ABSTRACT

Title of Dissertation:	PARAMETERIZED AND MACHINE LEARING REGRESSION METHODS FOR ESTIMATING EVAPOTRANSPIRATION FROM SATELLITE DATA
	Corinne Minette Carter, Doctor of Philosophy, 2019
Dissertation directed by:	Professor Shunlin Liang, Department of Geographical Sciences

The studies in this dissertation present evaluation of and improvement to parametric and machine learning regression methods for estimating evapotranspiration from remote sensing. It includes three main parts. The first part is an assessment of parametric regression methods for obtaining evapotranspiration from vegetation index and other variables. It was found that including more variables tends to improve results, but the form of the regression formula does not make a large difference. Algorithm performance is not as good for wetland and agricultural sites as for other land cover types. Re-training of algorithms for those surface type results in some improvement. The second part consists of an evaluation of ten machine learning techniques for retrieval of evapotranspiration from surface radiation and several other variables. It is found that the best results are obtainable using all available input

variables to train the bootstrap aggregation tree, random kernel, and two- and three-hidden layer neural network algorithms. Performance is again found to be weaker for wetland and agricultural surface types than for other surface types. However, separate training of the machine learning algorithms with data from those surface types does not significantly improve performance. The third part consists of further refinement to the machine learning algorithms and application of the bootstrap aggregation tree method to generate evapotranspiration maps of the continental United States for 2012. It is found that separating snow and non-snow data points improves performance. Performance for all tested algorithms was similar against the validation data set, but best for the bootstrap aggregation tree using an independent test data set. Monthly mean maps of the continental United States are generated for the drought year 2012 using the bootstrap aggregation tree. Evapotranspiration levels are lower than those shown in comparison data sets for the growing season in the eastern United States, resulting from a low bias at high evapotranspiration values. Retraining with the training data set weighted towards higher evapotranspiration values reduces this discrepancy but does not eliminate it. It is clear that machine learning evapotranspiration algorithm results have a significant dependence on training data set composition.

PARAMETERIZED AND MACHINE LEARNING REGRESSION METHODS FOR ESTIMATING EVAPOTRANSPIRATION FROM SATELLITE DATA

by

Corinne Minette Carter

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Shunlin Liang, Chair
Dr. Martha Anderson
Professor Christopher Justice
Professor Zhanquing Li – Dean's Representative
Dr. Eric Vermote
Dr. Xiwu Zhan

Dedication

To my husband, Jonathan Mawdsley, and my parents, Jack and Shelley Carter

Acknowledgements

I am indebted to my advisor, Professor Shunlin Liang, for his direction of and close attention to my work. He has shown me great kindness and patience.

My committee members, Drs. Martha Anderson, Xiwu Zhan, and Eric Vermote, and Professors Chris Justice and Zhanquing Li, have made many good suggestions that have improved the quality of this work.

I have received valuable assistance from members of Professor Liang's research group, especially Kaicun Wang, Meredith Brown, Yuhan Rao, Yi Zhang, and Dongdong Wang.

I am also grateful for the encouragement of my family and friends.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents iv
List of Tables vi
List of Figures vii
Chapter 1: Introduction 1
Motivation and background 1
Measurement of evapotranspiration 2
Remote sensing of LE via empirical methods 5
Summary of questions to be addressed 7
Chapter 2: Comprehensive evaluation of empirical algorithms for estimating land surface evapotranspiration 9
Introduction 9
Background and motivation 9
Description of VI-based algorithms to be evaluated 11
Data 20
Ground-based 20
Remote sensing 23
Methods 23
Results analysis 29
Global comparison of algorithms and coefficient tuning 29
Evaluation of algorithms by land cover type 33
Re-training of coefficients by surface type 35
Test of effect of linear interpolation of vegetation indices 39
Test of effect of rapid VI changes at agricultural sites 41
Discussion 43
Conclusions 46
Supplementary Material 48
Chapter 3: Development of machine learning methods for estimating terrestrial evapotranspiration from remote sensing 68
Introduction 68
Methods 71
Description of machine learning algorithms 71
Procedure for testing machine learning methods 81
Data 83
Results 88
Initial time trials of ML algorithms 88
Combinations of input variables 91
Tuning of ML algorithms 95
Discussion 120
Conclusions 123
Chapter 4: Machine learning applied to remote sensing of evapotranspiration in the continental United States 126

Introduction 126
Data 128
Methods 132
Machine learning techniques used 132
Trials with different data sets 134
Mapping over continental United States 135
Results 138
Tuning trials 138
Mapping over the continental United States 151
Monthly mean comparison with flux towers 159
Discussion 166
Conclusions 170
Chapter 5: Conclusion 173
Bibliography 183

List of Tables

Table 2-1: Vegetation index-based algorithms reviewed and compared ……11
Table 2-2: Land cover type categories used for algorithm evaluation……..…22
Table 2-3: Re-derived coefficients for each algorithm using all data from all sites….24
Table 2-4: Stations used for comparison of all dates to day of composite only …….39
Table 2-5: Results of comparison between all dates and day of composite only…….40
Table 2-6: Median site statistics of 23 agricultural sites……42
Table 2-7: List of Fluxnet sites included in this study.....48
Table 2-8: Original published coefficients for regression algorithms…….....57
Table 3-1: Algorithms used in this study, with corresponding abbreviations...……..…72
Table 3-2: Input and validation data used in this study……83
Table 3-3: Time in seconds for one iteration for each algorithm……...…89
Table 3-4: Accuracy and timing tests for different combinations of input variables….92
Table 3-5: Adjustable parameters optimized using the small training data set…….96
Table 3-6: Adjustable parameters optimized using the large training data set…….....99
Table 3-7: RMSE for each optimized ML algorithm……..…102
Table 3-8: Optimized RMSE for ML methods with GLASS DSR and PAR……...120
Table 4-1: Validation and test statistics for GLASS data only, snow cases…..…139
Table 4-2: Validation and test statistics for GLASS data only, non-snow cases.……141
Table 4-3: Validation and test statistics for MCD43C4 data only..……...…143
Table 4-4: Statistics for GLASS and MCD43C4 data combined.....147
Table 4-5: Summary statistics for BAGTREE tests……...…150

List of Figures

Figure 2-1: Global distribution of flux tower sites……...……21
Figure 2-2: RMSE for each algorithm by site for all cover types……... 31
Figure 2-3: R^2 values by site with the original and re-derived coefficients……..……32
Figure 2-4: Bias by site for all algorithms and land cover types……...…... 33
Figure 2-5: Bias and RMSE for those surface types where performance differed from all sites……..……. 35
Figure 2-6: RMSE for agricultural sites ……...36
Figure 2-7: RMSE for wetland sites……...……..37
Figure 2-8: Bias for agricultural sites……..……38
Figure 2-9: Bias for wetland sites……..……..39
Figure 2-10: Boxplots of RMSE by site for each cover type……...……62
Figure 2-11: Boxplots of R^2 by site for each cover type……...……64
Figure 2-12: Boxplots of bias by site for each cover type……...… 66
Figure 3-1: RMSE for validation data set for RLR against parameter λ.....104
Figure 3-2: RMSE for validation data set for LASSO regression against parameter λ.....104
Figure 3-3: RMSE for ELASTIC regression with small training data set.....105
Figure 3-4: RMSE for ELASTIC regression with large training data set…...……106
Figure 3-5: RMSE for KRR algorithm tuned with small training data set……..… 106
Figure 3-6: RMSE for RKS tuned with small training data set…….....…… 108
Figure 3-7: VRMSE for RKS tuned with large training data set…….110
Figure 3-8: Variation in validation RMSE with subset split and pruning parameters for TREE algorithm, small training data set….110
Figure 3-9: Variation in validation RMSE with subset split and pruning parameters for TREE algorithm, large training data set.....111
Figure 3-10: Validation RMSE vs number of trees used in BOOST algorithm…...…112
Figure 3-11: BAGTREE algorithm RMSE values, small training data set……..….. 113
Figure 3-12: BAGTREE algorithm RMSE values, large training data set……...…113
Figure 3-13: Validation RMSE for 1 hidden layer neural networks…..……114
Figure 3-14: RMSE for 2 hidden layer neural network using small training data set.....115
Figure 3-15: RMSE for 2 hidden layer neural network using large training data set……..…..116
Figure 3-16: RMSE for three-layer NN trained with small training data set…...…...117
Figure 3-17: RMSE for three-layer NN trained with large training data set…...…...118
Figure 3-18: Validation RMSE for support vector regression……...…120
Figure 4-1: Histograms of LE in weighted training data sets……....……138
Figure 4-2: Derived versus observed LE for GLASS only snow test data set…….....141
Figure 4-3: Derived versus observed LE for GLASS only non-snow test data set….143
Figure 4-4: Derived versus observed LE for NBAR only test data sets…...……..146
Figure 4-5: Derived versus observed LE for GLASS plus NBAR test data sets…....150
Figure 4-6: January monthly mean LE maps……...……...153
Figure 4-7: April monthly mean LE maps……...……...155
Figure 4-8: July monthly mean LE maps……...……157

Figure 4-9: October monthly mean LE maps……...……..158
Figure 4-10: Closer match of higher weighted training data sets to ground observations……..... 160
Figure 4-11: Delayed detection of decreasing LE……...……...160
Figure 4-12: Weaker match of flat histogram results to flux tower LE……...…...160
Figure 4-13: Cases where including vegetation index in BAGTREE had little effect…….....…… 161
Figure 4-14: Cases where including VI in BAGTREE reduces LE dropoff sensitivity…….....162
Figure 4-15: Cases where including VI produces higher LE values……...……...162
Figure 4-16: Cases where NARR has high bias…….....……...163
Figure 4-17: Cases where higher NBAR is better match to flux tower LE…….....164
Figure 4-18: Cases of differences in sensitivity to falling LEs……...……165
Figure 4-19: Cases of differing performance of MODIS LE…….....165

Chapter 1: Introduction

Motivation and background

Some parts of the world are sufficiently dry that reductions in the usual amount of rainfall can have severe consequences, including shortages of drinking water (Muller, 2018), increased severity of wildfires (Holden et al., 2018), and degraded fertility of the land (Wickens, 1997). As a result of predicted changes in global climate and increased population and consumption pressures, water resource constraints are expected to become increasingly severe (Schewe et al., 2014; Vorosmarty, 2000). Present and future water resource limitations are an important motivation to understand regional and global water cycles and how they may be changing.

The water balance for the land surface can be represented as

$$P = ET + \Delta S + R \tag{1-1}$$

where P is the precipitation rate, ET is the evapotranspiration rate, ΔS is the rate of change of subsurface water storage, and R is the rate at which water flows away at the surface or underground. Evapotranspiration is the combination of direct evaporation from land surfaces and transpiration of water to the atmosphere by plants. Evapotranspiration is expressed in units of water depth per time, such as mm/ day.

Evapotranspiration is equivalent to latent heat transfer between the surface and atmosphere, and as such is a component of the surface energy balance,

$$LE = R_n - H - \Delta G \quad (1\text{-}2)$$

where R_n is the net radiative heating of the surface, H is sensible heat exchange between the surface and atmosphere, and ΔG is the rate of change of subsurface heat storage. One mm/ day of evapotranspiration is equivalent to about 26 W/m^2 of latent heat transfer, although this relationship varies with temperature.

Measurement of evapotranspiration

Of the components of the surface water and energy balance equations, evapotranspiration is one of the more difficult to measure. At a small scale, it is possible to obtain ET using weighing lysimeters. These instruments consist of a container holding soil and plants attached to a scale. Changes in the weight of the soil and plants can be compared to measured precipitation rates to obtain ET or LE. These measurements are considered accurate enough to be used as a standard for comparison to models or other instruments (eg. Upreti and Ojha, 2018, Feng et al., 2018, Moorhead et al., 2017), but are also highly localized.

The other ground-based measurements of LE that are commonly available are those from flux towers. These towers measure LE, often via eddy correlation, over a larger area of up to a few square km. If the area upwind of the flux tower is representative of the surrounding landscape, the LE estimate can be considered valid over a larger area. However, it is common that

inhomogeneities in the landscape put the representativeness of flux tower measurements into question. These inhomogeneities may be part of the cause of a persistent surface energy balance closure problem with flux tower measurements, where Equation (1-2) is violated by the measured surface energy balance components. The imbalance of flux tower measurements was found by Wilson et al. (2002) to be about 20%.

Flux tower measurements have other significant limitations. It would be impractical to fully cover a nation or continent with a sufficient number of flux towers to determine trends on that scale. The distribution of flux towers is also biased towards the northern hemisphere midlatitudes, meaning that tropical and polar climates are underrepresented relative to temperate climates. LE measurements on regional and global scales are only possible through remote sensing. The flux tower measurements are the most useful we have for validation of remote sensing measurements, since flux tower footprints are at sufficiently large scales.

Remote sensing methods for retrieval of LE have been reviewed by Kalma et al. (2008), Wang and Dickinson (2012), and Zhang et al. (2016) Broadly speaking, most of these methods fall into a few categories. Energy balance residual models such as SEBAL (Bastiaanssen et al., 1998) and ALEXI (Anderson, 1997) and their descendants obtain LE by deriving the other terms in Equation (1-2) with LE as the remainder. Other methods, such as that used for the MODIS evapotranspiration product MOD16 (Mu et al., 2011) make use of the Penman-Monteith formulation of LE as a function of an energy-controlled term and an atmospheric stability-controlled term

$$LE = \frac{\Delta(R_n - G) + \rho_a c_p (e_s - e_a)/r_a}{\Delta + \gamma(1 + r_s/r_a)} \quad (1\text{-}3)$$

(Monteith, 1965; Penman, 1948). In this formulation, Δ is the derivative of saturation water vapor pressure with respect to temperature, ρ_a is the density of air, c_p is the specific heat of air at constant pressure, e_s is the saturation vapor pressure, e_a is the actual vapor pressure, γ is the psychrometric constant, and r_a and r_s are bulk resistance coefficients for water transfer through the atmosphere and from the surface, respectively.

Still other remote sensing methods for LE retrieval make use of the Priestley-Taylor approximation:

$$LE = \alpha \left(\frac{\Delta}{\Delta+\gamma}\right) f(e) * (R_n - G) \quad (1\text{-}4)$$

(Priestley and Taylor, 1972). In this formula, α is an empirical constant with a value often given as 1.26, and f(e) is a function ranging from 0 to 1 representing how close surface conditions are to an ideal, well-watered state. Two examples of Priestley-Taylor type LE retrievals are found in Yao et al. (2013) and Yao et al. (2015)

There are also methods that make use of a combination of vegetation index and temperature, or albedo and temperature, defining a "dry edge" where LE is zero and a "wet edge" where LE is at its theoretical maximum for a well-watered surface, in the space of those variables and determining LE according to where an observation falls between those edges (eg. Merlin et al., 2014, Long and Singh, 2012, Price, 1990). Less often used methods include maximum entropy production (MEP) based on thermodynamic principles, models that link water consumption to carbon cycling, and estimation of ET as a residual of the water balance (Equation 1-1), all of which are reviewed in Zhang et al. (2016). The methods that are used in the present study are

empirical methods, in which the parameters, or even the formulation, of the relation between LE and the input observations is determined through training data sets where both the input observations and LE are known.

Remote sensing of LE via empirical methods

Empirical methods, in which a data set with known LE is used to train algorithms, can take different forms. The simplest methods linearly correlate LE with one or more variables (Wang and Liang, 2008; Yebra et al., 2013) Other empirical formulas are more complex, based on the Penman-Monteith (Glenn et al., 2010; Kamble et al., 2013) or Priestley-Taylor (Yao et al., 2015) equations. Yao et al. (2013) use a multi- source formulation, with Priestley-Taylor parameterizations of evaporation from the soil surface and from leaf surfaces as well as of canopy transpiration. There are many other empirical formulas for LE in the published literature, several of which are reviewed in Carter and Liang (2018) (Chapter 2 of this work is based on this paper.) All of the formulations reviewed there include a vegetation index as one of the parameters, and virtually all of them also include net or incoming surface radiation. The more complex formulations also include surface temperatures or other meteorological variables such as relative humidity and wind speed.

Empirical methods with specified formulas have advantages and disadvantages. The primary advantage is their simplicity. In many cases, they can be used with a small number of input variables. The computational and user knowledge base demands of these algorithms are low compared to many others. On the other hand, the derived coefficients are most applicable under similar conditions to those represented in the training data set. The formulas are also inflexible in

their input variable requirements, which could be an issue in situations of limited data availability.

Machine learning is an alternative approach to regression that has proved its utility in many areas both inside and outside of remote sensing. Unlike the classical regression methods that specify the relationship between variables in a formula, machine learning methods do not require a priori assumptions concerning the relation between input and output. Instead, the algorithm adjusts its state to fit a training data set where input and output values are both specified. The neural network, support vector machine, and regression tree methods are well known and are described in Hastie et al. (2009) Other methods such as the extreme learning machine (Huang et al., 2006) and random kernel (RKS) (Rahimi and Recht, 2009) methods are more recent inventions. These and several other machine learning methods are reviewed in Carter and Liang (2019b). (Chapter 3 is a more detailed version of Carter and Liang 2019b).

The machine learning methods have their strengths and weaknesses. Their primary advantage is their flexibility, since any set of input variables can be used to train the algorithm and the relationship between input and output is not specified in advance. However, they can be computationally demanding to train and also usually constitute a “black box” where the inner workings of the algorithm are not readily accessible to the user. Similar to the specified-formula regression algorithms, the best performance can be expected under the range of conditions represented in the training data. The potential and limitations of machine learning algorithms are explored in Carter and Liang (2019a) and Carter and Liang (2019b). (Chapter 4 of this work is based on Carter and Liang 2019a.)

Summary of questions to be addressed

There are three sets of questions addressed by this work.

The second chapter of this work addresses the question of which vegetation index-based regression algorithms with specified formulas perform the best for retrievals of LE. Whether performance of these algorithms varies by surface type, and how much performance can be improved by retraining the algorithm coefficients, are also addressed in that chapter.

The third chapter contributes to an understanding of how several machine learning methods perform when applied to the problem of LE retrieval from remote sensing in terms of accuracy and computational efficiency. Tests are made of performance for different combinations of input data and for different surface types.

The fourth chapter presents further development of ML algorithms for remote sensing of evapotranspiration and their use for mapping the continental United States during the drought year of 2012. Three different algorithms are tested for snow and non-snow conditions, and one algorithm, the bootstrapping aggregation (bagging) regression tree, is selected for making the maps. Monthly mean maps generated using this algorithm are compared to three other LE data sets. Based on the results of the map comparison, trials are made using reconfigured training data sets in an effort to reduce a low bias at high LE values that is especially apparent during the growing season in the eastern US.

Taken together, these studies contribute to our understanding of a range of statistical methods for determining LE from satellite data. It is hoped that these studies will contribute towards both further research and operational algorithm development in this area.

Chapter 2: Comprehensive evaluation of empirical algorithms for estimating land surface evapotranspiration

Introduction

Background and motivation

A broad review of LE measurement methods has been performed by Wang and Dickinson (2012). Measurements from lysimeters characterize LE on scales of meters, and LE measurements from eddy correlation flux towers such as the Fluxnet network (Baldocchi et al., 2001) typically have footprints on the order of hundreds of meters. However, ground-based measurements are limited in their applicability due to their small scale and restricted areal coverage, as well as by the significant overrepresentation of northern hemisphere midlatitude sites. In addition, there are many sites with temporal records of a few years or less, and where there is no ongoing data collection. As a result, there is a great deal of interest in remote sensing of ET at larger spatial scales and in more remote areas.

There are many simple regression formulas that have been developed for estimation of ET. Many of these regression formulas are based on vegetation indices (VI), as reviewed by Glenn et al. (2010). The most frequently used vegetation indices in ET algorithms are the normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI). These ratios between near infrared, red, and blue band reflectances (ρ_{NIR}, ρ_{red}, and ρ_{blue} respectively) are as follows:

$$NDVI = \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red}} \quad (2\text{-}1)$$

$$EVI = G_{EVI} \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + C_1 \cdot \rho_{red} + C_2 \cdot \rho_{blue} + L} \quad (2\text{-}2)$$

The standard EVI product calculated from MODIS data has the constants G_{EVI}, C_1, C_2, and L set to values of 1.0, 6.0, 7.5, and 2.5 respectively.

Vegetation indices have several advantages for use in evapotranspiration algorithms. They are available from multiple instruments and at resolutions down to tens of meters. They have a high degree of consistency between instruments (Brown et al., 2006; Steven et al., 2003) Vegetation indices typically change on time scales of weeks to months, so interpolation can be used between observations separated by multiple days with some confidence. Algorithms that include a dependence on surface temperature are likely to be more responsive on shorter time scales, but the faster rate of change of surface temperature makes interpolation between observations more problematic. Overall, vegetation index-based methods have the advantages of simplicity, utility under a wide range of conditions, and resilience in the presence of data gaps. However, they may not respond well under conditions of stress.

Little work has been done evaluating these vegetation index-based algorithms under different conditions or comparing them to each other or to LE values derived through other methods. The goal of this paper is to provide a comprehensive evaluation of a range of VI- based evapotranspiration algorithms, identifying their strengths and weaknesses relative to each other.

Description of VI-based algorithms to be evaluated

A number of authors have proposed formulas for LE based on vegetation indices, ranging from highly simplified, depending only on the VI value with no additional data, to more complex formulas requiring ancillary data such as net radiation, surface and atmospheric temperatures, and other meteorological variables. All formulas to be evaluated in this paper are summarized in Table 2-1

Table 2-1: Vegetation index- based algorithms reviewed and compared, with full algorithm names and short names used to identify the algorithms in the figures. Key to variables: NDVI- Normalized difference vegetation index, EVI- Enhanced vegetation index, R_n- Net radiation at surface, G- Ground heat storage, T_{a_avg} – Daily average atmospheric temperature, T_{a_max}- Daily maximum atmospheric temperature, T_{a_dTr}- Daily atmospheric temperature range, T_{s_avg}- Daily average surface temperature, T_{s_max}- Daily maximum surface temperature, T_{s_dTr}- Daily surface temperature range, LE_0- Potential evapotranspiration, R_s- Incoming solar radiation at surface, RH- relative humidity, e_s- Saturation water vapor pressure, ws- Wind speed, VPD- vapor pressure deficit.

Algorithm	Short name	Reference	Required input data
Yebra direct (ET)	YET	Yebra et al. (2013)	NDVI or EVI
Yebra evaporative fraction (EF)	YEF	Yebra et al. (2013)	NDVI or EVI, R_n, G
Helman exponential	HEx	Helman et al. (2015)	NDVI or EVI
Helman scaled	HSc	Helman et al. (2015)	EVI, T_{s_avg}

Wang 2007	W07	Wang et al. (2007)	NDVI or EVI, R_n, one of T_{a_avg}, T_{a_max}, T_{s_avg}, or T_{s_max}
Wang/ Liang	WL	Wang and Liang (2008)	NDVI or EVI, R_n, T_{s_dTr}, one of T_{a_avg}, T_{a_max}, T_{s_avg}, or T_{s_max}
Choudhury/ FAO56	Ch	Choudhury et al. (1994) Allen et al. (1998)	EVI, LE_0
Kamble/ FAO56	Kmb	Kamble et al (2013) Allen et al. (1998)	NDVI, LE_0
Wang 2010	W10	Wang et al. (2010a)	NDVI or EVI, R_s, RH, e_s, ws, T_{a_avg}
Yao 2011	Y11	Yao et al. (2011)	NDVI, R_n, T_{a_avg}, T_{a_dTr}
Yao 2013	Y13	Yao et al. (2013)	NDVI, R_n, G, T_{a_avg}, T_{a_dTr} or T_{s_dTr},
Yao 2015	Y15	Yao et al. (2015)	NDVI, R_n, G, T_{a_avg}, RH, VPD

A total of 12 algorithms, based on 11 separate publications, are reviewed and evaluated in this paper. For each algorithm, Table 2-1 gives a short name, the source publication(s), and required

input data. Some of the publications listed also include other algorithms that depend on remote sensing parameters other than NDVI or EVI, but only the VI-based algorithms are included here.

Two of the algorithms, Yebra ET (Yebra et al., 2013) and Helman exponential (Helman et al. 2015), depend on the vegetation index alone. These algorithms were trained using 16 Fluxnet sites each. The Yebra algorithm sites were distributed over six different land cover types with forest and cropland sites most common, while the Helman algorithms were developed specifically for Mediterranean ecosystems with cropland and grassland sites most represented. The Yebra ET formula

$$LE_{YET} = a + b * VI \tag{2-3}$$

is a linear function of a vegetation index VI (NDVI or EVI), while the Helman exponential formula

$$LE_{HEx} = a * \exp(b * VI) \tag{2-4}$$

is an exponential function of either NDVI or EVI. For each of these algorithms, regression coefficients were found for NDVI and EVI separately.

The Yebra EF formula (Yebra et al., 2013) treats the evaporative fraction

$$EF = LE/(R_n - G) \tag{2-5}$$

as a linear function of NDVI or EVI, resulting in

$$LE_{YEF} = (R_n - G)(a + b * VI) \tag{2-6}$$

The Helman scaled algorithm (Helman et al. 2015), trained with the same data set as the Helman exponential algorithm, depends on a EVI and daily mean surface temperature T_{s_avg}, scaled according to:

$$EVI_{scl} = EVI - b \tag{2-7}$$

$$LST_{scl} = c - \left(d * T_{s\_avg}\right)$$

$$if\ (LST/e) < LST_{scl}, \qquad LST_{scl} = LST/e$$

then obtaining LE as the product of these scaled parameters:

$$LE_{Hsc} = a * EVI_{scl} * LST_{scl} \tag{2-8}$$

Wang et al. (2007) and Wang and Liang (2008) have published two empirical algorithms:

$$LE_{W07} = R_n * (a_1 + a_2 * VI + a_3 * T) \tag{2-9}$$

and

$$LE_{WL} = R_n * (a_1 + a_2 * VI + a_3 * T + a_4 T_{s\_dTr}) \tag{2-10}$$

respectively. Eight sets of coefficients were derived for each of these formulas, for each possible combination of MODIS NDVI or EVI, and average or maximum daily surface temperature (T_{s_avg}, T_{s_max}), or average or maximum atmospheric temperature (T_{a_avg}, T_{a_max}). The Wang and Liang (2008) formula also includes daily surface temperature range (T_{s_dTr}) as a proxy for moisture availability. These formulas are based on the maximum correlations between LE and other variables measured at eight Bowen ratio tower sites in the US Southern Great Plains, and, in the case of Wang and Liang (2008), four additional eddy correlation tower sites also in the US. In both studies, the strongest correlation was with net radiation, with VI and temperature variables following.

Two of the published formulas parameterize evapotranspiration as a function of the potential evapotranspiration ET_0, or the equivalent latent heat transfer LE_0, defined as the ET that would occur from a standardized, well-watered ground cover given a set of atmospheric conditions. LE_0 is often derived from the standard surface conditions and the Penman-Monteith formula for LE (Monteith, 1965):

$$LE = \frac{\Delta(R_n - G) + \rho_a c_p \frac{VPD}{r_a}}{\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)} \tag{2-11}$$

where Δ is the derivative of saturation vapor pressure with temperature, ρ_a is the density of air, c_p the specific heat of air at constant pressure, VPD the vapor pressure deficit ($e_s - e_a$, where e_s is the saturation vapor pressure and e_a is actual vapor pressure), γ the psychrometric constant, and r_s and r_a are bulk aerodynamic resistance factors characterizing surface and atmospheric conditions

respectively. A frequently-used formula for estimation of ET_0 is given in FAO56 (Allen et al. 1998) After conversion to units of LE, the FAO56 formula becomes

$$LE_0 = 26.3 * \left[\frac{0.408\Delta(R_n - G) + \gamma\left(\frac{900}{T+273}\right)ws * VPD}{\Delta + \gamma(1 + 0.34ws)}\right] \quad (2\text{-}12)$$

where ws represents wind speed.

Choudhury et al. (1994) combined observations of agricultural fields in an arid climate with surface and radiative transfer modeling to obtain a transpiration coefficient as a function of vegetation index. Glenn et al. (2010) proposed neglecting the bare soil evaporation in this formula, resulting in a formula for LE in terms of LE_0:

$$LE_{Ch} = LE_0\left(1.0 - \frac{EVI_{max} - EVI}{EVI_{max} - EVI_{min}}\right) \quad (2\text{-}13)$$

Choudhury et al. (1994) suggested using $EVI_{max} = 0.95$ and $EVI_{min}=0.05$.

Kamble et al. (2013) suggested a linear function of NDVI for obtaining LE based on LE_0, and derived coefficients based on agricultural sites in the US Great Plains:

$$LE_{Kmb} = LE_0(a * NDVI - b) \quad (2\text{-}14)$$

Wang et al. (2010a) developed their formula based on the approach of Penman (1948), estimating LE as consisting of two components, one controlled by available energy and another by atmospheric resistance. They developed the regression formula

$$LE_E = \frac{\Delta}{\Delta+\gamma} R_s[a_1 + a_2 VI + RHD(a_3 + a_4 VI)] \quad (2\text{-}15)$$

$$LE_A = \frac{\gamma}{\Delta+\gamma} ws * VPD[a_5 + RHD(a_6 + a_7 VI)]$$

$$LE_{W10} = a_8(LE_E + LE_A) + a_9(LE_E + LE_A)^2$$

with an energy control component LE_E dependent on incoming shortwave flux R_s and an atmospheric transmission control component LE_A. RHD represents the relative humidity deficit (as a function of relative humidity RH in percent: (100 – RH) / 100). This regression formula was trained using 64 eddy correlation and Bowen ratio ground stations, with the goal of obtaining globally-applicable coefficients. Unlike many of the other formulas, which contain an R_n or R_n – G term as a measure of available energy at the surface, the Wang formula uses the incoming solar radiation at the surface R_s. R_s may be measured directly, or estimated based on R_n, albedo, temperature, and relative humidity through the formula given in Wang and Liang (2009).

The three Yao et al. formulas considered here (2015, 2013, 2011), like the Wang et al (2010) model, are regressions based on pre-existing physical LE models. The Yao 2011 formula, developed for drought monitoring from a two-source LE model and data from 22 flux tower sites and global radiation and NDVI products, takes the form

$$LE_{Y11} = R_n^2(a_1 NDVI - a_2) + R_n\left(a_3 + a_4 T_{a_{avg}} + \frac{a_5}{T_{a_{dTr}}}\right) + R_n NDVI\left(a_6 + a_7 T_a + \frac{a_8}{T_{a\_dTr}}\right) \quad (2\text{-}16)$$

where T_{a_dTr} is the daily range of near-surface atmospheric temperature.

The Yao 2013 and Yao 2015 formulas are both based on the Priestley-Taylor (Priestley and Taylor 1972) parameterization, where r_s and r_a are combined into an empirically determined coefficient α with a value of 1.26 representing a well-covered and watered surface and a value f ranging from 0 to 1 representing constraints on LE:

$$LE = \alpha\left(\frac{\Delta}{\Delta+\gamma}\right) f * (R_n - G) \quad (2\text{-}17)$$

The Yao 2013 formula represents each of four separate components of LE through individual Priestley-Taylor parameterizations. These are a canopy transpiration component LE_c, a soil evaporation component LE_s, and components for evaporation from wet canopy and soil surfaces, LE_{ic} and LE_{ws}:

$$LE_{Y13} = LE_c + LE_s + LE_{ic} + LE_{ws} \quad (2\text{-}18)$$

$$LE_c = \alpha\left(\frac{\Delta}{\Delta+\gamma}\right)(1 - f_{wet}) f_v f_T R_{nc}$$

$$LE_s = \alpha\left(\frac{\Delta}{\Delta+\gamma}\right)(1 - f_{wet}) f_{sm}(R_{ns} - G)$$

$$LE_{ic} = \alpha\left(\frac{\Delta}{\Delta+\gamma}\right) f_{wet} R_{nc}$$

$$LE_{ws=} \alpha \left(\frac{\Delta}{\Delta + \gamma}\right) f_{wet}(R_{ns} - G)$$

The parameters f_{sm} and f_T represent soil moisture and temperature constraints respectively, f_v is fractional vegetation cover, f_{wet} is relative surface wetness parameterized as the relative humidity deficit to the fourth power, R_{nc} is net radiation to the vegetation canopy, and R_{ns} is net radiation to the soil. These variables are in turn parameterized in terms of vegetation index, daily average temperature, and daily temperature range. Separate sets of coefficients were derived using atmospheric and surface daily temperature ranges.

The Yao (2015) formula, which is similar in its basis to that of Fisher et al. (2008), is also based on the Priestley-Taylor equation, in this case with constraints on all sources of LE combined into one formulation. It was also developed for global applications, and the coefficients were trained with data from 240 Fluxnet sites.

$$LE_{Y15} = \phi \frac{\Delta}{\Delta+\gamma}(R_n - G)\left[a_1 + a_2 T_{a\_avg} + a_3 \left(\frac{RH}{100}\right)^{VPD} + VPD(a_4 NDVI - a_5)\right] \quad (2\text{-}19)$$

In summary, a range of formulas for obtaining LE from VI exist with different theoretical bases, degrees of complexity, and other input variables required. Some have forms that have a physical basis, but all ultimately depend on empirical regression for training of coefficients. In most cases they were trained with a limited number of ground sites, so it is desirable to test whether improvements can be made to their performance by using a larger training data set.

Vegetation index-based regression formulas are of interest because vegetation index and many of the other variables used are readily available over periods of years to decades, so they can be used to diagnose global and regional trends and anomalies. They are especially useful in situations of limited data availability. However, they are likely less useful than other LE algorithm classes for rapid drought identification, due to the relatively slow response of vegetation index to drying conditions. The focus of this study is to identify patterns of stronger or weaker performance within the class of vegetation index- based regression algorithms so that users of algorithms within this class will have guidance regarding which of these algorithms will be likely to produce the best results.

Data

Ground-based

A total of 184 flux tower sites were used, 119 from the Ameriflux network (http://ameriflux.lbl.gov) and 65 from the Fluxnet2015 data set (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/). All available sites with at least 3 continuous years of data were included. No screening of the sites was performed to exclude low quality or high spatial heterogeneity sites, keeping a large number of sites but possibly including some lower quality data. Most of the Ameriflux sites were within the United States, with good representation of the latitude range and land cover types of the continental US and Alaska. Eleven of the Ameriflux sites are Canadian, one Mexican, and one Brazilian. The Fluxnet2015 sites are mostly in Europe with some in Asia and Africa, cover a wide range of surface types and climates, but have the northern midlatitude bias typical of flux tower records. A total of 1166

site-years of data from 184 sites was used. The global distribution of these sites is shown in Figure 2-1. The IGBP surface types represented in the combined Ameriflux and Fluxnet2015 data, the categories used for further analysis here, and the number of sites and total site-years in each category are listed in Table 2-2. A list of all of the sites used is given in Supplementary Material, Table 2-7.

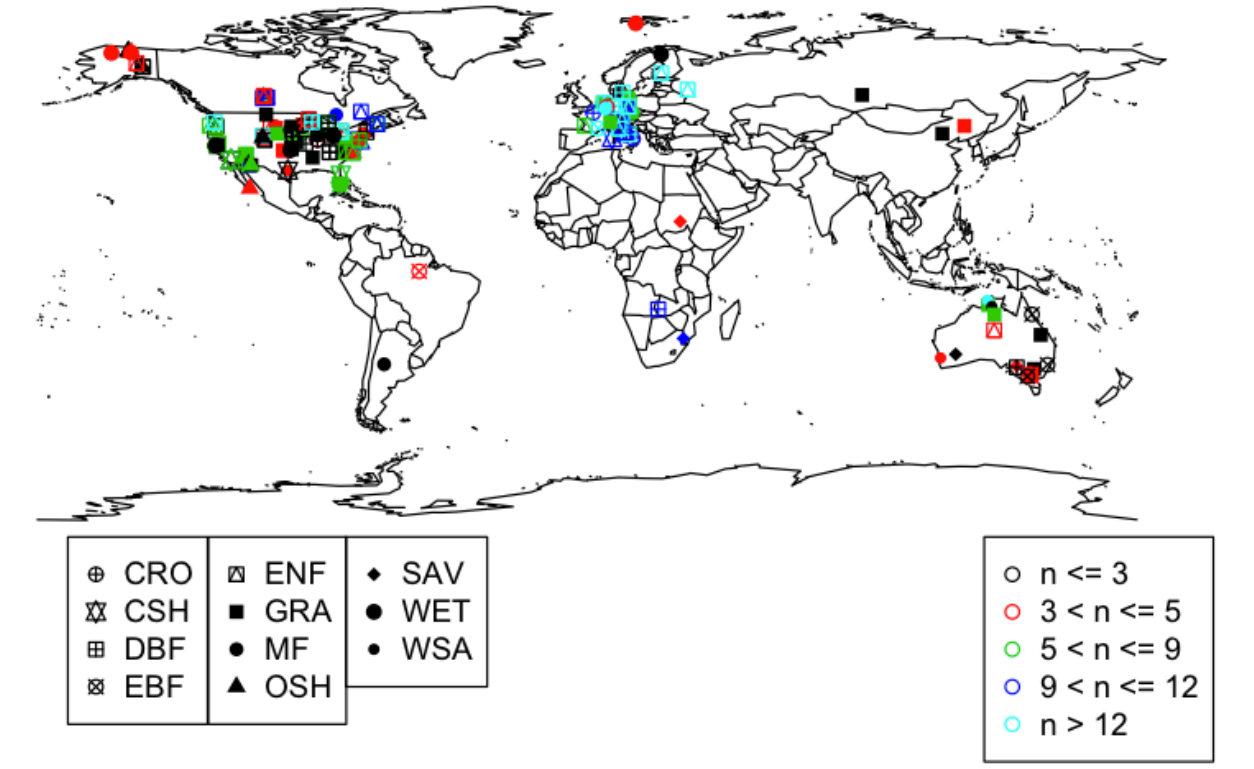


Figure 2-1: Global distribution of flux tower sites used in this study. Colors of points indicate number of years of data used from each site. Shapes of points indicate IGBP ecosystem type: CRO- crop, CSH- closed shrubland, DBF- deciduous broadleaf forest, EBF- evergreen broadleaf forest, ENF- evergreen needleleaf forest, GRA- grassland, MF- mixed forest, OSH- open shrubland, SAV- savannah, WET- wetland, WSA- woody savannah

Table 2-2: Land cover type categories used for algorithm evaluation, with IGBP classes included, number of sites available, and total site-years of data used for each.

Category	**Included IGBP classes**	**Number of sites**	**Total site-years**
Agricultural	CRO	23	115
Grassland	GRA	35	181
Deciduous	DBF, DNF, MF	29	228
Evergreen	EBF, ENF	50	392
Savannah	SAV, WSA	13	80
Shrub	CSH, OSH	18	76
Wetland	WET	13	94

The flux tower observations were preprocessed to obtain daily values of LE and all parameters required by the algorithms except for vegetation indices and albedo. For those days with at least 40 of 48 half hourly observations available for all variables, daily mean values of all required meteorological and energy balance variables were calculated. Adjustment was made to LE values to compensate for lack of energy balance closure by assuming a constant Bowen ratio, but no correction was made for ground heat storage above the flux plate. No modeled or gap-filled data were used, so days with insufficient flux tower data are not represented in our analysis. For atmospheric and surface temperatures, daily maximum and minimum values were also found and daily temperature ranges calculated. Days with and without snow cover were included in the data set.

Remote sensing

MODIS Collection 5 Terra 250m NDVI and EVI products (MOD13Q1, Didan 2015) and Terra/ Aqua 500m combined albedo (MCD43A, Schaaf and Wang 2015) time series were obtained for each site, for the same time period as the available flux tower data where it overlaps with the MODIS record. Subsets of each product were obtained from the Oak Ridge National Laboratory DAAC (https://daac.ornl.gov/MODIS/modis.shtml) Standard QC screening was applied. A 1km subset size was used, and all pixels that passed QC screening were included in calculations of mean NDVI, EVI, and albedo. (Preliminary testing with 0 km (same pixel), 1 km, and 3 km subset sizes indicated very little difference in LE algorithm results. Restricting included pixels to those with the same surface type as the central pixel also had a negligible effect.) Under ideal conditions VI is available every 16 days and albedo every 8 days, but longer data gaps exist in some locations due to insufficient high-quality pixels. VI and albedo were both linearly interpolated to generate daily time series.

Methods

Each model was first used to calculate LE (LE_{mod}) for each day where sufficient flux tower data was available at every site with the original published coefficients then compared against the ground observation LE (LE_{obs}). The coefficients for each algorithm were then re-derived using Levenberg-Marquardt fitting initialized with the published coefficient values. For purposes of algorithm evaluation, the last year of each site time series was reserved for testing and coefficients were trained with the remaining data. The algorithm evaluation results shown below all use this division of training and test data. In addition, a set of coefficients for each algorithm was derived using all available data, with results shown in Table 2-3. The coefficients for each algorithm from its original publication are given in Table 2-7 in the Supplementary Material.

Table 2-3: Re-derived coefficients for each algorithm using all available data from all sites. For the Yao (2013) and Yao (2015) algorithms, a set of coefficients was derived using a variable value of the Priestley-Taylor coefficient α and a constant α of 1.26.

Algorithm	Short name	Version	Re-derived coefficients
Yebra ET	YET	NDVI	a = -0.4589, b = 81.7987
		EVI	a = -1.2841, b = 149.9876
Yebra EF	YEF	NDVI	a = 0.02867, b = 0.6131
		EVI	a = 0.04879, b = 1.0316
Helman exponential	HEx	NDVI	a = 13.3611, b = 2.0344
		EVI	a = 17.0592, b = 2.8873
Helman scaled	HSc		a = -1518.3715, b = 0.001387, c = 33.6520, d = -1.1212, e = -4807.2619
Wang 2007	W07	EVI, T_{a_avg}	a_1 = -0.04417, a_2 = 0.9481, a_3 = 0.006516
		EVI, T_{a_max}	a_1 = -0.06821, a_2 = 0.9715, a_3 = 0.005585
		EVI, T_{s_avg}	a_1 = -0.02849, a_2 = 1.0189, a_3 = 0.004237
		EVI, T_{s_max}	a_1 = 0.0004923, a_2 = 1.0416, a_3 = 0.001707
		NDVI, T_{a_avg}	a_1 = -0.09575, a_2 = 0.5815, a_3 = 0.007896
		NDVI, T_{a_max}	a_1 = -0.1300, a_2 = 0.5995, a_3 = 0.006939
		NDVI, T_{s_avg}	a_1 = -0.09734, a_2 = 0.6438, a_3 = 0.005862
		NDVI, T_{s_max}	a_1 = -0.05442, a_2 = 0.6493, a_3 = 0.002534
Wang/Liang	WL	EVI, T_{a_avg}	a_1 = 0.07223, a_2 = 0.6681, a_3 = 0.009505 a_4 = -0.009441
		EVI, T_{a_max}	a_1 = 0.03066, a_2 = 0.6862, a_3 = 0.008800, a_4 = -0.009861
		EVI, T_{s_avg}	a_1 = 0.08232, a_2 = 0.7360, a_3 = 0.008243, a_4 = -0.01089

		EVI, T_{s_max}	$a_1 = 0.08224$, $a_2 = 0.7293$, $a_3 = 0.008534$, $a_4 = -0.01610$
		NDVI, T_{a_avg}	$a_1 = 0.05191$, $a_2 = 0.3879$, $a_3 = 0.01077$, $a_4 = -0.01048$
		NDVI, T_{a_max}	$a_1 = 0.0005417$, $a_2 = 0.4030$, $a_3 = 0.0101$, $a_4 = -0.01097$
		NDVI, T_{s_avg}	$a_1 = 0.04231$, $a_2 = 0.4534$, $a_3 = 0.009886$, $a_4 = -0.01223$
		NDVI, T_{s_max}	$a_1 = 0.04353$, $a_2 = 0.4484$, $a_3 = 0.01015$, $a_4 = -0.01837$
Choudhury/ FAO56	Ch		$EVI_{min} = 0.02355$, $EVI_{max} = 0.6117$
Kamble/ FAO56	Kmb		$a = 1.0452$, $b = -0.08478$
Wang 2010	W10	NDVI	$a_1 = -0.1387$, $a_2 = 1.9938$, $a_3 = 0.1542$, $a_4 = -2.1872$, $a_5 = 54.5977$, $a_6 = -79.8249$, $a_7 = 67.8465$, $a_8 = 0.6891$, $a_9 = -0.001150$
		EVI	$a_1 = -0.06988$, $a_2 = 3.1684$, $a_3 = 0.05535$, $a_4 = -3.2777$, $a_5 = 60.6141$, $a_6 = -99.1790$, $a_7 = 194.5842$, $a_8 = 0.6498$, $a_9 = -0.0009489$

Yao 2011	Y11		a_1 = -0.0009580, a_2 = -0.0004328, a_3 = 0.03625, a_4 = -0.003210, a_5 = 2.0066, a_6 = 0.5167, a_7 = 0.02503, a_8 = -2.7852
Yao 2013	Y13	T_{s_dTr}	α = 0.7888, $NDVI_{max}$ = 0.7052, $NDVI_{min}$ = -0.08551, T_{opt} = 32.8330, dTr_{max} = 30.9849
		T_{a_dTr}	α = 0.9987, $NDVI_{max}$ = 0.9198, $NDVI_{min}$ = -0.3712, T_{opt} = 25.5854, dTr_{max} = 22.9378
		T_{s_dTr}, α constant	$NDVI_{max}$ = 0.6486, $NDVI_{min}$ = -0.2723, T_{opt} = 141.0440, dTr_{max} = 10.9068
		T_{a_dTr}, α constant	$NDVI_{max}$ = 1.1234, $NDVI_{min}$ = -0.4696, T_{opt} = 25.7667, dTr_{max} = 15.7136
Yao 2015	Y15		α = 1.6445, a_1 = -0.002953, a_2 = 0.007440, a_3 = 0.4299, a_4 = 0.05653, a_5 = 0.01933
		α constant	a_1 = -0.003854, a_2 = 0.009711, a_3 = 0.5611, a_4 = 0.07379, a_5 = 0.02523

For each site and algorithm, RMSE, R^2, and bias were calculated based on LE_{mod} and LE_{obs}, where n is the number of days with valid data available:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(LE_{mod\_i} - LE_{obs\_i}\right)^2}{n}} \tag{2-20}$$

$$Bias = \frac{\sum_{i=1}^{n}\left(LE_{mod\_i} - LE_{obs\_i}\right)}{n} \tag{2-21}$$

$$R^2 = \left\{\frac{n\sum_{i=1}^{n}\left(LE_{mod\_i}LE_{obs\_i}\right) - \sum_{i=1}^{n}LE_{mod\_i}\sum_{i=1}^{n}LE_{obs\_i}}{\left[n\sum_{i=1}^{n}LE_{mod\_i}^2 - \left(\sum_{i=1}^{n}LE_{mod\_i}\right)^2\right]\left[n\sum_{i=1}^{n}LE_{obs\_i}^2 - \left(\sum_{i=1}^{n}LE_{obs\_i}\right)^2\right]}\right\}^2 \tag{2-22}$$

These results were then used to generate boxplots by algorithm. Boxplots were generated using all available sites, separately for the initial published and re-derived coefficients.

Similar statistical comparisons between algorithms were also conducted for the individual surface types specified in Table 2-2. Based on the results from the analyses with all surface types, four relatively well-performing algorithms with different theoretical bases (Yebra EF, Wang and Liang, Wang 2010, and Yao 2013) were selected for this evaluation. Coefficients were re-derived for each surface type using only data from sites with that type, again reserving the last year of each site for testing. Boxplots similar to those for all types were generated with the surface type specific coefficients and compared to results from the coefficients previously

derived from all available sites in order to evaluate whether use of data from only the same surface type improved algorithm performance.

Two additional tests were made of algorithm performance. In order to test whether linear interpolation of vegetation index and albedo was artificially improving algorithm statistics by introducing large numbers of non-independent data points, a subset of sites was selected and only data from the vegetation index composite dates were considered. Statistics from only the composite dates were compared to results including all days with sufficient flux tower data for each algorithm. An analysis was also performed for agricultural sites to assess whether interpolation over periods with sudden changes in vegetation index introduces error. To test for this effect, algorithm performance for agricultural sites was evaluated with dates with steep vegetation index slope (> 0.015/ day in NDVI or > 0.01/ day in EVI) excluded, then compared to agricultural site performance without this exclusion.

Results analysis

Global comparison of algorithms and coefficient tuning

Boxplots of RMSE, R^2, and bias by site for all surface types and for the original and re-derived coefficients are shown in Figures 2-2, 2-3, and 2-4. The algorithms are arranged left to right roughly in order of increasing complexity and number of input variables required. Figure 2-2a shows that the Yebra ET and Helman scaled algorithms have the highest median RMSEs. It is notable that these algorithms are the only ones that do not have any dependence on R_n. The best

performing algorithms have median RMSEs that cluster around 25-30 W/m^2 with the original coefficients.

Figure 2-2b shows the RMSE for all sites with the re-derived coefficients. All algorithms except Yao 2011 had similar or improved RMSE performance, with the best-performing models again having median RMSE in the 25-30 W/m^2 range. The most significant changes were for the Yebra and Helman algorithms, which have the simplest form and fewest required inputs. Most of the other algorithms had little change in median RMSE values, but RMSE tended to decrease for those algorithms that had higher RMSE using the original coefficients.

There are a significant number of outlier sites in the RMSE (Figure 2-2) and bias (Figure 2-4) results. Further investigation showed that different sites were outliers for different algorithms with the original coefficients (Figure 2-2a, 2-4a), with no systematic patterns apparent. With the re-derived coefficients (Figure 2-2b, 2-4b), six sites were responsible for most of the outliers. These sites either had 1 km subset areas that were unrepresentative of the area immediately surrounding the flux tower or were wetland sites. Wetland sites have greater bias and RMSE than other sites, as shown in Figure 2-5. The difference in performance between wetland sites and others is discussed in greater detail below.

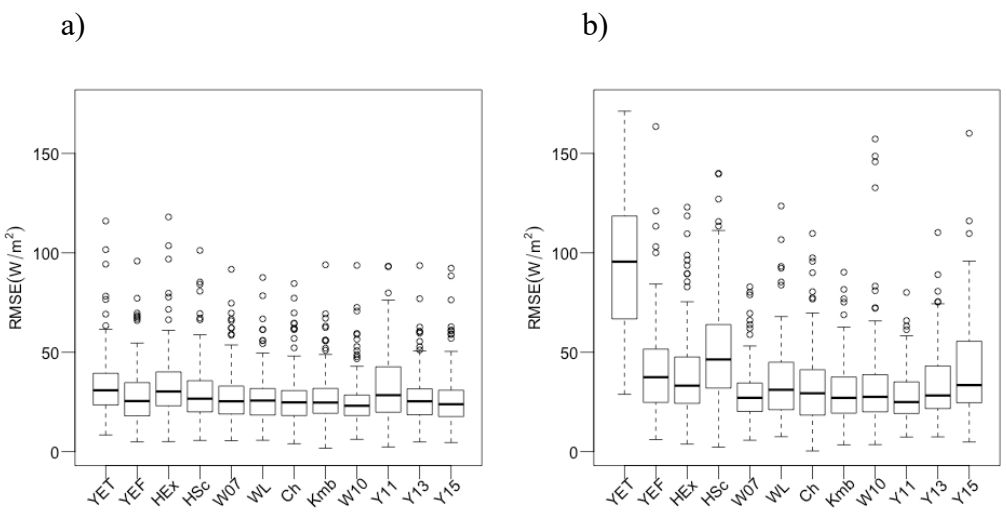


Figure 2-2: RMSE for each algorithm by site for all cover types. a) Using original published coefficients. b) Using re-derived coefficients. Key to algorithms: YET - Yebra ET, YEF - Yebra EF, HEx - Helman exponential, HSc - Helman scaled, W07 - Wang 2007, WL - Wang and Liang, Ch - Choudhury/ FAO56, Kmb - Kamble/ FAO56, W10 - Wang 2010, Y11 - Yao 2011, Y13 - Yao 2013, Y15 - Yao 2015

R^2 values for each site and algorithm are shown in Figure 2-3, with results for the original coefficients shown in Figure 2-3a and for the re-derived coefficients in Figure 2-3b. The median R^2 values for the best performing algorithms are between 0.6 and 0.7, with others, usually the simpler algorithms, having significantly lower values. Unlike the results for RMSE, re-fitting the coefficients did not have a strong impact on median R^2 or its distribution for any of the algorithms.

a) b)

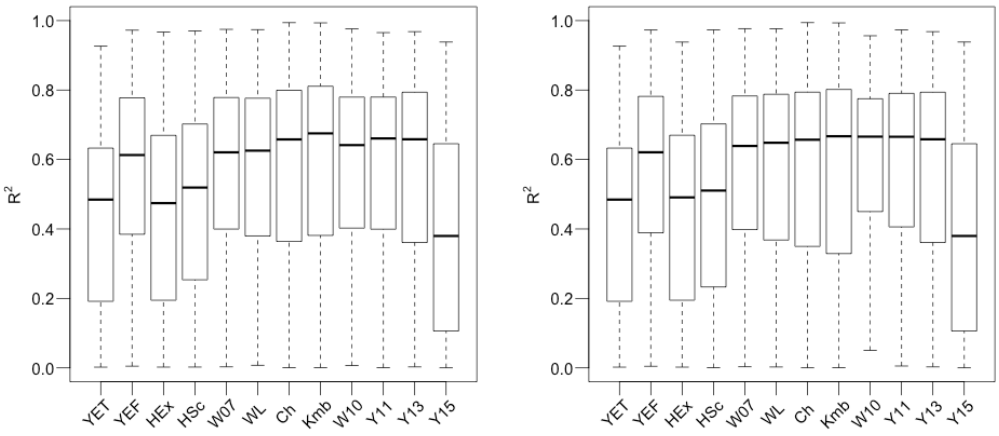


Figure 2-3: R^2 by site for all algorithms and land cover types. Results for original coefficients are shown in Figure 2-4a, and for re-derived coefficients in Figure 2-4b. Algorithm legend on horizontal axis is the same as for Figure 2-2.

Bias values for all sites and algorithms are shown in Figure 2-4, with results for the original coefficients in Figure 2-4a and for the re-derived coefficients in Figure 2-4b. The patterns here are similar to those seen for RMSE, with the simpler algorithms, especially Yebra ET, usually having the greatest absolute values of median bias with the original coefficients. Figure 2-4b shows that re-fitting the coefficients reduced the absolute value of median bias for many of the algorithms and reduced the range of bias values in many cases as well.

a) b)

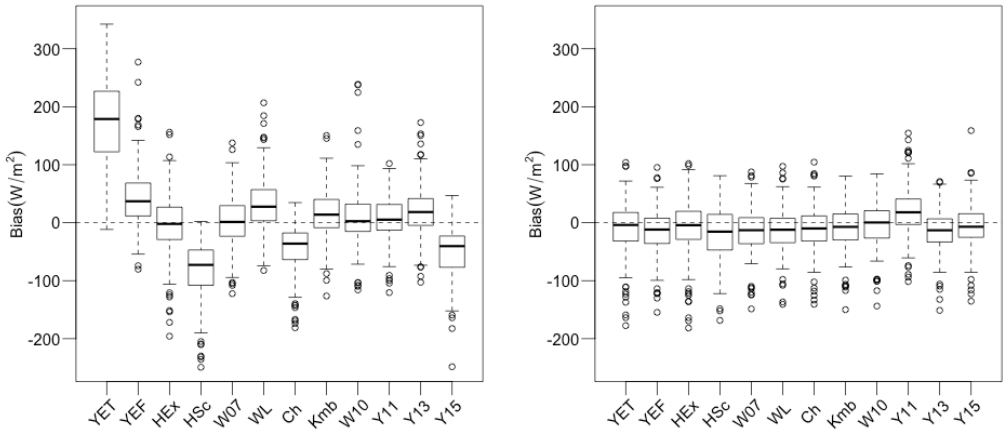


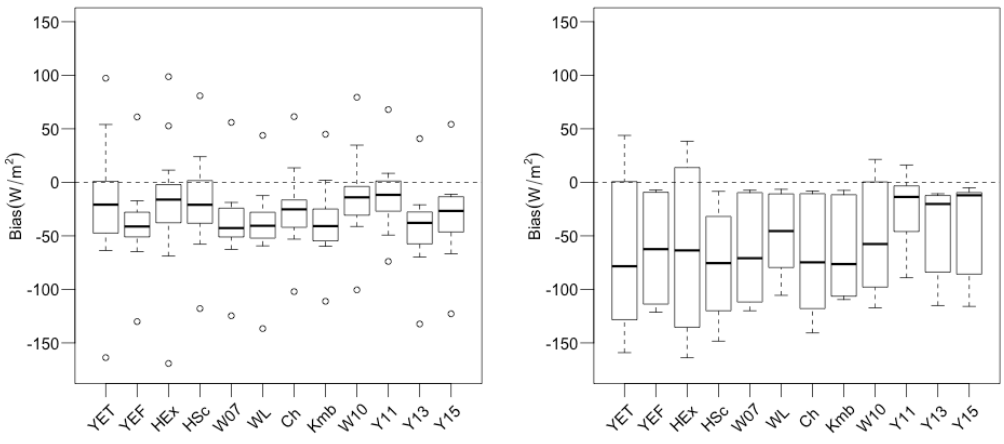
Figure 2-4: Bias by site for all algorithms and land cover types. Results for original coefficients are shown in Figure 2-4a, and for re-derived coefficients in Figure 2-4b. Algorithm legend on horizontal axis is the same as for Figure 2-2.

Evaluation of algorithms by land cover type

In general, there was little difference in the patterns of RMSE, R^2, and bias performance when the re-derived coefficients were used between surface types considered individually and what was shown in the previous section for all sites together. Exceptions to this overall pattern include higher R^2 values for agricultural, deciduous, evergreen, and grassland sites than for all sites considered together, and lower R^2 values for savannah, shrub, and wetland sites. There are also differences in bias and RMSE for agricultural and wetland sites.

Bias differences for agricultural and wetland sites, and RMSE differences for wetland sites, are shown below in Figure 2-5. Wetland sites (Figure 2-5a), and to a lesser degree agricultural sites (Figure 2-5b), showed a consistent low bias across algorithms, with typical bias values of around -25 W/m^2 for agricultural sites and -50 W/m^2 for wetland sites. The Yao 2011, Yao 2013, and Yao 2015 algorithms had a less pronounced bias than the others for wetland sites, but not for agricultural sites. In addition, RMSE for wetland sites was significantly higher than was typical for other surface types, with values of around 40 W/m^2 or more not being unusual (Figure 2-5c). The Yao algorithms had lower median RMSE, but RMSE was still relatively high for the sites where it was greatest.

a) b)



c)

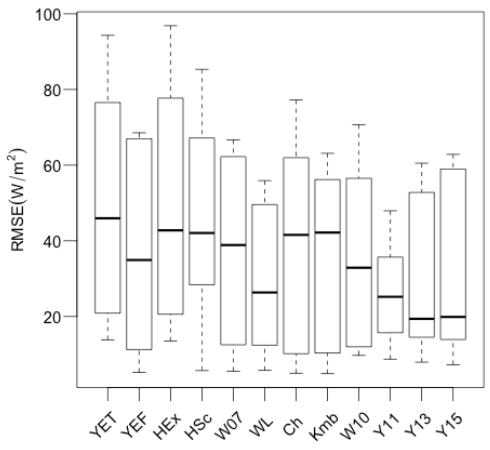


Figure 2-5: Bias and RMSE by site for those surface types where performance differed significantly from all sites with globally-derived coefficients. Figure 2-5a: Bias for agricultural sites. Figure 2-5b: Bias for wetland sites. Figure 2-5c: RMSE for wetland sites. Algorithm legend on horizontal axis is the same as for Figure 2-2.

Re-training of coefficients by surface type

For the four algorithms tested (Yebra EF, Wang and Liang, Wang et al. 2010, and Yao et al. 2013), training with data from sites from only one surface type did not result in much change from globally-trained coefficients for most surface types in most cases. (See Figures 2-10 through 2-12 in the Supplementary Material). The most pronounced exceptions occurred for bias and RMSE for agricultural and wetland sites, paralleling the results when comparing those surface types to the global results as described above. There were also modest improvements in RMSE for deciduous, grassland, and savannah sites (Figures 2-10b, 2-10d, and 2-10e), some modest increase in R^2 for savannah and decrease in R^2 for deciduous sites (Figures 2-11e and 2-11b) and modest reductions in absolute bias values for deciduous, grassland, and shrub sites

(Figures 2-12b, 2-12d, and 2-12f). For evergreen sites, bias values became somewhat more negative (Figure 2-12c). In all other cases, there was little change to the statistics, or performance improved for some algorithms and was reduced for others.

The results of surface type specific training for agricultural and wetland sites are shown in Figures 2-6 to 2-9. Figures 2-6 and 2-7 show a decrease in RMSE for agricultural sites and a reduction in the maximum RMSE by site for wetland sites, Figure 2-8 shows a decrease in bias for agricultural sites, and Figure 2-9 shows a decrease in bias for wetland sites.

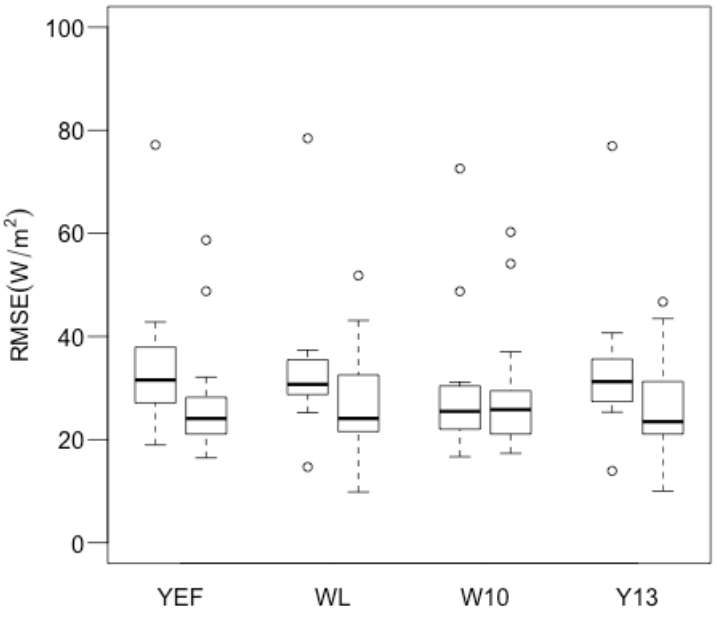
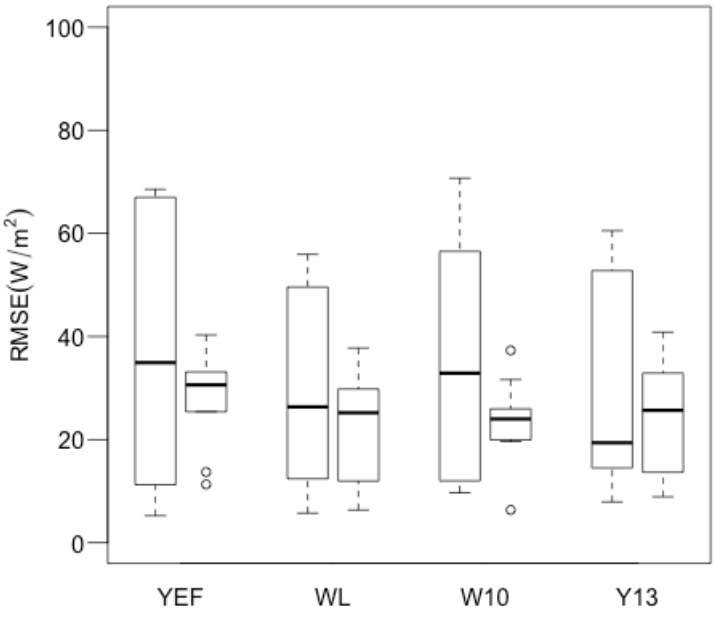


Figure 2-6: RMSE for agricultural sites for Yebra EF (YEF), Wang and Liang (WL), Wang et al. 2010 (W10) and Yao et al. 2013 (Yao13) algorithms. For each algorithm, left box is for training with data from all sites, and right box is for training with agricultural sites only.



f

Figure 2-7: RMSE for wetland sites. Algorithm labels on X axis are the same as for Figure 2-6. For each algorithm, left box is for training with data from all sites, and right box is for training with wetland sites only.

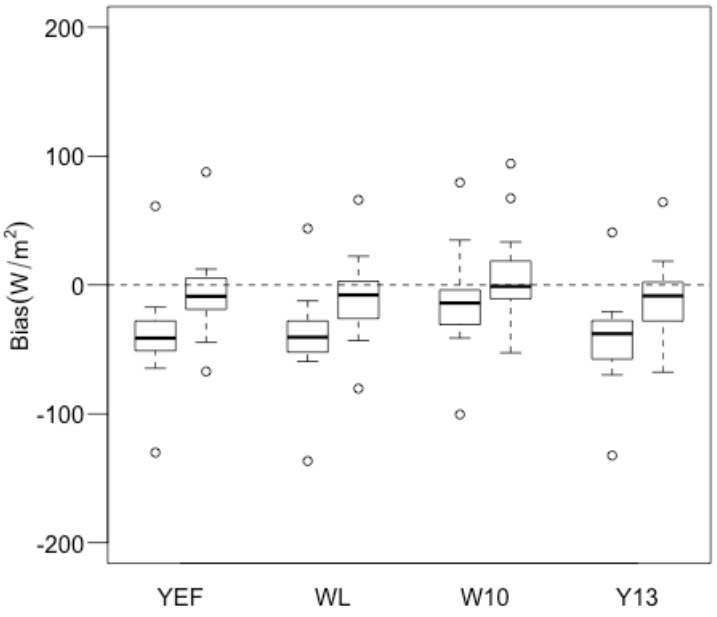


Figure 2-8: Bias for agricultural sites. Algorithm labels on X axis are the same as for Figure 2-6. For each algorithm, left box is for training with data from all sites, and right box is for training with agricultural sites only.

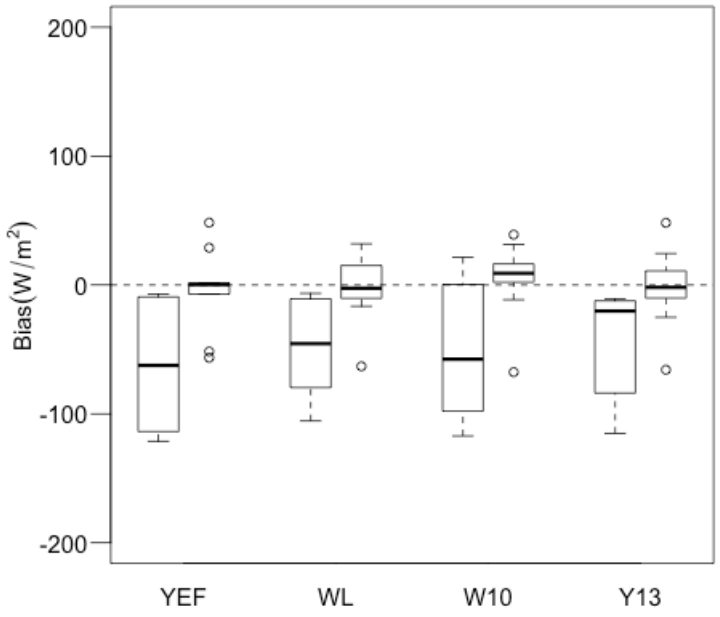


Figure 2-9: Bias for wetland sites. Algorithm labels on X axis are the same as for Figure 2-6. For each algorithm, left box is for training with data from all sites, and right box is for training with agricultural sites only.

Test of effect of linear interpolation of vegetation indices

The possibility that the statistical results of this analysis are being affected by the large number of non-independent data points introduced by linear interpolation of vegetation indices was tested. This was done using seven stations that each had a long data record, in order to obtain a significant number (659) station-days where that were both a composite date and had sufficiently complete Fluxnet records. These stations, listed in Table 2-4, also represent seven different land cover types. The analysis was conducted for seven of the best-performing algorithms.

Table 2-4: Stations used for comparison of results from all dates to day of composite only.

Station	Site ID	IGBP class
Audubon Ranch	US-Aud	Grassland (GRA)
Blodgett Forest	US-Blo	Evergreen needleleaf forest (ENF)
Lost Creek	US-Los	Wetland (WET)
Rosemount G21 conventional corn/ soy	US-Ro1	Cropland (CRO)
Santa Rita mesquite	US-SRM	Woody savannah (WSA)
Soroe	DK-Sor	Deciduous broadleaf forest (DBF)
Walnut Gulch Lucky Hills Shrub	US-Whs	Open shrub (OSH)

The results of this analysis are shown in Table 2-5. It was found that R^2 was higher and RMSE lower when only the composite days were used. The bias was a few W/m^2 more negative in most cases. These results could be because accuracy was lost through interpolation, or because composites were taken on clear weather days and the algorithms performed better under those conditions. It appears not to be the case that the interpolation artificially improved the apparent performance of the algorithms.

Table 2-5: Results of comparison between all dates and day of composite only.

Algorithm	RMSE all days (W/m^2)	RMSE composite days (W/m^2)	Bias all days (W/m^2)	Bias composite days (W/m^2)	R^2 all days	R^2 composite days
Yebra EF (YEF)	32.628	28.871	-5.555	-8.038	0.474	0.619

Choudhury (Ch)	38.761	37.958	-17.583	-21.059	0.473	0.559
Wang 2010 (W10)	31.243	27.028	-5.581	-7.665	0.523	0.673
Wang and Liang (WL)	33.279	29.033	-6.809	-8.952	0.454	0.618
Yao 2011 (Y11)	33.850	29.000	6.432	4.137	0.432	0.586
Yao 2013 (Y13)	32.213	28.789	-6.776	-9.805	0.502	0.656
Yao 2015 (Y15)	31.830	26.886	-2.258	-2.914	0.489	0.657

Test of effect of rapid VI changes at agricultural sites

At agricultural sites, there are periods where vegetation indices change rapidly, notably at harvest but also during greenup at the beginning of the growing season. The possibility that the vegetation index interpolation might not be as accurate at those times and degrade algorithm performance as a result was examined. The significance of this effect was tested using the 23 agricultural sites and seven algorithms. The median site RMSE, bias, and R^2 were found excluding those times where absolute value of the slope of NDVI > 0.015/ day, or of EVI > 0.01/ day, and compared against the results when all days were included. The results of this analysis are shown in Table 2-6. The performance of the algorithms was not much different between the

cases, or slightly worse when the steep VI slope periods were excluded. It does not appear that periods with steep VI slope are introducing additional error to the results for agricultural sites.

Table 2-6: Median site statistics of 23 agricultural sites, comparing results with and without exclusion of steep slope in vegetation indices.

Algorithm	RMSE all days (W/m^2)	RMSE VI slope exclusion (W/m^2)	Bias all days (W/m^2)	Bias VI slope exclusion (W/m^2)	R^2 all days	R^2 VI slope exclusion
Yebra EF (YEF)	28.892	29.699	-38.533	-39.340	0.685	0.682
Choudhury (Ch)	36.017	36.651	-51.922	-54.332	0.622	0.616
Wang 2010 (W10)	23.459	24.557	-7.470	-9.063	0.645	0.647
Wang and Liang (WL)	30.560	31.386	-36.285	-37.540	0.694	0.692
Yao 2011 (Y11)	24.746	25.386	-22.921	-23.666	0.666	0.676
Yao 2013 (Y13)	29.944	31.098	-34.811	-35.823	0.664	0.664

Yao 2015 (Y15)	24.056	24.125	-25.712	-26.290	0.688	0.688

Discussion

There has been a significant amount of effort devoted to measurement of evapotranspiration at regional to global scales, due to the variable's importance for a wide range of applications. At these scales, remote sensing is required for at least some of the input data. A large number of remote sensing methods to obtain LE have been developed, and the empirical methods evaluated here are just a subset of those available. There has been a significant amount of work evaluating different LE data sets at global (Jiménez et al. 2011; Mueller et al. 2011), and regional scales (e.g. Mao and Wang 2017; Chen et al. 2014) The focus of these studies has usually been on comparing different "families" of data sets (models vs. reanalyses vs. different remote sensing techniques), but less work has been done comparing results within each "family". The work done here was performed to fill in this gap for the "family" of regression- and VI-based models.

We found that most of the regression methods yielded useful estimates of LE with errors of similar magnitude to the differences in LE values between a wide range of methods according to Mueller et al., 2011 and Jimenez et al., 2011. The error levels we found were also consistent with the results provided by the original developers of these algorithms (references given in Table 2-1) and with the evaluation of VI-based LE retrieval methods by Glenn et al. (2010). Aside from the effect of inclusion of net radiation as an input parameter, the differences in performance were relatively modest, consistent with Mueller et al. (2011), where the two regression-based models included in the comparison had similar results.

The finding that, while increasing the number of input variables included improved the results, the specific formulation of the regression formula did not, was somewhat surprising. However, this is consistent with the fact that a broad range of different LE algorithms with different theoretical bases are all able to work with some skill, with no particular formulation coming out ahead consistently. The finding that formulas that do not include net radiation as a forcing variable stand out as performing especially poorly is consistent with Badgley et al. (2015), who found that changing the source of net radiation data used by a Priestley-Taylor model resulted in a greater change to its results than changing the source of meteorological or vegetation index data. In addition, the finding of the high significance of the net radiation variable is also consistent with Wang et al. (2007), who found a greater correlation of flux tower LE measurements to net radiation than to temperatures or vegetation indices.

The effect of land surface type on the performance of a range of empirical algorithms has not been examined in detail before this study. We found that there was some variation in performance, which is not unexpected, since different land cover types have different degrees of annual variation in vegetation index, and probably different relationships between VI and LE. (The differences in ecosystem response of different surface types to moisture stress has been a focus of much recent work, including De Keersmaecker et al., 2015; Joiner et al., 2018; and Seddon et al., 2016) Performance was weaker for agricultural sites, possibly because individual fields with different crop cover and irrigation characteristics occur at sub-pixel scales.

A probable reason for the low bias in wetland sites is that evaporation from the surface makes a more significant contribution to LE than for other site types, while vegetation indices are more of

an indicator of transpiration. Multiple studies (Allen et al., 2017; Malone et al., 2014; Runkle et al., 2014) have shown that H is a much smaller component of the surface energy budget than LE for wetland sites, and at least one study (Beigt et al. 2008) indicates that sensible heating can make a positive contribution to available energy at a wetland site. High values of LE relative to H are also seen in the wetland flux tower energy balance measurements used in this study. In addition, S. T. Allen et al. (2017) have shown that release of stored energy from the surface can contribute to available energy in the autumn season for a wetland. These sources of energy are available for evaporation but not transpiration. Along with higher surface moisture availability, these effects can result in high evaporative fraction and high rates of evaporation relative to transpiration from wetlands. Vegetation indices are not a good indicator of surface evaporation, as in the limiting case of open water where VIs are very low but surface evaporation is high.

The relationship between VI and LE is probably different for wetland and agricultural sites than for other surface types, especially so in the case of wetland sites. It appears that in those cases re-tuning the algorithm coefficients with just data from the same surface type adjusts for these differing relationships and produces a better fit.

There are other variables, such as precipitation and soil moisture, that are strongly related to LE but not used directly in any of the regression formulas reviewed. (They are parameterized in terms of other variables in some cases.) It should be possible to include precipitation and soil moisture from surface or microwave measurements, but it would be important to consider scaling effects when using these data. Surface precipitation and soil moisture measurements are in effect point measurements, limiting the possibilities for upscaling. On the other hand, while the

footprint of microwave observations is typically greater than the resolution of vegetation indices. For example, the resolution of the microwave-based Global Precipitation Measurement (GPM) is about 5 km. Global microwave soil moisture observations are currently available at scales of around 25 km, although there are ongoing efforts to downscale remote sensing soil moisture data sets, as reviewed by Peng et al. (2017). If precipitation is used as an input variable, a lag effect must be considered as the moisture made available in a precipitation event may remain available for several days. By contrast, soil moisture is a more immediate measure of water availability and a lag effect would not be expected.

Overall, the performance of the VI algorithms is consistent with what has been seen in previous work with those algorithms and with other methods for obtaining ET from remote sensing. For example, the RMSE values found here are comparable to the differences between ET values obtained by various methods as found by Jimenez et al., 2011, Li et al., 2018 and Sörensson and Ruscica, 2018 and the magnitudes of sensitivity to different inputs of the MOD16 algorithm found by Zhang et al., 2019. Where possible, it is preferable to use algorithms with more input data parameters if selecting between the algorithms tested in this study, since the specifics of the underlying basis appear to matter little. Simpler algorithms can perform almost as well as more complex ones, but it is more important that they be tuned with appropriate training data. At a minimum, inclusion of R_n as a parameter along with VI is recommended wherever possible.

Conclusions

In this study, we have noted certain patterns in the performance of vegetation index- based LE algorithms. Increasing the number of variables included in regression formulas tends to improve

performance, although the specific form of model used is not as significant. Those algorithms in which net radiation was one of the input variables produced much less error than those that did not, as demonstrated by the difference between the Yebra (2013) ET (YET) algorithm, and Yebra (2013) EF (YEF) algorithms, which are very similar to each other except that YEF has net radiation as an input while YET does not. (Figures 2-2, 2-3, 2-4). Tuning of the regression coefficients to the global data set improved performance in most cases, which is also demonstrated in Figures 2-4. This improvement was most significant for those models with fewer input variables. For wetland and agricultural surface types, tuning with data specific to that surface type produced improved results (Figures 2-6 to 2-8), but this was not the case for other surface types. Any user of VI-based regression methods would be well advised to ensure that the algorithms (especially the simpler ones) are tuned appropriately to the data to which the algorithms will be applied, and to consider separate tuning by surface type, especially for wetland and agricultural sites.

There are multiple opportunities for adaptation and improvement of the methods evaluated here. All of the input variables to the regression formulas are potentially available through remote sensing (Liang 2007, Liang et al. 2012) or reanalysis, so there is the potential for removing all dependence on ground-based observations. The accuracy of these remote sensing only retrievals would have a strong dependency on how accurately the remotely sensed quantities are retrieved. It is also likely that new empirical formulations will continue to be developed. It would be advisable to evaluate them relative to existing formulas like those considered in this chapter before using them for regional or global trend and anomaly detection.

Supplementary Material

Table 2-7: List of Fluxnet sites included in this study

Site code	Site name	Latitude	Longitude	IGBP Type
AR-SLu	San Luis	-33.4648	-66.4598	MF
AU-Ade	Adelaide River	-13.0769	131.1178	WSA
AU-ASM	Alice Springs	-22.283	133.249	ENF
AU-CPR	Calperum	-34.0021	140.5891	SAV
AU-Cum	Cumberland Plains	-33.6133	150.7225	EBF
AU-DaS	Daly River Cleared	-14.1593	131.3881	SAV
AU-DaP	Daly River Savanna	-14.0633	131.3181	GRA
AU-Dry	Dry River	-15.2588	132.3706	SAV
AU-Emr	Emerald Queensland	-23.8587	148.4746	GRA
AU-Gin	Gingin	-31.3764	115.7138	WSA
AU-GWW	Great Western Woodlands	-30.1913	120.6541	SAV
AU-How	Howard Springs	-12.4943	131.1523	WSA
AU-Lox	Loxton	-34.4704	140.6551	DBF
AU-RDF	Red Dirt Melon Farm	-14.5636	132.4776	WSA
AU-Rig	Riggs Creek	-36.6499	145.5759	GRA
AU-Rob	Robson Creek	-17.1175	145.6301	EBF
AU-Stp	Sturt Plains	-17.1507	133.3502	GRA
AU-Wac	Wallaby Creek	-37.4259	145.1878	EBF
AU-Whr	Whroo	-36.6732	145.0294	EBF
AU-Wom	Wombat	-37.4222	144.0944	EBF

AU-Ync	Jaxa	-34.9893	146.2907	GRA
BE-Bra	Brasschaat	51.3092	4.5206	MF
BE-Vie	Vielsalm	50.3051	5.9981	MF
BR-Sa3	Santarem Km83 Logged Forest	-3.018	-54.9714	EBF
CA-Gro	Groundhog River	48.2167	-82.1556	MF
CA-Oas	Saskatchewan Mature Aspen	53.6289	-106.1978	DBF
CA-Obs	Saskatchewan Mature Black Spruce	53.9872	-105.1178	ENF
CA-OjP	Saskatchewan Mature Jack Pine	53.9163	-104.692	ENF
CA-Qcu	Quebec Black Spruce/ Jack Pine Cutover	49.2671	-74.0365	ENF
CA-SF1	Saskatchewan Forest 1977 Burn	54.485	-105.8176	ENF
CA-SF2	Saskatchewan Forest 1989 Burn	54.2539	-105.8775	ENF
CA-SF3	Saskatchewan Forest 1998 Burn	54.0916	-106.0053	OSH
CA-TP1	Turkey Point 2002 Plantation White Pine	42.6609	-80.5595	ENF
CA-TP2	Turkey Point 1989 Plantation White Pine	42.7744	-80.4588	ENF

CA-TP3	Turkey Point 1974 Plantation White Pine	42.7068	-80.3483	ENF
CA-TP4	Turkey Point 1939 Plantation White Pine	42.7102	-80.3574	ENF
CA-TPD	Turkey Point Mature Deciduous	42.6353	-80.5577	DBF
CH-Cha	Chamau	47.2102	8.4104	GRA
CH-Dav	Davos	46.8153	9.8559	ENF
CN-Cng	Changling	44.5934	123.5092	GRA
CN-Du2	Duolon Grassland	42.0467	116.2836	GRA
CZ-wet	CZECHWET	49.0247	14.7704	WET
DE-Akm	Anklam	53.8662	13.6834	WET
DE-Geb	Gebesee	51.1001	10.9143	CRO
DE-Gri	Grillenberg	50.9495	13.5125	GRA
DE-Hai	Hainich	51.0792	10.453	DBF
DE-Kli	Klingenberg	50.8929	13.5225	CRO
DE-Obe	Oberbarenburg	50.7867	13.7213	ENF
DE-She	Selhausen	50.8706	6.4497	CRO
DE-Tha	Tharandt	50.9624	13.5652	ENF
DK-Fou	Foulum	56.4842	9.5872	CRO
DK-Sor	Soroe	55.4859	11.6446	DBF
FI-Hyy	Hyytiala	61.8475	24.295	ENF

FI-Lom	Lompolojänkkä	67.9972	24.2092	ENF
FR-Fon	Fontainebleau-Barbeau	48.4764	2.7801	DBF
FR-Gri	Grignon	48.8442	1.9519	CRO
FR-LBr	LeBray	44.7171	-0.7693	ENF
FR-Pue	Puechabon	43.7414	3.5958	EBF
IT-BCi	Borgo Cioffi	40.5238	14.9574	CRO
IT-CA1	Castel d'Asso 1	42.3804	12.0266	DBF
IT-CA2	Castel d'Asso 2	42.3772	12.026	CRO
IT-Cpz	Castelporziano	41.7052	12.3761	EBF
IT-Cp2	Castelporziano 2	41.7043	12.3573	EBF
IT-Col	Collelongo	41.8494	13.5881	DBF
IT-Lav	Lavarone	45.9562	11.2813	ENF
IT-MBo	Monte Bondone	46.0147	11.0458	GRA
IT-Noe	Arca di Noe	40.6061	8.1515	CSH
IT-PT1	Parco Ticino Forest	45.2009	9.061	DBF
IT-Ren	Renon	46.5869	11.4337	ENF
IT-Ro2	Roccarespampani 2	42.3903	11.9209	DBF
IT-SRo	San Rossore	43.7279	10.2844	ENF
IT-Tor	Torgnon	45.8444	7.5781	GRA
MX-Lpa	La Paz	24.12925	-110.43803	OSH
NL-Hor	Horstermeer	52.2404	5.0713	GRA
NL-Loo	Loobos	52.1666	5.7436	ENF

NO-Adv	Adventdalen	78.186	15.923	WET
RU-Fyo	Fyodorovskoye	56.4615	32.9221	ENF
RU-Ha1	Hakasia Steppe	54.7252	90.0022	GRA
SD-Dem	Demokeya	13.2829	30.4783	SAV
US-An1	Anaktuvuk River Severe Burn	68.99	-150.28	OSH
US-An2	Anaktuvuk River Moderate Burn	68.95	-150.21	OSH
US-An3	Anaktuvuk River Unburned	68.93	-150.27	OSH
US-AR1	ARM Woodward Switchgrass 1	36.4267	-99.42	GRA
US-AR2	ARM Woodward Switchgrass 2	36.6358	-99.5975	GRA
US-ARM	ARM SGP	36.6058	-97.4888	CRO
US-Aud	Audubon Ranch	31.59073	-110.51038	GRA
US-Bkg	Brookings	44.3453	-96.8362	GRA
US-Blk	Black Hills	44.158	-103.65	ENF
US-Blo	Blodgett Forest	38.8953	-120.6328	ENF
US-Bn1	Bonanza Creek, 1920 Burn	63.919813	-145.378178	ENF
US-Bn2	Bonanza Creek, 1987 Burn	63.919813	-145.3782	DBF
US-Bn3	Bonanza Creek, 1999 Burn	63.92268	-145.7442	OSH
US-Bo1	Bondville	40.0062	-88.2904	CRO
US-Bo2	Bondville Companion Site	40.009	-88.29	CRO
US-Br1	Brooks Field Site 10	41.9749	-93.6906	CRO
US-Br3	Brooks Field Site 11	41.97472	-93.69357	CRO

US-CaV	Canaan Valley	39.06333	-79.42083	GRA
US-Ced	Cedar Bridge	39.8379	-74.3791	CSH
US-ChR	Chestnut Ridge	35.9311	-84.3324	DBF
US-CPk	Chimney Park	41.067963	-106.118667	ENF
US-CRT	Curtice Walker Berger Crop	41.6285	-83.3471	CRO
US-Ctn	Cottonwood	43.95	-101.8466	GRA
US-Dia	Diablo	37.6773	-121.5296	GRA
US-Dix	Fort Dix	39.97123	-74.43455	MF
US-DK1	Duke Open Field	35.9712	-79.0934	GRA
US-Dk2	Duke Hardwood	35.9736	-79.1004	DBF
US-Dk3	Duke Loblolly	35.97817	-79.0942	ENF
US-Elm	Everglades Long Hydroperiod Marsh	25.5519	-80.7826	WET
US-Esm	Everglades Short Hydroperiod Marsh	25.4379	-80.5946	WET
US-Fmf	Flagstaff Managed Forest	35.1426	-111.7273	ENF
US-FPe	Fort Peck	48.3077	-105.1019	GRA
US-FR2	Freeman Ranch Mesquite Juniper	29.94949	-97.99623	WSA
US-FR3	Freeman Ranch Woodland	29.94	-97.99	CSH
US-Fuf	Flagstaff Unmanaged Forest	35.089	-111.762	ENF
US-Fwf	Flagstaff Wildfire	35.4454	-111.7718	GRA

US-GLE	GLEES	41.36653	-106.2399	ENF
US-GMF	Great Mountain Forest	41.96667	-73.23333	MF
US-Goo	Goodwin Creek	34.2547	-89.8735	GRA
US-Ha1	Harvard Forest	42.5378	-72.1715	DBF
US-Ho2	Howland Forest West Tower	45.2091	-68.747	ENF
US-Ho3	Howland Forest Harvest Site	45.2072	-68.725	ENF
US-IB1	Fermi Agricultural	41.8593	-88.2227	CRO
US-IB2	Fermi Prairie	41.8406	-88.241	GRA
US-ICh	Imnavait Creek Heath Tundra	68.6068	-149.2958	OSH
US-ICs	Imnavait Creek Wet Sedge Tundra	68.6058	-149.311	WET
US-ICt	Imnavait Creek Tussock Tundra	68.6063	-149.3041	OSH
US-Ivo	Ivotuk	68.4865	-155.7503	WET
US-KFS	Kansas Field Station	39.0561	-95.1907	GRA
US-Kon	Konza Prairie	39.0745	-96.5951	GRA
US-KS2	Kennedy Space Center Scrub Oak	28.6086	-80.6715	CSH
US-KUT	KUOM Turfgrass	44.994989	-93.18628	GRA
US-Los	Lost Creek	46.0827	-89.9792	WET
US-Me2	Metolius Intermediate Pine	44.4523	-121.5574	ENF
US-Me3	Metolius Second Young Pine	44.3154	-121.6078	ENF

US-Me4	Metolius Old Aged Ponderosa Pine	44.4992	-121.6224	ENF
US-Me5	Metolius First Young Aged Pine	44.43719	-121.56676	ENF
US-Me6	Metolius Young Pine Burn	44.3233	-121.6078	ENF
US-MOz	Missouri Ozark	38.7441	-92.2	DBF
US-MRf	Mary's River Fir	44.64649	-123.55148	ENF
US-Myb	Mayberry Wetland	38.0498	-121.7651	WET
US-NC1	NC Clearcut	35.8118	-76.7119	OSH
US-NC2	NC Loblolly Plantation	35.803	-76.6685	ENF
US-Ne1	Mead Irrigated Maize	41.1651	-96.4766	CRO
US-Ne2	Mead Irrigated Rotation	41.1649	-96.4701	CRO
US-Ne3	Mead Rainfed	41.1797	-96.4397	CRO
US-NR1	Niwot Ridge	40.0329	-105.5464	ENF
US-Oho	Ohio Oak Openings	41.5545	-83.8438	DBF
US-PFa	Park Falls/ WLEF	45.9459	-90.2723	MF
US-Pon	Ponca City	36.76667	-97.13333	CRO
US-Prr	Poker Flat Black Spruce	65.12367	-147.48756	ENF
US-Ro1	Rosemount G21	44.7143	-93.0898	CRO
US-Ro2	Rosemount G19	44.7217	-93.0893	CRO
US-SdH	Nebraska Sand Hills Dry Valley	42.0693	-101.4072	GRA

US-SFP	Sioux Falls Portable	43.2408	-96.902	CRO
US-Shd	Shidler Oklahoma	36.93333	-96.68333	GRA
US-Slt	Silas Little	39.9138	-74.596	DBF
US-SKr	Shark River Slough	25.362933	-81.077582	EBF
US-Snd	Sherman Island	38.0373	-121.7537	GRA
US-SO2	Sky Oaks Old Stand	33.3738	-116.6228	CSH
US-SO4	Sky Oaks New Stand	33.3845	-116.6406	CSH
US-SRC	Santa Rita Creosote	31.9083	-110.8395	OSH
US-SRG	Santa Rita Grassland	31.789379	-110.827675	GRA
US-SRM	Santa Rita Mesquite	31.8214	-110.8661	WSA
US-Syv	Sylvania Wilderness Area	46.242	-89.3477	MF
US-Ton	Tonzi Ranch	38.4316	-120.966	WSA
US-Tw1	Twitchell Wetland	38.1074	-121.6469	WET
US-Tw3	Twitchell Alfalfa	38.1159	-121.6467	CRO
US-Tw4	Twitchell East End Wetland	38.10298	-121.6414	WET
US-Twt	Twitchell Disturbance	38.10867	-121.653	CRO
US-UMB	University of Michigan Biological Station	45.5598	-84.7138	DBF
US-Umd	UMBS Disturbance	45.5625	-84.6975	DBF
US-Var	Vaira Ranch	38.4133	-120.9507	GRA
US-WBW	Walker Branch	35.9588	-84.2874	DBF
US-WCr	Willow Creek	45.8059	-90.0799	DBF

US-Wdn	Walden	40.7838	-106.2618	OSH
US-WHS	Walnut Gulch Lucky Hills Shrub	31.7438	-110.0522	OSH
US-Wi3	Wi3 Mature Hardwood	46.634722	-91.098667	DBF
US-Wi4	Wi4 Mature Red Pine	46.739333	-91.16625	ENF
US-Wkg	Walnut Gulch Kendall Grassland	31.7365	-109.9419	GRA
US-Wlr	Walnut River	37.5208	-96.855	GRA
US-WPT	Winous Point North Marsh	41.464639	-82.996157	WET
US-Wrc	Wind River Crane Site	45.8205	-121.9519	ENF
ZA-Kru	Skukuza	-25.0197	31.4969	SAV
Zm-Mon	Mongu	-15.4378	23.2528	DBF

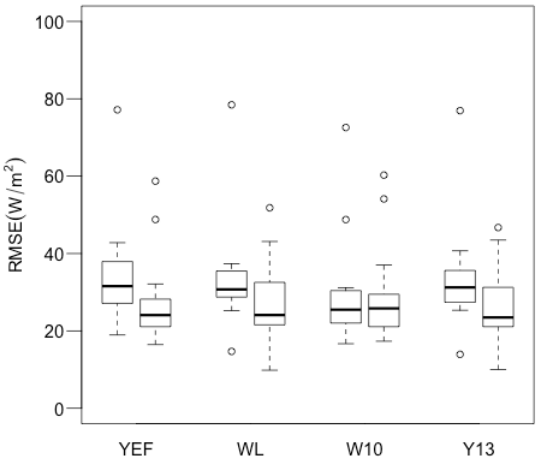
Table 2-8: Original published coefficients for regression algorithms.

Algorithm	Short name	Version	Original coefficients
Yebra ET	YET	NDVI	a = 37.39, b = 242.3
		EVI	a = 5.73, b = 347.77
Yebra EF	YEF	NDVI	a = 0.2, b = 1.03
		EVI	a = 0.087, b = 1.40
Helman exponential	HEx	NDVI	a = 6.735, b = 3.12
		EVI	a = 5.150, b = 6.31
Helman scaled	HSc		a = 2.0, b = 0.1, c = 2.5, d = 0.05, e = 30.0
Wang 2007	W07	EVI, T_{a_avg}	a_1 = 0.137, a_2 = 0.759, a_3 = 0.004
		EVI, T_{a_max}	a_1 = 0.114, a_2 = 0.778, a_3 = 0.0039
		EVI, T_{s_avg}	a_1 = 0.114, a_2 = 0.778, a_3 = 0.0039
		EVI, T_{s_max}	a_1 = 0.096, a_2 = 0.78, a_3 = 0.0039
		NDVI, T_{a_avg}	a_1 = 0.1505, a_2 = 0.45, a_3 = 0.004
		NDVI, T_{a_max}	a_1 = 0.106, a_2 = 0.49, a_3 = 0.0039
		NDVI, T_{s_avg}	a_1 = 0.106, a_2 = 0.49, a_3 = 0.0039
		NDVI, T_{s_max}	a_1 = 0.084, a_2 = 0.498, a_3 = 0.0039
Wang/Liang	WL	EVI, T_{a_avg}	a_1 = 0.3541, a_2 = 0.6257, a_3 = 0.0073 a_4 = -0.0134
		EVI, T_{a_max}	a_1 = 0.3315, a_2 = 0.6437, a_3 = 0.0073, a_4 = -0.0143
		EVI, T_{s_avg}	a_1 = 0.3637, a_2 = 0.6634, a_3 = 0.0062, a_4 = -0.0144
		EVI, T_{s_max}	a_1 = 0.3383, a_2 = 0.6698, a_3 = 0.0067, a_4 = -0.0159

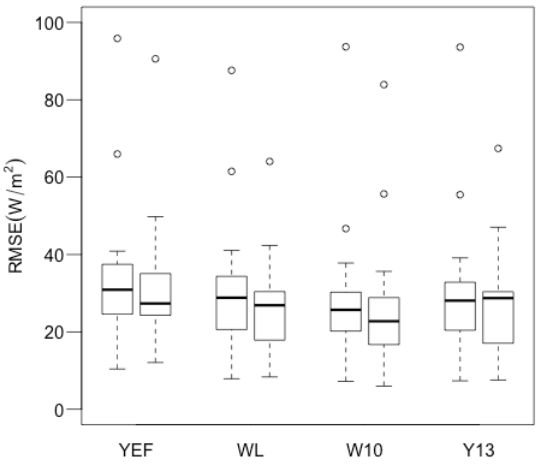
		NDVI, T_{a_avg}	a_1 = 0.3067, a_2 = 0.4425, a_3 = 0.0086, a_4 = -0.0141
		NDVI, T_{a_max}	a_1 = 0.2749, a_2 = 0.4668, a_3 = 0.0085, a_4 = -0.0150
		NDVI, T_{s_avg}	a_1 = 0.2925, a_2 = 0.4919, a_3 = 0.0075, a_4 = -0.0153
		NDVI, T_{s_max}	a_1 = 0.2816, a_2 = 0.4834, a_3 = 0.0079, a_4 = -0.0170
Choudhury/ FAO56	Ch		EVI_{min} = 0.05, EVI_{max} = 0.95
Kamble/ FAO56	Kmb		a = 1.4571, b = 0.1725
Wang 2010	W10	NDVI	a_1 = 0.476, a_2 = 0.284, a_3 = -0.654, a_4 = 0.264, a_5 = 3.06, a_6 = -3.86, a_7 = 3.64, a_8 = 0.819, a_9 = 0.0017
		EVI	a_1 = 0.504, a_2 = 0.364, a_3 = -0.760, a_4 = 0.855, a_5 = 2.99, a_6 = -3.25, a_7 = 7.73, a_8 = 1.00, a_9 = 0.0006
Yao 2011	Y11		a_1 = 0.00084, a_2 = -0.000978, a_3 = 0.3044, a_4 = 0.0029, a_5 = 0.284, a_6 = 0.1273, a_7 = 0.01, a_8 = 0.065
Yao 2013	Y13	T_{s_dTr}	α = 1.26, $NDVI_{max}$ = 0.95, $NDVI_{min}$ = 0.05, T_{opt} = 25.0, dTr_{max} = 60.0

		T_{a_dTr}	$\alpha = 1.26$, $NDVI_{max} = 0.95$, $NDVI_{min} = 0.05$, $T_{opt} = 25.0$, $dTr_{max} = 40.0$
Yao 2015	Y15		$\alpha = 1.26$, $a_1 = 0.1691$, $a_2 = 0.0073$, $a_3 = 0.4464$, $a_4 = 0.2122$, $a_5 = 0.4079$

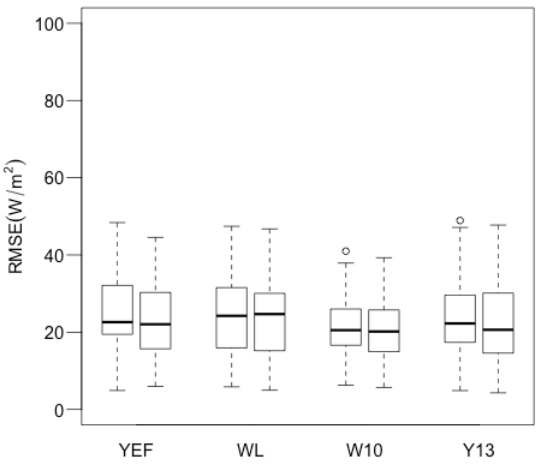
a)


RMSE(W/m²)
100
80
60
40
20
0
YEF
WL
W10
Y13

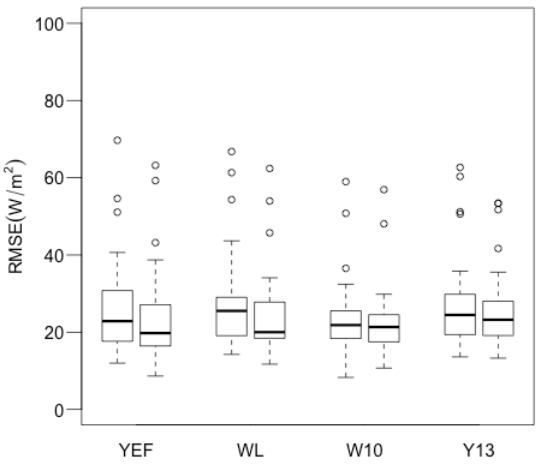
b)


RMSE(W/m²)
100
80
60
40
20
0
YEF
WL
W10
Y13

c)


RMSE(W/m²)
100
80
60
40
20
0
YEF
WL
W10
Y13

d)


RMSE(W/m²)
100
80
60
40
20
0
YEF
WL
W10
Y13

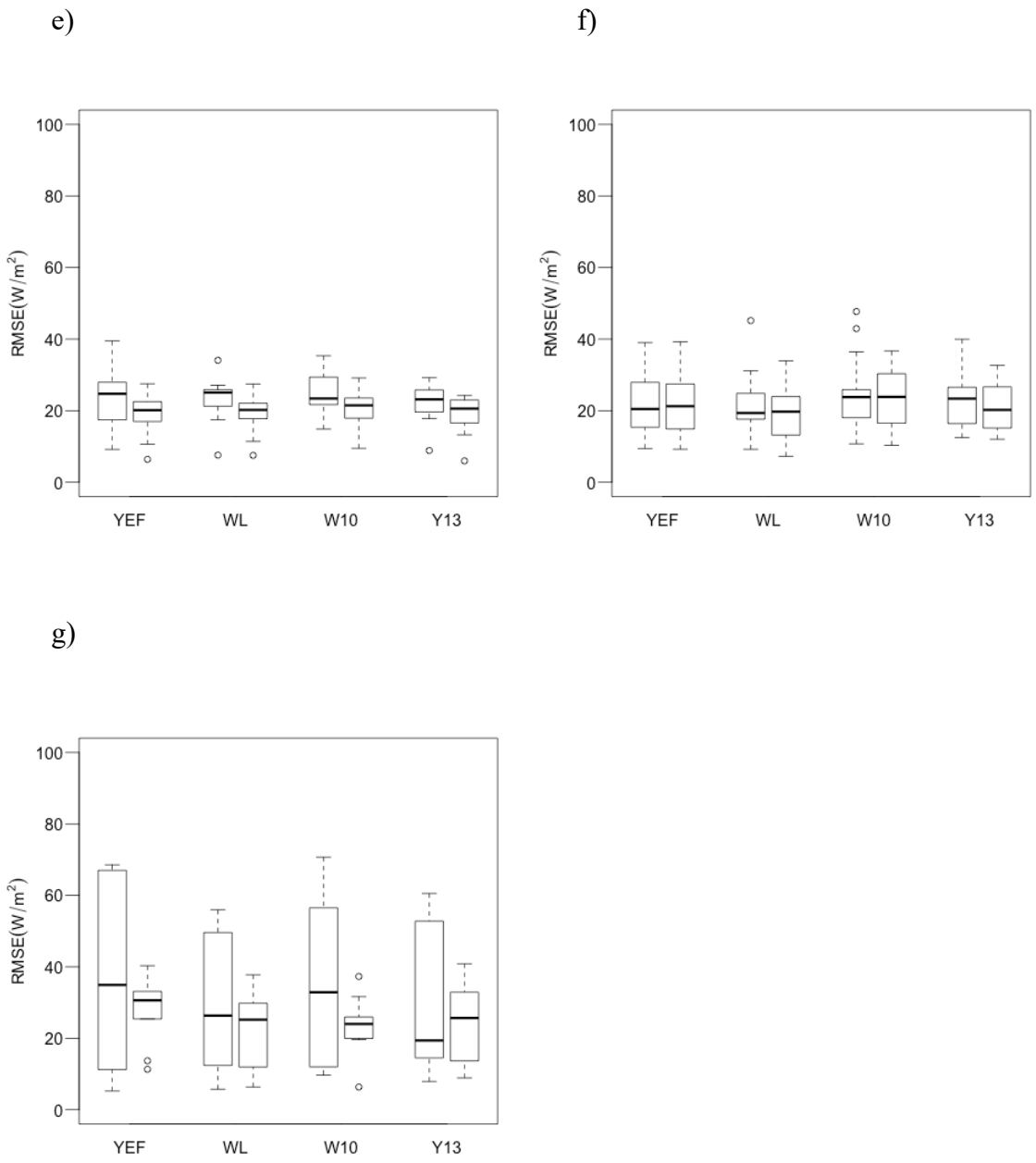
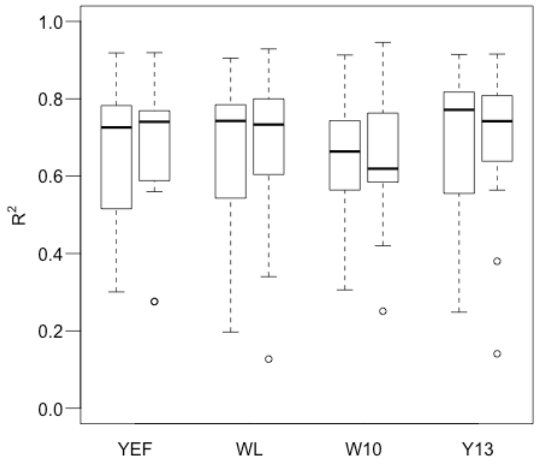


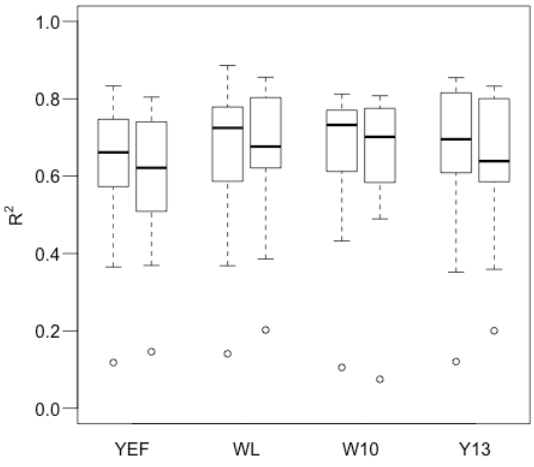
Figure 2-10: Boxplots of RMSE by site for each cover type, from coefficients derived from global data and from data from the same cover type only. For each algorithm, first box is for coefficients derived from all sites and second box is for coefficients derived from data from the

same cover type only. a) Agricultural. b) Deciduous. c) Evergreen. d) Grassland. e) Savannah. f) Shrub. g) Wetland.

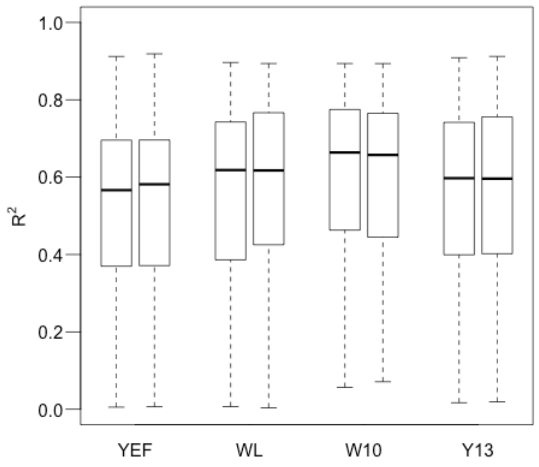
a)



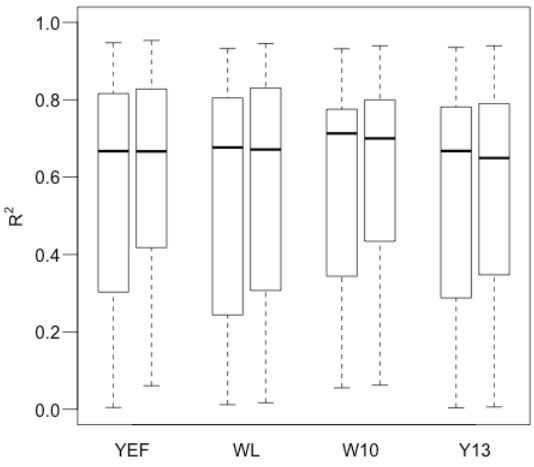
b)



c)



d)



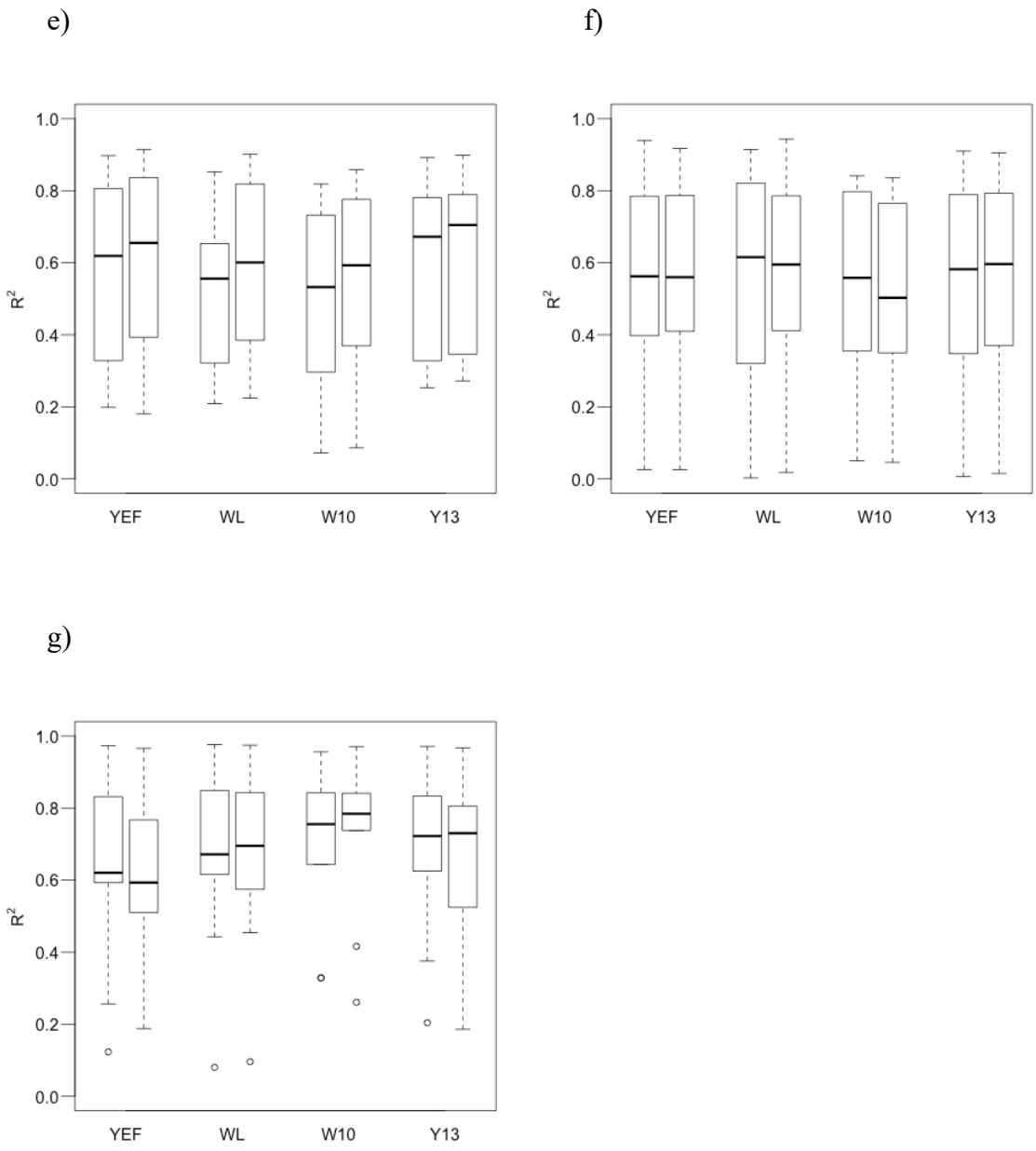
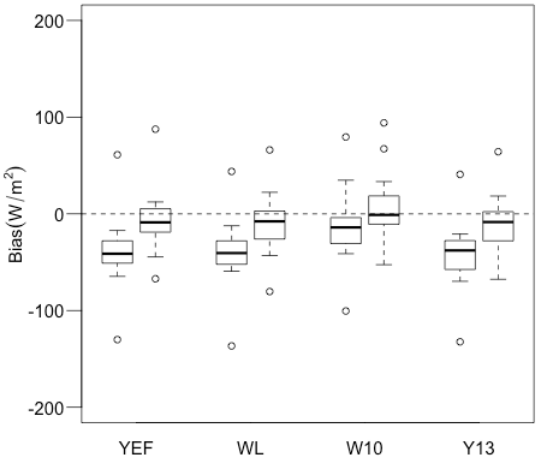


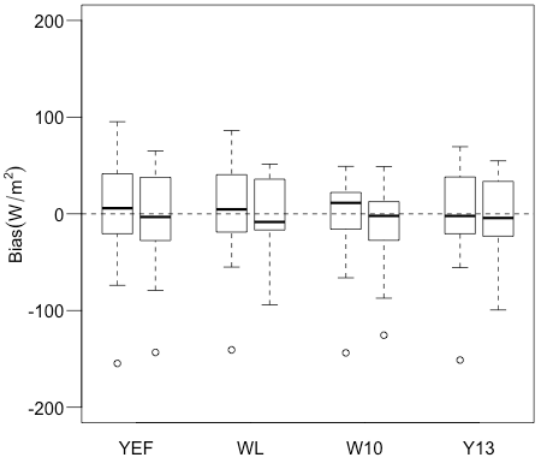
Figure 2-11: Boxplots of R^2 by site for each cover type, from coefficients derived from global data and from data from the same cover type only. For each algorithm, first box is for coefficients derived from all sites and second box is for coefficients derived from data from the

same cover type only. a) Agricultural. b) Deciduous. c) Evergreen. d) Grassland. e) Savannah. f) Shrub. g) Wetland.

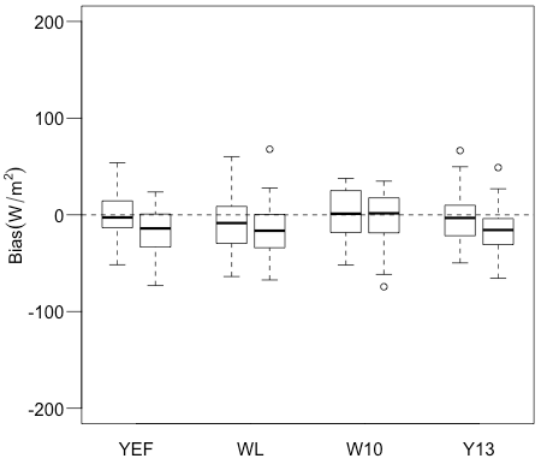
a)



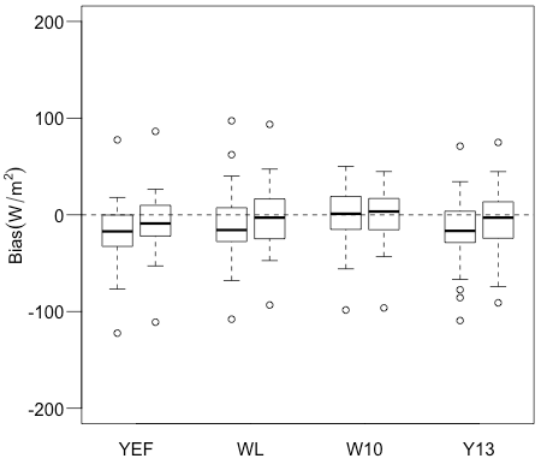
b)



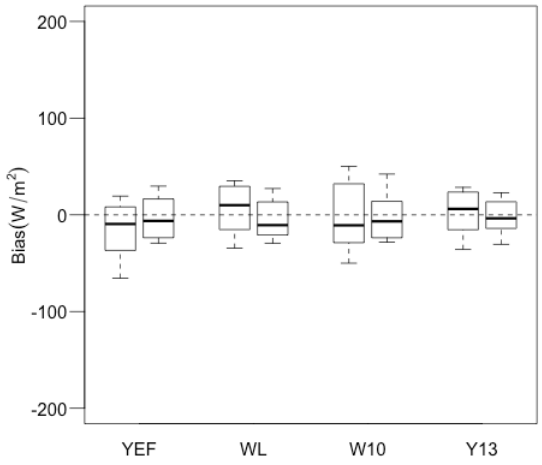
c)



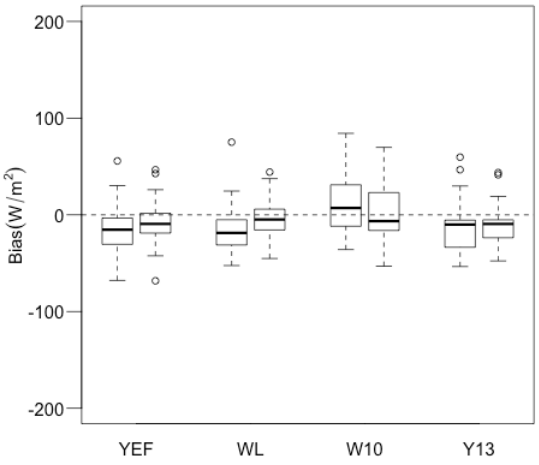
d)



e)



f)



g)

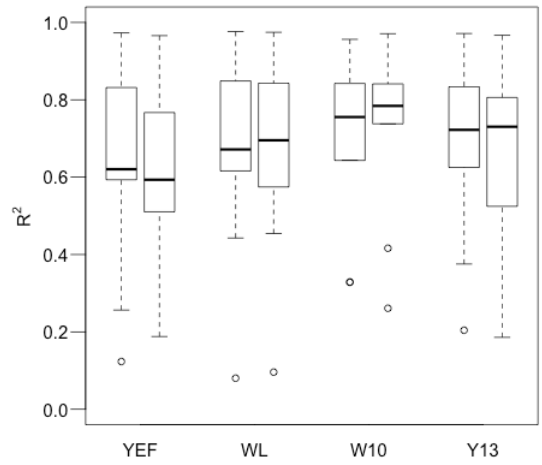


Figure 2-12: Boxplots of bias by site for each cover type, from coefficients derived from global data and from data from the same cover type only. For each algorithm, first box is for coefficients derived from all sites and second box is for coefficients derived from data from the same cover type only. a) Agricultural. b) Deciduous. c) Evergreen. d) Grassland. e) Savannah. f) Shrub. g) Wetland.

Chapter 3: Development of machine learning methods for estimating terrestrial evapotranspiration from remote sensing

Introduction

Compared to the radiative elements of the surface energy balance, LE can be difficult to measure, and there is more uncertainty in the measurements. Measurements of radiation balances and vegetation indices are made with global coverage on a daily or more frequent basis, and have been combined with other remote sensing data, reanalyses, and ground-based observations in a great variety of ways to generate LE data sets at various spatial and temporal resolutions. Methods for remote sensing of LE are being developed on an ongoing basis. There is a need to test and evaluate new means of obtaining LE as they become available. Machine learning is a relatively novel software technology that has been applied in many areas of remote sensing, including retrieval of LE. The goals of the present study are to evaluate the ability of machine learning methods to obtain LE from remote sensing data alone and to compare the accuracy and computational demand for different machine learning algorithms when applied to the LE retrieval problem.

Machine learning (ML) methods are means of extracting patterns from data sets with little prior knowledge of those patterns and can be used to address classification and regression problems. When used for regression, they are analogous to standard statistical methods, but more sophisticated in their ability to model complex relationships between input data sets with little a priori knowledge of the form of those relationships. The best-known ML methods include neural networks (NN), tree methods, and support vector machines (SVMs), which have all been used

extensively for a wide range of applications both inside and outside of the field of remote sensing (Camps-Valls et al. 2011; Hastie et al. 2009). Another method, the random kernel (RKS) algorithm (Rahimi and Recht, 2009), is less widely known, but has been shown to have great potential utility for remote sensing applications (Pérez-Suay et al. 2017).

Machine learning methods have been used a great deal recently for the determination LE for a variety of applications. The model tree ensemble machine learning technique has been applied to the problem of determining global trends in LE by Jung et al. (2010). Multiple studies have been conducted using machine learning techniques for downscaling LE (Kaheil et al., 2008; Ke et al., 2017, 2016) and drought detection and forecasting (Park et al., 2016; Rhee and Im, 2017). There are also a significant number of studies comparing the performance of different ML techniques for obtaining LE. Some of them (Dou and Yang 2018b, 2018a; Deo et al. 2016) found little difference between the results obtained using different ML algorithms. In other studies, some ML algorithms outperformed others, but no single ML method produced the best results consistently. In Deo and Samui (2017), a least square support vector machine produced more accurate results than three other techniques and Yang et al. (2006) obtained better results with a support vector machine than with a neural network or multiple regression. Pandey et al. (2017) had the best results obtaining reference evapotranspiration (ET_0) using a neural network compared to three other methods.

Virtually all of the previous work done with machine learning applications for LE involves training with input variable data sets that include ground-based station data only, or ground station data in addition to remote sensing. (An exception here is Lu and Zhuang (2010), who

obtained all of their input variable data from MODIS.) Most of these studies, with the notable exception of Jung et al. (2010), also involve training to measurements of LE from a relatively small number of locations (20 or fewer). In many studies just one ML technique was applied. In those where comparisons are performed four or fewer methods are compared. Here we endeavor to extend the range of algorithms and conditions used in comparative testing of ML methods for retrieval of LE, and also to demonstrate the utility of these methods for retrieval of LE from remote sensing data only. We focus on comparisons of different ML algorithms both in terms of accuracy and computational demand.

The process of training ML algorithms requires three separate training, validation, and test data sets, and also requires the tuning of one or more algorithm parameters. All algorithms tested here have at least one parameter that must be tuned. It is impossible to do this in an a priori manner, so the algorithms must be iteratively trained and validated to fit the parameters appropriately. Once the parameter fitting has been done, a final test must be done with a third data set that has been held separate from the training and validation data sets, in order to ensure that overfitting has not occurred in the process of tuning the parameters. The computational demands of performing this procedure vary, depending on the algorithm to be trained, the number of parameters within the algorithm that are adjusted, and the size of the training data set. For practical reasons, it is desirable to have some estimates of the computing resources required for training each ML algorithm as well as its performance.

This study is intended to evaluate the utility of a range of ML algorithms for obtaining LE from remote sensing data on a global basis. The algorithms' relative demands for computational

resources were evaluated as well as their accuracy. The initial step performed was to measure the length of time it took to perform a single train/ test iteration for each algorithm. This was done with a smaller and a larger training data set in order to determine which combinations it was feasible to work with further. Then the effects of including different combinations of variables in the input data set on accuracy and computational demand were analyzed, as it was considered possible that increasing the number of input variables could improve accuracy but also significantly increase training processing time. Finally, most of the ML methods considered were tuned with the small training data set, and a smaller number with the large training data set. The sensitivity of the algorithm results to changes in parameter values was evaluated in each case.

Methods

Description of machine learning algorithms

A total of 14 ML algorithms were subjected to initial timing tests with the smaller training data set. The algorithms considered fall into five "families" of related algorithms, each of which is discussed in a section below. Based on the results of this timing, 10 of the original algorithms were systematically tuned with the smaller training data set and given an initial timing test with the larger training data set. Of those 10, 8 were found to run with low enough computational demand for systematic tuning with the larger training data set to be feasible. The 14 ML algorithms considered are listed below in Table 3-1. Which of the algorithms were trained with the small and large data sets are also specified in Table 3-1.

Table 3-1: Algorithms used in this study, with corresponding abbreviations. Algorithms marked with an asterisk (*) are described in Hastie et al. (2008). Other references are provided with the algorithm names.

Family	Full name	Abbreviation	Tuned with small training data set	Tuned with large training data set
Linear	Regularized linear regression*	RLR	Yes	Yes
	Least absolute shrinkage and selection operator regression*	LASSO	Yes	Yes
	Elastic net regularization*	ELASTIC	Yes	Yes
Kernel	Gaussian process regression*(Murphy, 2012)	GPR	No	No
	Kernel ridge regression (Murphy, 2012)	KRR	Yes	No
	Random kernel (Rahimi and Recht 2009, Perez-Suay et al. 2017)	RKS	Yes	Yes
	Variational heteroscedastic Gaussian process regression(Lazaro-Gredilla et al., 2014; Lazaro-Gredilla and Titsias, 2011)	VHGPR	No	No
Tree	Regression tree*	TREE	Yes	Yes
	Bootstrap aggregation (bagging) tree*	BAGTREE	Yes	Yes
	Boosted regression tree*	BOOST	Yes	Yes
Neural network	Standard neural network (1, 2, and 3 hidden layers)*	NN	Yes	Yes
	Extreme learning machine (Huang et al. 2006)	ELM	No	No
Support vector	Support vector regression*	SVR	Yes	No
	Relevance vector machine(Thayananthan et al., 2006)	RVM	No	No

A general property of machine learning algorithms is that they are capable of giving results that are too independent of the training data set to yield good predictions with new data (underfitting, or high bias), or too closely fit to the training data set to give good predictions (overfitting, or high variance). The values of tunable parameters may result in overfitting or underfitting, which is reflected in the results when tested with a separate validation data set.

For this study, optimum values of the parameters are found by minimizing the root mean square error (RMSE) of the algorithm when applied to the validation data set. It is possible for the resulting algorithm parameters to represent overfitting to the combination of the training and validation data set. Whether or not this has occurred is checked by applying the trained and tuned algorithm to an independent test data set. The proportions of the originally available data which are allotted to training, validation, and test data sets are usually set so that most of the data, typically around 80%, is in the training data set, and the remainder divided between the validation and test data sets. For the purposes of this study, we used one training/ validation/ test split of approximately 80%/ 10%/ 10% proportions (69,752 training, 7910 validation, and 7910 test), and another of 10%/ 45%/ 45% proportions (7910 training, 35594 validation, and 35594 test). For some of the algorithms, a single iteration of training and validation could not be performed quickly enough to perform the iterative process of tuning in a reasonable length of time (10 minutes for a single train/ test cycle), even with the smaller training data set. Other algorithms could be tuned in a

reasonable amount of time with the smaller training data set, but not with the larger one.

Each of the algorithms considered (as listed in Table 3-1) will be described briefly below. The implementation in Matlab of all of these algorithms, with the exception of the random kernel (RKS), was obtained from package "simpleR" (Lazaro-Gredilla et al., 2014). The RKS algorithm code was obtained from http://isp.uv.es/code/rks2017.html and is supplemental material to Pérez-Suay et al. (2017) .

Linear regression variants

A standard linear regression to a scalar output fits a set of m coefficients $\Theta = (\theta_1 \ldots \theta_m)$ to a set of n input vectors ($X_n = x_{n,1} \ldots x_{n,m}$) and n output scalars ($y_1 \ldots y_n$). The coefficients Θ are found that minimize the mean squared error cost function:

$$J = \sum_{i=1}^{n} (y_i - \Theta X_i)^2 \quad (3\text{-}1)$$

For a simple linear regression, there is a single optimum set of coefficients and no additional parameters that need to be tuned.

There are three variants of linear regression that apply regularization parameters in an effort to prevent overfitting to the training data. In each case, a penalty is applied that increases as the magnitude of the regression coefficients increases. The first of these linear regression variants, the regularized linear regression, produces a set of coefficients Θ that minimize the following cost function:

$$J = \sum_{i=1}^{n}(y_i - \Theta X_i)^2 + \lambda \sum_{j=1}^{m} \theta_j^2 \quad (3\text{-}2)$$

This has the potential for underfitting or overfitting, depending on the value of λ chosen, with higher λ tending to produce greater bias and lower λ tending to produce greater variance.

The second of the linear regression variants, the Least Absolute Shrinkage and Selection Operator (LASSO) method, formulates the cost function with a penalty applied to the absolute value of the θ values rather than their square, resulting in

$$J = \sum_{i=1}^{n}(y_i - \Theta X_i)^2 + \lambda \sum_{j=1}^{m} |\theta_j| \quad (3\text{-}3)$$

Once again, there is a single tunable parameter λ.

The third linear regression variant, the elastic net algorithm, includes both a squared term and an absolute value term in the cost function:

$$J = \sum_{i=1}^{n}(y_i - \Theta X_i)^2 + \lambda\left(\sum_{j=1}^{m} \theta_j^2 + \alpha \sum_{k=1}^{m} |\theta_k|\right) \quad (3\text{-}4)$$

Here there are two parameters: λ, representing the overall degree of regularization, and α, representing the weighting between squared and absolute value regularization.

Kernel methods

The kernel methods are based on implicit mapping of the input data to a space where a regression problem is more tractable by applying a kernel function to each pair of input data points. Tuning was performed here for two kernel methods, the kernel ridge regression (KRR) and the random kernel algorithm (RKS).

The kernel ridge regression (KRR) performs a regularized linear regression procedure using the "kernel trick" to represent the inner product of each pair of input points X_i and Y_j in the training data set in a reprojected space. The radial basis function (RBF)

$$\langle \phi(X_i), \phi(Y_j) \rangle = exp\left(-\frac{\|X_i - Y_j\|^2}{2\sigma^2}\right) \quad (3\text{-}5)$$

is used as the kernel function, where ϕ represents the reprojection (which is not explicitly calculated). There are two parameters to be tuned, the weight λ assigned to the regularization term of the modified RLR and the width σ of the RBF kernel.

The random kernel (RKS) function is an approximation of KRR, where the kernel function is approximated by a set of randomly chosen functions drawn from Fourier transform of the kernel function before being applied to each combination of data points in the training data set. The RKS transformation is performed in order to obtain an improvement in computational efficiency over KRR while attaining similar accuracy. The RKS algorithm is described in detail by Pérez-Suay et al. (2017) , who also provide examples of its application to remote sensing problems.

Initial timing tests were made for two other kernel methods, the Gaussian process regression (GPR) and variational heteroscedastic Gaussian process regression (VHGPR). However, the computational resources required by these algorithms were such that tuning could not have been completed in a reasonable amount of time, so no further testing was done with them.

Regression tree methods

Regression trees are a frequently-used ML method in which the space of input variables is iteratively subdivided and an output value given for each subset. Three model tree methods are used in this study: the standard model tree, bootstrap aggregation (bagging) tree, and boosted tree algorithms.

The basic regression tree algorithm constructs a single tree by iteratively optimizing decision points within the range of each input variable. There are two tunable parameters in the version of the regression tree tested here. The number of data points required in a subdivision of the input space for it to be split can be varied, resulting in a tree that models variations of small numbers of input data points more or less closely. The other parameter is the degree of pruning, in which some branch points of the decision tree are removed and replaced by their parent, in an effort to prevent overfitting. Testing variations in these parameters of model trees with the LE data set showed very little effect, approximately 0.5 W/m^2 RMSE difference between the best and worst fits over a range of 25-125 points required for splitting and a range of pruning levels from 0 to 20. Therefore, variations in these parameters were not tested in the other two more complex tree algorithms.

The second regression tree method tested here is the bootstrap aggregation, or bagging, tree algorithm. Bootstrap aggregation is the selection of subsets of the original training data set and using each of them to generate a tree, then combining

those trees. Other than the parameters internal to the individual trees, two parameters need to be tuned for the bagging tree: the fraction of the input data set used to generate the bootstrap data sets and the number of trees generated.

The third regression tree method used in this study is the boosting tree. In this method, regression trees are generated iteratively, with error information from each tree used to generate a refined version in the next iteration by adjusting weights given to the input data points. The parameters to be varied are the same as for the bagging tree: the number of trees to generate and the fraction of the input data to be used in their generation.

Regression trees, bootstrap aggregation, and boosting trees are all discussed in detail in Hastie et al. (2009).

Neural networks

The neural network, an algorithm modeled on the interactions of biological neurons, is one of the most commonly used ML techniques. The standard neural network (NN) consists of numerical values connected by nodes (neurons) arranged so that each node computes a weighted combination of each of the outputs from the previous layer, then uses an activation function (typically a sigmoid function) to produce its own output. Every neural network has an "input layer" corresponding to the input variable values and an "output layer" containing the output values, and one or more "hidden layers". The simplest form of a neural network has only a single hidden layer, while multiple

hidden layers can be connected to each other to form more complex networks. During the training procedure, the neuron connection weights are adjusted to yield the most accurate possible fit to the training data. More complexity can be added to a neural network by increasing the number of hidden layers or by increasing the number of nodes in each layer. In general, more complex NNs offer more detailed representations of the input data, but at the expense of increased computational demand and a greater possibility of overfitting. For a more detailed review of neural networks, see Hastie et al. (2009).

In order to produce a version of the neural network with less computational demand, Huang et al. (2006) developed the extreme learning machine (ELM). In the ELM procedure, the values of the weights leading from the input layer to the hidden layer of a one hidden layer NN are set randomly and held fixed during the NN training process. In our trials with the ELM, it actually required more computational power than the standard NN and was dropped from further consideration after initial timing tests.

Support vector machines

Support vector machines (SVMs) are another ML method that can be applied to regression problems. Like the kernel methods, a reprojection of the input data is performed implicitly. In the case of the SVM, a regularized linear regression in the reprojected space is performed subject to the condition that points with an error e less than a given error tolerance ε do not contribute to the cost function. Usually the cost

function is then linear in the absolute value of e - ε. There are three adjustable parameters: ε, a regularization parameter C, and, if the RBF kernel is used, a kernel width σ. The Matlab library used in the SVM computations in this study, LIBSVM, is available from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. Hastie et al. (2009) and Smola and Scholkopf (2004) provide more background about the SVM as applied to regression problems.

A variant of the SVM, the relevance vector machine (RVM), was removed from further consideration after initial timing trials showed that a single train/test iteration was too computationally demanding.

Procedure for testing machine learning methods

The following procedure was followed in order to test the performance of the ML algorithms with the LE data set. Initially, timing of one iteration of training and testing was made with default or arbitrary coefficient values. The results of this testing do not provide a precise quantification of the computing requirements for each algorithm, but they do indicate which of the algorithms are relatively more or less demanding. This timing is also a means of identifying those algorithms that are not suitable for further testing because the computation time required for enough iterations to tune the parameters would render the tuning impractical.

The timing was initially conducted using the smaller training data set. Algorithms that took more than ten minutes for one iteration of training and validation were removed

from further consideration. The remaining algorithms were then timed for one training/ validation iteration with the larger training data set. Two of the algorithms that were tractable with the smaller training data set became too computationally demanding with the larger training data set. For those algorithms (KRR and SVR), further analysis was carried out with the smaller training data set but not with the larger one. The results of the timing procedure are discussed in more detail in the first Results section.

The algorithms that took the least time to process the larger data set were the linear regression, boost tree, and RKS. Those algorithms were used to test the effects on accuracy and computation time of using different combinations of input variables. Since the best results were found using all of the input variables, but the computation time was found not to vary much with the number of input variables, all the input variables were included in the remaining training, validation, and testing. This analysis of input variable combinations is discussed more in the second Results section.

Once the most viable combinations of ML algorithms and input variables had been identified, a tuning procedure was performed for each algorithm. All of the parameters for each ML algorithm were varied independently, producing a results data set with the same dimensions as the number of parameters. For example, a two hidden layer neural network has two parameters: the numbers of neurons in the first and second hidden layers. The range and intervals of the tuning parameter values

tested were varied in several different trials for each algorithm, in order to come as near as possible to the global minimum RMSE value. The ranges, intervals, and results described here are from the final trial for each algorithm.

Once optimal tuning parameters with respect to the validation data set (lowest validation RMSE) were found for each algorithm, the algorithm trained with those parameters was applied to the test data set. This is done in order to ensure that the parameter determination with the training and validation data sets did not result in overfitting to those data sets. Algorithm tuning and final test results are given in the third Results section.

Data

The data used in this study come from three sources. The remote sensing data used are Global Land Surface Satellite (GLASS) radiation data and Moderate-Resolution Imaging Spectroradiometer (MODIS) high-level data products. Ground-based Fluxnet tower site data were used for validation and for comparison of the results using satellite radiation data to those from measuring net radiation at the surface. The data variables, sources, and spatial and time resolutions for each data set used are listed in Table 3-2.

Table 3-2: Input and validation data used in this study

Abbreviation	Variable	Source	Frequency	Spatial resolution
LE	Surface latent heat	Fluxnet	Half-hourly, averaged to daily	Flux tower footprint
R_n	Net radiation at surface	Fluxnet	Half-hourly, averaged to daily	Flux tower footprint
DSR	Downward surface radiation	GLASS	Daily	5 km
PAR	Photosynthetically active radiation	GLASS	Daily	5 km
NDVI	Normalized difference vegetation index	MODIS	16-day, interpolated to daily	250 meters
EVI	Enhanced vegetation index	MODIS	16-day, interpolated to daily	250 meters
LAI	Leaf area index	MODIS	8-day, interpolated to daily	500 meters
FPAR	Fraction of photosynthetically adjusted radiation	MODIS	8-day, interpolated to daily	500 meters

Albedo	Albedo	MODIS	8-day, interpolated to daily	500 meters
NBAR	Nadir BRDF-adjusted reflectance	MODIS	8-day, interpolated to daily	500 meters

The GLASS data set (Liang et al., 2014, 2013) consists of radiative and biophysical parameters generated using data from multiple satellite sensors. The products used here are the downward shortwave radiation (DSR) and photosynthetically active radiation (PAR). The algorithms and data used to generate the GLASS DSR and PAR data sets are described in Zhang et al. (2014). GLASS DSR and PAR are generated by combining Moderate-Resolution Imaging Spectroradiometer (MODIS) polar-orbiting sensor with data from four geostationary satellites using a look-up table (LUT) method. The LUT was generated after sensitivity analyses using the MODTRAN radiative transfer model determined that surface elevation, atmospheric water vapor, and surface BRDF had the most significant impact on the accuracy of the radiation products. The GLASS data are available at daily time resolution and 5 km spatial resolution. The nearest neighbor values were used for each station location. GLASS leaf area index (LAI) data were also used for trial runs of three of the ML algorithms to evaluate the feasibility of using other GLASS data sets in place of MODIS in the future.

Several parameters obtained from MODIS were also used in this analysis: Normalized-difference vegetation index (NDVI) and enhanced vegetation index (EVI), LAI, fraction of photosynthetically active radiation absorbed (FPAR), surface albedo, and nadir BRDF-adjusted reflectance (NBAR) in seven reflective bands. Subsets of all of these MODIS Collection 5 products used were generated centered on the coordinates of each flux tower site using the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) MODIS subset tool and the NASA/ USGS Land Processes Distributed Active Archive Center (LP DAAC) Application for Extracting and Exploring Analysis Ready Samples (AρρEEARS) All MODIS products were linearly interpolated to daily frequency for use in this study.

The MODIS MOD13 NDVI and EVI products (Didan 2015) are available at 250m resolution at 16-day intervals. They are derived from surface reflectance values from the formulas

$$NDVI = \frac{\rho_{NIR}-\rho_{red}}{\rho_{NIR}+\rho_{red}} \tag{3-7}$$

and

$$EVI = G_{EVI}\frac{\rho_{NIR}-\rho_{red}}{\rho_{NIR}+C_1\cdot\rho_{red}+C_2\cdot\rho_{blue}+L} \tag{3-8}$$

For the MODIS products, ρ_{red}, ρ_{NIR}, and ρ_{blue} represent band 1, 2, and 3 reflectances respectively, and the EVI formula constants G_{EVI}, C_1, C_2, and L are set to values of 1.0, 6.0, 7.5, and 2.5 respectively.

MODIS LAI and FPAR are available from the MOD15 data sets (Myneni and Knyazikhin 2015) as 8-day composite values at 500m resolution. They are determined via a LUT based on Sun/ sensor viewing geometry, surface vegetation characteristics, and ρ_{red} and ρ_{NIR} for a primary algorithm and NDVI for a backup algorithm.

The MODIS broadband albedo and NBAR at 500m are elements of the MCD43A3 and MCD43A4 data products (Schaaf and Wang 2015a, 2015b) They are produced from 16-day composite data sets by generating models of the surface BRDF from the available observations during the composite period. The 0.3-5.0μm broadband albedo and bands 1-7 NBAR are used in this study. The NBAR is an estimate of the surface reflectance in each band at a nadir view angle and local solar noon sun angle, producing some correction for anisotropy in surface radiation reflection, but retaining more information about spectral variation than the broadband albedo.

Flux tower data were used for validation of the LE values calculated from the ML algorithms against ground measurements, and also for testing the effects of using remote sensing vs. ground-based radiation data as input. A total of 184 flux tower sites were used, 119 from the Ameriflux network (http://ameriflux.lbl.gov) and 65 from the Fluxnet2015 data set (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/). The half-hourly LE and net radiation (R_n) variables from these data were pre-processed by removing all data days for which there were not at least 40 of 48 possible observations present, then averaging the remaining observations. A map of

the site locations and information about their distribution across ecosystem types is given in Carter and Liang (2018) and shown in Figure 2-1.

In all, a total of 79098 site-days of data were used. They were randomly partitioned twice. First, a small training data set of 7910 site-days was drawn from the full sample, with the remaining data divided into a validation data set of 35594 site-days and a test data set of 35594 site-days. The second partition generated a larger training data set of 63278 site-days and validation and test data sets of 7910 site-days each.

Results

Initial time trials of ML algorithms

The time in seconds for each algorithm to run a single iteration of training and validation with all input variables from the small and large training data sets is shown in Table 3-3. If an algorithm took longer than 10 minutes to run a single iteration, it is labeled "prohibitive" and no further testing was done for that combination of algorithm and training data set. Any algorithm that took longer than 10 minutes to run a training/ validation iteration with the small training data set was not timed with the large training data set.

The fastest run time for both training sets was found for the RLR. The RKS, BOOST, and smaller NN algorithms also took less than 10 seconds for one iteration with the smaller training data set. The remaining linear and tree algorithms all completed

within 25 seconds. The two algorithms that did complete within 10 minutes but took longest to do so with the smaller training data set, KRR and SVR, took over 10 minutes to complete with the larger training data set. These algorithms were only tuned with the smaller training data set in later stages. The RKS timing increased at a faster than linear rate with increasing numbers of random functions. Not surprisingly, the NN took longer to complete both with increasing numbers of neurons in each hidden layer and for the larger training data set, although the additional computing time required when adding a second or third hidden layer was relatively modest.

Most algorithms took approximately 5-10 times as long to run with the large training data set than with the small training data set, representing a roughly linear increase in time required with the number of training data points. The primary exceptions to this pattern were the NN trials, which took about 20-25 times as long with the large training data set. The TREE trial took over 17 times as long with the larger training set, while the BOOST method took less than three times as long.

Table 3-3: Time in seconds for one iteration of training and validation for each algorithm for small and large training data sets. If time for one iteration exceeded 10 minutes, it was labeled “prohibitive” and no further testing was done for that combination of algorithm and training data set. No timing or further testing was done with the large training data set for those algorithms where the computational demands were prohibitive with the small training data set. For RKS, “D” represents the number

of random functions used in the trial. The number of hidden layers (HL) and neurons in each of the NN trials is also indicated.

Algorithm	Small training data set (7910 data points)	Large training data set (69,752 data points)
RLR	0.0014	0.0087
LASSO	19.3208	118.1654
ELASTIC	21.8523	109.0231
GPR	prohibitive	
KRR	219.6609	prohibitive
RKS, D = 100	0.0945	1.1625
D = 400	0.3950	4.7896
D = 1000	1.1794	13.026
D = 4000	9.1198	91.2732
TREE	20.2009	351.7759
BAGTREE	15.9202	114.309
BOOST (200 trees)	3.1619	9.1393
NN, 1 HL, 5 neurons	4.0422	102.3333
1 HL, 30 neurons	6.2271	207.9746
2 HL, 5 x 5	5.5793	108.4992
2 HL, 10 x 10	6.0898	131.7211
2 HL, 30 x 30	10.0308	436.8679

3 HL, 5 x 5 x 2	4.6199	128.9298
3 HL, 10 x 10 x 10	7.8482	153.7699
3 HL, 50 x 5 x 2	12.849	prohibitive
3 HL, 150 x 30 x 10	249.671	prohibitive
ELM	prohibitive	
SVR	41.6029	prohibitive
RVM	prohibitive	
VHGPR	prohibitive	

Combinations of input variables

In order to test the effects on speed and accuracy of using different combinations of input variables, trials were done with the linear regression, boost tree, and RKS methods using the small training data set. The linear regression required no tuning. The timing of the boost tree method was conducted for 100 and 1000 trees. The RKS method was used with 100 random functions, but the λ and σ parameters were tuned, and timing was conducted for the optimum case.

Several sets of cases were run, and the results are summarized in Table 3-4. Some overall patterns are notable. The linear regression ran the most quickly but was the least accurate. The RKS ran more quickly than the boost tree method and was more accurate. Generally, including more input variables produced results of similar or improved accuracy at little additional computational cost, although many of the other

variables appear to have redundancy with NBAR. Using radiation information from surface measurements produced results of similar accuracy to using the GLASS radiation variables. This shows that using the radiation variables from the satellite data is a viable alternative to using ground-based radiation data.

Table 3-4: Accuracy and timing tests for different combinations of input variables, using the linear regression, boost tree, and RKS algorithms. Boost tree RMSE was found after optimizing the number of trees. All RKS parameters were optimized except for the number of functions, which was set at 100. All timing test results are for a single iteration of training and testing. All trials were made with the smaller training data set. RMSE is given in W/m^2, and timing is given in seconds.

Variables	Linear regression RMSE	Linear regression timing	BOOST RMSE	BOOST timing (100 trees)	BOOST timing (1000 trees)	RKS RMSE	RKS timing
R_n + NDVI	34.68	6.74×10^{-4}	32.52	2.71	28.25	31.71	0.17
R_n + NBAR	31.80	0.0032	29.18	3.15	31.11	28.10	0.20
R_n + NDVI + EVI + LAI + FPAR + NBAR + Albedo	31.78	0.0098	28.03	4.33	38.08	28.10	0.20
DSR + NDVI	33.83	8.69×10^{-4}	33.07	2.70	28.51	31.73	0.17
PAR + NDVI	33.67	6.72×10^{-4}	32.89	2.77	27.74	31.84	0.18
DSR + PAR + NDVI	33.66	0.0016	32.84	2.69	28.88	31.19	0.18

DSR + PAR + EVI	33.29	0.0013	32.23	2.94	28.96	30.41	0.18
DSR + PAR + NDVI + EVI	33.21	0.0014	31.62	3.04	28.27	29.78	0.18
DSR + PAR + FPAR	33.62	0.0014	32.75	2.99	27.83	31.09	0.17
DSR + PAR + LAI	34.47	0.0016	32.46	2.94	28.31	31.10	0.20
DSR + PAR + LAI + FPAR	33.63	0.0014	31.84	2.80	27.85	30.36	0.20
DSR + PAR + NDVI + EVI + LAI + FPAR	32.90	0.0030	30.53	3.28	31.55	29.14	0.20
DSR + PAR + Albedo	36.96	0.0015	35.85	2.70	26.92	35.49	0.17
DSR + PAR + NDVI + EVI + LAI + FPAR + Albedo	32.89	0.0028	30.26	3.26	31.31	28.95	0.18
DSR + PAR + NBAR	31.27	0.0044	29.44	3.46	33.81	28.44	0.19
DSR + PAR + NDVI + EVI + LAI + FPAR + NBAR	31.21	0.0058	28.53	3.74	35.94	28.32	0.17

The first set of three trials was made with the R_n taken from the ground-based measurements. These trials showed that using NBAR as an input produced better results than using NDVI. When all of the input variables were included, the boost tree algorithm produced better results than with R_n and NBAR alone, but the linear regression and RKS did not.

The second set of trials tested the effects of using DSR or PAR or both in combination with NDVI and EVI. Using both radiation variables with NDVI produced better results than using either of them separately. When using both radiation variables, performance using EVI was better than that using NDVI. Using all four variables produced the lowest RMSEs at little additional computational cost. For all of the subsequent trials, both DSR and PAR were included.

The third set of comparisons tested the use of LAI and/ or FPAR as input variables. Using LAI as an input yielded better results than using FPAR for the boost tree, worse results with the linear regression, and similar results with RKS. Using both LAI and FPAR produced similar results to just using FPAR for the linear regression, but better results than either LAI or FPAR alone for the other two methods. Again, there was little computational cost associated with using more input variables. Finally, combining both vegetation indices with FPAR, LAI, DSR, and PAR produced better results yet, with some additional time cost to the linear regression and boost tree algorithms but little for the RKS.

The fourth set of trials included albedo as one of the input data sets. The first of these trials included albedo as the only input other than DSR and PAR and produced the highest RMSEs for any of the combinations of input variables considered. Including NDVI, EVI, LAI, and FPAR along with albedo improved the results to be similar to, or slightly better than, the trial with those four variables but without albedo.

The final set of trials included NBAR as an input, either alone or with all of the rest of the input variables. These trials including NBAR produced the lowest RMSEs of any combination of variables, with slightly better results for the boost and RKS algorithms when all of the other variables were included along with NBAR.

Based on the overall patterns in the results of these trials, further tuning of all of the algorithms was conducted using all of the remote sensing input data variables: DSR, PAR, NDVI, EVI, LAI, FPAR, albedo, and NBAR.

Tuning of ML algorithms

Each of the algorithms that ran sufficiently quickly to be iterated for purposes of tuning with the small or large training data set was optimized with that data set. In some cases, there is only a single parameter and optimization is straightforward. In other cases, repeated searches had to be made of portions of the tuning parameter space before the optimum parameter values were found. There were also cases where increasing a parameter beyond a certain point made the run time for the algorithm prohibitive, so the testing was cut off at that point even though it might not be the

global minimum. A list of the parameters tuned for each algorithm, the parameter values tested in the final set of iterations, and the parameter values that produced the lowest RMSE in the validation data set using the small training data set are shown in Table 3-5. The equivalent information is shown for those algorithms that could be tuned with the larger data set in Table 3-6.

Table 3-5: Adjustable parameters optimized for each ML method using the small training data set. Values within ranges were distributed logarithmically unless otherwise noted. In many cases individual values are specified.

Algorithm	Parameter	Value range	Optimum value
RLR	λ (regularization weight)	0.0 to 10^{10}	1.0
LASSO	λ (regularization weight)	0.0 to 10^{-3}	7.54×10^{-5}
ELASTIC	λ (regularization weight)	0.0 to 0.1	3.29×10^{-4}
	α (square vs absolute value weighting)	Values {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 1.0}	0.001
KRR	λ (regularization weight)	10^{-3} to 10.0	0.251
	σ (kernel width)	10^{-2} to 100.0	1.0
RKS	λ (regularization weight)	0.0 to 10,000	1000.0

	σ (kernel width)	Values {0.1, 0.25, 0.5, 1.0, 1.2, 1.5, 2.0, 3.0, 4.0, 8.0, 12.0}	3.0
	D (number of random functions)	Values {100, 200, 450, 700, 1000, 1200, 1700, 2000, 2200, 2700, 3500}	2000
TREE	Degree of pruning	0 to 12, intervals of 2	8
	Minimum points for subset split	25 to 200, intervals of 25	150
BAGTREE	Number of trees	Values {1, 3, 6, 12, 25, 50, 100, 150, 200, 400, 1000, 1500, 2000, 2500, 3000, 4000, 5000}	5000
	Fraction of training data in each tree	Values {0.1, 0.3, 0.5, 0.7, 0.9, 1.0}	1.0
BOOST	Number of trees	Values {10, 20, 50, 100, 150, 200,	1000

		300, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000}	
NN, 1 hidden layer	Number of neurons in hidden layer	Values {4, 7, 10, 13, 16, 20, 25, 30, 35, 40, 50, 75, 100, 125, 150, 200, 250, 300}	200
NN, 2 hidden layers	Number of neurons in first hidden layer	Values {2, 5, 10, 25, 50, 75, 100, 125, 150, 200}	150
	Number of neurons in second hidden layer	Values {1, 3, 5, 10, 15, 25, 40, 75}	3
NN, 3 layers	Number of neurons in first hidden layer	Values {1, 3, 5, 10, 15, 25, 40, 65, 100, 150}	100
	Number of neurons in second hidden layer	Values {2, 5, 10, 15, 25, 40}	10
	Number of neurons in third hidden layer	Values {1, 3, 5, 10, 15, 25, 40}	5
SVR	C (regularization factor)	0.32 to 10.0	1.78

	σ (kernel width)	0.32 to 3.16	1.0
	ε (error tolerance)	0.0032 to 10.0	0.18

Table 3-6: Adjustable parameters optimized for each ML method using the large training data set. Values within ranges were distributed logarithmically unless otherwise noted. In many cases individual values are specified.

Algorithm	Parameter	Value range	Optimum value
RLR	λ (regularization weight)	0.0 to 10^{10}	0.0
LASSO	λ (regularization weight)	0.0 to 0.001	9.10×10^{-6}
ELASTIC	λ (regularization weight)	0.0 to 0.1	9.24×10^{-6}
	α (square vs absolute value weighting)	Values {0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 1.0}	1.0
RKS	λ (regularization weight)	Values {0.0, 10^{-6}, 10^{-5}, 10^{-4}, 0.001, 0.01, 0.1, 1.0, 2.0, 4.0}	10^{-5}
	σ (kernel width)	Values {0.1, 0.25, 0.5, 1.0, 1.2, 1.5,	1.0

		2.0, 3.0, 4.0, 8.0, 12.0?	
	D (number of random functions)	Values {100, 200, 450, 700, 1000, 1200, 1700, 2000, 2200, 2700, 3500	3500
TREE	Degree of pruning	25 to 125 by 25	75
	Minimum points for subset split	0 to 20 by 2	16
BAGTREE	Number of trees	Values: {1, 3, 6, 12, 25, 50, 100, 150, 200, 400, 1000, 1500, 2000, 2500, 3000, 4000, 5000}	5000
	Fraction of training data in each tree	Values: {0.1, 0.3, 0.5, 0.7, 0.9, 1.0}	1.0
BOOST	Number of trees	Values: {10, 20, 50, 100, 150, 200, 300, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000}	3000

NN, 1 hidden layer	Number of neurons in hidden layer	Values {4, 7, 10, 13, 16, 20, 25, 30, 35, 40, 50, 75, 100, 125, 150, 200, 250, 300}	125
NN, 2 hidden layers	Number of neurons in first hidden layer	Values {2, 5, 10, 25, 50, 75, 100, 125, 150, 200}	150
	Number of neurons in second hidden layer	Values {1, 3, 5, 10, 15, 25, 40, 75}	3
NN, 3 hidden layers	Number of neurons in first hidden layer	Values {1, 3, 5, 10, 15, 25, 40, 65, 100, 150}	100
	Number of neurons in second hidden layer	Values {2, 5, 10, 15, 25, 40}	40
	Number of neurons in third hidden layer	Values {1, 3, 5, 10, 15, 25, 40}	5

The overall minimum RMSE results for the validation and test data sets for all algorithms tested are shown in Table 3-7.

Table 3-7: RMSE for each optimized ML algorithm against validation and test data sets when trained with small and large training data sets. Algorithms that are too computationally demanding for training with the large data set are labeled "Prohibitive".

	Small training data set		Large training data set	
Algorithm	Validation	Test	Validation	Test
RLR	30.55	29.84	31.22	30.23
LASSO	30.55	29.84	31.22	30.23
ELASTIC	30.55	29.84	31.22	30.23
KRR	23.85	23.41	Prohibitive	
RKS	25.35	25.52	22.22	22.10
TREE	29.19	28.71	25.14	25.45
BAGTREE	24.50	23.91	19.91	20.15
BOOST	28.86	28.33	28.21	27.86
NN, 1 HL	26.42	26.78	23.18	23.48
NN, 2 HL	25.76	25.20	21.58	22.69
NN, 3 HL	25.59	25.51	20.94	21.79
SVR	24.13	23.63	Prohibitive	

Linear regression variants

Except for the RLR with the smaller training data set, the optimum regularization parameters for all of the linear regression variants were small, or equal to zero. For the RLR with the smaller training data set, increasing the regularization parameter

made little difference to the results until a value of 100 was reached, at which point increasing the regularization parameter further degraded the results (Figure 3-1). The small degree of regularization applied in the optimum cases for the linear regression variants may explain why the RMSE values using these optimum parameters vary so little between methods. These results show that there is little advantage in adding regularization terms to the standard linear regression for either the small or large training data set. It is also notable that the linear regression variants performed better when using the small training data set, possibly because the linear trend is better defined in the small data set. However, this difference is only about 1 W/m^2 or less.

Figure 3-2 shows the results for the LASSO regression. Similar to the RLR, the results do not improve with increasing λ, and the performance is better with the smaller training data set. The primary difference is that changing the regularization parameter made little difference to the results.

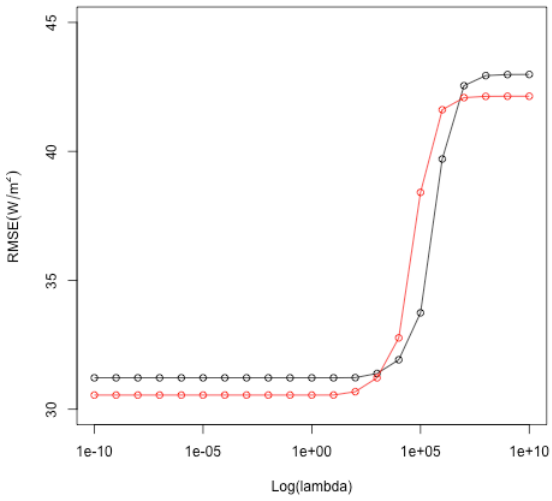


Figure 3-1: RMSE for validation data set for RLR against regularization parameter λ. Red: Small training data set. Black: Large training data set.

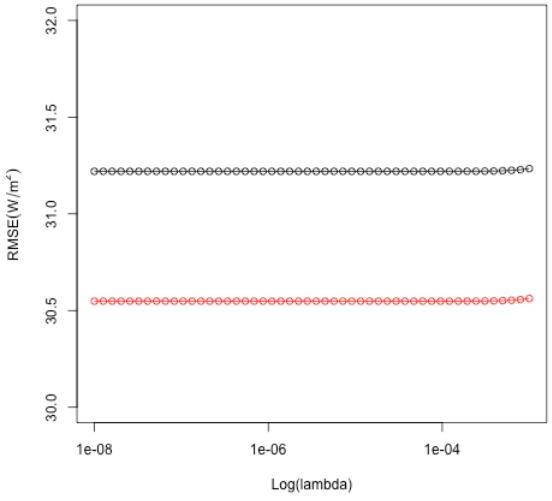


Figure 3-2: RMSE for validation data set for LASSO regression against regularization parameter λ. Red: Small training data set. Black: Large training data set.

The ELASTIC regression tuning results are shown below in Figures 3-3 and 3-4. The results again show that the lowest RMSE values occur for low values of λ. There is also a modest trend towards better performance at lower values of α, which represent a regression formula closer to RLR than to the LASSO formula. Once again, performance is better with the smaller training data set.

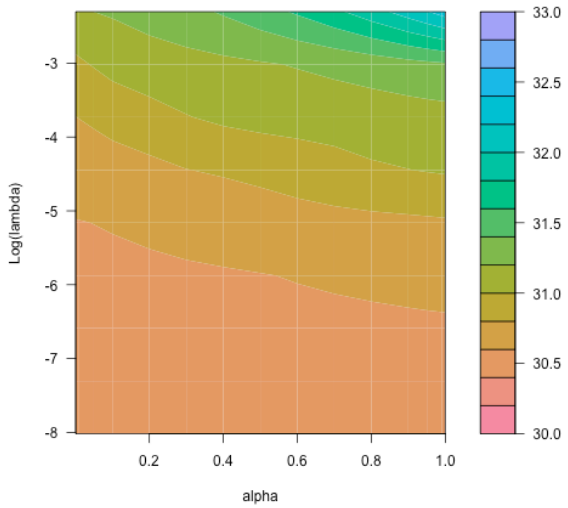


Figure 3-3: Validation RMSE for varying values of λ and α when ELASTIC regression is tuned with the small training data set.

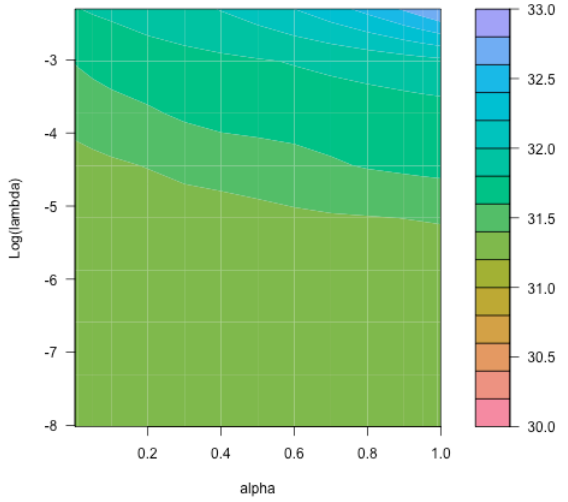


Figure 3-4: Validation RMSE for varying values of λ and α when ELASTIC regression is tuned with the large training data set.

Kernel methods

The KRR was only optimized for the smaller training data set since it could not complete iterations with the larger data set sufficiently quickly. As shown in Figure 3-5, it has much more sensitivity to the σ parameter than to the λ parameter.

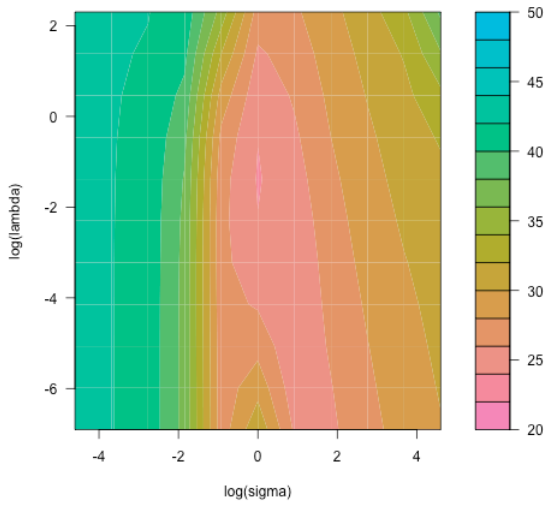


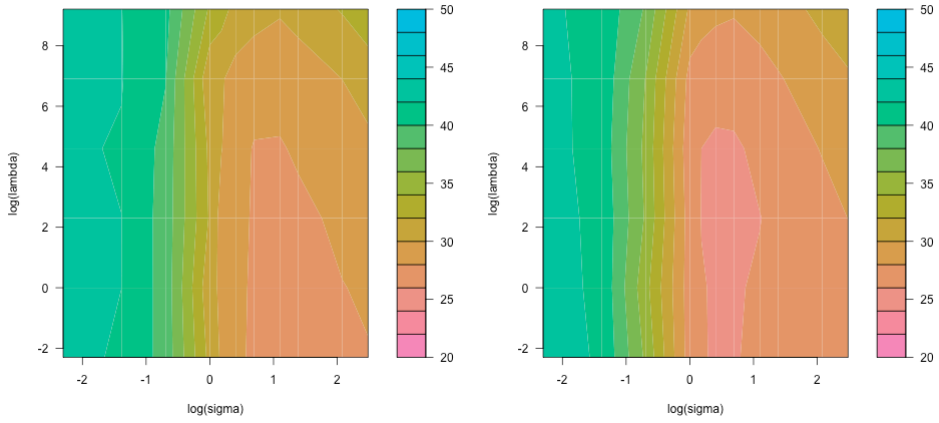
Figure 3-5: RMSE in W/m^2 for validation data set for KRR algorithm tuned with small training data set.

The RKS algorithm could be run with both the small and large training data sets, with fitting required for three parameters. In addition to λ and σ parameters analogous to those of the KRR, there is also a parameter (D) representing the number of random functions used to approximate the kernel. For the small training data set, RMSE was

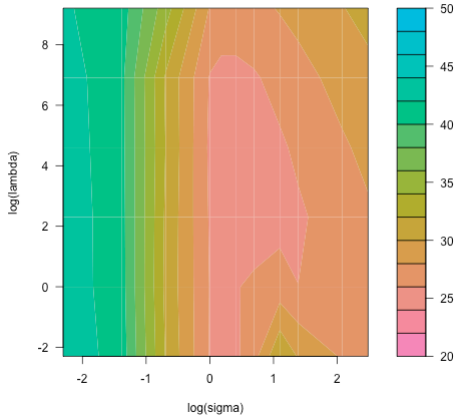
minimized at a D value of 2000, but for the large training data set a smaller minimum RMSE value was found with increasing D up to the maximum value tested of 3500. Due to longer computation times for higher D, the iteration was not carried out for D values over 3500.

The validation RMSE values for the RKS trained with the small training data set are shown in Figure 3-6, and with the large training data set in Figure 3-7. A smaller minimum RMSE was found using the large training data set than the small training data set. For both the small and large training data set, for low D values the results were much more sensitive to the σ parameter than to the λ parameter. At higher D, sensitivity to the λ parameter started to show at higher σ, with the pattern becoming more pronounced as D increased.

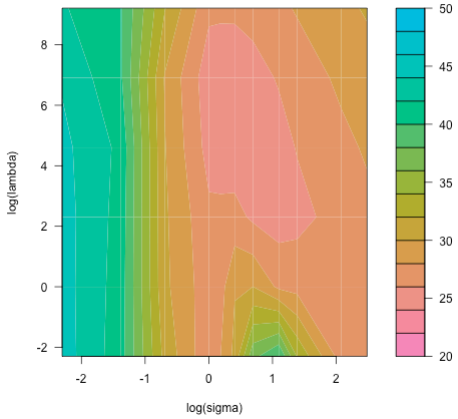
a) b)



c)



d)



e)

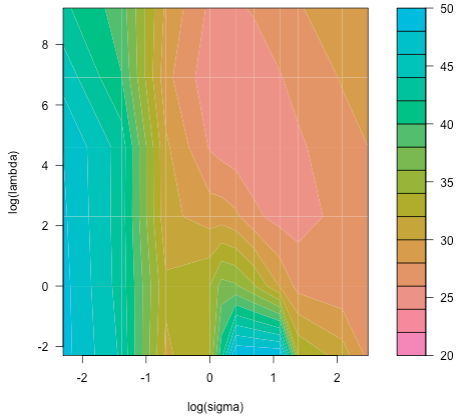
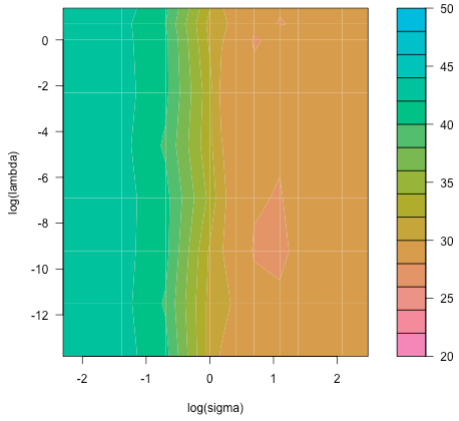
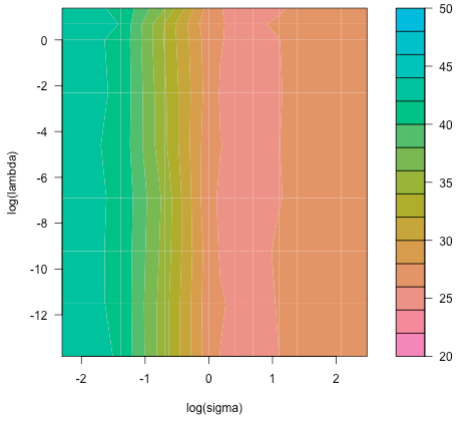


Figure 3-6: Validation RMSE for RKS tuned with small training data set. a) D = 100. b) D = 450. c) D = 1000. d) D = 2000. e) D = 3500

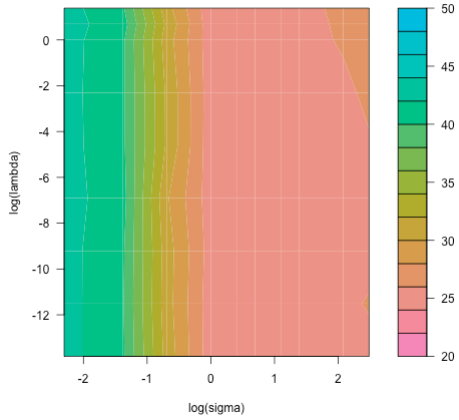
a)


log(lambda)
0
-2
-4
-6
-8
-10
-12
-2
-1
0
1
2
log(sigma)
50
45
40
35
30
25
20

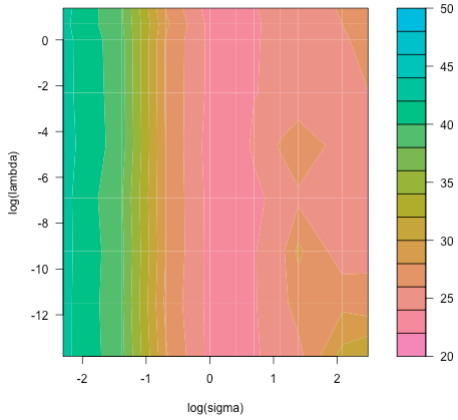
b)


log(lambda)
0
-2
-4
-6
-8
-10
-12
-2
-1
0
1
2
log(sigma)
50
45
40
35
30
25
20

c)


log(lambda)
0
-2
-4
-6
-8
-10
-12
-2
-1
0
1
2
log(sigma)
50
45
40
35
30
25
20

d)


log(lambda)
0
-2
-4
-6
-8
-10
-12
-2
-1
0
1
2
log(sigma)
50
45
40
35
30
25
20

e)

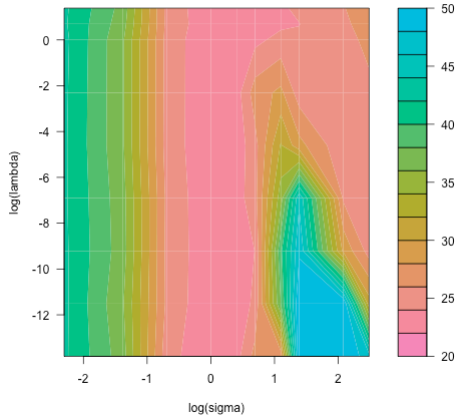

log(lambda)
0
-2
-4
-6
-8
-10
-12
-2
-1
0
1
2
log(sigma)
50
45
40
35
30
25
20

Figure 3-7: Validation RMSE for RKS tuned with large training data set. a) D = 100. b) D = 450. c) D = 1000. d) D = 2000. e) D = 3500.

Tree methods

The tuning parameters for the basic regression tree algorithm are the degree of pruning and the number of points required for a subset to be split. As shown in Figures 3-8 and 3-9, the results of the tree algorithm are relatively insensitive to these parameters, especially the degree of pruning. The performance of the tree algorithm was significantly better for the large training data set than for the small training data set (minimum RMSE of 25.45 vs. 28.71 W/m^2).

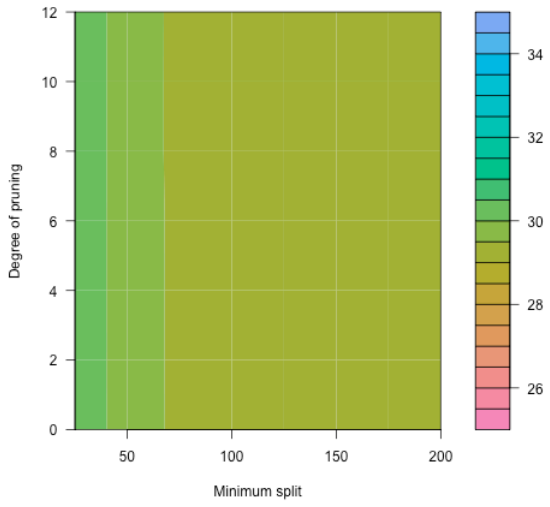


Figure 3-8: Variation in validation RMSE with subset split and pruning parameters for TREE algorithm, small training data set

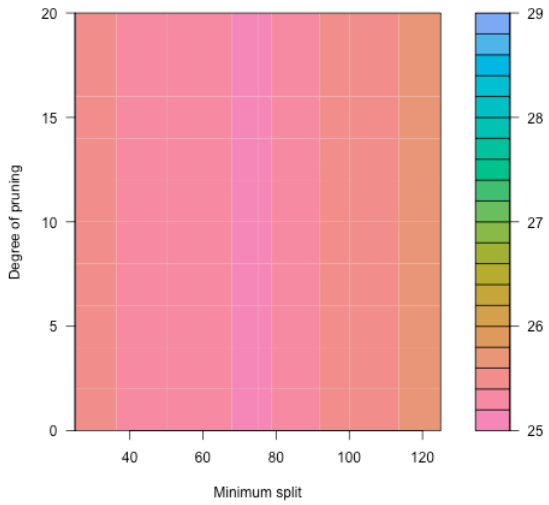


Figure 3-9: Variation in validation RMSE with subset split and pruning parameters for TREE algorithm, large training data set

Due to the insensitivity of the tree algorithm to the pruning and splitting parameters, the default settings for those parameters were used when training the more complex BAGTREE and BOOST algorithms. The only parameter left to tune for the BOOST algorithm is the number of trees to use. Figure 3-10 shows the validation RMSE for the small (red) and large (black) training data sets. The RMSE for the small training data set reaches a minimum when trained with 1000 trees. With the large training data set, the minimum RMSE was reached with 3000 trees and the RMSE did not increase much beyond that with higher numbers of trees.

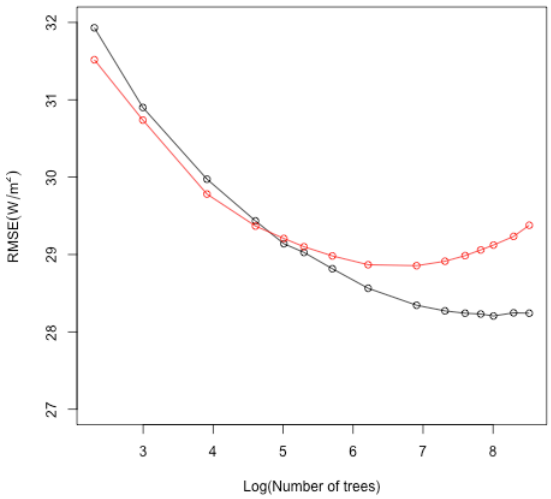


Figure 3-10: Validation RMSE vs number of trees used in BOOST algorithm. Red line is for small training data set and black line is for large training data set.

The BAGTREE algorithm performance improved with the number of trees used for both the small and large training data sets, as shown in Figures 3-11 and 3-12. The BAGTREE algorithm also performed somewhat better when larger subsets were drawn in the bootstrap aggregation, with the best performance occurring when the whole input data set was used for each tree. The BAGTREE algorithm produced the lowest validation RMSE out of all the algorithms tested when trained with the large training data set and optimized.

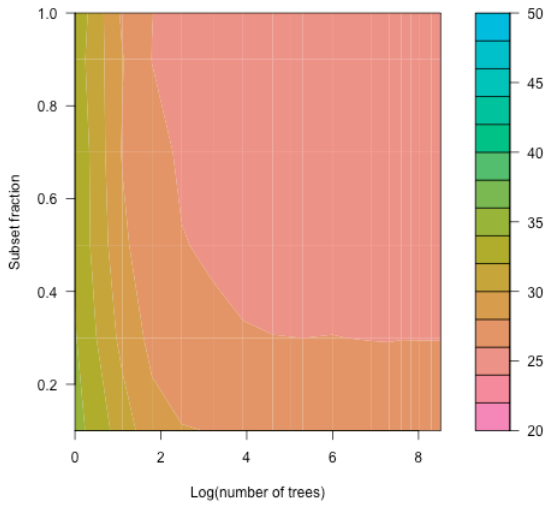


Figure 3-11: BAGTREE algorithm RMSE values vs number of trees and fraction of training data set used in bootstrap aggregation, small training data set

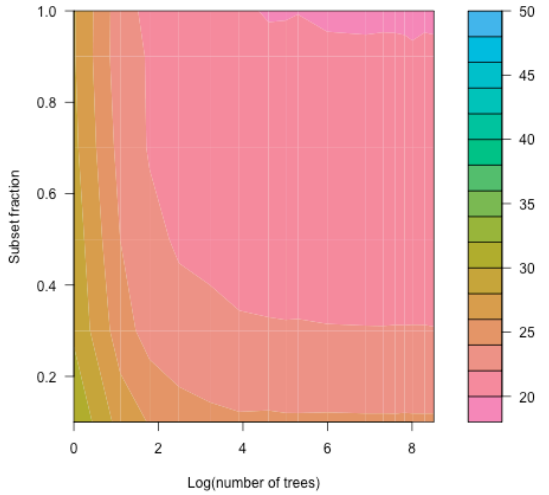


Figure 3-12: BAGTREE algorithm RMSE values vs number of trees and fraction of training data set used in bootstrap aggregation, large training data set

Neural networks

The validation RMSE is shown for one hidden layer neural networks trained with the small and large training data sets in Figure 3-13. When a one hidden layer neural

network is applied to the small training data set, validation RMSE fluctuates between about 27 and 29 W/m^2, with no clear trend with increasing numbers of neurons. With the large training data set, the performance of the neural network exhibits a general trend towards improvement out to about 100 neurons, then levels off. It is clear that adding more neurons to a one hidden layer network is likely not to improve performance with either data set once the number of neurons exceeds 100.

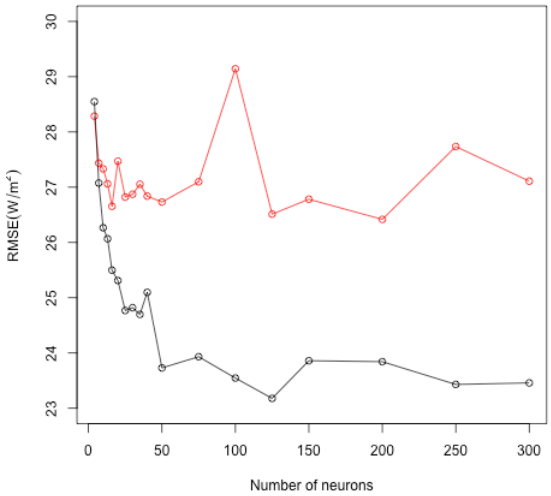


Figure 3-13: Validation RMSE for small (red) and large (black) 1 hidden layer neural networks.

The pattern of the two hidden layer neural network RMSE variation with the number of neurons in the first layer was similar to the one hidden layer NN results, but with greater accuracy achieved. With the small training data set, there are few to no gains in accuracy when the number of neurons in the second layer exceeds 10. With the large training data set, performance plateaus at higher numbers of neurons in the first

layer than for the small training data set, especially when the number of neurons in the second layer is relatively small. The validation RMSE for the two hidden layer NNs are shown in Figures 3-14 and 3-15 for the small and large training data sets, respectively.

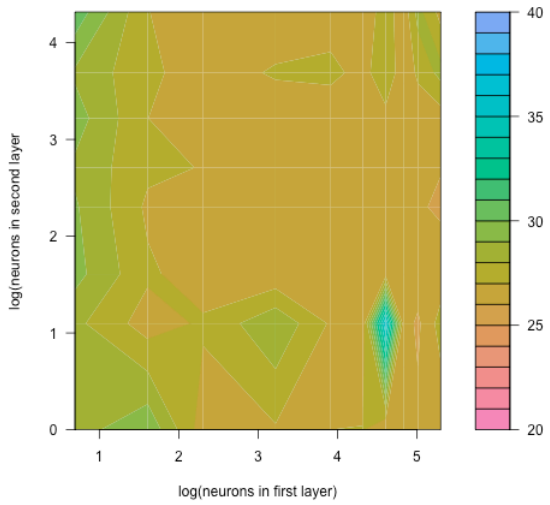


Figure 3-14: Validation RMSE for 2 hidden layer neural network trained using small training data set.

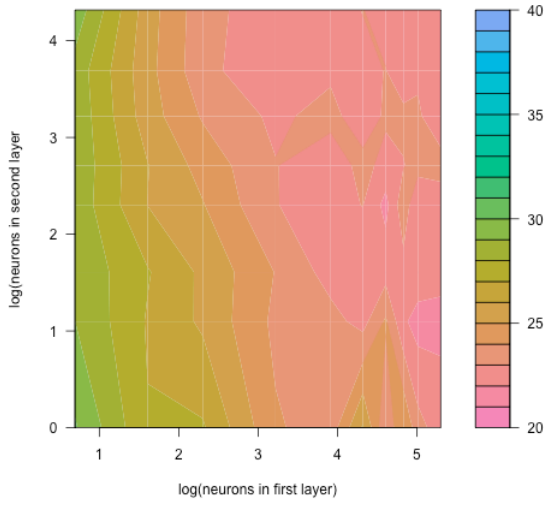


Figure 3-15: Validation RMSE for 2 hidden layer neural network trained using large training data set.

The addition of a third hidden layer to the neural network did not greatly improve the performance with the small training data set. Validation RMSE tended to decrease with increasing numbers of neurons in the first hidden layer, but the results were relatively insensitive to the number of neurons in the second and third hidden layers. The minimum RMSE achieved in the three hidden layer NN trials with the small training data set was 25.59 W/m^2, which is only somewhat lower than the 25.76 W/m^2 minimum value for the two hidden layer NN.

Training with the large training data set led to higher accuracy results than with the small training data set, but the patterns in performance with parameter variation were similar. When trained using the large training data set, the three hidden layer NN resulted in an optimized validation RMSE value of 20.94 W/m^2, compared to 21.58 W/m^2 for the two hidden layer NN. Performance improved with the number of neurons in the first hidden layer, but as the number of neurons in the first hidden layer increased, the sensitivity to the numbers of neurons in the second and third hidden layers decreased. Figure 3-17 shows the validation RMSE for different numbers of neurons in the hidden layers when trained with the large training data set and a three hidden layer NN.

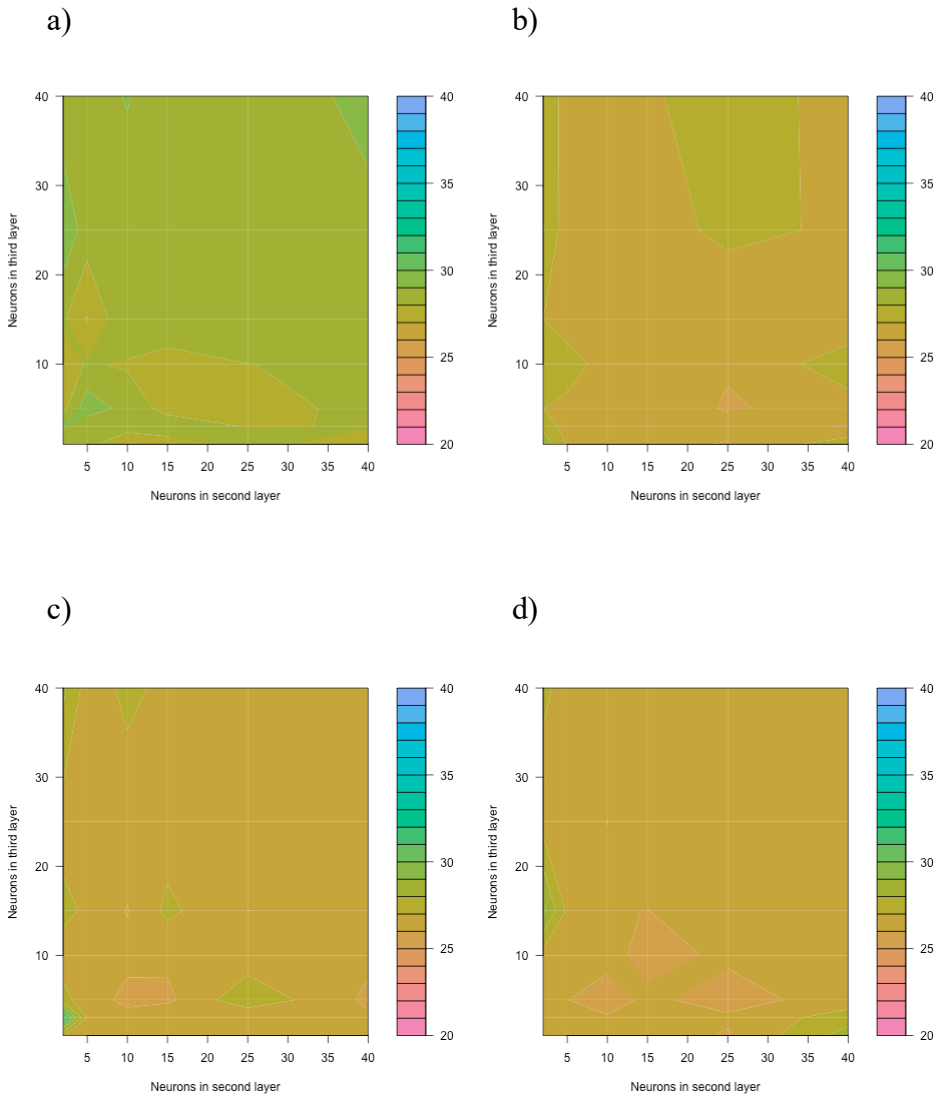


Figure 3-16: Validation RMSE for three-layer NN trained with small training data set. a) 3 neurons in first hidden layer. b) 10 neurons in first hidden layer. c) 25 neurons in first hidden layer. d) 100 neurons in first hidden layer.

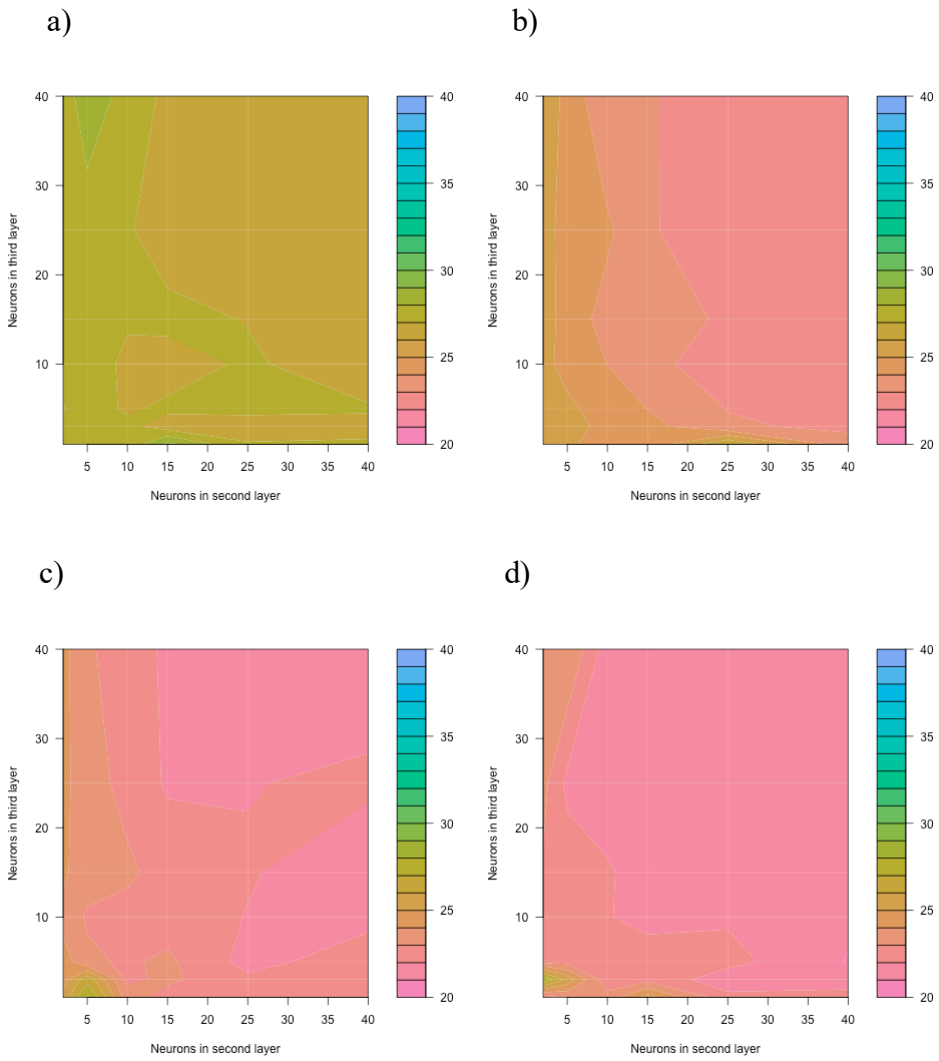
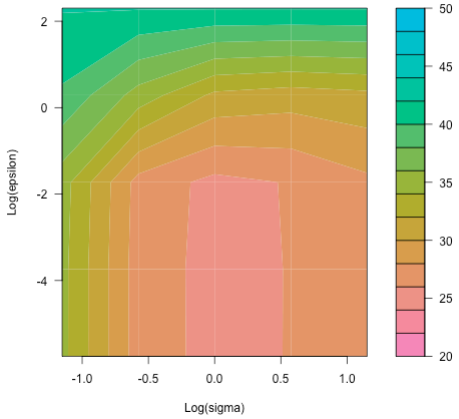


Figure 3-17: Validation RMSE for three-layer NN trained with large training data set. a) 3 neurons in first hidden layer. b) 10 neurons in first hidden layer. c) 25 neurons in first hidden layer. d) 100 neurons in first hidden layer.

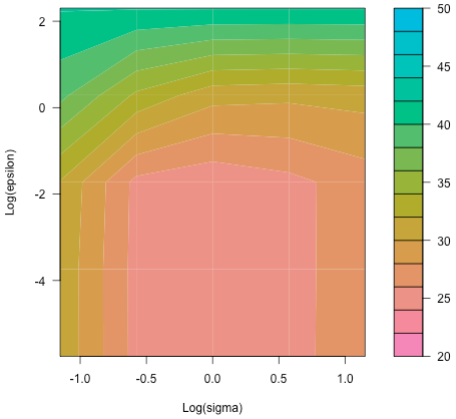
Support vector regression

The support vector regression was only performed with the smaller training data set. Validation results show most sensitivity to the ε parameter, and relatively little to the σ or C parameters.

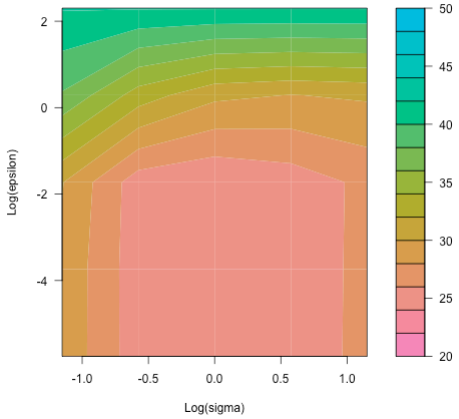
a)


Log(epsilon)
2
0
-2
-4
-1.0
-0.5
0.0
0.5
1.0
Log(sigma)
50
45
40
35
30
25
20

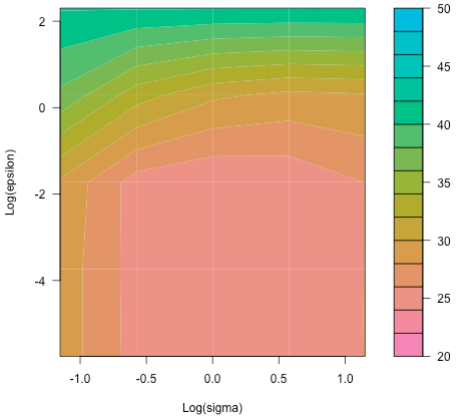
b)


Log(epsilon)
2
0
-2
-4
-1.0
-0.5
0.0
0.5
1.0
Log(sigma)
50
45
40
35
30
25
20

c)


Log(epsilon)
2
0
-2
-4
-1.0
-0.5
0.0
0.5
1.0
Log(sigma)
50
45
40
35
30
25
20

d)


Log(epsilon)
2
0
-2
-4
-1.0
-0.5
0.0
0.5
1.0
Log(sigma)
50
45
40
35
30
25
20

e)

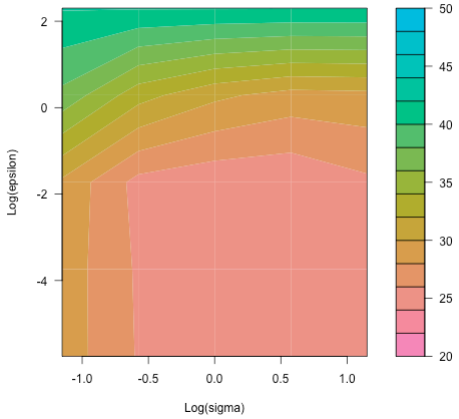

Log(epsilon)
2
0
-2
-4
-1.0
-0.5
0.0
0.5
1.0
Log(sigma)
50
45
40
35
30
25
20

Figure 3-18: Validation RMSE for support vector regression. a) C = 0.316 b) C = 0.750 c) C = 1.778 d) C = 4.217 e) C = 10.0

Trials with GLASS LAI

Three of the ML techniques, the bagging tree, the 2-layer NN, and the RKS with 1000 random functions, were trained with just the GLASS radiation and either GLASS or MODIS LAI, in order to test whether using GLASS data in the place of MODIS is potentially viable. The results indicate that similar RMSEs relative to flux tower data were obtained through use of GLASS LAI or of MODIS LAI, as shown in Table 3-8. There was also little difference in the RMSEs obtained from the three ML methods tested.

Table 3-8: Optimized validation RMSEs with respect to flux tower data for ML methods trained with GLASS DSR and PAR, plus either GLASS or MODIS LAI.

Method and training data set used	**GLASS LAI**	**MODIS LAI**
2-layer neural network, small training set	31.20	30.71
Bagging tree, large training set	30.13	30.58
RKS, D =1000, large training set	30.79	30.63

Discussion

Machine learning has been used frequently for estimation of LE from both ground-based and remote sensing data, but usually only a single method is used or only a few

methods are compared in each study. There is not much discussion of the computational demands of the various ML techniques in the literature, or of the potential tradeoffs between accuracy and computational demand. It is also not common to use a global data set or to derive LE from remote sensing data only. Here we systematically compared several machine learning methods for obtaining LE from a remote sensing only input data set, representing a range of the most commonly used types of ML algorithms. We also compared the results of training with a smaller or larger training data set in terms of both accuracy and required computing power.

The machine learning method performance differed between the small and large training data set. The best results for the small training data set were with the kernel ridge regression (KRR), which ran too slowly to be viable with the large training set. Three of the other algorithms (RKS, BAGTREE, and multi-layer neural networks) were able to produce a lower RMSE with the large training data set than the lowest RMSE attained with the small training data set. This implies that although better results are likely with more training data, it is also likely that the optimal method to use will differ between smaller and larger training data sets. The cloud-detection example given in Pérez-Suay et al. (2017) also demonstrated this dynamic between the KRR and RKS methods. Here we also had good performance with the RKS, but even better performance with the bagging tree and multi-layer neural network, demonstrating the necessity of testing a range of algorithms.

The linear regression variants did not demonstrate much advantage over linear regression without any regularization. The RMSE values decreased very little, if at

all, when regularization was added and increased dramatically in the case of high values of a squared coefficient value regularization term (the RLR algorithm). Adding an absolute value regularization term (the LASSO method) increased the computation time required with no reduction in RMSE. Combining a squared and absolute value regularization (the ELASTIC method) yielded no further advantage. All of the other ML methods produced more accurate results than the linear regression variants, which is unsurprising, due to the greater ability of the other methods to respond to nonlinear signals in the information available in the data. In addition, it is clear that the linear regression variants are not capable of converting the greater amount of information available from a larger training data set into a more accurate trained algorithm. Their use is not recommended unless computational power availability is highly restricted.

Other than the linear regression variants, no single family of methods clearly outperformed the rest in speed or accuracy. RMSE values less than 25 W/m^2 were achieved with kernel, tree, neural network, and SVM methods with the smaller training data set. However, for those families with more than one method tested in this study, some of those methods outperformed others. It is difficult to say why one method outperforms another in the same family. Overall, the best results were obtained with the large training data set and a tree algorithm, which may be at least partially due to the fact that the tree algorithms perform an implicit division of the data set into different regimes where the relationships between the variables may differ.

It has been shown here that machine learning can be used profitably to extract information from remote sensing data alone to obtain LE. Use of GLASS DSR, PAR, or both resulted in similar performance to use of ground station R_n. It has also been shown that, while some of the ML algorithms perform well in terms of both accuracy and computational demand, there is also some tradeoff between training efficiency and performance. This is seen most clearly in the results with the large training data set, where the BAGTREE algorithm produced the lowest RMSE but required more run time than the RKS, boost tree, or smaller neural networks. When trials were made with just radiation and LAI variables as input and the LAI coming from GLASS or MODIS, the results were similar between GLASS and MODIS LAI, indicating that use of the ML methods with GLASS data sets alone is likely to be a viable approach.

Conclusions

A comparison of ten ML methods for obtaining LE from a combination of remote sensing data (GLASS and MODIS) was performed in terms of accuracy and speed. The results showed wide variation in the time required to perform a single train/test iteration, both between algorithms and between a smaller and larger training data set, as shown in Table 3-3. Experimentation with different combinations of input variables showed that including more variables generally improved the results with little or no additional computational cost. This experimentation also showed that using the remote sensing GLASS radiation variables produced results comparable to using ground-based net radiation measurements. In addition, it was shown that inclusion of NBAR as one of the parameters made for a substantial improvement to

the results, reducing validation RMSE by 1.5 W/m^2 or more over the results with the same parameters except NBAR.

Optimization was performed for one or more algorithm parameters for each combination of methods and training data sets tested. The best performance with a smaller training data set was obtained using the kernel ridge regression (KRR), which was too computationally demanding for use with the larger data set. The best performance with the larger data set was achieved by the bootstrap aggregation tree (BAGTREE) method, followed by the random kernel (RKS) and multiple hidden layer neural network (NN) methods. Other than relatively weak performance by all of the linear regression variants considered, all "families" of ML methods with similar theoretical bases had at least one method that produced validation RMSE values of less than 25 W/m^2 for daily LE with the smaller training data set. Regression trees, RKS, and neural networks all produced validation RMSE values of less than 23 W/m^2 with the larger training data set, with the bootstrap aggregation tree (BAGTREE) having the best performance at about 20 W/m^2. It is noted here that these performance statistics are based on including all sites in training, validation, and test data sets.

Additional work could be done to test the conclusions of this study. It would be of interest to perform testing with data from sites independent of the training data set to see whether similar performance would be obtained. A more thorough assessment could also be made of the effects of using variables that probably have a high degree of redundancy, such as DSR and PAR, or NBAR and vegetation indices. In this study,

sometimes there were apparent gains when using such redundant variables, but it would be of interest to further characterize the effects of using or not using them.

Due to the lack of a clear pattern in which techniques had better performance, and the difference in results between the small and large training data sets, use of a still larger training data set or data from different sources would require further experimentation. Trials with additional variables, such as soil moisture or precipitation, that have weak redundancy with those already considered could be especially advantageous. However, the methods with the best performance here should be included in this experimentation. For a training data set of <10,000 data points, the KRR should be considered, as should the BAGTREE, RKS, and multi-layer neural network for training data sets of sizes >10,000.

There are implications from this study for future work in machine learning for evapotranspiration. The performance of the ML algorithms varied, even when using the same data set and after tuning for optimal performance. This indicates that multiple methods should be tested for any particular application. The performance of the algorithms with redundant data inputs varied, with improved performance when including multiple radiation variables but little change when adding vegetation index to NBAR as inputs, showing that different combinations of input variables should continue to be tested in future development of ML methods for evapotranspiration retrieval.

Chapter 4: Machine learning applied to remote sensing of evapotranspiration in the continental United States

Introduction

With the world facing a changing climate and increasing demands on a limited supply of fresh water, monitoring of the hydrological cycle on all scales from local to global has great utility. Efforts have been made to deduce regional and global trends in LE from flux tower measurements and remote sensing data. Jung et al. (2010) used a model tree to upscale LE from the global Fluxnet tower network and combine the ground-based data with remote sensing, concluding that there was a global trend towards increasing LE between 1982 and 1997 that reversed direction to become a decreasing trend from 1998 to 2008. Wang et al. (2010a, 2010b) made use of meteorological data, vegetation indices, and a regression formula based on Penman (1948) separating LE into energy controlled and atmospherically controlled components, reaching the conclusion that global LE increased at a rate of 0.6 W/m^2 per decade between 1982 and 2002. Yao et al. (2013) estimated LE trends in China between 2001 and 2010 based on an algorithm derived from the formula of Priestley and Taylor (1972) and surface net radiation, vegetation index and surface temperature data. They found that LE decreased over most of China during that period, although increasing LE was indicated in some regions.

In addition to monitoring of global trends, detection and monitoring of regional drought is an important application of LE measurements. There is much recent

research into the development and use of remote sensing-based indices for detection and monitoring of meteorological and agricultural drought (e. g. Anderson et al., 2016, Amani et al., 2017, Meng et al., 2016, Roundy and Santanello, 2017) These indices sometimes do not represent any physical quantity, although vegetation indices that consist of ratios of red and near-IR bands that are sensitive to the presence of photosynthesizing vegetation, moisture indices that include SWIR bands sensitive to canopy moisture, and temperatures derived from thermal IR measurements are often included in the indices in various combinations. Validation of these indices is usually against ground-based meteorological measurements or against crop yield data. A time sequence of LE maps can be used as an alternative indicator of regional drought. This approach has the advantage of representing variations in a physical quantity that can be directly validated against ground measurements. However, any comparison with drought maps generated by other methods will not be comparing the same quantity.

Machine learning, the use of nonparametric algorithms that change their internal state in response to a training data set, has been used in a great many applications both inside and outside of the field of remote sensing. Machine learning techniques have distinct advantages for application to the problem of remote sensing of evapotranspiration. Virtually all existing models or parametric formulas require particular variables as input, while any data set can be used as input to a machine learning routine. Machine learning formulations make no assumptions regarding the form of the relationship between the input and output variables. Some caution is in order, though, because the ML methods cannot be expected to perform well outside

of the range of conditions represented in their training data. ML methods are also a "black box", meaning that the relationships found between the input and output variables are not easily characterized in a way that is meaningful to a human user. A discussion of the application of machine learning to the problem of detecting evapotranspiration is made in Carter and Liang (2019b). The global data sets recently produced by Jung et al. (2018) are especially notable because data sets from an ensemble of machine learning methods have been made available along with data produced by several individual machine learning methods.

Regional mapping provides an opportunity to test many aspects of LE retrieval with machine learning. It is of interest whether retrievals validated using point locations are applicable over an extended area, and how algorithms trained with global data work on a regional scale. Recently, much research has focused on mapping of evapotranspiration at local to regional scales (Elnmer et al. 2019, Khand et al. 2019, Yi et al. 2018), but little work exists comparing evapotranspiration maps produced by different machine learning methods and with different machine learning input data weights, a gap which we attempt to address in this study. We also perform some examination of the ability of different methods to respond to drought signals by performing mapping for the year 2012, a drought year in much of the United States.

Data

Two sources of remote sensing data are used to derive LE from machine learning methods in this study: Global Land Surface Satellite (GLASS) (Liang et al., 2013,

2014) and standard MODIS products. The standard MODIS products used are combined Terra and Aqua nadir BRDF-adjusted reflectance (NBAR) (the MCD43A4 product) and Terra normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) (the MOD13C1 product). Data are available from both of these sources for the full year 2012. GLASS data have been produced using data from both the Advanced Very High Resolution Radiometer (AVHRR) and MODIS instruments. GLASS data for the year 2012 were produced from MODIS data. The GLASS products used are the downward shortwave radiation (DSR) and photosynthetically active radiation (PAR) (Zhang et al., 2014) , leaf area index (LAI) (Xiao et al., 2017, 2016), fraction of absorbed photosynthetically active radiation (FAPAR) (Shi et al., 2016), and albedo (Qu et al., 2014, Liu et al., 2013). These data are on a 0.05 degree by 0.05 degree global grid, with dimensions of 7200 x 3600 for global coverage and 1300 x 525 to cover the continental United States. The GLASS DSR and PAR data are at daily time resolution, the MODIS vegetation indices are available once every 16 days, while the other variables are available for every 8 days. The variables with an 8-day and 16-day time resolution were linearly interpolated to daily values for this study.

The MODIS NBAR product used in this study, MCD43C4 (Schaaf, 2015b), combines data from the MODIS sensors aboard the Terra and Aqua platforms. It is also produced globally at a resolution of 0.05 degrees by 0.05 degrees and is available daily. The product for each day is produced using surface reflectance data from the 16-day period centered at that day. The algorithm used to produce the MCD43C4

product makes use of BRDF models combined with pixels selected to optimize representativeness for the time period, adjusting the observed surface reflectance values to approximate the reflectance that would be seen from a nadir view. In the MCD43C4 product, NBAR values are given for MODIS bands 1-7, which range in wavelength from 0.6 to 2.1 μm (blue to SWIR).

The expectation that including vegetation index as an input would make little difference to the results because they would be expected to have high correlation to NBAR was tested by conducting trials with and without vegetation index as an input. The MODIS vegetation index product used in this study, MOD13C1 (Didan 2015), contains normalized-difference vegetation index (NDVI) and enhanced vegetation index (EVI) at 0.05 degree by 0.05 degree resolution and is produced every 16 days. The MOD13C1 product is a spatial and time composite product, where the highest quality observations are selected from the 250m resolution pixels that fall within the spatial range and time composite period represented by each MOD13C1 grid cell.

The ground-based LE data used for training and validation in this study are flux tower measurements taken from the Ameriflux (http://ameriflux.lbl.gov) and Fluxnet 2015 (http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/) networks. Sites with at least three continuous years of data available between 2001 and 2015 were selected. There were a substantial number of gaps of varying lengths within these data records, so the data were preprocessed to include only those dates for which at least 40 of 48 possible half-hourly observations were available. Daily mean LE values were

determined from these observations. As flux tower sites are known to have an energy balance closure issue, the LE values were adjusted by assuming a constant Bowen ratio and the energy balance equation

$$R_n = H + LE + G \tag{4-1}$$

where R_n is the surface net radiation, H is sensible heat flux, and G is net heat storage. If G was not available, its value was assumed to be zero. Typically, flux tower energy balance is within 80% of closure (Wilson et al., 2002), so days with a significantly greater imbalance (imbalance was greater than 0.4 times R_n) were excluded. This reduced the total number of site days available by about 20 percent. For more details about the flux tower sites included in the initial training and validation data sets, refer to the "Data" section of Carter and Liang (2018).

The resulting training and validation data set consists of 79098 site-days from 184 sites that include 12 IGBP surface types. The snow and non-snow data from each site were randomly distributed to training, validation, and test data sets with an 80%/ 10%/ 10% split. LE data from an additional 14 Ameriflux sites was used to provide an independent check on the LE maps.

LE from five data sets were used for purposes of comparison to the 2012 monthly mean maps. The North American Regional Reanalysis (NARR) (Mesinger et al., 2006) is a reanalysis product generated by the US National Center for Environmental Prediction (NCEP) covering a portion of the Northern Hemisphere including the

United States, Canada, and Mexico at 32 km resolution. Two FluxCom products from ensembles of machine learning methods (Jung et al., 2018) were also used for comparison. The FluxCom "RS" data set uses input from MODIS only with global coverage at 0.0833°. The "RS + METO" data set uses a combination of MODIS and meteorological data and provides global coverage at a resolution of 0.5°. A global version of the standard MODIS LE product (MOD16) was obtained from https://www.ntsg.umt.edu/project/modis/mod16.php (Zhang et al., 2015, 2010). The MOD16 product is based on the Penman-Monteith formula and relies on MODIS LAI along with meteorological variables as parameters of a biome property look-up table (Mu et al., 2011, 2007). Monthly MOD16 data for the year 2012 were obtained at 0.0833 degree by 0.0833 degree. Finally, an LE retrieval is included in the GLASS data sets, at 0.05 degree by 0.05 degree spatial and 8-day time resolution. The GLASS data set was produced by Bayesian model averaging applied to five process-based algorithms, all of which are based on the Penman-Monteith, Priestley-Taylor, or semiempirical Penman methods (Yao et al., 2014).

Methods

Machine learning techniques used

Ten machine learning techniques were tested for the application of evapotranspiration remote sensing in Carter and Liang (2019b). The techniques that yielded the best results while also running quickly enough to make running many iterations of training and testing feasible were the bootstrap aggregation (bagging) tree, the random kernel

(RKS) algorithm, and the two and three hidden layer neural networks. Each of these methods will be described briefly below.

Bootstrap aggregation tree

Variants of regression trees, in which the space of input variables is iteratively partitioned into sections and each section assigned a value, are used frequently in machine learning. Often multiple regression trees are trained and the results from all of them combined. Bootstrap aggregation (bagging) is the selection of subsets of the input data for use in construction of the model trees. There are two parameters that must be tuned, the number of trees and the fraction of the input data used in training each tree. The bootstrap aggregation tree is described by Hastie et al. (2009) as part of their discussion of regression trees.

Random kernel function (RKS)

The random kernel method is a variant of the kernel ridge regression in which the input data are implicitly reprojected using a kernel function while performing a ridge regression. In the RKS algorithm, the kernel is simulated using randomly selected functions from its Fourier transform. The radial basis function kernel (RBF)

$$\langle \phi(X_i), \phi(Y_j) \rangle = exp\left(-\frac{\|X_i - Y_j\|^2}{2\sigma^2} \right) \quad (4\text{-}2)$$

is the reprojection formula that is simulated here. There are three tunable parameters in the RKS method using the RBF kernel: the number of random functions, a regularization parameter λ, and the σ parameter specifying the width of the RBF

kernel to be approximated. The RKS algorithm is described in detail by Pérez-Suay et al. (2017) , who also provide examples of its application to remote sensing problems.

Neural networks

The neural network is a well-known machine learning technique in which the input data are passed through successive layers of nodes, with each node performing a weighted sum of its inputs and passing that sum into an activation function to determine its output. The most commonly used form of neural network has only one "hidden layer" of nodes between the input and output, but Carter and Liang (2019b) found that increasing the number of these hidden layers improved performance on the evapotranspiration problem significantly while the computational demand was not so great as to make training these deeper networks intractable. A neural network is specified by a number of hidden layers and a number of nodes in each hidden layer. In this study, the optimum number of nodes in two and three hidden layer networks was found for each trial. Neural networks are described in detail in Hastie et al. (2009).

Trials with different data sets

Preliminary experimentation showed that separation of snow and non-snow days according to an albedo threshold of 0.4 reduced typical RMSE with the test data sets by several W/m^2 in the case of non-snow data and over 10 W/m^2 for snow data. It should be noted that correlation coefficients were very low for the snow data set, as

the data were compressed near the low end of the LE range, as illustrated in the Results section below.

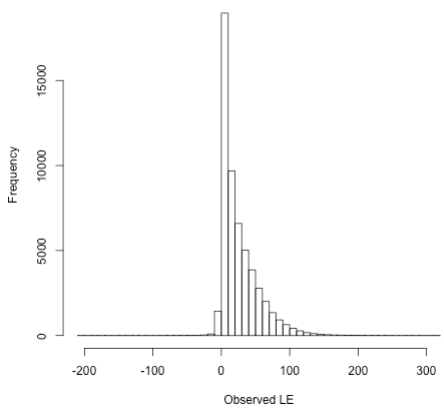
Three initial sets of trials were conducted with different combinations of training data: the GLASS data alone, the MODIS NBAR data alone, and both GLASS and MODIS NBAR included. The parameters of each ML model were tuned for snow and non-snow days for each of these training data combinations. Those parameters which minimized RMSE with the training and validation data sets were checked for overfitting with the corresponding test data sets.

Mapping over continental United States

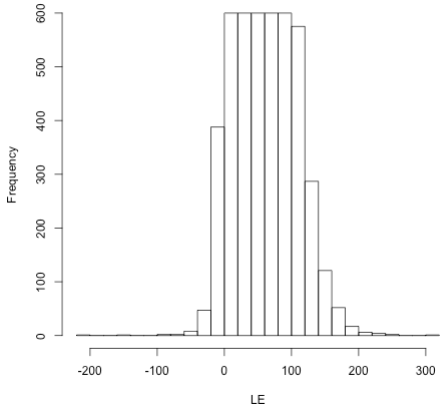
The BAGTREE algorithm used with both the GLASS and MODIS NBAR data sets as input usually produced the lowest error values with the flux tower site test data set. Therefore, this combination as optimized in the trials was used to construct monthly mean LE maps of the continental United States from 23.75º to 50.0º N and 127.5º to 62.5º W. Tests were also conducted including MODIS vegetation indices along with GLASS and MODIS NBAR data sets. Initial testing showed that inclusion of the VIs reduced the RMSE for the global station test data set by about 1 W/m^2 (see Results section below). Maps were then made with the combination of GLASS, MODIS NBAR, and MODIS VIs as input. Monthly mean data from the same subset region were also mapped for each of the comparison data sets, mapped onto the same geographic (latitude/ longitude) projection at 0.05 degree resolution.

Initial results of the BAGTREE and comparison data set mapping indicated the likelihood of a low LE bias in the eastern United States during the growing season in the BAGTREE data. In an attempt to address this, training data sets were generated with different weightings of the station-measured LE values. Histograms of the LE values in the original training data sets without and with vegetation index included in the input are shown in Figure 4-1. Both with and without VI as an input, low LE values of less than 50 W/m^2 are most represented in the training data set (Figures 4-1a and 4-1e). Four other training data sets were generated by excluding some of the points with lower LE values. Three of these were generated with vegetation index excluded from the training data, and one with vegetation index included. The training data sets without vegetation index included are one generated to produce a flat histogram of LE (Figure 4-1b), a "high weighted" data set including only one in ten of the points with LE under 100 W/m^2 included but all points with LE over 100 W/m^2 included (Figure 4-1c), and a "highest weighted" data set in which most of the points have LE > 50 W/m^2 (Figure 4-1d). A "highest weighted" training data set where most of the points have LE > 50 W/m^2 was also generated with vegetation index as an input variable (Figure 4-1f). Monthly mean LE maps were generated with the BAGTREE algorithm and each of the weighted training data sets.

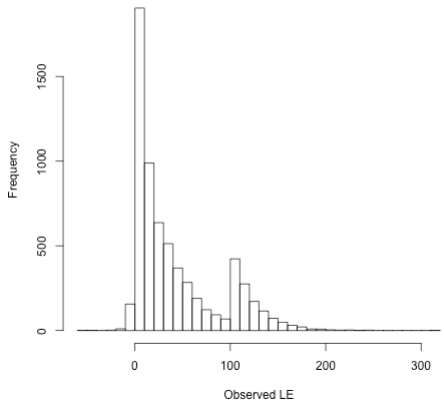
a)


Frequency
0
5000
10000
15000
-200
-100
0
100
200
300
Observed LE

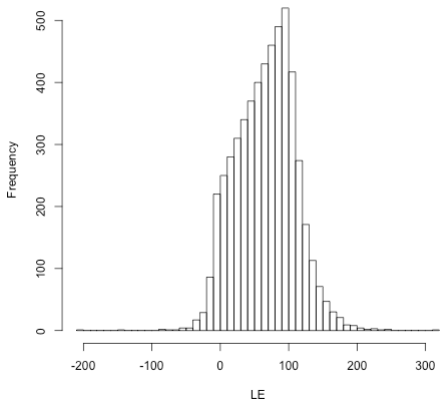
b)


Frequency
0
100
200
300
400
500
600
-200
-100
0
100
200
300
LE

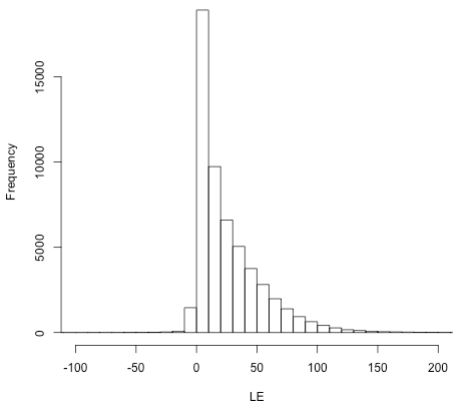
c)


Frequency
0
500
1000
1500
0
100
200
300
Observed LE

d)


Frequency
0
100
200
300
400
500
-200
-100
0
100
200
300
LE

e)


Frequency
0
5000
10000
15000
-100
-50
0
50
100
150
200
LE

f)

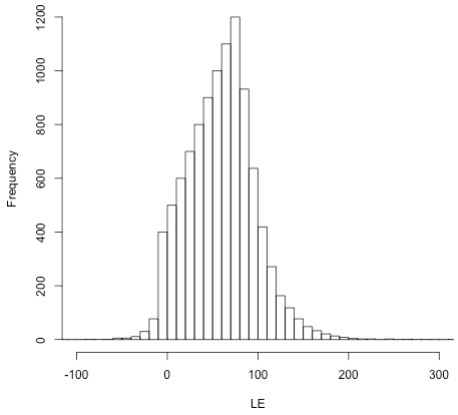

Frequency
0
200
400
600
800
1000
1200
-100
0
100
200
300
LE

Figure 4-1: Frequency of flux tower LE values occurring in each training data set. a) through d) are for training data without vegetation indices. a) Original. b) Flat histogram. c) High weight. d) Highest weight. e) and f) are for training data with vegetation indices. e) Original with VI. f) Highest weight, with VI.

In order to gain more insight into the discrepancies that were seen between the different BAGTREE generated maps and between all of the LE comparison data sets, monthly mean values were extracted from all of the LE maps at the nearest neighbor point to 23 Ameriflux stations. These monthly mean values were plotted against daily Ameriflux LE, since gaps in the Ameriflux data make it unfeasible to generate monthly mean Ameriflux LE. Even with daily Ameriflux LE, generalizations can be made about the quality of the fit to each monthly mean time series derived from the maps. Three sets of these monthly mean plots were made: one comparing the results of the BAGTREE algorithm and training data sets without vegetation index included, one comparing BAGTREE results excluding and including vegetation index in the training data, and one comparing the LE data from different sources.

Results

Tuning trials

Three sets of algorithm tuning trials were performed, using the GLASS data only, using the MCD43C4 data only, and using both the GLASS and MCD43C4 data. For each case, algorithm training was performed with data separated into snow and non-

snow with a threshold albedo value of 0.4. The snow and non-snow data were each divided using an 80% training/ 10% validation/ 10% test data split.

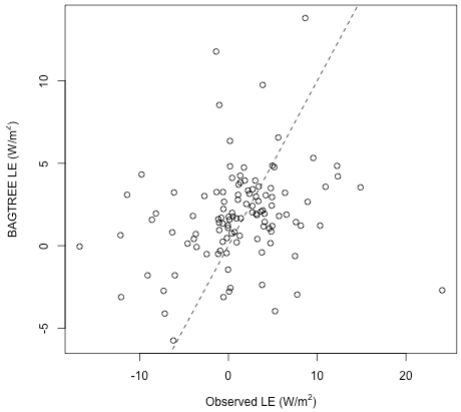
The statistical summary results of the snow cases using the GLASS data only are shown in Table 4-1. Validation RMSE values for all algorithms tested are below 4.2 W/m^2, and test RMSE values are below 5.7 W/m^2. However, R^2 values are very low. This is due to the compression of LE to low values for the data classified as snow, as illustrated in Figure 4-2, which shows the retrieved versus measured LE values for the test GLASS data snow cases for the BAGTREE (a), RKS (b), 2 layer NN (c), and 3 layer NN (d), respectively. Figures 4-1b and 4-1d also show that the ML algorithms sometimes produce outliers. The best performance was with the BAGTREE algorithm, and the RKS algorithm showed the weakest performance.

Table 4-1. Validation and test statistics for GLASS data only, snow cases (albedo > 0.4)

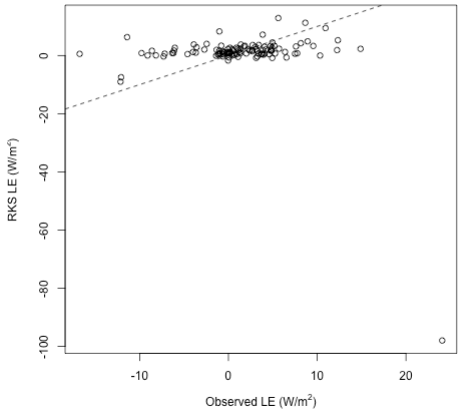
	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	3.958	0.201	-0.283	5.342	0.033	0.308
RKS	4.108	0.119	-0.164	5.358	0.015	-0.007
NN, 2 layer	4.135	0.108	-0.122	5.605	0.002	-0.007

NN, 3 layer	4.077	0.150	-0.317	5.618	0.001	0.309

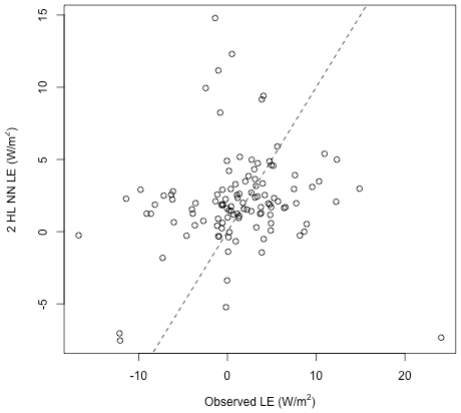
a)



b)



c)



d)

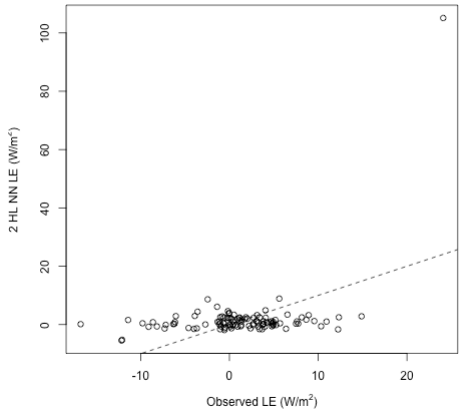


Figure 4-2: Derived versus observed LE for GLASS only snow test data set. a) BAGTREE algorithm. b) RKS algorithm. c) Two hidden layer neural network. d) Three hidden layer neural network. Dashed line is 1:1 line.

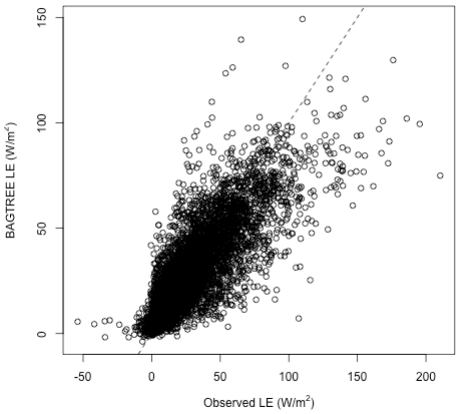
For the GLASS only non-snow cases using the full training data set, test RMSE values range from 15.727 W/m^2 to 18.808 W/m^2, and R^2 values range from 0.637 to 0.739. Test RMSE values exceed the validation RMSE values by about 1 W/m^2 or less and test R^2 values are slightly lower than validation R^2 values, indicating that some overfitting may be occurring, but the effects are slight. The best performing algorithm for the GLASS only non-snow case is the bagging tree (BAGTREE), and the weakest is the 2-layer NN. Summary statistics for the GLASS non-snow data are shown in Table 4-2, and plots of derived versus measured LE for the GLASS non-snow test data set are shown in Figure 4-3. In addition to the occurrence of outliers, it is also notable that most retrieved and observed values in the test data set are less than 100 W/m^2, and there is some indication of a low bias at high LE values.

Table 4-2. Validation and test statistics for GLASS data only, non-snow cases (albedo <= 0.4)

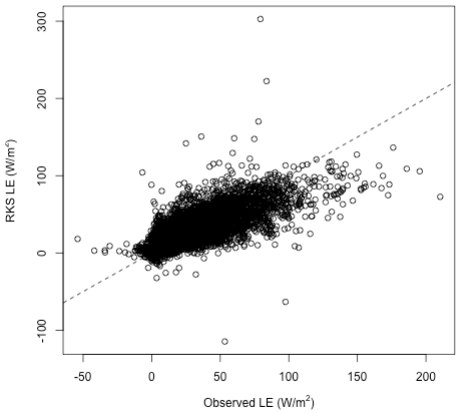
	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	15.374	0.681	-0.066	15.727	0.663	-0.076

RKS	16.536	0.632	-0.201	16.990	0.611	-0.118
NN, 2 layer	16.783	0.619	-0.056	18.808	0.527	-0.084
NN, 3 layer	16.161	0.647	-0.094	16.411	0.634	-0.171

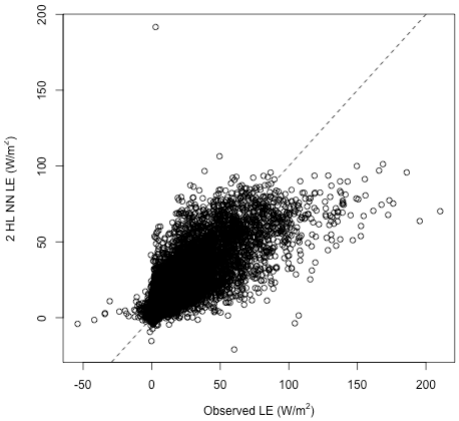
a)



b)



c)



d)

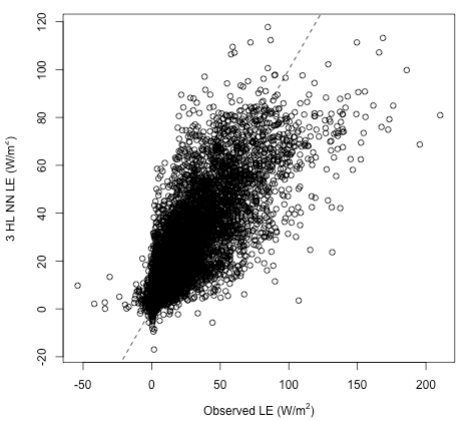


Figure 4-3: Derived versus observed LE for GLASS only non-snow test data set. a) BAGTREE algorithm. b) RKS algorithm. c) Two hidden layer neural network. d) Three hidden layer neural network. Dashed line is 1:1 line.

With the MCD43C4 data only, results are similar to those with the GLASS data only, except that algorithm performance is weaker for the snow case, with test RMSEs of about 5 to 6 W/m^2, and stronger for the non-snow case, with test RMSEs of about 16 W/m^2 or less for all algorithms. For the non-snow data, R^2 values for the MCD43C4 data were on the order of 0.1 greater than for the GLASS data. The strongest algorithm performance for the MCD43C4 data only was with BAGTREE, and the three hidden layer NN was the weakest. Summary statistics for the MCD43C4 data only are shown in Table 4-3, and plots of derived versus measured LE for the MCD43C4 only test data set are shown in Figure 4-4.

Table 4-3. Validation and test statistics for MCD43C4 data only

a) Snow data (albedo > 0.4)

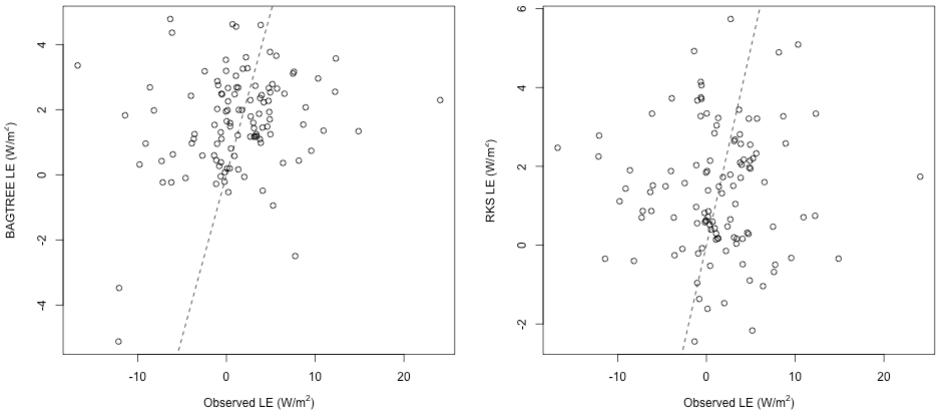
	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	5.819	0.077	-0.108	5.853	0.070	-0.215
RKS	5.706	0.111	-0.006	6.724	0.068	-0.040

NN, 2 layer	5.911	0.054	-0.414	6.272	0.000	0.394
NN, 3 layer	5.801	0.086	-0.160	6.048	0.011	0.217

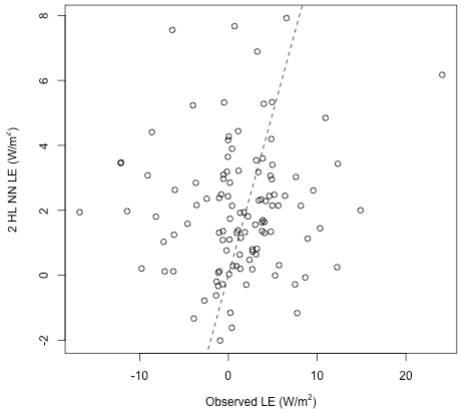
b) Non-snow data (albedo <= 0.4)

	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	13.584	0.741	0.355	13.632	0.739	0.337
RKS	14.788	0.694	0.207	14.891	0.690	0.038
NN, 2 layer	15.454	0.664	0.361	15.694	0.654	0.200
NN, 3 layer	15.016	0.683	-0.051	16.071	0.637	0.362

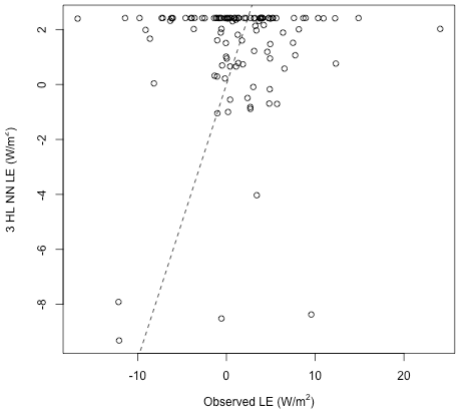
a) b)



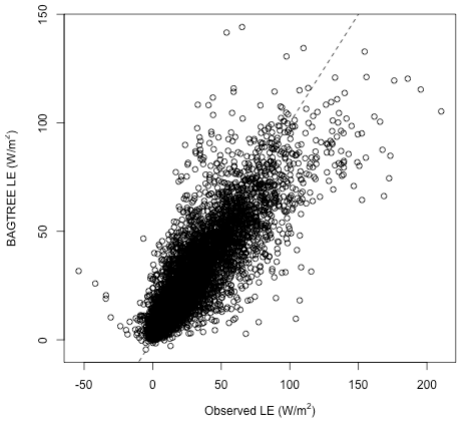
c)


2 HL NN LE (W/m²)
Observed LE (W/m²)

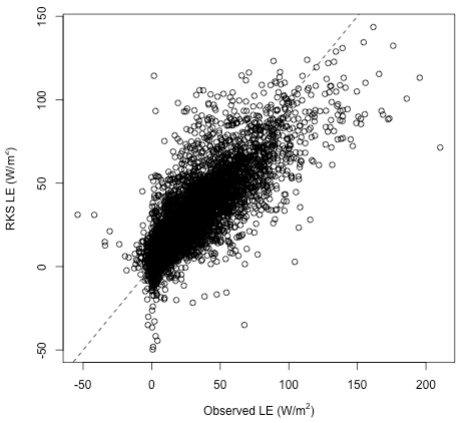
d)


3 HL NN LE (W/m²)
Observed LE (W/m²)

e)


BAGTREE LE (W/m²)
Observed LE (W/m²)

f)


RKS LE (W/m²)
Observed LE (W/m²)

g) h)

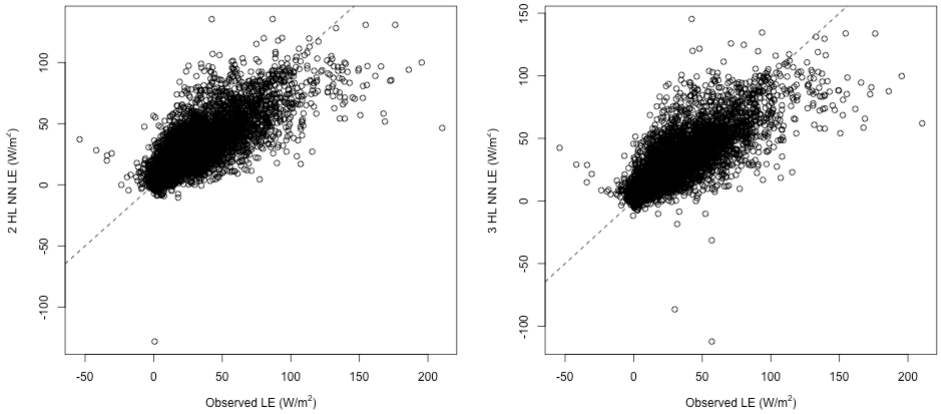


Figure 4-4: Derived versus observed LE for NBAR only test data sets. a) BAGTREE algorithm, snow. b) RKS algorithm, snow. c) Two hidden layer neural network, snow. d) Three hidden layer neural network, snow. e) BAGTREE algorithm, no snow. f) RKS algorithm, no snow. g) Two hidden layer neural network, no snow. h) Three hidden layer neural network, no snow. Dashed line is 1:1 line.

When both the GLASS and MCD43C4 NBAR data were used, results were similar to the other cases for the snow data, except that the 3-layer NN had a higher test RMSE of 9.378 W/m^2. R^2 values were again very low for all cases. For the non-snow data, test RMSE ranged from 12.376 to 14.089 W/m^2 and R^2 from 0.728 to 0.792. The BAGTREE algorithm demonstrated the best performance and the two-layer NN the weakest. Summary statistics for the GLASS and NBAR combined data are shown in Table 4-4, and plots of derived versus measured LE for these data in Figure 4-5.

Table 4-4. Validation and test statistics for GLASS and MCD43C4 data combined

a) Snow data (albedo > 0.4)

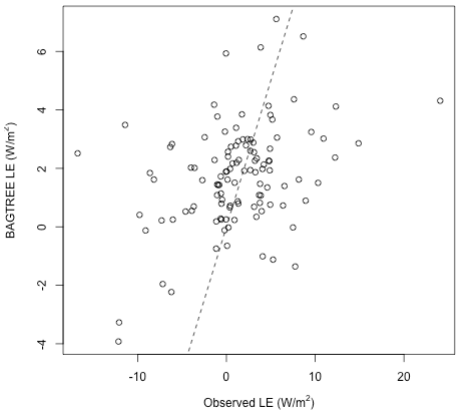
	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	5.702	0.113	-0.201	5.346	0.144	0.518
RKS	5.435	0.194	-0.322	6.010	0.032	0.359
NN, 2 layer	5.894	0.073	0.448	5.417	0.118	1.270
NN, 3 layer	5.715	0.117	-0.114	9.378	0.031	1.420

b) Non-snow data (albedo <= 0.4)

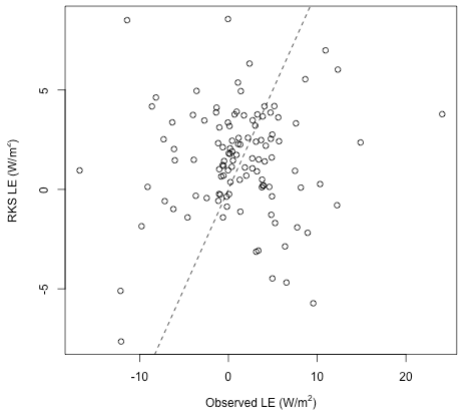
	Validation			Test		
Algorithm	RMSE (W/m^2)	R^2	Bias (W/m^2)	RMSE (W/m^2)	R^2	Bias (W/m^2)
BAGTREE	12.108	0.795	0.173	12.376	0.792	-0.011
RKS	13.086	0.759	0.249	13.469	0.752	0.061
NN, 2 layer	13.150	0.758	0.290	14.089	0.728	-0.509

NN, 3 layer	12.894	0.767	0.227	13.710	0.744	0.138

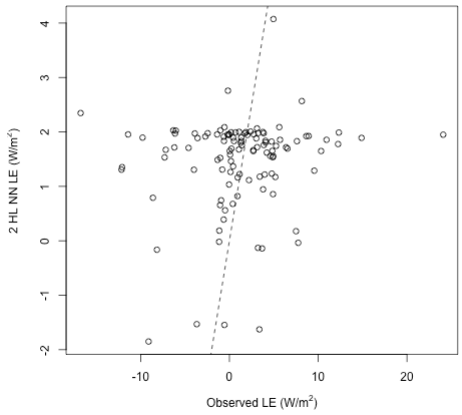
a)



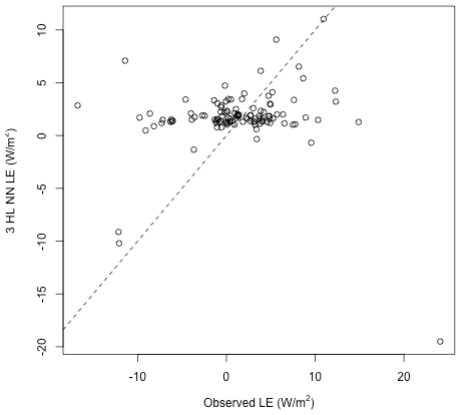
b)



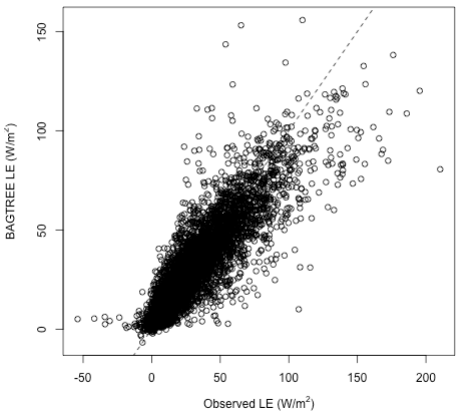
c)



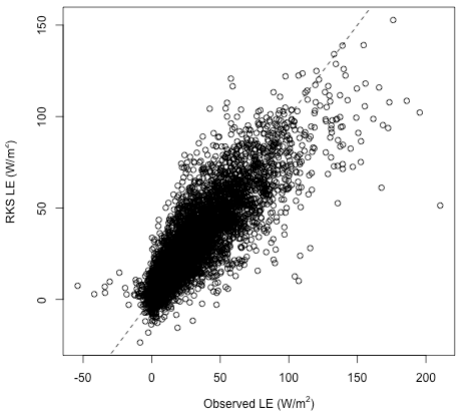
d)



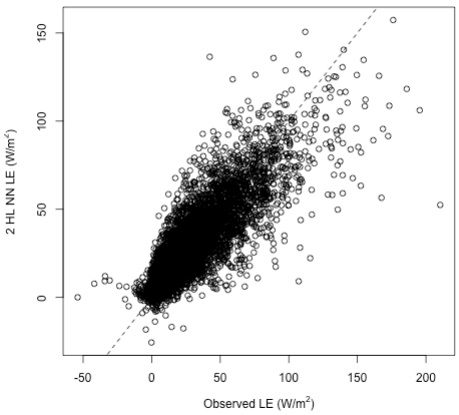
e)


BAGTREE LE (W/m²)
Observed LE (W/m²)
-50
0
50
100
150
200
0
50
100
150

f)


RKS LE (W/m²)
Observed LE (W/m²)
-50
0
50
100
150
200
0
50
100
150

g)


2 HL NN LE (W/m²)
Observed LE (W/m²)
-50
0
50
100
150
200
0
50
100
150

h)

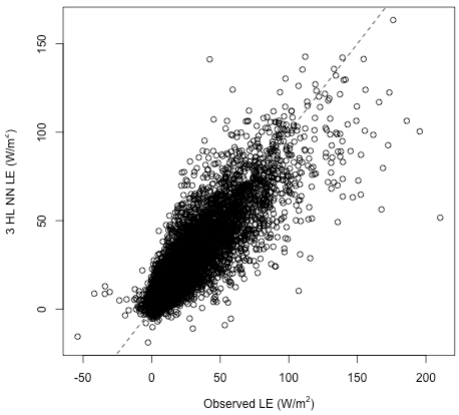

3 HL NN LE (W/m²)
Observed LE (W/m²)
-50
0
50
100
150
200
0
50
100
150

Figure 4-5: Derived versus observed LE for GLASS plus NBAR test data sets. a) BAGTREE algorithm, snow. b) RKS algorithm, snow. c) Two hidden layer neural network, snow. d) Three hidden layer neural network, snow. e) BAGTREE algorithm, no snow. f) RKS algorithm, no snow. g) Two hidden layer neural network, no snow. h) Three hidden layer neural network, no snow. Dashed line is 1:1 line.

Due to its stronger performance in the preceeding tests, the BAGTREE algorithm was selected for use in further testing. The BAGTREE algorithm was run with each of the training data sets with different weightings described in the Methods section: original, flat histogram, high weight, and highest weight without vegetation index, and original and highest weight with vegetation index. The validation and test statistics for these tests are reported in Table 4-5. These results indicate that on a global basis, the BAGTREE results most closely matched the flux tower measurements for the original training data set that included all available training data. Including vegetation index as an input reduced RMSE by about 1 W/m^2 for the original weighting and by over 2.5 W/m^2 for the highest weighting. The higher weighted training data sets also produced validation and test results with positive biases that are not present with the original training data set.

Table 4-5: Summary statistics for BAGTREE tests with differently weighted training data sets

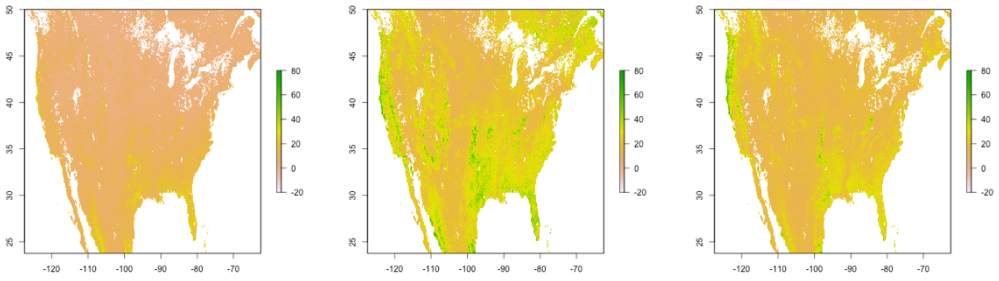
	Validation			Test		
	RMSE	R^2	Bias	RMSE	R^2	Bias
Original	12.131	0.794	0.187	12.368	0.792	-0.005
Flat histogram	17.835	0.640	9.015	17.793	0.648	8.829
High weight	15.888	0.689	4.329	15.971	0.692	3.946
Highest weight	17.976	0.628	10.367	17.983	0.635	10.270
Original with VI	11.243	0.824	-0.024	11.337	0.824	-0.095
Highest weight with VI	15.056	0.696	8.330	15.245	0.693	8.374

Mapping over the continental United States

Monthly mean BAGTREE LE values for each of the training data sets and monthly mean LE values from each of the comparison data sets are shown for January (Figure 4-6), April (Figure 4-7), July (Figure 4-8), and October (Figure 4-9). (Note that the scale on the figures is from -20 to 80 W/m^2 for January, -20 to 160 W/m^2 for April, -20 to 200 W/m^2 for July, and -20 to 100 W/m^2 for October.) Some patterns are immediately obvious from these figures. Other than the general patterns of higher LE in the eastern US and lower LE in the western US and higher LEs in July, at the

height of the growing season, compared to the other months, there is not much commonality between the data sets in general. The BAGTREE LEs tend to have lower LEs than any of the other products, a pattern which is especially notable in July (Figure 4-7). The BAGTREE results with training data sets weighted towards higher LEs (subfigures b, c, d, and f in each of Figure 4-6 through 4-9) contain higher LEs than the BAGTREE results with the original training data sets (subfigures a and e). The BAGTREE cases without VI in the training data sets (subfigures a through d) are mostly consistent in geographic pattern, but with higher or lower values overall. The same is true of the BAGTREE results with VI in the training data (subfigures e and f). The maps generated from the flat histogram training data set (subfigure b) show higher LE values in the April and October intermediate seasons than the other BAGTREE data sets. Those BAGTREE results with VI included in the training data have less contrast between the wetter eastern US and the drier western US than those without VI.

a) b) c)



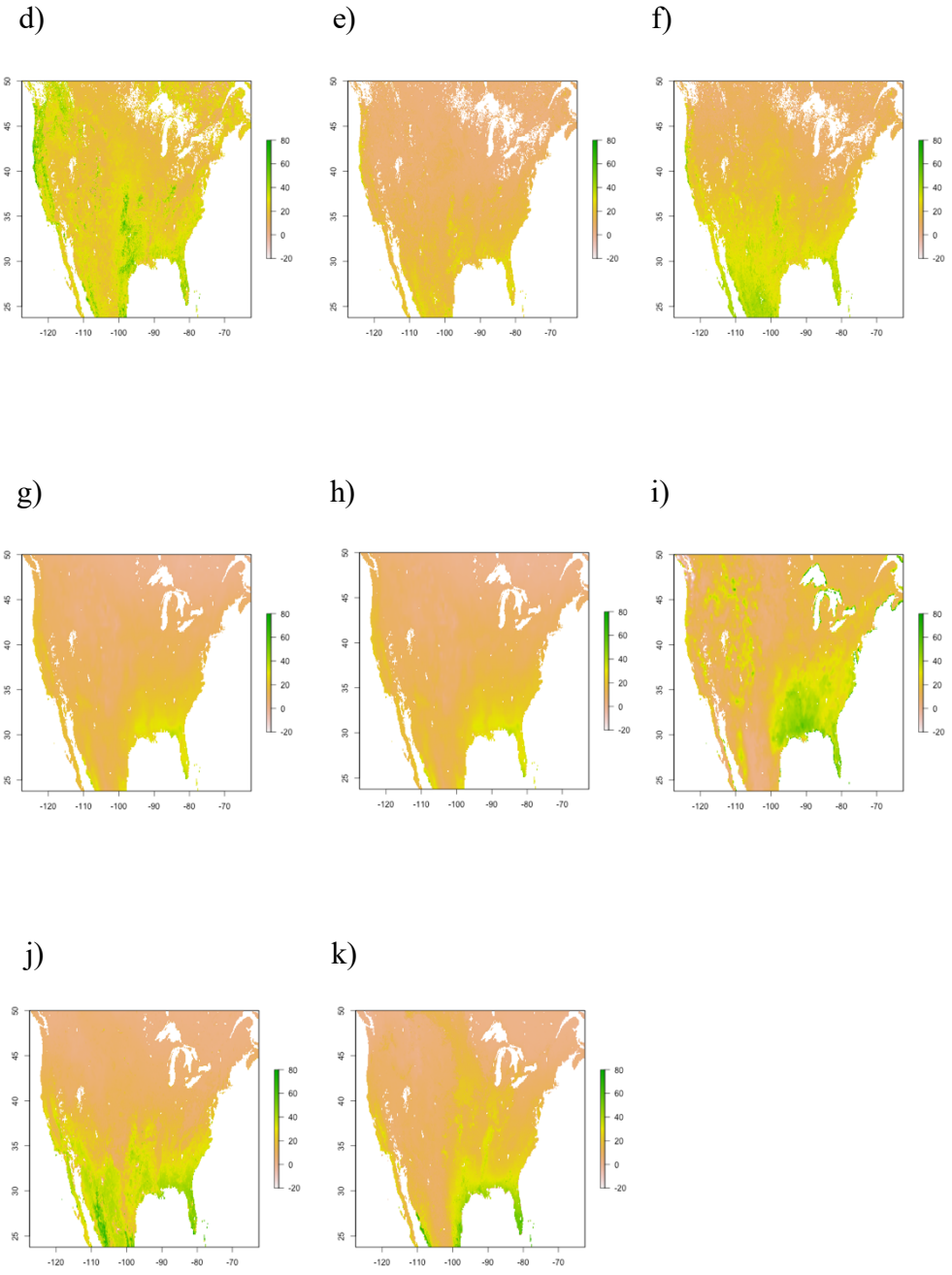
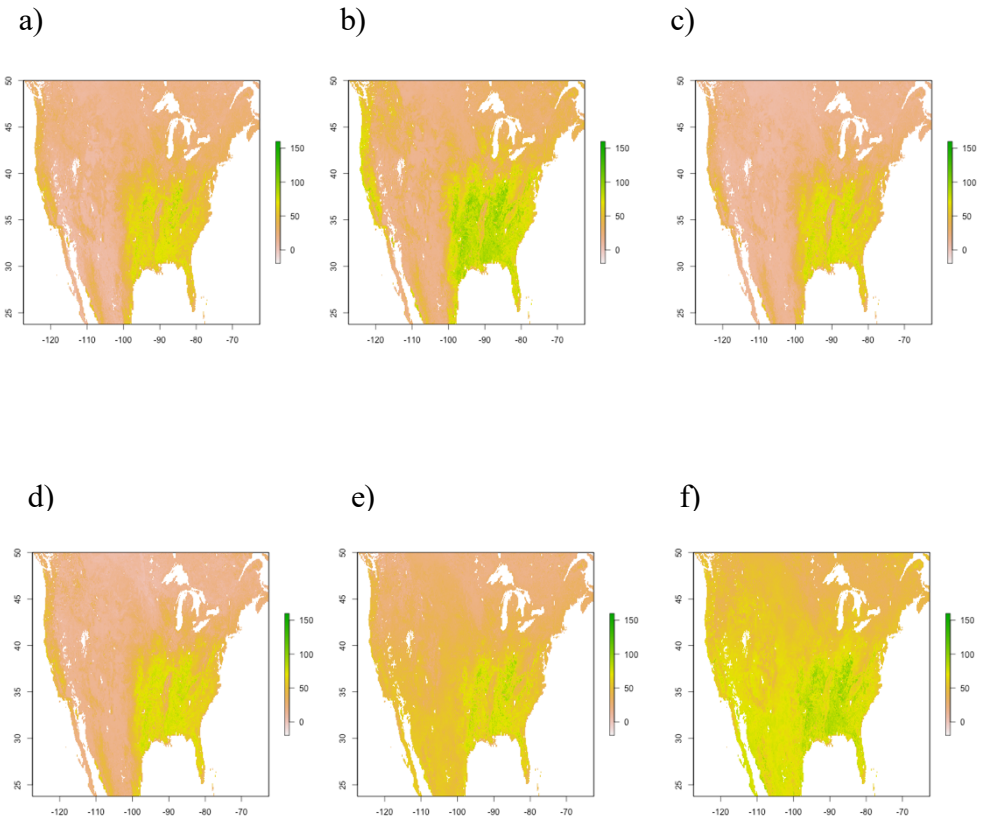


Figure 4-6: January monthly mean LE. a) through f) are BAGTREE algorithm results. a) Original training data set without VI. b) Flat histogram data set without VI. c) High weighted data set without VI. d) Highest weighted data set without VI. e) Original training data set with VI. f) Highest weighted training data set with VI. g) through k)

are comparison data sets. g) FluxCom remote sensing only. h) FluxCom with remote sensing and meteorology input. i) NARR. j) MODIS. k) GLASS.

When comparing the maps generated from LE data sources other than the BAGTREE results, other patterns become apparent. Usually, the North American Regional Reanalysis (NARR) LE values are higher than those of the other data sets, especially in the eastern US, and the MODIS LE values tend to be lower, more similar to those from the BAGTREE algorithm. The spatial patterns of the LE values differ significantly between data sets as well.



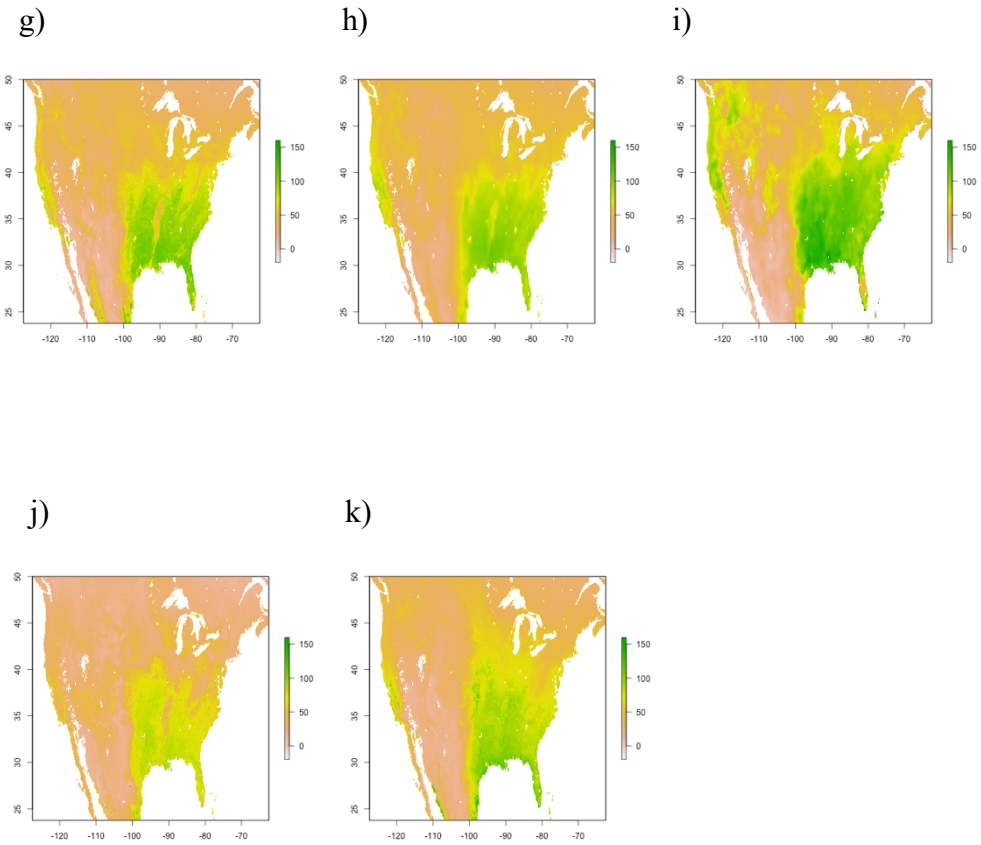
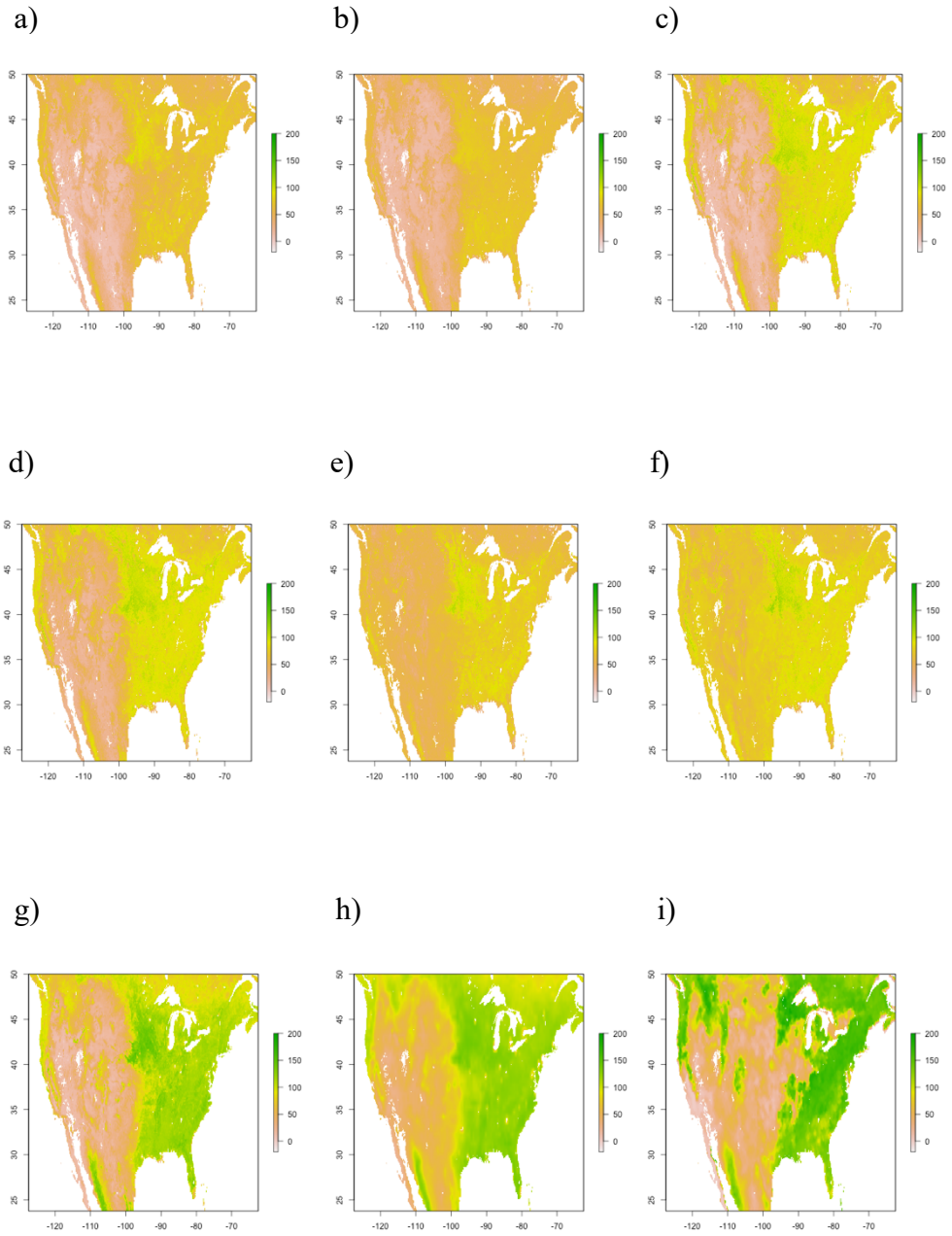


Figure 4-7: April monthly mean LE. a) through f) are BAGTREE algorithm results. a) Original training data set without VI. b) Flat histogram data set without VI. c) High weighted data set without VI. d) Highest weighted data set without VI. e) Original training data set with VI. f) Highest weighted training data set with VI. g) through k) are comparison data sets. g) FluxCom remote sensing only. h) FluxCom with remote sensing and meteorology input. i) NARR. j) MODIS. k) GLASS.

The July maps (Figure 4-8) are notable as they characterize the peak of the growing season in much of the US, and also a time of high drought intensity in the Midwestern US. Here the BAGTREE and MODIS results typically have low LE values overall,

while the NARR values are typically higher. However, there is a region just south and west of the Great Lakes where the NARR and MODIS LE values are both lower than those of the other data sets, including the BAGTREE.



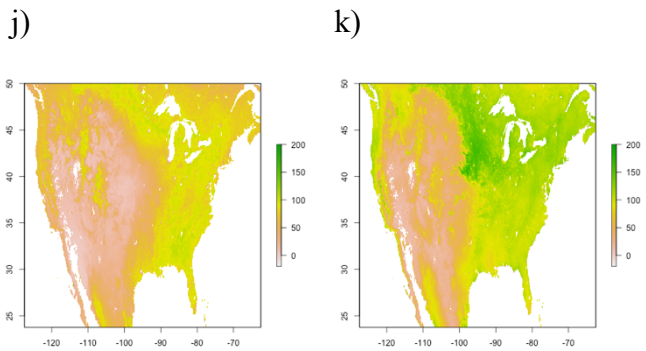
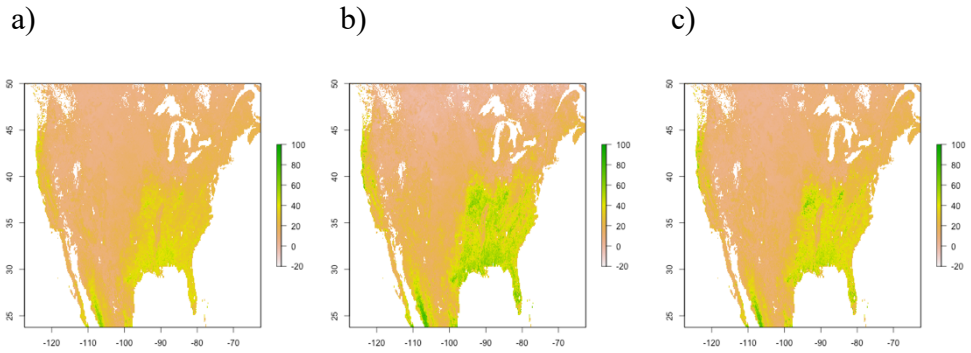


Figure 4-8: July monthly mean LE. a) through f) are BAGTREE algorithm results. a) Original training data set without VI. b) Flat histogram data set without VI. c) High weighted data set without VI. d) Highest weighted data set without VI. e) Original training data set with VI. f) Highest weighted training data set with VI. g) through k) are comparison data sets. g) FluxCom remote sensing only. h) FluxCom with remote sensing and meteorology input. i) NARR. j) MODIS. k) GLASS.



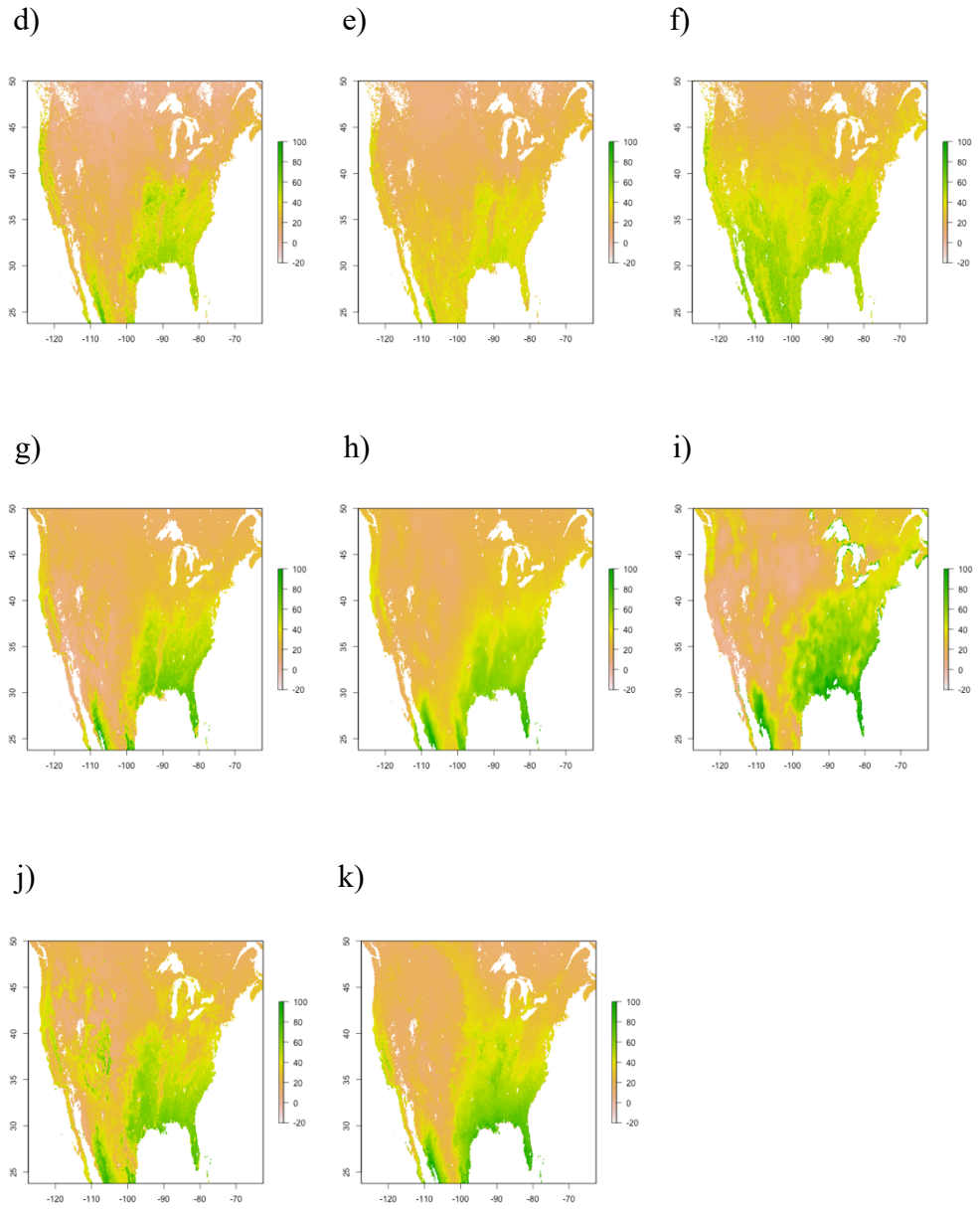


Figure 4-9: October monthly mean LE. a) through f) are BAGTREE algorithm results. a) Original training data set without VI. b) Flat histogram data set without VI. c) High weighted data set without VI. d) Highest weighted data set without VI. e) Original training data set with VI. f) Highest weighted training data set with VI. g)

through k) are comparison data sets. g) FluxCom remote sensing only. h) FluxCom with remote sensing and meteorology input. i) NARR. j) MODIS. k) GLASS.

Monthly mean comparison with flux towers

A selection of monthly mean mapped LE and daily flux tower LE plots are shown below to illustrate what was found in this analysis. Plotting LE obtained from the differently weighted training data sets without vegetation index in the input tended to show that the higher-weighted data sets matched the flux tower measurements more closely because those values tended to be higher overall. Examples of this phenomenon are shown in Figure 4-10. There were sometimes discrepancies in the timing of mid to late year drops in LE between the BAGTREE and flux tower measurements, as shown in Figure 4-11, with the BAGTREE apparently having a delayed detection of decreased LE. It is notable that in the case of the Missouri Ozark site, the original weighted training data produces a dropoff timing that more closely matches the Ameriflux observations. Figure 4-12 shows multiple cases where the "flatter" seasonal variation of the flat histogram trained results does not appear to match the temporal variation of the ground observations as well as the results with the other training data sets.

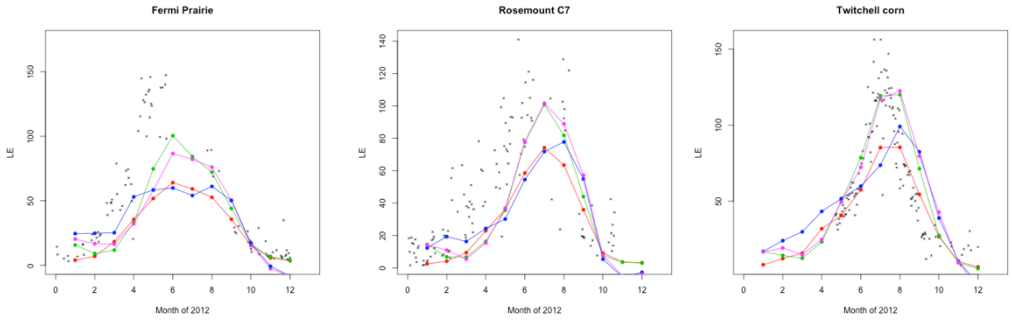


Figure 4-10: Monthly mean time series of bagtree LE from differently weighted training data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating closer match of higher-weighted training data sets to ground observations. Red: Original weighting. Blue: Flat histogram. Green: High weighting. Magenta: Highest weighting.

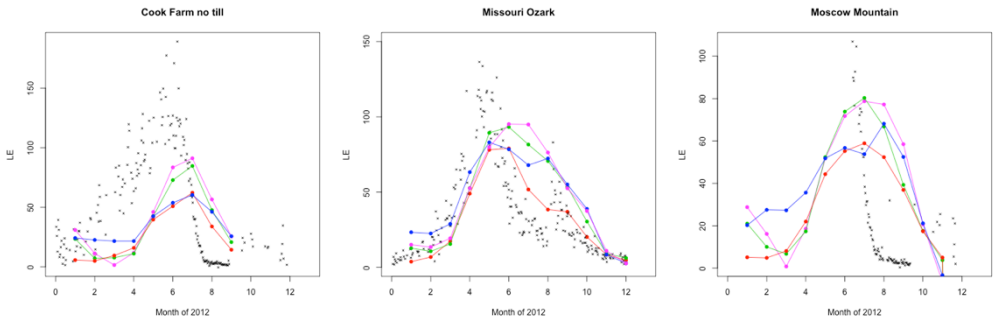


Figure 4-11: Monthly mean time series of bagtree LE from differently weighted training data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating delayed detection of decreasing LE. Red: Original weighting. Blue: Flat histogram. Green: High weighting. Magenta: Highest weighting.

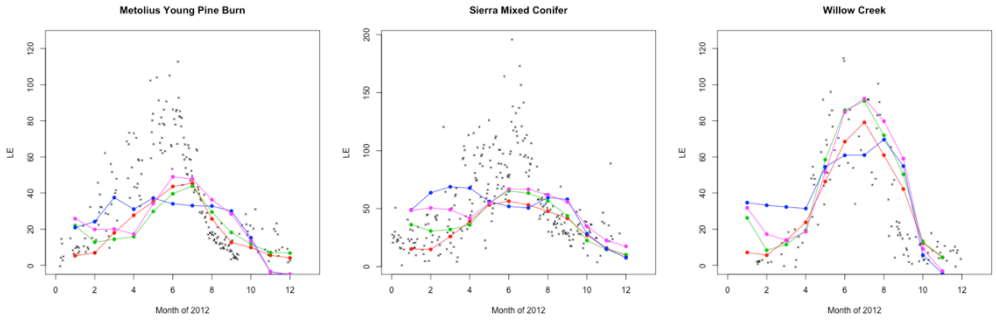


Figure 4-12: Monthly mean time series of bagtree LE from differently weighted training data sets (colored lines) and daily flux tower LE observations (black crosses),

illustrating weaker match of flat histogram training set results to temporal pattern of flux tower LE. Red: Original weighting. Blue: Flat histogram. Green: High weighting. Magenta: Highest weighting.

When comparing BAGTREE LEs derived with and without VIs as an input at different sites, certain patterns occurred. In several cases, including vegetation index made little difference unless high weighting was also applied (Figure 4-13). In other cases, including vegetation index appeared to reduce the sensitivity to drop-offs in LE (Figure 4-14). There were also sites in the western US where including vegetation index resulted in substantially higher LEs (Figure 4-15). The three sites shown in Figure 4-15 are all at about -105 to -106 degrees latitude. In these cases, the ground-measured LEs tended to be intermediate between the results with and without vegetation index as an input.

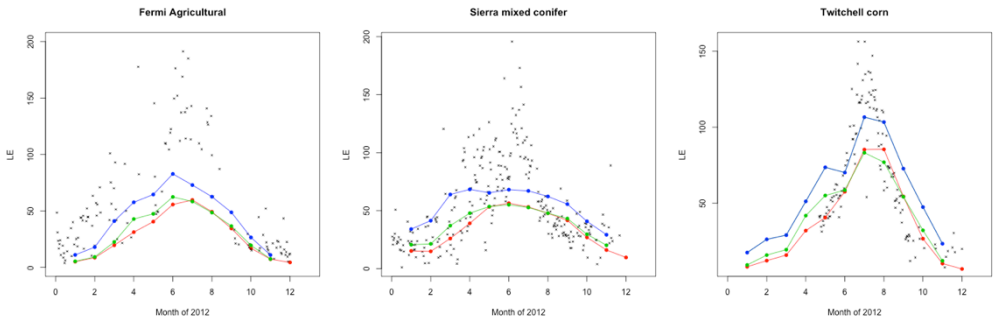


Figure 4-13: Monthly mean time series of bagtree LE trained with and without vegetation index as an input (colored lines) and daily flux tower LE observations (black crosses), illustrating cases where including vegetation index had little effect on

results. Red: Original weighting, no vegetation index input. Green: Original weighting, vegetation index input. Blue: Highest weighting, vegetation index input.

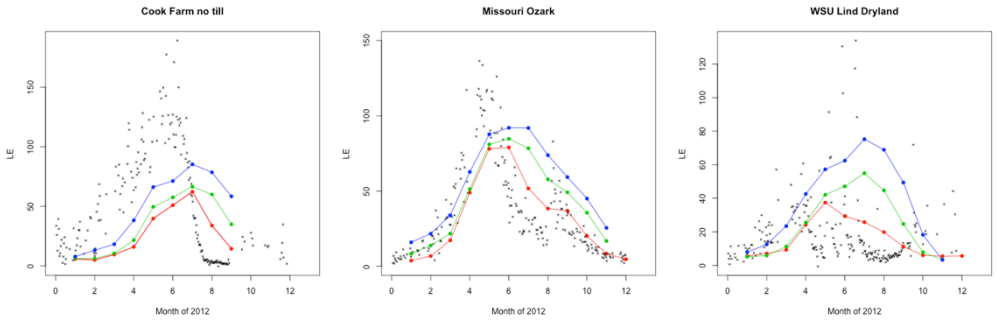


Figure 4-14: Monthly mean time series of bagtree LE trained with and without vegetation index as an input (colored lines) and daily flux tower LE observations (black crosses), illustrating cases where including vegetation index appears to reduce sensitivity to a dropoff in LE. Red: Original weighting, no vegetation index input. Green: Original weighting, vegetation index input. Blue: Highest weighting, vegetation index input.

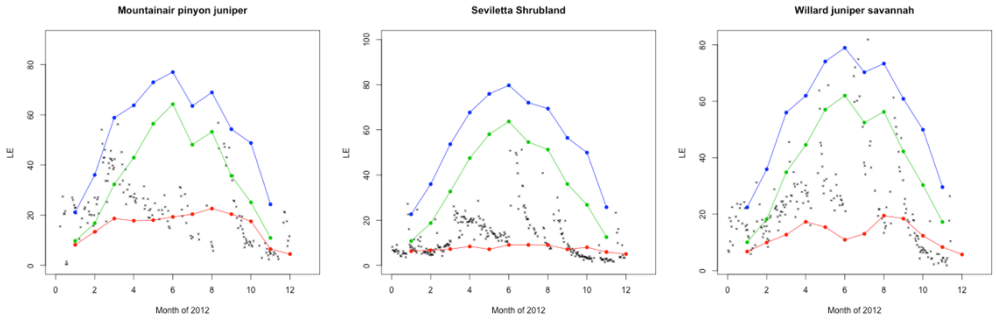


Figure 4-15: Monthly mean time series of bagtree LE trained with and without vegetation index as an input (colored lines) and daily flux tower LE observations (black crosses), illustrating cases where including vegetation index produces higher

LE values. Red: Original weighting, no vegetation index input. Green: Original weighting, vegetation index input. Blue: Highest weighting, vegetation index input.

The most common pattern observed when comparing monthly means from all map data sources to the Ameriflux LEs was that most of the data were relatively consistent with each other and with the Ameriflux LE, except for the NARR, which had a high bias (Figure 4-16). Exceptions occurred at the Cook Farm and Sierra Conifer sites, where NARR had the highest values but was also most consistent with the Ameriflux data (Figure 4-17). When comparing the BAGTREE data with vegetation index as an input to the other map data sets, the other data sets appeared to catch drops in LE more quickly than the BAGTREE data, but all of the other data had about the same lag (Figure 4-18). The MODIS data were a better match to an observed dropoff in the Missouri Ozark data than the other map data sets but were significantly biased low for the Fermi sites (Figure 4-19). Overall, none of the mapped data sets performed consistently better than the others.

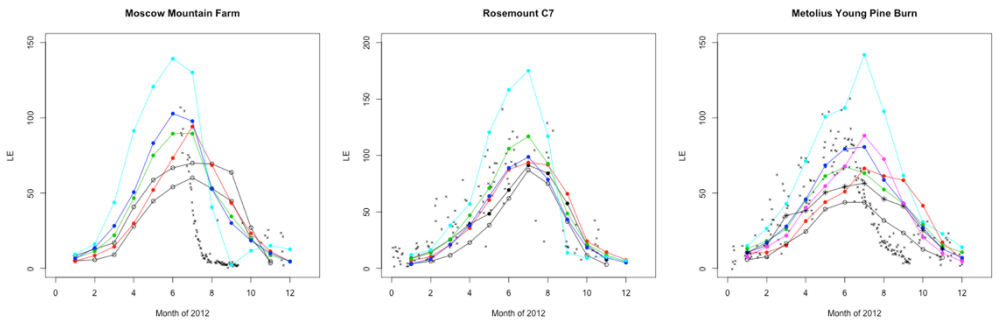


Figure 4-16: Monthly mean time series of LE from different map data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating cases where

NARR data are biased high relative to other map retrievals and flux tower observations. Black line, open circles: BAGTREE, with VI, original weight. Black line, closed circles: BAGTREE, with VI, highest weight. Red: MODIS. Green: FluxCom, remote sensing only. Blue: FluxCom, remote sensing and meteorology input. Cyan: NARR. Magenta: GLASS.

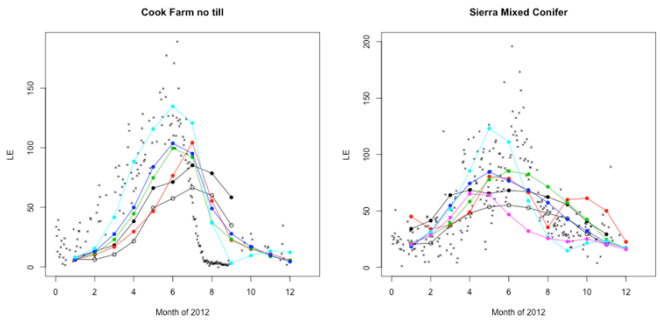


Figure 4-17: Monthly mean time series of LE from different map data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating cases where NARR data are higher than other map observations and closer to Ameriflux LEs. Black line, open circles: BAGTREE, with VI, original weight. Black line, closed circles: BAGTREE, with VI, highest weight. Red: MODIS. Green: FluxCom, remote sensing only. Blue: FluxCom, remote sensing and meteorology input. Cyan: NARR. Magenta: GLASS.

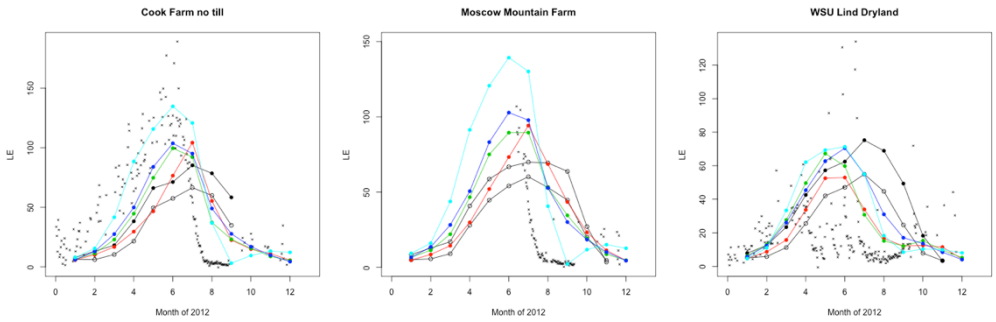


Figure 4-18: Monthly mean time series of LE from different map data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating differences in sensitivity to falling LEs. Black line, open circles: BAGTREE, with VI, original weight. Black line, closed circles: BAGTREE, with VI, highest weight. Red: MODIS. Green: FluxCom, remote sensing only. Blue: FluxCom, remote sensing and meteorology input. Cyan: NARR. Magenta: GLASS.

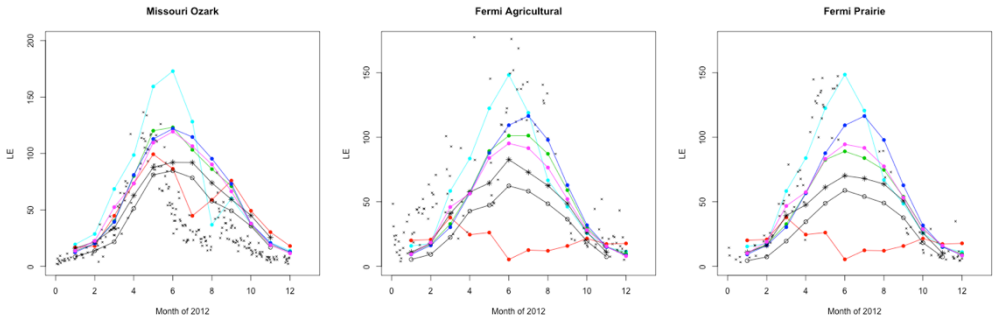


Figure 4-19: Monthly mean time series of LE from different map data sets (colored lines) and daily flux tower LE observations (black crosses), illustrating closer MODIS match to flux tower LEs at Missouri Ozark site and closer match of all other data sets at Fermi sites. Black line, open circles: BAGTREE, with VI, original weight. Black line, closed circles: BAGTREE, with VI, highest weight. Red: MODIS. Green:

FluxCom, remote sensing only. Blue: FluxCom, remote sensing and meteorology input. Cyan: NARR. Magenta: GLASS.

Discussion

This study has shown that the retrievals of LE using ML algorithms are quite sensitive to the composition of the training data set. In the case shown here, initial training with the BAGTREE algorithm produced a low bias at higher LE values (above 100 W/m^2) (Figures 4-3 to 4-5). This leads to the algorithm indicting relatively low LE values during the growing season in wet regions, as shown in Figure 4-8, because that is where the highest LE values occur. In an attempt to address this low bias, trials were made with higher LE values more represented in the training data set relative to lower values. Summary statistics indicated better performance globally for the original, full training data sets (Table 4-5), with lower RMSE and bias values closer to zero than those for the higher weighted training data sets. However, the higher weighting of the input data sets and the resulting high bias did not appear to resolve the persistent underestimation of LE in the eastern US during the growing season when compared to other data sets.

Monthly mean maps were made of the BAGTREE results for the continental United States using each of the differently weighted training data sets. Use of training data sets with relatively more observed LE values over 100 W/m^2 versus under 100 W/m^2

produced higher LE values during the growing season in the eastern US than training with all available data (Figures 4-6 through 4-9). The closer match to the comparison data sets shows that it is likely that the higher weighted training data are producing more realistic results during the growing season in the eastern United States, despite the weaker performance against global validation and test data according to summary statistics (Table 4-5). Yet it appears that the low bias during the growing season in the eastern US persisted even with the higher-weighted input data sets. The smaller range of values of the results of the algorithm trained with the original training data set may be inaccurate, but at the same time results at lower LE values are probably more accurate. However, it is clear that the seasonal cycle pattern often produced with the flat histogram training data set, with relatively higher values during the transition seasons and relatively lower values at the peak of the growing season, is not a good representation of the seasonal cycle as measured at individual sites (Figure 4-12).

Including vegetation index as an input resulted in maps with decreased contrast between the wetter eastern US and drier western US. It is unclear whether this represents an improvement or degradation in performance. When individual sites in the western US where the BAGTREE algorithm produced higher LEs with vegetation index as an input were checked against ground station measured LE, the ground station LEs were intermediate between the BAGTREE LEs with and without vegetation index as an input (Figure 4-15). The vegetation index trials show that including input data that are expected to be highly correlated with other input data, as the vegetation indices are expected to be correlated with NBAR, can still produce

significant differences in the output results. It is unclear why this occurs, and counterintuitive as vegetation index maps tend to demonstrate contrast between a greener eastern US and a less green western US. Further sensitivity testing could be performed by generating BAGTREE LEs with test vegetation index data to see how much LE is affected by differences in vegetation index. However, determining why any given pattern occurs is difficult when using machine learning methods due to the opaque nature of the algorithms.

The higher weighted training data set produced higher LE values during the growing season in the eastern US than the full training data set, but those LE values were still lower than those from the comparison data sets. This could be because even the higher weighted training data sets contain only a few data days with observed LE values over 150 W/m^2, resulting in little representation of these higher LE values in the training data. It is likely that this limited range of the training data is producing a limited range to the output. This could be tested by obtaining and using more training data with daily flux tower station measured LE values greater than 150 W/m^2.

The comparison data sets considered here differ from each other about as much as they do from the BAGTREE results, except for the apparent low bias of the BAGTREE results at the height of the growing season in the eastern US. This lack of consensus between the data sets especially holds for the Midwestern US in July 2012, indicating the lack of a consistent drought signal that might be expected for that time and location. There is also not a consistent result when checking which of the data sources produces a more accurate time series when checked against individual ground

sites. When the BAGTREE results without vegetation index are considered alone, the highest weighted training data sets usually produce the closest match to the flux tower measurements during the growing season, because they demonstrate less low bias (Figure 4-10). All of the BAGTREE products show a lag in sensitivity to drops in LE during the growing season (Figure 4-11). However, the higher weighted BAGTREE results sometimes show less sensitivity to these drops (Missouri Ozark site in Figure 4-11).

When considering all of the comparison data sets included in this analysis, no single data set more consistently matches the ground-based observations. The NARR data set LEs were usually higher than those from the other map products, and often demonstrated a high bias relative to the ground observations (Figure 4-16), but sometimes the higher NARR values were a closer match to the flux tower LEs (Figure 4-17). The MODIS data exhibited a closer match to falling LEs in one case, but a significant low bias in others (Figure 4-19).

Overall, the bagging tree mapping of the continental United States presented here has substantial room for improvement. It is quite possible that including more variables in the input data sets or segregating the input data by land cover type could improve performance. The FluxCom data set production (Jung et al., 2018) included more input variables and performed separate trainings of ML algorithms by surface type, and produced what appears to be a better representation of LE during the growing season in the eastern US (as shown in the Results section above). However, the

changes in results observed here with different input data set weighting indicates the possibility that all ML results, including FluxCom, have similar sensitivity to input data set characteristics.

Conclusions

This study has shown that the results of ML algorithms for the retrieval of evapotranspiration have a significant dependency on the properties of the data set used for training. Because of this dependency, these algorithms should be used with caution. Overall global statistics are not necessarily a good indication of performance in particular situations. The higher weighted input data sets produced a better representation of the growing season in the eastern US despite their higher RMSEs, lower correlation coefficients, and higher biases on a global basis, although the higher weighted inputs still produced low biases during the eastern US growing season. Adding vegetation index to the set of input variables resulted in an improvement in global statistics that did not indicate clear improvements in performance when regional patterns were examined.

Tests including NBAR as input in all cases but with and without vegetation indices as input indicate that seemingly duplicative input variables may still have a significant effect on outcomes. When vegetation indices were included, the overall geographic pattern of the results changed, with decreased contrast between the wetter eastern US and drier western US. Including vegetation indices improved global statistics but the results for some individual sites went from underestimation without VIs to overestimation with VIs. Further experimentation could shed more light on the effect

of including or excluding VIs. The NDVI, EVI, and other data fields for individual days should be examined before use to generate maps using the BAGTREE algorithm for multiple sample days, to see if patterns apparent in the input data are also occurring in the output. Sensitivity testing could also be performed by systematically altering the VI data fields before generating the LE maps.

The inconsistencies of the comparison data sets examined with each other and with the BAGTREE results suggest caution in the use of any LE mapping data. More intercomparison work could yield more insight into the patterns of variation between different data sources and the causes of those differences. It would be useful to examine results for high precipitation as well as drought years.

Direct diagnosis of the reasons for patterns in machine learning output is difficult due to the "black box" nature of the algorithms. For application to the evapotranspiration problem, further investigation of sensitivity to input data set characteristics for different input data types, regions, and climates outside of the midlatitude range represented by the continental US could be useful. In addition, the effects of input data set selection on drought sensitivity and regional and global trend analysis would be of interest.

The most significant finding of this study is the sensitivity of the mapped LEs to the weighting of the training data sets. It also appears that limitations in the range of values represented in an input data set imposes similar limitations on the output.

Further investigation of the properties of LE retrievals generated with input data sets with different properties is highly recommended.

Chapter 5: Conclusion

The work carried out in the preceding studies addressed several issues in the remote sensing of evapotranspiration. As discussed in the Introduction, measurement of evapotranspiration is desirable because ET is an important component of the land surface water and energy balances, but ET can be difficult to measure. Remote sensing of ET has the capacity to provide information on regional and global scales that are impractical to reach via ground-based measurements. Here we have evaluated the utility of a number of simple regression algorithms for this purpose, assessed the capacity of a range of machine learning methods to address this problem, and applied machine learning methods to produce monthly mean ET maps for the continental United States for the drought year of 2012. The results show that globally-averaged statistics at individual sites indicate performance comparable to that of other methods from both regression formula and machine learning algorithms, and that selection of the best machine learning method and refinements to ground data use can produce improved results, with validation and test RMSEs of around 12 W/m^2 on a daily basis. However, good performance on global statistical measures does not mean that the results are reliable at all places and times. This is especially clear from the results of the third part of this study, the regional mapping, which showed a pattern of likely underestimation of LE during the growing season in the eastern United States for all training data cases considered. Much further work will be required before consistent and reliable retrievals of LE are regularly available on a global basis.

The first section of this work (Chapter 2) was an evaluation of twelve simple regression formulas for determination of LE from remote sensing. These sorts of formulas are often used when data availability or computational power are limited. It was found that the more complex formulas with more input data variables tended to perform better than the others, regardless of the specific form of the regression formula. This was most true when using the original published coefficients for these formulas, as re-tuning with training data drawn from the same global data sets that were used for testing produced more improvement in RMSE and bias for the simpler formulas. However, R^2 values were not much improved by this tuning in most cases.

Tests were also done to evaluate algorithm performance for different surface types. For most surface types, performance was similar to the global case. The most notable differences were a low bias for cropland and wetland types, and higher RMSE for wetland. In the case of wetlands, it is likely that LE is higher than indicated by the vegetation indices used in the regression formulas, since in the extreme case open water has a low vegetation index and a high LE. Irrigation may be producing similar, though less pronounced, effects for croplands. Four of the algorithms were re-tuned with training data with the same surface type only. Testing with these re- tuned algorithms improved performance for croplands and wetlands.

The results of the study of simple regression formulas indicate some practices that users of these formulas should follow. It is best to use a formula which makes as much use as possible of the available data. Especially if a formula with a small

number of inputs is used, it is advisable to perform tuning to the particular data set being used. Results for wetland and cropland areas should be used with greater caution, and separate tuning for those surface types is recommended if possible. Due to the study described in Chapter 2, the utility and limitations of simple regression formulas for retrieval of LE from remote sensing are better understood and can be used with better results.

Evaluation of machine learning algorithms for obtaining LE was carried out in Chapter 3 of this work. Ten algorithms of five different types were evaluated. Some of the algorithms produced good results with a small training data set of about 8000 data points but were too demanding of computational resources for tuning to be practical with a large training data set of about 70,000 data points. The algorithms that were shown to have the best combination of accuracy and computational efficiency where the bootstrap aggregation (bagging) tree, the random kernel (RKS), and the 2 and 3 hidden layer neural networks. These algorithms all produced RMSEs of at least 3 W/m^2 lower when trained with the larger training data set than with the smaller training data set.

One advantage of the machine learning algorithms is that any combination of variables can be used as the input data set. Tests with different combinations of variables showed that using more input variables can produce lower RMSE and higher R^2 values than using smaller numbers of input variables, although NBAR appeared to have significant redundancy with the other non-radiation variables.

Computational demand was usually not significantly affected by using more input variables. Therefore, for the remainder of the study all available input variables (downward shortwave radiation, PAR, albedo, NDVI, EVI, LAI, and FAPAR) were used. All of these variables were obtained from MODIS or GLASS data sets, showing that the machine learning algorithms can be trained, tested, and used with all remote sensing data except for the ground- based LE values used for training and validation.

Using all available input variables and the large training data set, the RMSE of the best algorithms when checked with a test data set was 19.91 W/m^2 for the bagging tree, 20.94 W/m^2 for the 3 hidden layer neural network, and 22.22 W/m^2 for the RKS. This represents good performance relative to the simpler regression formulas tested in Chapter 2, with the additional advantage of less dependence on ground-based data.

Further evaluation of the machine learning algorithms was performed by testing the globally-tuned algorithms with test data of individual surface types, and then tuning to the individual surface types. The results of this testing showed weaker performance for wetland and cropland sites, and stronger performance for evergreen, grassland, savannah, and shrub sites. Unlike the simple regression formulas, tuning to individual surface types did not result in significant improvements to algorithm performance. It is recommended that, in general, machine learning algorithms be trained with all available data rather than restricting to particular input variables or surface types. Ideally, experiments should be made with different combinations of input variables to test the effect of their inclusion or exclusion.

The third study in this work (Chapter 4) is an investigation of the utility of machine learning algorithms for generating a time series of regional LE maps. After some further tuning of the algorithms, GLASS and MODIS data representing the continental United States during the drought year of 2012 were processed using the bootstrap aggregation regression tree (BAGTREE) algorithm, then compared to five other LE maps of the continental United States, two using other machine learning methods, one from the North American Regional Reanalysis, one from a standard MODIS product, and one from a GLASS ET product.

Some preliminary refinements were conducted before applying the machine learning technique to the continental US data. Tests were conducted using GLASS radiation, albedo, LAI, and FAPAR data only as input, using MODIS NBAR only as input, and using both the GLASS and MODIS NBAR data. The BAGTREE, RKS, and 2 and 3 hidden layer NN methods were all tested, but the best results were obtained with the BAGTREE method. Snow and non- snow data were separated according to a threshold albedo of 0.4, and those data points where the energy balance closure adjustment was greater than 0.4 of net radiation were discarded. With these adjustments, non-snow RMSE values of about 16-18 W/m^2 were obtained for the GLASS data alone, about 13-16 W/m^2 for the NBAR data alone, and 12-14 W/m^2 for the GLASS and MODIS NBAR data combined. This represents additional improvement in global algorithm performance over the results described in Chapter 3. Based on these results, monthly mean continental US maps were produced with the

BAGTREE algorithm and the GLASS and MODIS NBAR data as input for the year 2012. Maps were also produced using MODIS vegetation indices as input along with the MODIS NBAR and GLASS data.

The most notable result from the initial maps generated using the BAGTREE algorithm was that a low bias at high LE values resulted in a weaker signal of the growing season in the eastern US than in the comparison data sets. In an attempt to address this bias, the BAGTREE algorithm was retrained with training data sets containing a higher proportion of observed LE values over 100 W/m^2 than the original training data set. This resulted in increased RMSE and bias and lower R^2 when tested with the original test data set but produced a closer match to the comparison data sets for the growing season in the eastern US. These results show that ML algorithms can have significant sensitivity to the characteristics of the input training data set, leading to results that are more or less accurate under different circumstances. Comparisons to monthly mean LE maps for the year 2012 from other data sources showed little consensus between the data sets. Comparisons with ground measurements showed that no single data source was consistently the most accurate.

The differences between the maps generated when vegetation index was included as an input in addition to NBAR versus those generated with NBAR only raise significant questions. It was not expected that including vegetation indices would result in significant differences due to the probable redundancy in information content between NBAR and the vegetation indices. When examining results from sites where

LE obtained with vegetation index is very different from LE obtained without, the tower LE results tend to be intermediate between the two. As a result, it is not clear whether the results with or without vegetation index as an input are more accurate. Comparison of input vegetation index and NBAR fields to output maps may shed some further light on this issue.

The maps generated using the BAGTREE algorithm and from other sources indicate the need for extreme caution in the use and interpretation of LE maps, especially those generated through machine learning techniques. Good performance in global statistical terms does not guarantee good performance at all places and times. Machine learning generated maps have significant sensitivity to the characteristics of the data sets used for training. In addition, LE maps generated by different methods were not shown to converge on a consensus pattern, and evaluation of those maps against ground station data did not identify any of the retrievals as clearly superior to the others.

In order to resolve these discrepancies between LE retrievals, more comparison to ground-based data would be desirable, especially in the regions where the discrepancies are the most pronounced. Flux towers should be selected that are known to produce high-quality data and are not in areas of high spatial heterogeneity. Ground-based LE measurements that are independent in location or time from the original training, validation, and test data sets would be especially useful. Daily LE values during the growing season in both the eastern and western US should be

examined from both the ground-based measurements and the ML data, to identify times where agreement is good and where it is poor. Possibilities such as systematic underestimation on days where LE is especially high or when precipitation occurs should be investigated. If LE retrievals on days with especially high LE are often underestimates, retraining with a data set where high LE conditions are more represented could improve the results.

This work has produced refinements in simple formula and machine learning techniques used for remote sensing of evapotranspiration. It has shown that machine learning techniques can be used to derive maps on a continental scale, but also that the reliability of these maps is questionable. It has also demonstrated the potential for machine learning to detect drought signals in a few cases. However, the results of the mapping study also indicate that detection of growing season and drought trends are likely to be influenced by the characteristics of the training data set used. These results have implications for the potential operational use of these regression techniques and also for future research efforts.

The work performed here used data acquired between the years of 2000 and 2015, but generation of the data sets used for input in these studies is ongoing. The accuracy of the results obtained here is comparable to those of currently operational LE retrievals such as the MODIS MOD16 product. Our results show that a range of implementations of operational regression retrieval of LE are possible. If only a small number of input variables are available or if computational power is limited, simple

regression formulas of the type investigated in Chapter 2 can be used. Some of these formulas require meteorological data which could be acquired from stations, reanalysis, or other remote sensing products. Use of reanalysis or remote sensing for this purpose would make retrieval on regional and global scales possible, although the mapping study done here indicates that the quality of those retrievals, like those from all other methods evaluated, would be open to question.

If data for a large number of input variables and sufficient computing resources are available, use of machine learning has advantages over simple formulas. Any combination of input variables may be used with machine learning algorithms. Once the algorithms are trained, retrieval of LE is usually fast and can be expected to perform well on global statistical measurements. To obtain the best possible accuracy under the most different conditions, the quality, size and breadth of the training data set should first be maximized before any operational use. The characteristics of the training data sets should then be evaluated and potentially adjusted due to the sensitivity of the results to the input data set used. It is also necessary to be aware that performance may vary widely regionally or seasonally, so further testing over extended areas is recommended. An improved network of validation sites representing locations with high spatial homogeneity across a range of climate and ecosystem types would be especially useful for regional coarse resolution product evaluation.

Since any combination of input data can be used with machine learning algorithms, it is theoretically possible to obtain LE from a vast number of combinations of input data. However, the relationship between LE and the different input variables is likely to vary in strength and type. Some implementations of machine learning algorithms can be used to indicate which of the input variables have the most effect on the output once the algorithm is trained. Use of this kind of test could provide an indication of which available variables are most closely related to LE. These tests could be done for data from different kinds of instruments, such as soil moisture sensors and sounders, and for data from geostationary as well as polar-orbiting platforms. These sorts of experiments, along with other testing similar to what has been done in this work, would be useful as efforts towards improving retrievals of LE from remote sensing. In turn, these improved retrievals should be more useful as indicators of water consumption, crop and ecosystem health, and hydrological cycle trends at scales ranging from local to global.

Bibliography

Allen, R.G., Food and Agriculture Organization of the United Nations (Eds.), 1998. Crop evapotranspiration: guidelines for computing crop water requirements, FAO irrigation and drainage paper. Food and Agriculture Organization of the United Nations, Rome.

Allen, S.T., Reba, M.L., Edwards, B.L., Keim, R.F., 2017. Evaporation and the subcanopy energy environment in a flooded forest. Hydrol. Process. 31, 2860–2871. https://doi.org/10.1002/hyp.11227

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., Dehnavi, S., 2017. Temperature-Vegetation-soil Moisture Dryness Index (TVMDI). Remote Sens. Environ. 197, 1–14. https://doi.org/10.1016/j.rse.2017.05.026

Anderson, M., 1997. A Two-Source Time-Integrated Model for Estimating Surface Fluxes Using Thermal Infrared Remote Sensing. Remote Sens. Environ. 60, 195–216. https://doi.org/10.1016/S0034-4257(96)00215-5

Anderson, M.C., Zolin, C.A., Sentelhas, P.C., Hain, C.R., Semmens, K., Tugrul Yilmaz, M., Gao, F., Otkin, J.A., Tetrault, R., 2016. The Evaporative Stress Index as an indicator of agricultural drought in Brazil: An assessment based on crop yield impacts. Remote Sens. Environ. 174, 82–99. https://doi.org/10.1016/j.rse.2015.11.034

Badgley, G., Fisher, J.B., Jiménez, C., Tu, K.P., Vinukollu, R., 2015. On Uncertainty in Global Terrestrial Evapotranspiration Estimates from Choice of Input Forcing Datasets*. J. Hydrometeorol. 16, 1449–1455. https://doi.org/10.1175/JHM-D-14-0040.1

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, 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, 2415–2434. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Bastiaanssen, W.G.M., Pelgrum, H., Wang, J., Ma, Y., Moreno, J.F., Roerink, G.J., van der Wal, T., 1998. A remote sensing surface energy balance algorithm for land (SEBAL). J. Hydrol. 212–213, 213–229. https://doi.org/10.1016/S0022-1694(98)00254-6

Beigt, D., Piccolo, M.C., Perillo, G.M.E., 2008. Surface heat budget of an estuarine tidal flat (Bah´ıa Blanca Estuary, Argentina). Cienc Mar 34, 1–15.

Brown, M.E., Pinzon, J.E., Didan, K., Morisette, J.T., Tucker, C.J., 2006. Evaluation of the consistency of long-term NDVI time series derived from AVHRR,SPOT-vegetation, SeaWiFS, MODIS, and Landsat ETM+ sensors. IEEE Trans. Geosci. Remote Sens. 44, 1787–1793. https://doi.org/10.1109/TGRS.2005.860205

C. Schaaf, Z.W., 2015a. MCD43A3 MODIS/Terra+Aqua BRDF/Albedo Daily L3 Global - 500m V006. https://doi.org/10.5067/MODIS/MCD43A3.006

C. Schaaf, Z.W., 2015b. MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted RefDaily L3 Global - 500m V006. https://doi.org/10.5067/MODIS/MCD43A4.006

Camps-Valls, G., Tuia, D., Gomez-Chova, L., Jimenez, S., Malo, J. (Eds.), 2011. Remote sensing image processing. Morgan & Claypool.

Carter, C., Liang, S., 2019a. Application of machine learning to mapping of evapotranspiration over the continental United States. Prep.

Carter, C., Liang, S., 2019b. Evaluation of ten machine learning methods for estimating terrestrial evapotranspiration from remote sensing. Int. J. Appl. Earth Obs. Geoinformation 78, 86–92. https://doi.org/10.1016/j.jag.2019.01.020

Carter, C., Liang, S., 2018. Comprehensive evaluation of empirical algorithms for estimating land surface evapotranspiration. Agric. For. Meteorol. 256–257, 334–345.

Chen, Y., Xia, J., Liang, S., Feng, J., Fisher, J.B., Li, Xin, Li, Xianglan, Liu, S., Ma, Z., Miyata, A., Mu, Q., Sun, L., Tang, J., Wang, K., Wen, J., Xue, Y., Yu, G., Zha, T., Zhang, L., Zhang, Q., Zhao, T., Zhao, L., Yuan, W., 2014. Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China. Remote Sens. Environ. 140, 279–293. https://doi.org/10.1016/j.rse.2013.08.045

Choudhury, B., Ahmed, N., Idso, S., Reginato, R., Daughtry, C., 1994. Relations Between Evaporation Coefficients and Vegetation Indexes Studied by Model Simulations. Remote Sens. Environ. 50, 1–17. https://doi.org/10.1016/0034-4257(94)90090-6

De Keersmaecker, W., Lhermitte, S., Tits, L., Honnay, O., Somers, B., Coppin, P., 2015. A model quantifying global vegetation resistance and resilience to short-term climate anomalies and their relationship with vegetation cover: Global vegetation resistance and resilience. Glob. Ecol. Biogeogr. 24, 539–548. https://doi.org/10.1111/geb.12279

Deo, R.C., Samui, P., 2017. Forecasting Evaporative Loss by Least-Square Support-Vector Regression and Evaluation with Genetic Programming, Gaussian Process, and Minimax Probability Machine Regression: Case Study of Brisbane City. J. Hydrol. Eng. 22, 05017003. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001506

Deo, R.C., Samui, P., Kim, D., 2016. Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. Stoch. Environ. Res. Risk Assess. 30, 1769–1784. https://doi.org/10.1007/s00477-015-1153-y

Dou, X., Yang, Y., 2018a. Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. Comput. Electron. Agric. 148, 95–106. https://doi.org/10.1016/j.compag.2018.03.010

Dou, X., Yang, Y., 2018b. Modeling Evapotranspiration Response to Climatic Forcings Using Data-Driven Techniques in Grassland Ecosystems. Adv. Meteorol. 2018, 1–18. https://doi.org/10.1155/2018/1824317

Feng, Y., Burian, S., Pardyjak, E., 2018. Observation and Estimation of Evapotranspiration from an Irrigated Green Roof in a Rain-Scarce Environment. Water 10, 262. https://doi.org/10.3390/w10030262

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

Glenn, E.P., Nagler, P.L., Huete, A.R., 2010. Vegetation Index Methods for Estimating Evapotranspiration by Remote Sensing. Surv. Geophys. 31, 531–555. https://doi.org/10.1007/s10712-010-9102-2

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science + Business Media, LLC.

Helman, D., Givati, A., Lensky, I.M., 2015. Annual evapotranspiration retrieved from satellite vegetation indices for the eastern Mediterranean at 250 m spatial resolution. Atmospheric Chem. Phys. 15, 12567–12579. https://doi.org/10.5194/acp-15-12567-2015

Holden, Z.A., Swanson, A., Luce, C.H., Jolly, W.M., Maneta, M., Oyler, J.W., Warren, D.A., Parsons, R., Affleck, D., 2018. Decreasing fire season precipitation increased recent western US forest wildfire activity. Proc. Natl. Acad. Sci. 115, E8349–E8357. https://doi.org/10.1073/pnas.1802316115

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

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., Fisher, J.B., Jung, M., Kanamitsu, M., Reichle, R.H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., Wang, K., 2011. Global intercomparison of 12 land surface heat flux estimates. J. Geophys. Res. 116. https://doi.org/10.1029/2010JD014545

Joiner, J., Yoshida, Y., Anderson, M., Holmes, T., Hain, C., Reichle, R., Koster, R., Middleton, E., Zeng, F.-W., 2018. Global relationships among traditional reflectance vegetation indices (NDVI and NDII), evapotranspiration (ET), and soil moisture variability on weekly timescales. Remote Sens. Environ. 219, 339–352. https://doi.org/10.1016/j.rse.2018.10.020

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., Reichstein, M., 2018. The FLUXCOM ensemble of global land-atmosphere energy fluxes. Submitt. Sci. Data.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S.I., Sheffield, J., Goulden, M.L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A.J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B.E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A.D., 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. https://doi.org/10.1038/nature09396

K. Didan, 2015. MOD13A1 MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MOD13A1.006
Kaheil, Y.H., Rosero, E., Gill, M.K., McKee, M., Bastidas, L.A., 2008. Downscaling and Forecasting of Evapotranspiration Using a Synthetic Model of Wavelets and Support Vector Machines. IEEE Trans. Geosci. Remote Sens. 46, 2692–2707. https://doi.org/10.1109/TGRS.2008.919819
Kalma, J.D., McVicar, T.R., McCabe, M.F., 2008. Estimating Land Surface Evaporation: A Review of Methods Using Remotely Sensed Surface Temperature Data. Surv. Geophys. 29, 421–469. https://doi.org/10.1007/s10712-008-9037-z
Kamble, B., Irmak, A., Hubbard, K., 2013. Estimating Crop Coefficients Using Remote Sensing-Based Vegetation Index. Remote Sens. 5, 1588–1602. https://doi.org/10.3390/rs5041588
Ke, Y., Im, J., Park, S., Gong, H., 2017. Spatiotemporal downscaling approaches for monitoring 8-day 30 m actual evapotranspiration. ISPRS J. Photogramm. Remote Sens. 126, 79–93. https://doi.org/10.1016/j.isprsjprs.2017.02.006
Ke, Y., Im, J., Park, S., Gong, H., 2016. Downscaling of MODIS One Kilometer Evapotranspiration Using Landsat-8 Data and Machine Learning Approaches. Remote Sens. 8, 215. https://doi.org/10.3390/rs8030215
Lazaro-Gredilla, M., Titsias, M.K., 2011. Variation heteroschedastic Gaussian process regression. 28th Int. Conf. Mach. Learn. ICML 2011.
Lazaro-Gredilla, M., Titsias, M.K., Verrelst, J., Camps-Valls, G., 2014. Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes. IEEE Geosci. Remote Sens. Lett. 11, 838–842. https://doi.org/10.1109/LGRS.2013.2279695
Li, S., Wang, G., Sun, S., Chen, H., Bai, P., Zhou, S., Huang, Y., Wang, J., Deng, P., 2018. Assessment of Multi-Source Evapotranspiration Products over China Using Eddy Covariance Observations. Remote Sens. 10, 1692. https://doi.org/10.3390/rs10111692
Liang, S., Zhang, X., Xiao, Z., Cheng, J., Liu, Q., Zhao, X., 2014. Global LAnd Surface Satellite (GLASS) Products Algorithms, Validation and Analysis.
Liang, S., Zhao, X., Liu, S., Yuan, W., Cheng, X., Xiao, Z., Zhang, X., Liu, Q., Cheng, J., Tang, H., Qu, Yonghua, Bo, Y., Qu, Ying, Ren, H., Yu, K., Townshend, J., 2013. A long-term Global LAnd Surface Satellite (GLASS) data-set for environmental studies. Int. J. Digit. Earth 6, 5–33. https://doi.org/10.1080/17538947.2013.805262
Liu, Q., Wang, L., Qu, Y., Liu, N., Liu, S., Tang, H., Liang, S., 2013. Preliminary evaluation of the long-term GLASS albedo product. Int. J. Digit. Earth 6, 69–95. https://doi.org/10.1080/17538947.2013.804601
Long, D., Singh, V.P., 2012. A Two-source Trapezoid Model for Evapotranspiration (TTME) from satellite imagery. Remote Sens. Environ. 121, 370–388. https://doi.org/10.1016/j.rse.2012.02.015
Lu, X., Zhuang, Q., 2010. Evaluating evapotranspiration and water-use efficiency of terrestrial ecosystems in the conterminous United States using MODIS and AmeriFlux data. Remote Sens. Environ. 114, 1924–1939. https://doi.org/10.1016/j.rse.2010.04.001

Malone, S.L., Staudhammer, C.L., Loescher, H.W., Olivas, P., Oberbauer, S.F., Ryan, M.G., Schedlbauer, J., Starr, G., 2014. Seasonal patterns in energy partitioning of two freshwater marsh ecosystems in the Florida Everglades: ENERGY DYNAMICS IN EVERGLADES ECOSYSTEMS. J. Geophys. Res. Biogeosciences 119, 1487–1505. https://doi.org/10.1002/2014JG002700

Mao, Y., Wang, K., 2017. Comparison of evapotranspiration estimates based on the surface water balance, modified Penman-Monteith model, and reanalysis data sets for continental China: Terrestrial Evapotranspiration in China. J. Geophys. Res. Atmospheres. https://doi.org/10.1002/2016JD026065

Meng, L., Dong, T., Zhang, W., 2016. Drought monitoring using an Integrated Drought Condition Index (IDCI) derived from multi-sensor remote sensing data. Nat. Hazards 80, 1135–1152. https://doi.org/10.1007/s11069-015-2014-1

Merlin, O., Chirouze, J., Olioso, A., Jarlan, L., Chehbouni, G., Boulet, G., 2014. An image-based four-source surface energy balance model to estimate crop evapotranspiration from solar reflectance/thermal emission data (SEB-4S). Agric. For. Meteorol. 184, 188–203. https://doi.org/10.1016/j.agrformet.2013.10.002

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., 2006. North American Regional Reanalysis. Bull. Am. Meteorol. Soc. 87, 343–360. https://doi.org/10.1175/BAMS-87-3-343

Monteith, J., 1965. Evaporation and environment. Symp. Soc. Exp. Biol. 19, 205–234.

Moorhead, J., Marek, G., Colaizzi, P., Gowda, P., Evett, S., Brauer, D., Marek, T., Porter, D., 2017. Evaluation of Sensible Heat Flux and Evapotranspiration Estimates Using a Surface Layer Scintillometer and a Large Weighing Lysimeter. Sensors 17, 2350. https://doi.org/10.3390/s17102350

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. https://doi.org/10.1016/j.rse.2007.04.015

Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115, 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019

Mueller, B., Seneviratne, S.I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J.B., Guo, Z., Jung, M., Maignan, F., McCabe, M.F., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Teuling, A.J., Wang, K., Wood, E.F., Zhang, Y., 2011. Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations: GLOBAL LAND EVAPOTRANSPIRATION DATASETS. Geophys. Res. Lett. 38, n/a-n/a. https://doi.org/10.1029/2010GL046230

Muller, M., 2018. Cape Town's drought: don't blame climate change. Nature 559, 174–176. https://doi.org/10.1038/d41586-018-05649-1

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT Press.

Pandey, P.K., Nyori, T., Pandey, V., 2017. Estimation of reference evapotranspiration using data driven techniques under limited data conditions. Model. Earth Syst. Environ. 3, 1449–1461. https://doi.org/10.1007/s40808-017-0367-z
Park, S., Im, J., Jang, E., Rhee, J., 2016. Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. Agric. For. Meteorol. 216, 157–169. https://doi.org/10.1016/j.agrformet.2015.10.011
Peng, J., Loew, A., Merlin, O., Verhoest, N.E.C., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture: Downscale Satellite-Based Soil Moisture. Rev. Geophys. 55, 341–366. https://doi.org/10.1002/2016RG000543
Penman, H.L., 1948. Natural Evporation from Open Water, Bare Soil and Grass. R. Soc. 193, 120–145.
Pérez-Suay, A., Amorós-López, J., Gómez-Chova, L., Laparra, V., Muñoz-Marí, J., Camps-Valls, G., 2017. Randomized kernels for large scale Earth observation applications. Remote Sens. Environ. 202, 54–63. https://doi.org/10.1016/j.rse.2017.02.009
Price, J.C., 1990. Using spatial context in satellite data to infer regional scale evapotranspiration. IEEE Trans. Geosci. Remote Sens. 28, 940–948. https://doi.org/10.1109/36.58983
Priestly, C., Taylor, R., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Mon. Weather Rev. 100, 81–92.
Qu, Y., Liu, Q., Liang, S., Wang, L., Liu, N., Liu, S., 2014. Direct-Estimation Algorithm for Mapping Daily Land-Surface Broadband Albedo From MODIS Data. IEEE Trans. Geosci. Remote Sens. 52, 907–919. https://doi.org/10.1109/TGRS.2013.2245670
R. Myneni, Y.K., 2015. MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MOD15A2H.006
Rahimi, A., Recht, B., 2009. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning., in: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 1313–1320.
Rhee, J., Im, J., 2017. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. Agric. For. Meteorol. 237–238, 105–122. https://doi.org/10.1016/j.agrformet.2017.02.011
Roundy, J.K., Santanello, J.A., 2017. Utility of Satellite Remote Sensing for Land–Atmosphere Coupling and Drought Metrics. J. Hydrometeorol. 18, 863–877. https://doi.org/10.1175/JHM-D-16-0171.1
Runkle, B.R.K., Wille, C., Gazovic, M., Wilmking, M., Kutzbach, L., 2014. The surface energy balance and its drivers in a boreal peatland fen of northwestern Russia. J. Hydrol. 511, 359–373. https://doi.org/10.1016/j.jhydrol.2014.01.056
Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N.W., Clark, D.B., Dankers, R., Eisner, S., Fekete, B.M., Colón-González, F.J., Gosling, S.N., Kim, H.,

Liu, X., Masaki, Y., Portmann, F.T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., Kabat, P., 2014. Multimodel assessment of water scarcity under climate change. Proc. Natl. Acad. Sci. 111, 3245–3250. https://doi.org/10.1073/pnas.1222460110

Seddon, A.W.R., Macias-Fauria, M., Long, P.R., Benz, D., Willis, K.J., 2016. Sensitivity of global terrestrial ecosystems to climate variability. Nature 531, 229–232. https://doi.org/10.1038/nature16986

Shi, H., Xiao, Z., Liang, S., Zhang, X., 2016. Consistent estimation of multiple parameters from MODIS top of atmosphere reflectance data using a coupled soil-canopy-atmosphere radiative transfer model. Remote Sens. Environ. 184, 40–57. https://doi.org/10.1016/j.rse.2016.06.008

Smola, A.J., Scholkopf, B., n.d. A tutorial on support vector regression. Stat. Comput. 14, 199–222.

Sörensson, A.A., Ruscica, R.C., 2018. Intercomparison and Uncertainty Assessment of Nine Evapotranspiration Estimates Over South America. Water Resour. Res. 54, 2891–2908. https://doi.org/10.1002/2017WR021682

Steven, M.D., Malthus, T.J., Baret, F., Xu, H., Chopping, M.J., 2003. Intercalibration of vegetation indices from different sensor systems. Remote Sens. Environ. 88, 412–422. https://doi.org/10.1016/j.rse.2003.08.010

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P.H.S., Cipolla, R., 2006. Multivariate relevance vector machines for tracking. Comput. Vis.- ECCV 2006 Pt 3 Proc., Lecture Notes in Computer Science 3953, 124–138.

Upreti, H., Ojha, C.S.P., 2018. Evaluation of the Vapor Pressure Models in the Estimation of Actual Vapor Pressure and Evapotranspiration. J. Irrig. Drain. Eng. 144, 05018007. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001346

Vorosmarty, C.J., 2000. Global Water Resources: Vulnerability from Climate Change and Population Growth. Science 289, 284–288. https://doi.org/10.1126/science.289.5477.284

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability: GLOBAL TERRESTRIAL EVAPOTRANSPIRATION. Rev. Geophys. 50. https://doi.org/10.1029/2011RG000373

Wang, K., Dickinson, R.E., Wild, M., Liang, S., 2010a. Evidence for decadal variation in global terrestrial evapotranspiration between 1982 and 2002: 1. Model development. J. Geophys. Res. 115. https://doi.org/10.1029/2009JD013671

Wang, K., Dickinson, R.E., Wild, M., Liang, S., 2010b. Evidence for decadal variation in global terrestrial evapotranspiration between 1982 and 2002: 2. Results. J. Geophys. Res. 115. https://doi.org/10.1029/2010JD013847

Wang, K., Liang, S., 2009. Estimation of Daytime Net Radiation from Shortwave Radiation Measurements and Meteorological Observations. J. Appl. Meteorol. Climatol. 48, 634–643. https://doi.org/10.1175/2008JAMC1959.1

Wang, K., Liang, S., 2008. An Improved Method for Estimating Global Evapotranspiration Based on Satellite Determination of Surface Net

Radiation, Vegetation Index, Temperature, and Soil Moisture. J. Hydrometeorol. 9, 712–727. https://doi.org/10.1175/2007JHM911.1
Wang, K., Wang, P., Li, Z., Cribb, M., Sparrow, M., 2007. A simple method to estimate actual evapotranspiration from a combination of net radiation, vegetation index, and temperature. J. Geophys. Res. 112. https://doi.org/10.1029/2006JD008351
Wickens, G.E., 1997. Has the Sahel a future? J. Arid Environ. 37, 649–663. https://doi.org/10.1006/jare.1997.0303
Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B.E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., Verma, S., 2002. Energy balance closure at FLUXNET sites. Agric. For. Meteorol. 113, 223–243. https://doi.org/10.1016/S0168-1923(02)00109-0
Xiao, Z., Liang, S., Jiang, B., 2017. Evaluation of four long time-series global leaf area index products. Agric. For. Meteorol. 246, 218–230. https://doi.org/10.1016/j.agrformet.2017.06.016
Xiao, Z., Liang, S., Wang, J., Xiang, Y., Zhao, X., Song, J., 2016. Long-Time-Series Global Land Surface Satellite Leaf Area Index Product Derived From MODIS and AVHRR Surface Reflectance. IEEE Trans. Geosci. Remote Sens. 54, 5301–5318. https://doi.org/10.1109/TGRS.2016.2560522
Yang, F., White, M.A., Michaelis, A.R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.-X., Nemani, R.R., 2006. Prediction of Continental-Scale Evapotranspiration by Combining MODIS and AmeriFlux Data Through Support Vector Machine. IEEE Trans. Geosci. Remote Sens. 44, 3452–3461. https://doi.org/10.1109/TGRS.2006.876297
Yao, Y., Liang, S., Cheng, J., Liu, S., Fisher, J.B., Zhang, X., Jia, K., Zhao, X., Qin, Q., Zhao, B., Han, S., Zhou, Guangsheng, Zhou, Guoyi, Li, Y., Zhao, S., 2013. MODIS-driven estimation of terrestrial latent heat flux in China based on a modified Priestley?Taylor algorithm. Agric. For. Meteorol. 171–172, 187–202. https://doi.org/10.1016/j.agrformet.2012.11.016
Yao, Y., Liang, S., Li, X., Chen, J., Wang, K., Jia, K., Cheng, J., Jiang, B., Fisher, J.B., Mu, Q., Grünwald, T., Bernhofer, C., Roupsard, O., 2015. A satellite-based hybrid algorithm to determine the Priestley–Taylor parameter for global terrestrial latent heat flux estimation across multiple biomes. Remote Sens. Environ. 165, 216–233. https://doi.org/10.1016/j.rse.2015.05.013
Yao, Y., Liang, S., Li, X., Hong, Y., Fisher, J.B., Zhang, N., Chen, J., Cheng, J., Zhao, S., Zhang, X., Jiang, B., Sun, L., Jia, K., Wang, K., Chen, Y., Mu, Q., Feng, F., 2014. Bayesian multimodel estimation of global terrestrial latent heat flux from eddy covariance, meteorological, and satellite observations. J. Geophys. Res. Atmospheres 119, 4521–4545. https://doi.org/10.1002/2013JD020864
Yao, Y., Liang, S., Qin, Q., Wang, K., Zhao, S., 2011. Monitoring global land surface drought based on a hybrid evapotranspiration model. Int. J. Appl. Earth Obs. Geoinformation 13, 447–457. https://doi.org/10.1016/j.jag.2010.09.009

Yebra, M., Van Dijk, A., Leuning, R., Huete, A., Guerschman, J.P., 2013. Evaluation of optical remote sensing to estimate actual evapotranspiration and canopy conductance. Remote Sens. Environ. 129, 250–261. https://doi.org/10.1016/j.rse.2012.11.004

Zhang, K., Kimball, J.S., Nemani, R.R., Running, S.W., 2010. A continuous satellite-derived global record of land surface evapotranspiration from 1983 to 2006: GLOBAL RECORD OF LAND SURFACE EVAPOTRANSPIRATION. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008800

Zhang, K., Kimball, J.S., Nemani, R.R., Running, S.W., Hong, Y., Gourley, J.J., Yu, Z., 2015. Vegetation Greening and Climate Change Promote Multidecadal Rises of Global Land Evapotranspiration. Sci. Rep. 5, 15956. https://doi.org/10.1038/srep15956

Zhang, K., Kimball, J.S., Running, S.W., 2016. A review of remote sensing based actual evapotranspiration estimation: A review of remote sensing evapotranspiration. Wiley Interdiscip. Rev. Water 3, 834–853. https://doi.org/10.1002/wat2.1168

Zhang, K., Zhu, G., Ma, J., Yang, Y., Shang, S., Gu, C., 2019. Parameter Analysis and Estimates for the MODIS Evapotranspiration Algorithm and Multiscale Verification. Water Resour. Res. 55, 2211–2231. https://doi.org/10.1029/2018WR023485

Zhang, X., Liang, S., Zhou, G., Wu, H., Zhao, X., 2014. Generating Global LAnd Surface Satellite incident shortwave radiation and photosynthetically active radiation products from multiple satellite data. Remote Sens. Environ. 152, 318–332. https://doi.org/10.1016/j.rse.2014.07.003