by R^2 by ~0.15 per model. For the purposes of our analysis, we assume that this improvement in performance is due largely to the averaging out of random variation that is more pronounced at the finer time scales, and that the monthly averaging therefore allows us to interrogate the underlying model structure mode effectively. The data used to calculate monthly scale ET predictions were pre-processed as follows: (1) First, fluxes and input variables were partitioned into daytime and nighttime values using global incoming solar radiation (R_g) measurements. If $R_g > 10$ W m^{-2} and the quality flag was equal to 0 (i.e. no quality problems), an observation was classified as daytime. If $R_g \leq 10$ and the quality flag was equal to 0 an observation was classified as nighttime. Because of problems eddy covariance instrumentation errors that are aggravated during the nighttime, we only selected daytime values for our analysis; (2) If there were more than 10 daytime observations a daytime value was calculated from the data as a simple average. Otherwise the daytime value was assigned a fill; (3) We then removed all site/month combinations for which there were not at least 15 days of data. If this criterion was satisfied, we calculated monthly values of fluxes and model inputs as simple averages; (4) Since PT-JPL requires an annual series of data, we removed all site/year combinations for which there was not at least one observation for six months of the year; (5) For each missing month of data that had valid adjacent months of data on either side we interpolated the missing values as simple averages. E.g. if for March–April–May April is missing, April = (March + May)/2; (6) We linearly interpolated the values for missing month pairs, conditional on the missing pair being bordered by complete adjacent months; (7) We removed any site/year combinations that did not contain at least 10 months of data. (8) Finally, we only used observations that contained the requisite data to produce both the MOD16 and PT-JPL products (Table 1).

This filtering procedure reduced the size of the data set by approximately 32 percent—from 2071 potential site/month observations to 1400 usable observations. Notwithstanding, the final sample contained 40 sites distributed throughout the diverse climate regions and biomes of North America. See Table 2 for a list of the land cover types, AmeriFlux ID, and sample representation of the sites used in this study.

We supplemented the AmeriFlux data with 9 km^2 square gridded MODIS data from the Oak Ridge National Laboratory Distributed Active Archive Center. The gridded MODIS data provides the necessary land cover class (Friedl et al., 2002) and absorbed photosynthetically active radiation (fPAR), LAI (Myeni et al., 2002), EVI, and NDVI (Huete et al., 2002) inputs necessary to calculate PT-JPL, MOD16_2007, and MOD16_2011.

Table 1
List of inputs to the Fisher et al. (2008) and Mu et al. (2007, 2011) models. MOD16_2007 and MOD16_2011 are grouped together. All values represent daytime averages, unless stated otherwise.

Input	Details	Units	Value	Model
Ta_{min}	Minimum monthly air temperature	°C	[−40,25]	MOD16
G_{day}	Soil heat flux	W m^{-2}	[−50,90]	MOD16
LAI	Leaf area index	unitless	[0,7]	MOD16
g_{cu}	Minimum leaf cuticular conductance	m s^{-1}	0.00001	MOD16
ë	Latent heat of evaporation	J kg^{-1}	2,260,000	MOD16
M_a	Molar mass of air	g mol^{-1}	28.97	MOD16
M_w	Molar mass of water	g mol^{-1}	18.015	MOD16
LC	UMD land cover classification	unitless	1, 2, . . ., 12	MOD16,PT-JPL
ó	Stefan-Boltzman constant	W m^{-2} K^{-4}	5.67×10^{-8}	MOD16,PT-JPL
ñ	Density of air	kg m^{-3}	1.234	MOD16,PT-JPL
c_P	Specific heat capacity of air	J kg^{-1} K^{-1}	1003.5	MOD16,PT-JPL
P	Unit of atmospheric pressure	Pa	101,325	MOD16,PT-JPL
ã	Psychrometric constant: $(M_a/M_w)(c_P \times P/\tilde{a})$	Pa K^{-1}	72.35	MOD16,PT-JPL
Ta_{day}	Average day time air temperature	°C	[−40,40]	PT-JPL
Ta_{max}	Maximum daily air temperature	°C	[−25,40]	PT-JPL
EVI	Enhanced vegetation index	unitless	[−1,1]	PT-JPL
Ta_{night}	Average night time air temperature	°C	[−40,40]	PT-JPL, MOD16
RH	Relative humidity	%	[0,100]	PT-JPL, MOD16
R_n	Incoming net radiation	W m^{-2}	[−35,250]	PT-JPL, MOD16
VPD_{day}	Vapor pressure deficit	Pa	[0,∞]	PT-JPL,MOD16

Table 2
Table of sites included in study.

Site ID	Land cover classification	n	Site ID	Land cover classification	n
CANS2	Evergreen Needle-leaf Forest	12	USMe5	Woody Savannas	36
CANS3	Woody Savannas	9	USMMS	Deciduous Broad-leaf Forest	55
CANS5	Evergreen Needle-leaf Forest	24	USMOz	Deciduous Broad-leaf Forest	36
CANS6	Water	24	USNe2	Croplands	47
CANS7	Woody Savannas	12	USNe3	Croplands	47
USARM	Grassland	36	USNR1	Evergreen Needle-leaf Forest	59
USAud	Open Shrublands	36	USRo1	Croplands	24
USBar	Deciduous Broad-leaf Forest	24	USRo3	Croplands	35
USBkg	Croplands	23	USSO2	Woody Savannas	12
USBo1	Croplands	83	USSO3	Closed Shrublands	12
USDix	Urbanand Built-up	24	USSO4	Closed Shrublands	24
USFPe	Grassland	84	USSP1	Evergreen Broad-leaf Forest	47
USFR2	Grassland	12	USSP2	Woody Savannas	60
USFuf	Woody Savannas	12	USSP3	Evergreen Broad-leaf Forest	48
USFwf	Croplands	12	USSRM	Open Shrublands	36
USGoo	Cropland/Natural Vegetation Mosaic	36	USTon	Mixed Savannas	72
USIB1	Cropland/Natural Vegetation Mosaic	12	USVar	Woody Savannas	72
USIB2	Cropland/Natural Vegetation Mosaic	24	USWkg	Grassland	24
USKS2	Evergreen Broad-leaf Forest	24	USWlr	Grassland	24
USMe2	Evergreen Needle-leaf Forest	59	USWrc	Evergreen Needle-leaf Forest	48

n indicates per-site monthly sample size.
For more detail about each site, visit http://public.ornl.gov/ameriflux/site-select.cfm.

4. Methods

4.1. Estimation of resistances

To estimate canopy resistance we inverted the Penman–Monteith equation (4) as follows:

$$r_a = r_a \left(\frac{\Delta R_n + \rho c_P(\mathrm{VPD}/r_a)}{\gamma \lambda E} - \frac{\Delta}{\gamma} - 1 \right) \quad (6)$$

All inputs (including ET) to items on the right hand of this formula are measured empirically at half-hour intervals at each flux site. We then calculate their monthly daytime averages to use as inputs for calculating monthly average fluxes. For the purposes of calculating Ω we assume that $r_c = r_s$. This is not a valid assumption in all cases, and we discuss the implications of this in Section 6.

Aerodynamic resistance was estimated from the surface friction velocity u^* (Thom, 1975) measured by eddy covariance (Hasler and Avissar, 2006; Lee and Black, 1993):

$$r_a = \frac{u}{u^{*2}} \quad (7)$$

where u is wind speed and u^* is the friction velocity. Though more sophisticated parameterizations of aerodynamic resistance are possible (Liu et al., 2007), these require additional data inputs that must often be empirically derived, e.g. coefficients of the integral stability functions for wind and temperature. For the purposes of this study, we assume that Eq. (7) represents an adequate simplification of physics of aerodynamic resistance.

We calculated resistances using monthly scale parameters to estimate monthly values of Ω. These monthly values were then used in our analysis of the structure of the residual predicted data series, as outlined below.

4.2. Statistical analysis

In attributing ET model error to the covariates in the data set we have two objectives: (1) to describe the relationship between the input variables and model error; and (2) to quantify the significance of that relationship in terms of explaining variance in the residual series.

A classic approach to this problem is through ordinary least squares (referred to here as OLS) regression. For OLS regression to be valid we require independently, identically, and normally distributed prediction residuals with mean zero and constant variance over. We also need the dependent variable (e.g. ET) to be linearly related to its covariates. In practice these assumptions are often not satisfied. For example, since Ω has strict lower and upper boundaries of 0 and 1 the relationship of Ω to model error will become increasingly nonlinear as Ω approaches its upper and lower boundaries. Furthermore, interaction effects between the variables must be specified manually, meaning that unless the researcher has good prior knowledge about how the variables are interrelated it may be easy to miss subtle yet significant relationships between them. In spite of these challenges, we choose to include OLS (i) because of its ease of interpretation, (ii) its widespread use as an analytical tool in the field of environmental science, and (iii) it sets a benchmark for the performance of the more sophisticated machine learning algorithms we will use to explore the data.

The second class of models that we apply here are sum-of-trees models based on Breiman's classification and regression trees (CARTs) (Breiman et al., 1984). These are random forests (Breiman, 2001) and Bayesian additive regression trees (BARTs) (Chipman et al., 2010). In contrast to OLS, these models are able to account for complex interactions and additive effects between model variables without explicit training. They are therefore extremely useful for exploratory analyses where the researcher has little prior knowledge about how the prediction variables relate to the dependent variable, as is the case with this study. For a detailed explanation of the theoretical basis of CART, random forests, and BART, refer to Appendix A.

The final statistical technique that we use to model the error structure of our data is the neural network. We use the R Stuttgart Neural Network Simulator (Bergmeir and Benítez, 2010) to train an Elman network (Elman, 1990). The simplest examples of neural networks utilize a structure whereby the input layer connects to the hidden layers that connect to the output layer in a feed-forward manner. An Elman network follows this same basic structure, but with the modification that the hidden layer is linked to a "context layer" via a 1:1 connection. For each iteration i of network training, the context layer inherits the values of the hidden layer, and these values are then used to train the hidden layer in iteration $i+1$ along with the input layer (Cruse, 2006).

To quantify the relative importance of each variable in explaining residual series variance we use variable importance rankings. For the neural network and random forest models we calculate variable importance metrics by randomly scrambling the entries for the variable of interest, then predicting the data using the same model but with the scrambled inputs (Liaw and Wiener, 2002). This random permutation introduces additional error into model predictions, quantified as the sum of squared residuals (SSR). Those variables whose permutations result in the largest reductions in the models' ability to predict the training set, i.e. the largest increase in SSR, are classified as relatively more important. This technique is well established for random forests (Breiman, 2001; Liaw and Wiener, 2002; Strobl et al., 2007), but not typically used for neural networks (Olden et al., 2004). Random permutation of the input layers to a neural network can be considered an adaptation of the input perturbation method (Gevrey et al., 2003), whereby a small amount of white noise is added to inputs and the resulting change in mean squared error is analyzed. Our view is that the permutation is preferable to perturbation for the purposes of variable selection, since it introduces an even greater degree of randomness into model inputs. Though we could repeat a similar procedure for BART, the maintainers of the R BART package (Chipman and McCulloch, 2010) provide a function that uses the number of times that a variable appears in the selection node of the final suit of Bayesian trees calculated from the Markov chain runs described above. Those variables most frequently selected as node splits are classified as relatively more important. For ease of comparison we normalize the scales of all variables important by dividing by either the largest RSS or proportion value such that all importance ranking range between 0 and 1. A value of 1 indicates that, relative to the other variables in the data set, a variable is 'very important.' A value of 0 indicates 'not important at all.'

Machine learning algorithms like random forests are essentially "black box" techniques—they tend to have good prediction performance, but the relationship between the dependent and explanatory variables is not nearly as explicit as in OLS. Fortunately, a technique called "partial dependence plotting" allows us to estimate good visual representations of these relationships (Friedman, 2001). The purpose of partial dependence plots is to illustrate how a marginal change in a variable of interest influences the expected value of the dependent variable. Friedman (2001) suggest the average function for cases where the functional form of the model used for partial dependence plotting does not depend too heavily on the specific subset of data used for model training. The average function for partial dependence plotting is defined as:

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}) \tag{8}$$

where x is the variable for which partial dependence is being examined, and x_{iC} is the rest of the data set (Liaw and Wiener, 2002). In practice, we estimate the partial dependence function by selecting k values of x. For each of these k values we then recalculate model predictions, but with all entries for x only containing the kth value. When this process is repeated for a sufficiently fine-grained subset of x, a partial dependence curve can be plotted.

5. Results

In an intercomparison of PT-JPL, MOD16_2007, and MOD16_2011, PT-JPL explained 71 percent of the variance in the sample data, compared to 54 percent and 48 percent for MOD16_2011 and MOD16_2007, respectively (Fig. 1). All models tended to over-predict the measured ET (or, conversely, measured ET was "under-measured" relative to predicted ET). A slope coefficient of less than one indicates that we obtain the expected value of an observed data point by shrinking that observation's predicted value. Both of the MOD16 models over-predicted less than PT-JPL. The 6 percent performance improvement from MOD16_2007 to MOD16_2011 confirms that the authors' updates to the 2007 model in their 2011 paper did indeed improve performance.

We recognize the unresolved problem of energy balance closure in the observed ET data, which has the potential to introduce significant error into observed flux values, skewing model performance diagnostics (Fisher et al., 2007). Energy balance closure refers to the difference between measured incoming energy (R_n) and the sum of measured fluxes LE, H, and G. In theory these should be equal, but in practice they often are not. The mean of the energy balance closure ratio—$\mathrm{EBC} = (\mathrm{ET}+H+G)/R_n$—across the monthly sample for the 40 sites was 0.91, a similar though slightly higher level of closure compared to what is commonly observed in the literature (Aubinet et al., 2005; Scott, 2010; Shuttleworth, 2007; Wilson et al., 2002). However, a sample mean of 0.91 does not rule out the possibility of there being significant variation in the level of closure across sites. We expressed energy balance closure as a difference, $\mathrm{EBC} = R_n - (\mathrm{ET}+H+G)$, and examined its sample distribution (Fig. 2). The mean of the differenced EBC sample distribution here is 18.4 $\mathrm{W\,m^{-2}}$, a relatively small figure, but with a standard deviation of 43.8 $\mathrm{W\,m^{-2}}$. We control for the effect of this

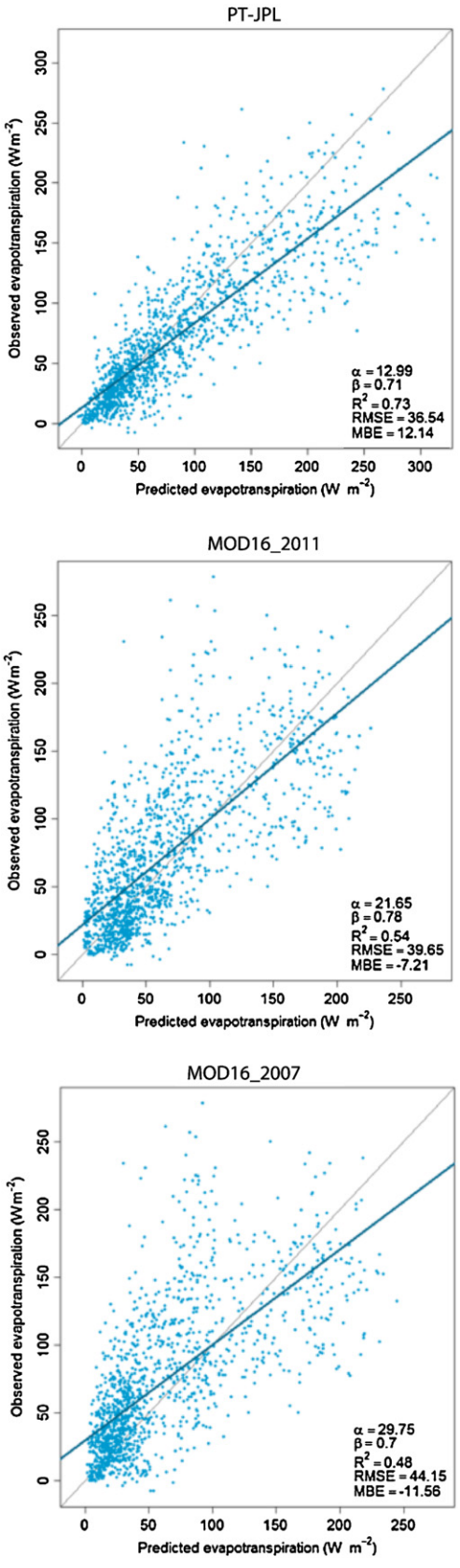

Fig. 1. Performance diagnostic plots for each of the three evapotranspiration products used in this study. The constant coefficient and slope of the regression lines are represented by α and β, respectively. R^2 represents portion of variance explained by the dependent variable, RMSE represents "root mean squared error," and MBE represents "mean biased error.".

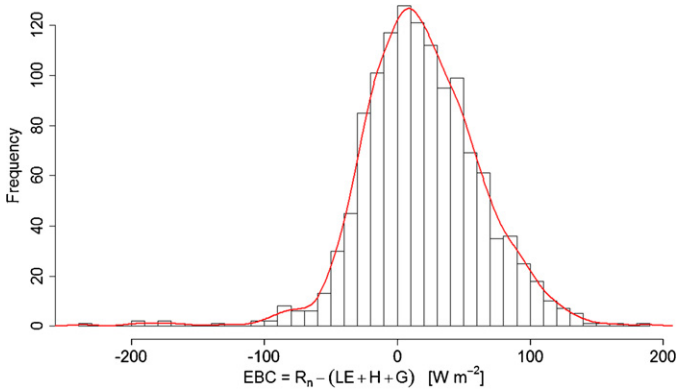

Fig. 2. Histogram of energy balance closure at the monthly level across all site, expressed as the difference between daytime net radiation and the sum of observed ET, *H*, and *G*. Estimated density curve drawn in red.

wide variation in EBC by including differenced EBC as a variable in the error structure analysis.

Three of the variables used in the error structure analysis are functions of other variables in the data set. The decoupling coefficient is heavily driven by r_c, which has observed ET as a significant component. Similarly, r_a is a function of wind speed and friction velocity. We therefore used principal component analysis (PCA) to identify how the explanatory variables were related with respect to the two primary axes of variation in the data (Fig. 3). The first two principal components captured 53 percent of the total variation in the explanatory data set, with the remaining 12 components explaining the other 47 percent. The first component explained approximately 30 percent of the variation, and the second explained approximately 23 percent. Because these two components describe a large share of the variation in the data, a two-dimensional projection of the data onto them is a good way to visualize patterns of similarity among variables in this study. We can see that Ω and r_c are strongly, inversely, related along the second principal component. We therefore expect these two variables to capture similar sources of variation in the residual series. This is to be expected given the construction of Ω—when

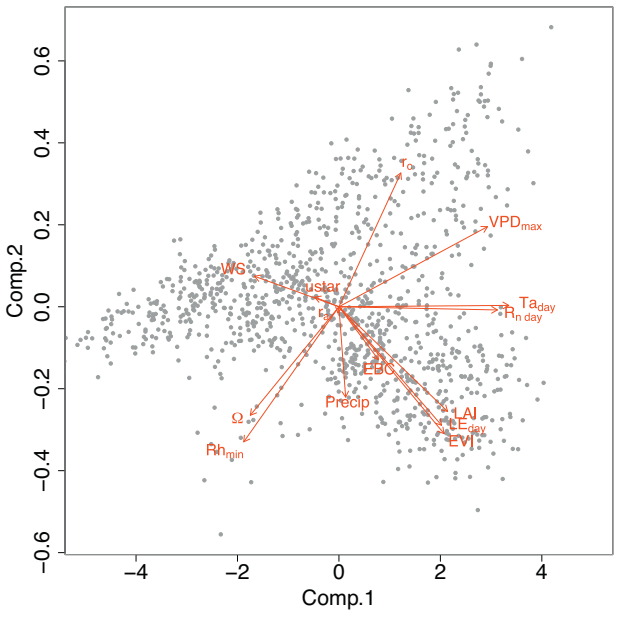

Fig. 3. Labeled biplot diagram of projection of all flux product/predictor variables onto first and second principle components.

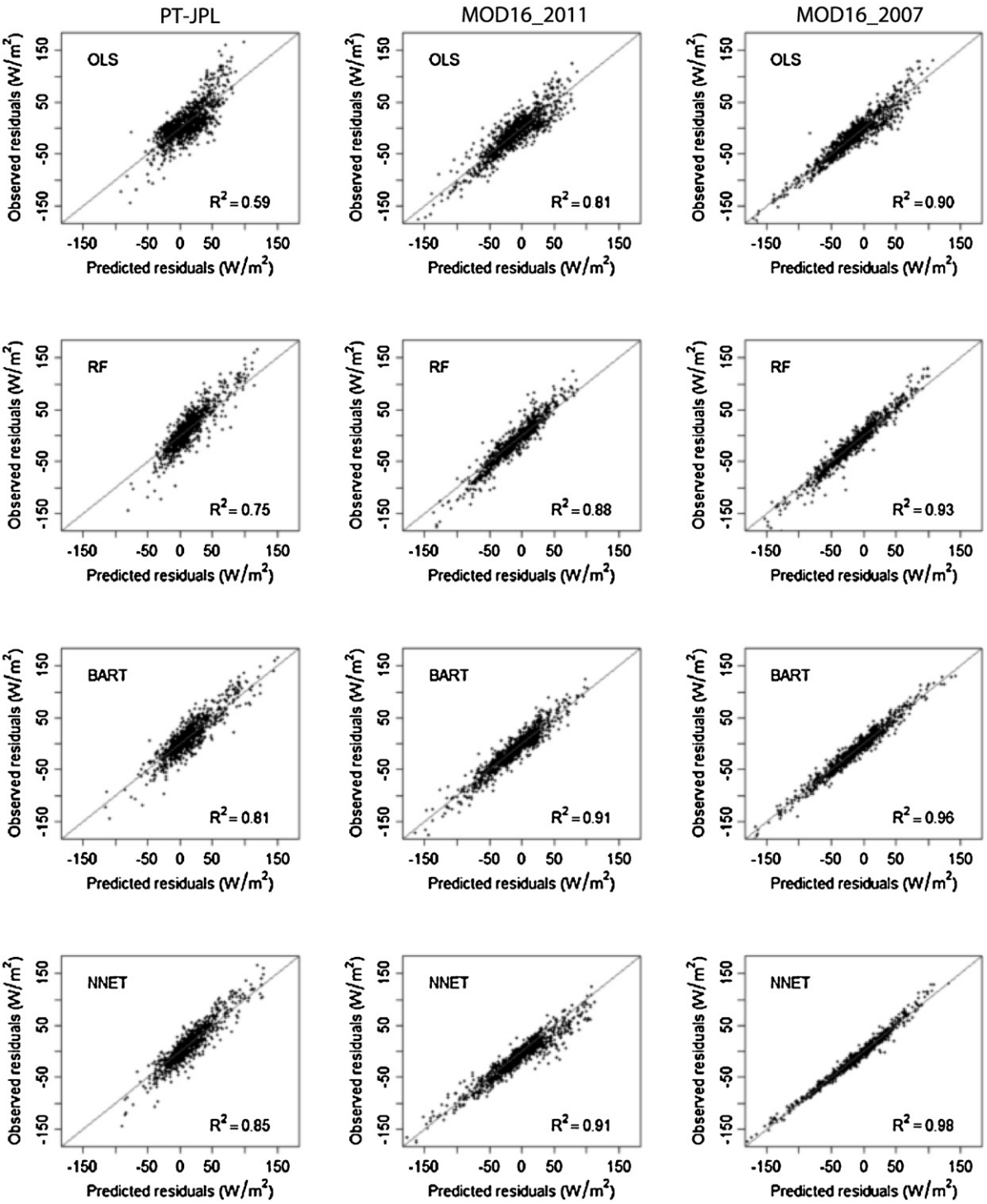

Fig. 4. Intercomparison of the performance of four statistical models (rows) for explaining residual series data for the three evapotranspiration products used in this study (columns).

canopy resistances are large (high r_c) we will generally expect to observe decoupled canopies (Ω close to 0). It is encouraging to see r_c and Ω varying more or less orthogonally to *LE_day*, *Ta_day*, and *Rn_day*, all of which are used to derive r_c via the Penman–Monteith equation. In spite of being composite variables, Ω and r_c do not seem overwhelmingly driven by any one of their inputs, though there is a strong apparent relationship between Ω and *Rh_min*. That precipitation—which relates to water availability—Ω, and r_c have principal component loadings that are distinct from the majority of other model input variables suggests that they describe distinct sources of variation. We hypothesize that the first principal component, where *Ta_day* and *Rn_day* are primarily loaded, relates primarily to incoming energy, while the second principal component, where *Precip* is primarily loaded, relates mainly to water availability.

The dependent variable in the error structural analysis is the prediction residual, $e_{st} = P_{st} - O_{st}$, where s represents site and t represents month. We use the statistical models described above to

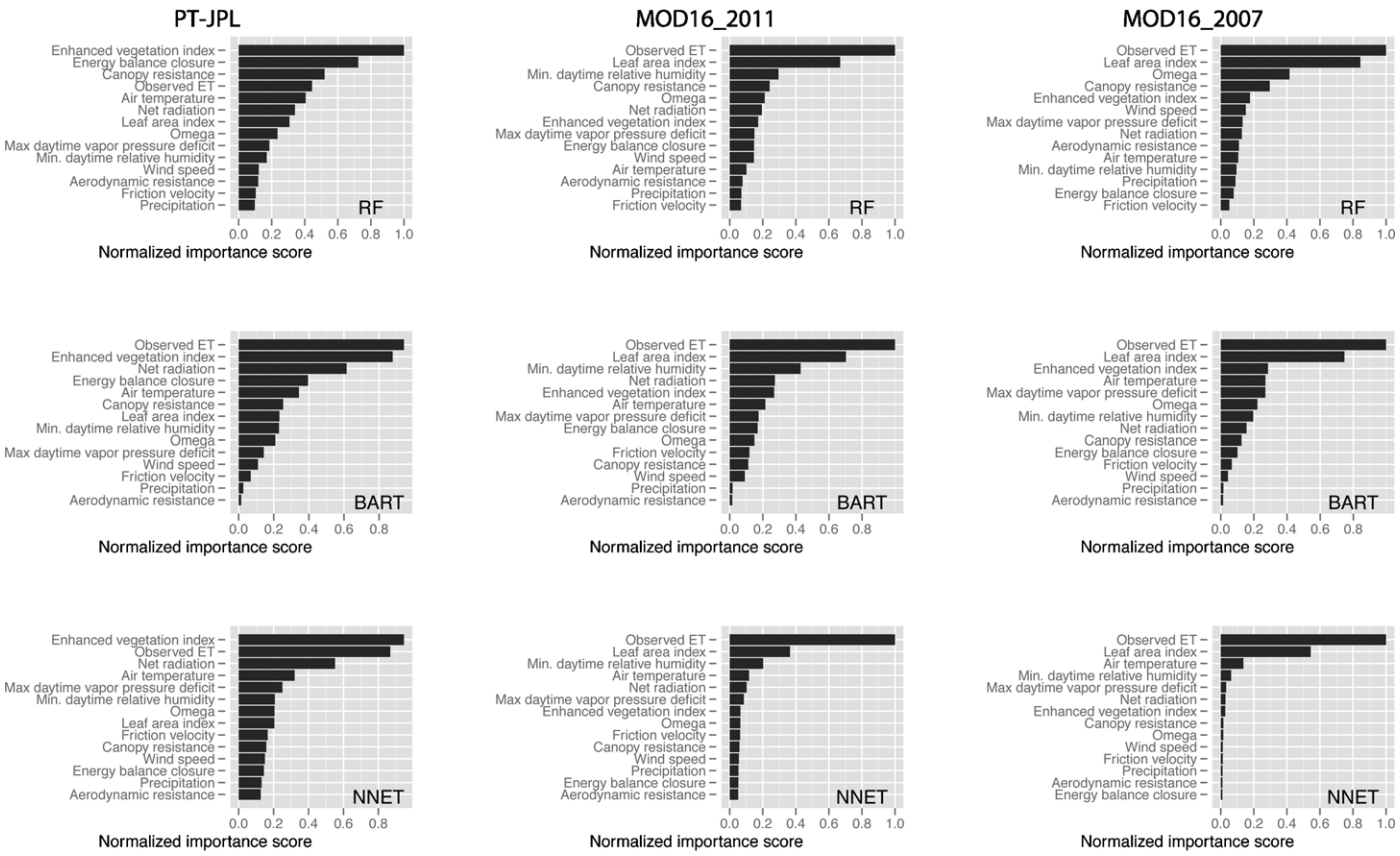

Fig. 5. Variable importance rankings for the random forest (RF), BART, and neural network (NNET) statistical models.

predict this residual as a function of model inputs, along with u^*, wind speed, precipitation, Ω, and r_c such that:

$$e_{st} = M_i(X_{st}) + \varepsilon_{st}, \quad i = 1,\ 2,\ 3,\ 4 \tag{9}$$

where M_i represents one of the four statistical models, X_{st} represents the predictor set, indexed by site and time, and ε_{st}, the residual of the model M_i prediction on the X_{st} predictor set. Fig. 4 shows how OLS, random forest, BART, and neural network models, respectively, performed in modeling the monthly residual series of PT-JPL, MOD16_2011, and MOD16_2007 predictions. Error predictive performance was weakest across all statistical models for PT-JPL; because PT-JPL had the best overall performance in explaining the observed data, there is less "left over" variation to be explained in the residual series. If we were only implementing the OLS statistical model, one could object that the error structure of the residuals is simply more complicated for MOD16. However, we implement several machine learning techniques that effectively model non-linear relationships and variable interactions. This gives us confidence that the weaker performance of the statistical models for PT-JPL is related to the smaller amount of unexplained variation in the original data relative to MOD16_2007 and MOD17_2011. The random forest and BART models improved error prediction performance over OLS by 16–22 percent for PT-JPL, but only 7–10 percent and 3–6 percent for MOD16_2011 and MOD16_2007, respectively. This suggests that simple, linear effects dominate the error structures of the MOD16 models. For PT-JPL, the difference in prediction performance between the linear and non-linear models suggests a relatively stronger presence of variable interactions and non-linearities.

Fig. 5 shows the variable importance plot partitioned by ET product and residual analysis model (M). We should be careful about the inferences we make from these variable importance rankings. Unlike OLS, each of the three models illustrated here incorporates random processes in the training stage, and therefore can be expected to yield slightly different results each time the model is trained. To control for this we ran the analysis 41 times, and generated a single variable importance ranking for each product and model combination as a simple average of the rankings across all 41 iterations. In statistical analysis, the number 30 is a threshold often used as the sample size for which the distribution of the mean of a random variable will be essentially normal according to the Central Limit Theorem. This is not a quantitative threshold, but rather a qualitative default used for convenience. Since we present partial dependence plot lines as means of multiple analysis runs, we would like these means to have approximately normal distributions for the sake of estimating confidence intervals. The fact that we ran the analysis 41 times is a result of the fact that this is where the computer used for this analysis encountered a memory constraint while executing the multiple-runs analysis. Overall, Ω and r_c did not score highly in the variable importance analysis. Though they were ranked the highest in the random forest analyses, they were still dwarfed by variables relating to incoming energy and canopy greenness. There was no clear trend across all statistical models and ET products as to whether Ω or r_c was more important.[2] In addition to energy- and canopy-related variables, daytime air temperature ranked highly in several of the analyses, particularly the neural network.

We are unable to infer from the results of this particular analysis whether canopy decoupling or bulk stomatal resistance is most

[2] Note that even though PT-JPL does not model r_c explicitly, the PT-JPL error structure could still be related to r_c if canopy resistance is indeed an important limiting factor of ET and the model does not adequately implicitly model this.

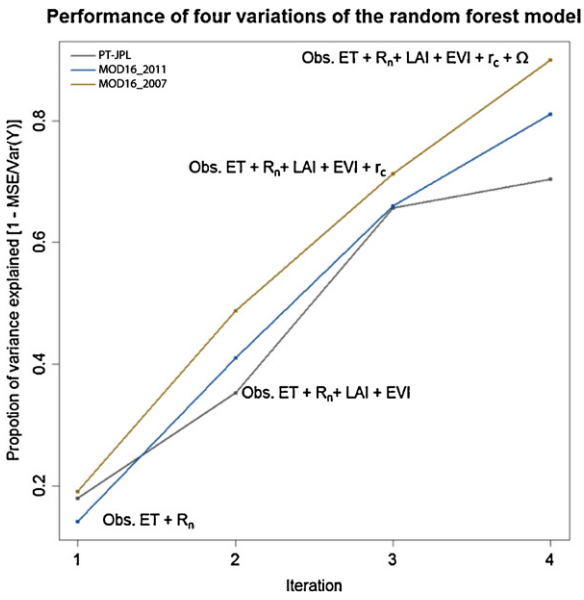

Fig. 6. Performance improvement associated with iterative inclusion of top-ranked important variables from figure ....., Iteration I: observed ET, R_n; Iteration 2: . . . + EVI, LAI; Iteration 3: . . . + r_c; Iteration 4: . . . + Ω.

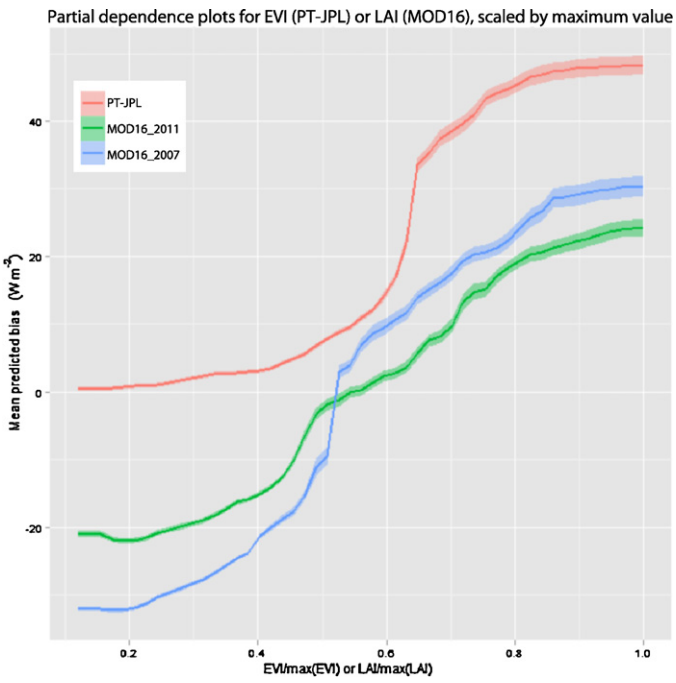

Fig. 7. Partial dependence of mean prediction bias on canopy indices EVI and LAI. Calculated using random forests. Shaded area represents that 95 percent confidence band for each partial dependence curve.

relevant for predicting model error. Breiman (2001) note that when two variables, x_1 and x_2, express the same information in a data set, the fact that they will each be selected an approximately equal number of times in the random forest leads them to have similar variable importance scores. However, when a model that includes x_1 is modified to include x_2, the change in prediction performance will be minimal since no new information is added. If x_1 and x_2 are Ω and r_c, we can map the performance change in a series of random forest models to see not only if each variable is important, but whether these variables add new information with respect to the residual series for each flux product.

Fig. 6 shows the proportion of variance explained using a random forest model calculated over four iterations. The variable ensemble used for model training was selected according to the variable importance data in Fig. 5. Iteration 1 used only *observed ET* and R_n, iteration 2 included *LAI* and *EVI*, the second-most important variables, iteration 3 included r_c, and iteration 4 included Ω. Up to iteration 3, we can see that the random forest model actually explains similar amounts of various across the PT-JPL and MOD16 products. However, there is a significant difference in how the random forest models respond to the incorporation of Ω at iteration 4: for PT-JPL the incorporation of Ω only contributes a 4 percent improvement in prediction performance whereas for MOD16_2011 and MOD16_2007 it contributes approximately 14 and 16 percent, respectively. For both of these models, the amount of new information that Ω added is comparable to the new information added by incorporating r_c. For the residual series of PT-JPL, it is the information that r_c and Ω have in common that drives the performance improvement for the random forest model, i.e. bulk stomatal resistance. This implies that PT-JPL is failing to describe important information related to r_c (namely, the role of plant stomata in restricting transpiration), but succeeding in describing the aspects of Ω related to r_a and VPD. For the residual series of both MOD16 models, we see that r_c and Ω each add new information. This suggests that while all three models struggle to adequately describe the biophysical constraints on transpiration associated with r_c, PT-JPL is better at controlling for air temperature and VPD, the other key inputs to Ω.

Observed ET occupied the position of importance in the residual analysis across all statistical models. This follows from the fact that *e* is a function of the observed and predicted data. As we see in the diagnostic plots in Fig. 1, all ET products tend to under-predict large observed values. This tight relationship between large observed ET values and large residuals explains why *observed ET* is consistently ranked the most important variable. The importance of vegetation indices *EVI* and *LAI*, particularly for PT-JPL, where *EVI* was ranked as the most important explanatory variable for all *M*, were consistent across models. *Observed ET* and *LAI* were jointly the most significant variables for both ET_MU products for all *M*. As seen in the step from iteration 1 to iteration 2 in Fig. 6, these variables contribute significant, unique information about prediction error to random forest model. Fig. 7 shows partial dependence plots of the relationship between mean predicted error and *EVI* or *LAI*, depending on which was ranked as the most important for each ET product. We see that for all models the relationship resembled a logarithmic curve, with relatively little change in expected model bias for high and low values of the vegetative index before sloping steeply upward in the middle of the range. This is evidence that the mechanisms for incorporating canopy data into the flux products tested here transition from overly restrictive (or in the case of PT-JPL, appropriately non-restrictive) to not restrictive enough. On average, large values of canopy indices correspond to larger prediction residuals, suggesting that the ET products are relaxing the canopy-based restrictions on transpiration too aggressively when canopy indices are large, leading to over-prediction.

Fig. 8 shows the partial dependence plots of the predicted residuals on Ω for all statistical models across the three flux products: we see that PT-JPL begins with a mean, positive predicted bias that diminishes as Ω increases, approaching zero as Ω approaches 1. The MOD16 products have approximately zero mean predicted bias for low levels of Ω, but tend to become increasingly negatively biased as Ω increases. This implies that for large-Ω conditions, where we generally expect to observe larger volumes of evapotranspiration, the ET_MU products are often overly restrictive. This finding varies across statistical models, and is less pronounced for the BART and random forest analyses, which incorporate non-linear effects that OLS and the neural network do not.

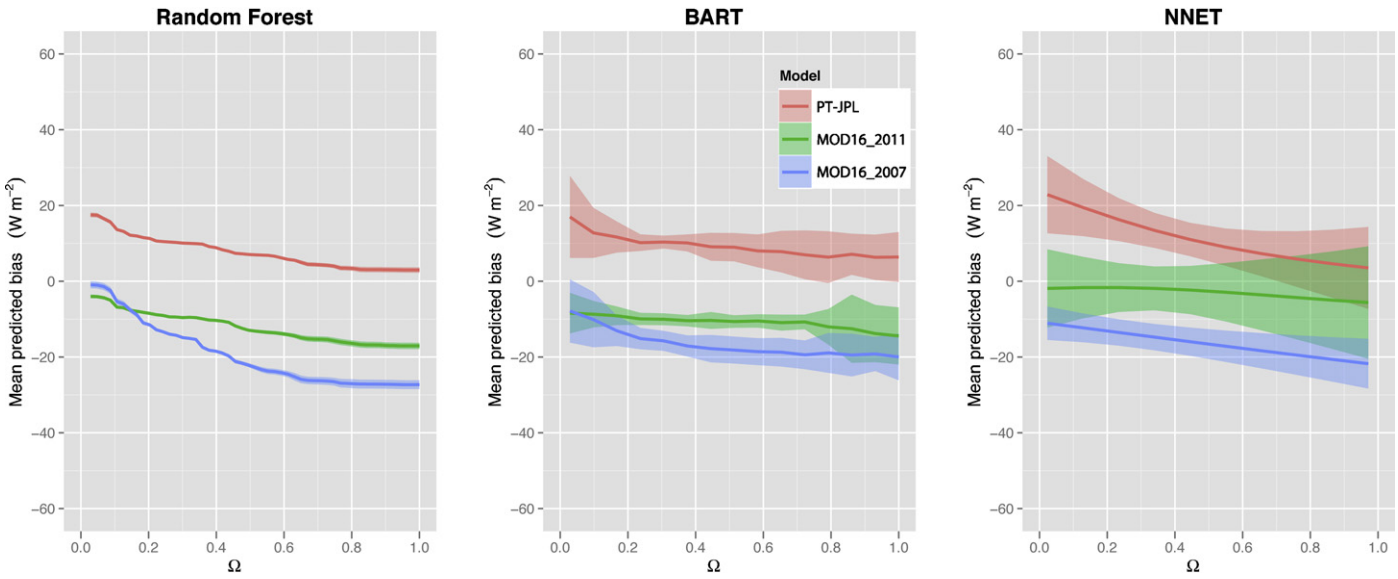

Fig. 8. Partial dependence of mean prediction bias on Ω, estimated using each of the study's four statistical models. Shaded area represents 95 percent confidence interval around each line. Shaded areas represents the 95 percent confidence band for each partial dependence curve.

6. Discussion

The goal of this paper was to study the structure of the residual error in ET model predictions, specifically with respect to Ω, a proxy for canopy control of transpiration. We tested the null hypothesis was that Ω would be unrelated to model error using variable importance sampling and partial dependence plotting based on several statistical models. Though the strongest predictors of ET model error were *LAI* (for MOD16), *EVI* (for PT-JPL) and *observed ET*, we found evidence of a non-trivial role for Ω and r_c. Recognizing that Ω is largely driven by r_c, we tested a random forest model for the residual series of each latent heat product that iteratively incorporated the top-ranked variables from the importance sampling, including r_c and Ω. We found that the importance of Ω for explaining the error of PT-JPL was due primarily to its functional relationship to r_c, whereas for MOD16_2011 and MOD16_2011 both variables had unique explanatory power with respect to the residual series. All models, therefore, could improve with respect to their description of biophysical constraints on transpiration—whether those constraints are modeled implicitly or explicitly—and the MOD16 models could further improve with respect to their modeling of the surface–atmosphere interactions captured by the decoupling coefficient.

What are the implications of this study's results for improving ET models going forward? As a first step, modelers need to identify why models consistently under-predict large values of ET, despite an aggregate over-prediction bias. *Observed ET* was the single strongest predictor of model error over the entire study for MOD16, closely followed by *LAI*. For PT-JPL, *EBC*, R_n, and *observed ET*, were all top-ranked variables, depending on the statistical model used. Since EBC is a function of R_n and *observed ET*, this is evidence the energy balance closure is a significant error factor for PT-JPL. Energy balance closure has the potential to insert large biases into the accuracy of observed ET estimates. This, in turn, could call into question the reliability of our residual estimates. Richardson et al. (2008) also found that flux magnitude is strongly related to uncertainty for CO_2. MOD16 models were affected primarily by the magnitude of observed ET; PT-JPL was more sensitive to the difference between R_n and ET, energy balance closure. The question of how to explicitly incorporate uncertainties arising from energy balance closure is beyond the scope of this paper but warrants further research.

The second step for modelers is to identify why models' positive prediction bias increases so sharply with LAI and EVI. This could be related to r_c: currently both classes of models are setup to relax the canopy's control over transpiration as greenness increases. The rationale is that greener the canopies indicate a greater the supply of water available for ET, and thus a convergence between potential and actual ET. If this relationship does not hold, relatively large values of LAI or EVI, could correlate to over-relaxed model constraints on transpiration and therefore model over-prediction. This seems to be the case here. If the divergence between water availability and LAI and EVI is indeed the underlying cause of these indices' positive correlation with ET model bias, then this may recommend the current work being done to remotely predict soil moisture, e.g. the SMAP mission (Chen et al., 2011). Estimating soil moisture directly, rather than as a proxy of canopy reflectance, could potentially boost prediction performance substantially.

Finally, not all canopies will respond similarly to the combined effects of water stress, atmospheric demand, and photosynthetic energy. For example, research has shown significant species-level differences in strategies for managing leaf water potential through maximum stomatal conductivity and sensitivity to VPD (Mackay et al., 2003). At the regional scale these differences become less relevant, but at the tower scale Mackay et al. (2003) found them to be highly significant. Therefore, even after resolving the above issues there may be scope to incorporate more accurate models of canopy resistance or decoupling into revised evapotranspiration products. The challenge is how to do so with remotely sensed data? One possibility would be to solve r_c using the Penman–Monteith equation, with the "true" observed ET value provided by another evapotranspiration product, e.g. PT-JPL. This back-solved r_c could then be an input to an ET product explicitly requiring parameterized stomatal resistance.

This paper makes the convenient assumption that surface resistance (r_s) is equal to canopy resistance (r_c). In fact, this is not often the case. In sparsely vegetated canopies with large areas of exposed soil, it will often be the case that direct evaporation, rather than transpiration, is the main contributor to ET. Where the evaporation component of ET is large relative to the transpiration component r_s estimates will contain little information about the level of stomatal resistance. This introduces bias into our analysis, though in this study we have not attempted to quantify this or select sites

above a particular threshold of canopy density. Further studies could improve on our method.

Some authors (Pereira, 2004; Fisher et al., 2005) have attempted to improve PT-based models by varying the α coefficient. We have not tested this strategy here, nor do we recommend it as a basis of physics-based model improvement because it typically relies on site-level, empirical calibration, rather than modeling an underlying physical reality. Fisher et al.'s (2008) approach is to calculate a series of multipliers that collectively adjust the potential ET from PT using the default value of $\alpha = 1.26$.

An analysis of the effect of specific biome type, species, and land cover on error structure is missing from this analysis. To the extent that land cover, r_c, *LAI*, and *EVI* are correlated with each other, this means that we are missing a potentially valuable control from the error structure analysis. To have treated this variable with sufficient rigor would have expanded the manuscript significantly, as well as have been much more computationally costly when training the machine learning models implemented. Furthermore, we were limited by the fact that for the AmeriFlux data set some biome types are much more heavily represented than others, potentially biasing the results of an analysis including biome type. A future, detailed analysis of the effect if biome type on ET model error would be a valuable addition to the results contained in this paper.

Though Ω and r_c were not the most significant explanatory variables of ET model error, this paper gives evidence that the underlying biophysical factors they describe do have a systematic relationship to the error structure of ET predictions. Improved techniques for estimating resistances (such as the double model approach), or innovations such as soil water content data, have the potential to contribute to the improvement of these models. As a first step toward model improvement, we identify a consistent under-prediction bias associated with large values of ET and over-prediction associated with large canopy index values.

Acknowledgements

We thank Qiaozhen Mu for the time and energy she invested in helping us code her model and develop this paper's analysis; Robert McCulloch, maintainer of the R package BayesTree (Chipman and McCulloch, 2010) for contributing the code used to generate the variable importance plots in the BART analysis section; and AmeriFlux Principal Investigators and for their diligent efforts to provide high-quality flux data to the environment science community. Two anonymous reviewers provided helpful suggestions for improving the manuscript. The Jet Propulsion Laboratory, California Institute of Technology carried out the research described in this paper, under a contract with the National Aeronautics and Space Administration.

Appendix A.

Two of the machine learning techniques that we apply here—random forests (Breiman, 2001) and Bayesian additive regression trees (BARTs)—are sum-of-trees methods for combining classification and regression trees (CARTs) (Breiman et al., 1984). Trees can take two forms, depending on the response variable that they attempt to describe (De'Ath and Fabricius, 2000; Venables and Ripley, 2002). If the response variable (as is ET in this study) is numeric then the tree will be a regression tree, which typically minimizes the sum of squared residuals within population subgroups. If the response variable is a class the tree will be a classification tree, which splits subgroups according to a measure of distributional impurity. Since the dependent variable in this study (model prediction residuals) is continuous, we only treat regression trees in this paper. A simple regression tree begins with a single node representing the entire sample population used for tree growing. The tree growing algorithm then searches across all predictor variables to select a binary split that minimizes the sum of the squared residuals (SSR). For regression trees this will typically be the sum of the squared differences between each observation in the subgroup and that subgroup's population mean.[3] For example, consider a sample of model residuals from PT-JPL with an SSR—based on an overall sample mean of 20 $W m^{-2}$—of 1000 $W^2 m^{-4}$. Now, say that we split this population into two groups on the basis of whether an observation's Ω value is greater or less than 0.5. The group corresponding to $\Omega \geq 0.5$ has a lower sample mean of 5 $W m^{-2}$, while the group corresponding to $\Omega < 0.5$ has a larger sample mean of 40 $W m^{-2}$. With these new sample means we calculate new SSRs within each population subgroup of 200 $W^2 m^{-4}$ and 250 $W^2 m^{-4}$, respectively, for an overall SSR of 450 $W^2 m^{-4}$. If this is the maximum possible reduction in SSR that we can achieve across all variable splits, then we select it as the first split for our regression tree model. We continue to further subdivide the two resulting groups using the same methodology. We refer to a node leading to a split as a *parent* node, and the nodes proceeding from that split as *children* (Breiman et al., 1984). Nodes that split into other nodes are *interior* nodes, while nodes that are associated with a prediction are *terminal nodes*. The algorithm continues to grow the tree by splitting the terminal node the results in the maximum possible reduction in residual sum of squares. Such algorithms are "greedy" in that they only look one step ahead. In theory it would be optimal to test all possible trees, but such a technique is computationally infeasible, generating a "combinatorial explosion" of possible trees (Venables and Ripley, 2002). Though trees can be grown to the extent that every observation inherits its own unique terminal node, this is almost guaranteed to over-fit the model to the data. Various techniques for pruning trees according to their performance on a set of training data exist, such as *k*-fold cross validation and bootstrapping (Kohavi, 1995; Venables and Ripley, 2002). Trees therefore have the useful property of invariance under monotonic transformation of the variables, which is not the case in linear regression. We will also describe how the sum-of-trees models implemented here can be used for variable importance ranking partial dependence plotting. Sum of trees models improve over single tree models in their ability to account for additive effects among the variables, while preserving trees' ability to incorporate complex interaction effects (Breiman, 2001; Chipman et al., 2010). Unlike linear models, which assume a linear relationship between the regressors and response variable, classification and regression tree are only sensitive to the rank ordering of variables (Breiman et al., 1984; De'Ath and Fabricius, 2000).

By introducing randomness in the training sets and covariate ensembles used for tree growing, Random Forests often out-perform single trees (Breiman, 2001) in terms of prediction accuracy. The basic random forest, which we use for this analysis, is trained by taking *B* bootstrap samples of the original data, and randomly sampling *M* covariates out of the full ensemble of explanatory variables for each of the *B* samples. For each of the *B* samples a tree is grown to maximum depth (each observation inherits a single terminal node) using its corresponding sample of covariates, and the out-of-bag samples are used to cast "votes" for the value of each observation. These predicted values are then aggregated across each tree in the forest and used to determine the final predicted values. By randomly permuting the values of a given variable and examining how this permutation changes

[3] There are other, more complex criteria for selecting node splits, i.e. via the sum of squared residuals from a linear regression of the subset on a suite of predictor variables.

prediction performance random forests also provide useful variable importance metrics.

Bayesian additive regression trees (Chipman et al., 2010) expand upon earlier work by Chipman et al. (1998) in the field of Bayesian CART model search. In the 1998 paper the authors consider possible CART selections as having posterior probabilities given by:

$$p(T|X, Y) \propto p(Y|X, T)p(T) \quad (1A)$$

The prior probability, $p(T)$, is initially expressed as a function of T and Θ such that:

$$p(\Theta, T) = p(\Theta|T)p(T) \quad (2A)$$

where T represents a tree with b terminal nodes and $\Theta = (\theta_1, \theta_2, \ldots, \theta_b)$ associates a parameter value(s) θ_i with the ith terminal node. The authors solve for this prior probability by assuming independence between Θ and T and solving $p(\Theta|T)$ and $p(T)$ separately. Note that as an index for a given tree T also refers to the structure of that tree in terms of its shape/depth and node splitting rules. The authors use the rules p_{SPLIT} and p_{RULE} to determine (a) and (b), respectively, where:

$$p_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta} \quad (3A)$$

where η is a node index and d_η represents the depth of that node. The choices of β and α may be informed by plotting prior probabilities as a function of terminal nodes, but being parameters of a prior distribution, are fundamentally subjective. This splitting rule serves to restrict the likelihood of observing complex, potentially over-fit trees. If a node is selected for splitting, p_{RULE} randomly selects a variable x from the set of predictors in X according to a uniform distribution, and randomly selects either a category (if x is qualitative) or a single value (if x is quantitative) as the split point, s, again from a uniform distribution. The probability of observing a tree T, $p(T)$, is therefore a function of the probabilities from p_{SPLIT} and p_{RULE}.

A detailed explanation of how the authors choose $p(\Theta, T)$ would be beyond the scope of this paper. In short, that they select Gaussian forms for Θ such that the problem:

$$p(Y|X, T) = \int p(Y|X, \Theta, T)p(\Theta|T)d\Theta \quad (4A)$$

can be solved analytically. Trees with high posterior probability per Eq. (2A) are then searched for stochastically using the Metropolis–Hastings search algorithm (Hastings, 1970; Metropolis et al., 1953) that generates a chain of trees T^0, T^1, $T^2, \ldots, T^N$. For each step from T^i to T^{i+1} a candidate tree T^* is generated from T^i by randomly choosing to either (i) GROW the tree according to p_{RULE}, (ii) PRUNE a tree by collapsing two terminal nodes into their parents, (iii) CHANGE the splitting rule of an internal node according to p_{RULE}, or (iv) SWAP the splitting rules of a parent-child pair that are both internal nodes. After the candidate tree is generated T^{i+1} is set to T^* with the probability given by the Metropolis–Hastings decision function, otherwise it is set to T^i. BART is essentially an algorithm that generates multiple chains of Bayes' tree (e.g. 200) and aggregates the information in these chains by using the mean of the posterior probability distribution for each final tree as its summation weighting (Chipman et al., 2010).

References

Allen, R., Pereira, L., Raes, D., Smith, M., 1998. Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements. FAO, Rome, Italy, Report nr 56.

Arnell, N., Bates, B., Lang, H., Magnuson, J., Mulholland, P., 1996. Hydrology and Freshwater Ecology. Cambridge University Press, New York, NY, (USA).

Aubinet, M., Berbigier, P., Bernhofer, C., 2005. Comparing CO_2 storage and advection conditions at night at different CARBOEUROFLUX sites. Bound.-Lay. Meteorol. 116 (1), 63–93.

Baldocchi, D.D., 2008. Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. Aust. J. Bot. 56, 1–26.

Bergmeir, C., Benítez, J.M., 2010. Neural Networks in R using the Stuttgart Neural Network [Computer Program]. R Package Version 0.3.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Chen, F., Crow, W.T., Starks, P.J., Moriasi, D.N., 2011. Improving hydrologic predictions of a catchment model via assimilation of surface soil moisture. Adv. Water Resour. 34, 526–536.

Chipman, H., McCulloch, R., 2010. BayesTree: Bayesian Methods for Tree Based Models [Computer Program].

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: bayesian additive regression trees. Ann. Appl. Stat. 4, 266–298.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. Bayesian CART model search. J. Am. Stat. Assoc. 93, 935–948.

Cleugh, H.A., Leuning, R., Mu, Q., Running, S.W., 2007. Regional evaporation estimates from flux tower and MODIS satellite data. Remote Sens. Environ. 106, 285–384.

Cruse, H., 2006. Neural Networks as Cybernetic Systems, 2nd ed. Brains, Minds & Media, Bielefeld, Germany.

De'Ath, G., Fabricius, K., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

De Bruin, H.A.R., Holtslag, A.A.M., 1982. A simple parameterization of the surface fluxes of sensible and latent heat during daytime compared with the Penman-Monteith concept. J. Appl. Meteorol. 21, 1610–1621.

Elman, J.L., 1990. Finding structure in time. Cognitive Sci. 14, 179–211.

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, validate at 16 FLUXNET sites. Remote Sens. Environ. 112, 901–919.

Fisher, J.B., Baldocchi, D.D., Misson, L., Dawson, T., Goldstein, A.H., 2007. What the towers don't see at night: nocturnal sap flow in trees and shrubs at two AmeriFlux sites in California. Tree Physiol. 27, 597–610.

Fisher, J.B., DeBiase, T.A., Qi, Y., Xu, M., Goldstein, A.H., 2005. Evapotranspiration models compared on a Sierra Nevada forest ecosystem. Environ. Model. Softw. 20, 783–796.

Fisher, J.B., Malhi, Y., Bonal, D., Da Rocha, H.R., De Arajuo, A.C., Gamo, M., 2009. The land-atmosphere water flux in the tropics. Global Change Biol. 15, 2694–2714.

Friedl, M.A., McIver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sens. Environ. 83, 287–302.

Friedman, J., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1180–1232.

Garcia, A., Andre, R., 2000. Analysis of the Priestley–Taylor alpha parameter for a bean crop. Acta Horticult. 537, 151–157.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model. 160, 249–264.

Hasler, N., Avissar, R., 2006. What Controls Evapotranspiration in the Amazon Basin?, vol. 8, pp. 380–395.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.

Jarvis, P.G., McNaughton, K.G., 1986. Stomatal control of transpiration: scaling up from leaf to region. Adv. Ecol. Res. 15, 1–49.

Jimenez, C., Prigent, C., Mueller, B., Seneviratne, S.I., McCabe, M.F., Wood, E.F., Rossow, W.B., Balsamo, G., Betts, A.K., Dirmeyer, P.A., et al., 2011. Global intercomparison of 12 land surface heat flux estimates. J. Geophys. Res., 116.

Jin, Y., Randerson, J.T., Goulden, M.L., 2005. Continental-scale net radiation and evapotranspiration estimated using MODIS satellite observations. J. Hydrol. 308, 2302–2319.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S.I., Sheffield, J., Goulden, M.L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., et al., 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature 467, 951–954.

Kohavi, Ron, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence.

Langensiepen, M., Fuchs, M., Bergamaschi, H., Moreshet, S., Cohen, Y., Wolff, P., Jutzi, S.C., Cohen, S., Rosa, L.M.G., Fricke, T., 2009. Quantifying the uncertainties of transpiration calculations with the Penman–Monteith equation under different climate and optimum water supply conditions. Agric. Forest Meteorol. 149, 1063–1072.

Lee, X., Black, T.A., 1993. Atmospheric turbulence within and above a douglas-fir stand, II, Eddy fluxes of sensible heat and water vapour. Bound.-Lay. Meteorol. 64, 369–390.

Leuning, R., Zhang, Y.Q., Rajaud, A., Cleugh, H., Tu, K., 2008. A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman–Monteith equation. Water Resour. Res. 44, 1–17.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Liu, S., Lu, L., Mao, D., Jia, L., 2007. Evaluating parameterizations of aerodynamic resistance to heat transfer using field measurements. Hydrol. Earth Syst. Sci. 11, 769–783.

Mackay, D.S., Ahl, D.E., Ewers, B.E., Samanta, S., Gower, S.T., Burrows, S.N., 2003. Physiological tradeoffs in the parameterization of a model of canopy transpiration. Adv. Water Resour. 26, 179–194.

Malhi, Y., Aragão, L.E.O.C., Galbraith, D., Huntingford, C., Fisher, R., Zelazowski, P., Sitch, S., McSweeney, C., Meir, P., 2009. Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest. In: Proceedings of the National Academy of Sciences USA, pp. 1–6.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Monteith, J.L., 1965. Evaporation and the environment. Symp. Soc. Explor. Biol. 19, 205–234.

Mu, Q., Heinsch, F.A., Zhao, M., Running, S.W., 2007. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. Remote Sens. Environ. 111, 519–536.

Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115, 1781–1800.

Myeni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J.L., Glassy, J., Tian, Y., 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. Remote Sens. Environ. 83, 214–231.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proc. R. Soc. Lond. Ser. A 193, 120–146.

Pereira, A., 2004. The Priestley–Taylor parameter and the decoupling factor for estimating reference evapotranspiration. Agric. Forest Meteorol. 125, 305–313.

Pereira, A., Villa Nova, N., 1992. Analysis of the Priestley–Taylor parameter. Agric. Forest Meteorol. 61, 1–9.

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large scale parameters. Mon. Weather Rev. 100, 81–92.

Richardson, A.D., Mahecha, M.D., Falge, E., Kattge, J., Moffat, A.M., Papale, D., Reichstein, M., Stauch, V.J., Braswell, B.H., Churkina, G., Kruijt, B., Hollinger, D.Y., 2008. Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. Agric. Forest Meteorol. 148, 38–50.

Scott, R.L., 2010. Using watershed water balance to evaluate the accuracy of eddy covariance evaporation measurements for three semiarid ecosystems. Using watershed water balance to evaluate the accuracy of eddy covariance evaporation measurements for three semiarid ecosystems. Agric. Forest Meteorol. 150, 219–225.

Sheffield, J., Wood, E.F., Munoz-Ariola, F., 2010. Long-term regional estimates of evapotranspiration for Mexico based on downscaled ISCCP data. J. Hydrometeorol. Res. 11, 253–275.

Shuttleworth, W.J., 2007. Putting the vap into evaporation. Hydrol. Earth Syst. Sci. 11, 210–244.

Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8, 1–21.

Sumner, D.M., Jacobs, J.M., 2005. Utility of Penman–Monteith, Priestley–Taylor, reference evapotranspiration, and pan evaporation methods to estimate pasture evapotranspiration. J. Hydrol. 308, 81–104.

Takahashi, K., 2008. The global hydrological cycle and atmospheric shortwave absorption in climate models under CO < SUB > 2</SUB > forcing. J. Clim. 22, 5667–5675.

Thom, A.S., 1975. Momentum, mass and heat exchange of plant communities. In: Monteith, J. (Ed.), Vegetation and the Atmosphere. Principles Academic Press, London, New York, San Francisco.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer Publishing, U.S.A.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., 2002. Energy balance closure at FLUXNET sites. Agric. Forest Meteorol. 113, 223–243.