Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment xx (2007) xxx – xxx
Remote Sensing of Environment
www.elsevier.com/locate/rse

Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach

Feihua Yang [a,b,*], Kazuhito Ichii [b,c], Michael A. White [d], Hirofumi Hashimoto [b,e], Andrew R. Michaelis [b,e], Petr Votava [b,e], A-Xing Zhu [a,f], Alfredo Huete [g], Steven W. Running [h], Ramakrishna R. Nemani [b]

[a] *University of Wisconsin, Madison, USA*
[b] *NASA Ames Research Center, USA*
[c] *San Jose State University, USA*
[d] *Utah State University, USA*
[e] *California State University, Monterey Bay, USA*
[f] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China*
[g] *University of Arizona, Tucson, USA*
[h] *University of Montana, Missoula, USA*

Received 19 October 2006; received in revised form 8 February 2007; accepted 10 February 2007

Abstract

Remote sensing is a potentially powerful technology with which to extrapolate eddy covariance-based gross primary production (GPP) to continental scales. In support of this concept, we used meteorological and flux data from the AmeriFlux network and Support Vector Machine (SVM), an inductive machine learning technique, to develop and apply a predictive GPP model for the conterminous U.S. In the following four-step process, we first trained the SVM to predict flux-based GPP from 33 AmeriFlux sites between 2000 and 2003 using three remotely-sensed variables (land surface temperature, enhanced vegetation index (EVI), and land cover) and one ground-measured variable (incident shortwave radiation). Second, we evaluated model performance by predicting GPP for 24 available AmeriFlux sites in 2004. In this independent evaluation, the SVM predicted GPP with a root mean squared error (RMSE) of 1.87 gC/m^2/day and an R^2 of 0.71. Based on annual total GPP at 15 AmeriFlux sites for which the number of 8-day averages in 2004 was no less than 67% (30 out of a possible 45), annual SVM GPP prediction error was 32.1% for non-forest ecosystems and 22.2% for forest ecosystems, while the standard Moderate Resolution Imaging Spectroradiometer GPP product (MOD17) had an error of 50.3% for non-forest ecosystems and 21.5% for forest ecosystems, suggesting that the regionally tuned SVM performed better than the standard global MOD17 GPP for non-forest ecosystems but had similar performance for forest ecosystems. The most important explanatory factor for GPP prediction was EVI, removal of which induced a GPP RMSE of 0.85 gC/m^2/day in a cross-validation experiment. Third, using the SVM driven by remote sensing data including incident shortwave radiation, we predicted 2004 conterminous U.S. GPP and found that results were consistent with expected spatial and temporal patterns. Finally, as an illustration of SVM GPP for ecological applications, we estimated maximum light use efficiency (e_{max}), one of the most important factors for standard light use efficiency models, for the conterminous U.S. by integrating the 2004 SVM GPP with the MOD17 GPP algorithm. We found that e_{max} varied from ~0.86 gC/MJ in grasslands to ~1.56 gC/MJ in deciduous forests, while MOD17 e_{max} was 0.68 gC/MJ for grasslands and 1.16 gC/MJ for deciduous forests, suggesting that refinements of MOD17 e_{max} may be beneficial.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Gross primary production; AmeriFlux; Support Vector Machines; Light use efficiency; Moderate Resolution Imaging Spectroradiometer (MODIS)

* Corresponding author. Department of Geography, University of Wisconsin-Madison, 426 Science Hall, 550 North Park Street, Madison, WI 53706, USA.
E-mail address: feihuayang@wisc.edu (F. Yang).

0034-4257/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2007.02.016

1. Introduction

Gross primary production (GPP), the integral of photosynthesis by all leaves, is a critical component in ecological systems. Carbon accumulated by ecosystem GPP and not used for plant growth and maintenance is returned to the atmosphere via respiration or disturbance (e.g. combustion), or is transported to other ecosystems through flow paths such as dissolved organic carbon. Autotrophic respiration consumes about half of GPP (Chapin et al., 2004); net primary production (NPP) is the residual. Quantitative estimates of the spatial and temporal distribution of GPP and NPP at regional to global scales are critical for the understanding of ecosystem response to increased atmospheric carbon dioxide (CO_2) level and are thus central to policy-relevant decisions (Metz et al., 2006).

Currently, GPP is estimated via process-based models and satellite-data based models. Process-based models such as BIOME-BGC (Thornton, 1998; White et al., 2000) can dynamically simulate vegetation physiology (e.g. rubisco activities and soil water stress), leading to potentially useful application at watershed scales (Zhu & Scott, 2001). However, process-based models are difficult to extend to large regions due to their complex structures and requirements for complete coverage of frequently poorly known land surface state variables such as specific leaf area and respiration coefficients. On the other hand, models based on or ingesting remote sensing data have two central advantages over purely process-based models: (1) satellite remote sensing offers broad spatial coverage and regular temporal sampling; and (2) requirements for spatial and temporal parameterization of vegetation physiological variables are reduced or eliminated. Remote sensing models are thus theoretically capable of accurately predicting actual carbon fluxes at regional to continental scales.

Methods using remote sensing data for carbon flux calculation are mostly based on the concept of light use efficiency (LUE). Monteith (1972) suggested that the NPP of well-watered and fertilized annual crop plants was linearly related to the absorbed photosynthetically active radiation (APAR). This formulation simplifies photosynthesis calculation over large region and is the basis of many remote sensing-based GPP algorithms such as the Moderate Resolution Imaging Spectroradiometer (MODIS) GPP/NPP algorithm (MOD17; Running et al., 2004), the Vegetation Photosynthesis Model (VPM; Xiao et al., 2005) and the Biosphere Model Integrating Eco-physiological and Mechanistic Approaches using Satellite Data (BEAMS; Sasai et al., 2005). In LUE-based GPP models, APAR (the product of photosynthetically active radiation (PAR) and the fraction of PAR absorbed by plant canopies, FPAR) is usually linearly converted to GPP using biome-specific maximum light use efficiency (e_{max}, gC/MJ) attenuated by temperature and water stress status. However, recent studies have suggested that this approach may lead to considerable errors in modeled GPP (Heinsch et al., 2006; Turner et al., 2003, 2005; Zhao et al., 2005). Likely causes of the error include uncertainties in PAR, FPAR, and the conversion from incident shortwave radiation to PAR. Perhaps most centrally, though, uncertainty in the spatiotemporal variation of e_{max} is a crucial limiting factor for LUE models.

While remote sensing has been used to extrapolate field-measured NPP since the late 20th century (Paruelo et al., 1997), the increasing availability of near-real time observations of water and carbon exchange from eddy covariance flux towers (e.g. AmeriFlux; Baldocchi et al., 2001) has sparked a renewed interest in extrapolative techniques, especially through the use of empirical modeling techniques using remote sensing explanatory variables. For example, Rahman et al. (2005) observed a strong correlation between across-site tower-GPP and enhanced vegetation index (EVI). Wylie et al. (2003) related coarse resolution normalized difference vegetation index (NDVI) to 14-day average daytime CO_2 fluxes in a sage–brush–steppe ecosystem. Xiao et al. (2005) integrated e_{max} derived from tower-based net ecosystem exchange (NEE) and PAR into the VPM model. Gilmanov et al. (2005) found that NDVI was statistically significantly correlated with tower-based GPP and ecosystem respiration leading to potential scaling-up of tower fluxes to larger areas. Drolet et al. (2005) reported strong correlations between MODIS-derived photochemical reflectance index and tower-based LUE. These studies established the potential of using statistical and machine learning methods to extrapolate tower-based GPP to a regional to continental scale.

The goal of this paper is to explore the application of Support Vector Machine (SVM) learning techniques for GPP prediction at a continental scale. To do so, we tuned and trained SVMs driven by ground measured and remotely sensed explanatory variables to predict AmeriFlux GPP, tested the SVMs using a withheld portion of the flux data, and applied the final model for GPP prediction over the conterminous U.S. In the following sections, we present: (1) a brief description of the SVM technique; (2) SVM tuning and training, including a description of the AmeriFlux GPP observations and the selection of explanatory variables; (3) results from independent testing of the SVMs; (4) a comparison of SVM GPP to MOD17 GPP; and (5) extrapolation of the SVMs to the conterminous U.S. Finally, to demonstrate the potential use of the SVM GPP for a broad modeling community, we present a method to estimate e_{max} for the conterminous U.S. by coupling the SVM GPP with the MOD17 GPP algorithm.

2. Methods

2.1. SVM for regression

Regression methods attempt to construct an approximate function which maps an input domain to a real valued output domain based on a set of data examples (Cristianini & Shawe-Taylor, 2000). Commonly used regression methods include conventional statistical methods, such as multiple regressions, and machine learning methods, such as neural network and SVM.

Multiple regression is a standard statistical method designed to predict the values of a target concept from two or more explanatory variables. It is conceptually simple but less suited for highly nonlinear problems, especially those outside a prescribed range of nonlinear approaches. Neural network is a computing system motivated by the function of a human brain (Haykin, 1998). It is widely used for regression due to its ability

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

to approximate any nonlinear function. However, neural network suffers from challenges in selecting proper network structure and finding optimal solutions due to its complexity and high nonlinearity (Haykin, 1998).

The problems inherent to neural network led researchers to look for alternatives such as SVM for nonlinear regressions. SVMs were first developed by Vapnik (Vapnik, 1998; Vapnik & Chervonenkis, 1991) for solving pattern classification problems, but they have been extended to the domain of regression approximation. For example, Zhan et al. (2003) used SVM for the nonlinear approximation of the relationships between ocean chlorophyll concentration and remotely sensed marine reflectance. Yang et al. (2006) used SVM for evapotranspiration prediction from satellite remote sensing.

SVMs transform nonlinear regression into linear regression by mapping the original low dimensional input space to a higher dimensional feature space using kernel functions satisfying Mercer's condition (i.e. the kernel function must be positive semi-definite) (Vapnik, 1998). A linear model is then constructed in the new feature space, leading to a convex quadratic programming (QP) problem guaranteed to have a global optimal solution. We used SVM in this study due to its simplicity, global optimality and great predictive power. For details of SVM regression, readers are referred to Cristianini and Shawe-Taylor (2000) and Yang et al. (2006). For model comparisons between multiple regressions, neural network and SVM, information can be found in Yang et al. (2006).

2.2. Explanatory variable selection

Vegetation photosynthesis is a complex process influenced by a large suite of edaphic, atmospheric, and physiological variables operating at different spatiotemporal scales. Spatially, individual leaf level photosynthesis is regulated by stomatal conductance and is limited by light, water and nutrient availability while regional GPP is affected by light, water and temperature (Chapin et al., 2004; Nemani et al., 2003). Temporally, hourly and daily photosynthesis reflect the influence of sunlight and temperature while monthly and seasonal photosynthesis anomalies are related to climate variations (Chapin et al., 2004; Goulden et al., 1996). Based on the availability of remote sensing data for real-time SVM implementation over large regions, we selected land surface temperature (LST), enhanced vegetation index (EVI), incident shortwave radiation (SWR), and land cover as explanatory variables.

2.3. Data

We required five types of data for our SVM analysis: (1) tower-based GPP from AmeriFlux sites (Baldocchi et al., 2001) for training and testing; (2) explanatory environmental datasets (LST, EVI, SWR, land cover) for both the AmeriFlux sites and the continental application; (3) MOD17 GPP for model comparison at AmeriFlux sites; (4) NDVI and FPAR for GPP predictive performance comparison with EVI at AmeriFlux sites; and (5) inputs for e_{max} estimation. Temporally, we used data from 2000–2004 for AmeriFlux research, 2004 for conterminous U.S. application, and 2004 for SVM comparison with MOD17. For the continental application, all inputs were resampled and/or reprojected to an 8 km resolution.

2.3.1. Tower-based GPP

We acquired CO_2 flux or NEE (CO_2 flux with storage CO_2 flux corrected) from hourly or half-hourly measurements at 36 available AmeriFlux sites with a wide diversity of vegetation structures (Baldocchi, 2003; Baldocchi et al., 2001) (Fig. 1 and Table 1). We used site-provided GPP when available and otherwise calculated hourly or half-hourly daytime GPP using NEE (CO_2 flux was used as a surrogate of NEE if NEE not available) in the following three steps.

First, we performed outlier removal for each flux site in 2000–2004. We grouped flux observations for each year in 2000–2004 into nine bins: [−∞, −100], [−100, −80], [−80, −60], [−60, −40], [−40, 40], [40, 60], [60, 80], [80, 100], and [100, +∞] (μmol/m^2/s; excluding missing data) and accepted all observations in central [−40, 40] range without outlier detection. Otherwise, we computed the percentage of records in each bin. If the percentage was smaller than 0.1%, we replaced the NEE observations in the corresponding bin as missing. This process detected an overall 0.02% records as outliers.

Second, we filtered nighttime flux measurements (PAR less than 10 W/m^2, ≈45.5 μmol/m^2/s) using friction velocity (u*) (Gu et al., 2005). We identified the upper u_H* and lower u_L* thresholds for all unique collections of nighttime fluxes (grouped by site, season, and year) and used the flux values with u* between u_H* and u_L* to develop a temperature response function (Gu et al., 2002),

$$R_e = c_1 e^{c_2[c_3 T_a + (1-c_3) T_s]} + d_1 e^{d_2 T_s} \quad (1)$$

where R_e is the ecosystem respiration, c_1, c_2, c_3, d_1 and d_2 are regression coefficients, T_a is air temperature and T_s is soil

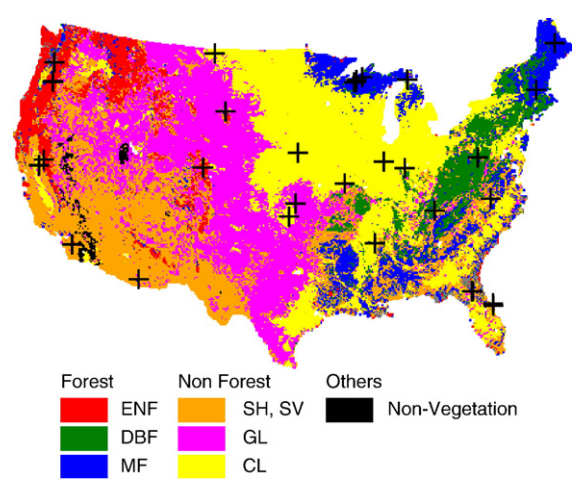

Fig. 1. Land cover and the distribution of AmeriFlux sites. Land cover was derived from 2001 MODIS land cover products (MOD12Q1) and regrouped into forest (Evergreen Needleleaf Forest, ENF; Deciduous Broadleaf Forest, DBF; Mixed Forest, MF), non-forest (Closed/Open Shrubland and Woody Savannas and Savannas, SH and SV; Grassland, GL; Cropland and Cropland/Natural vegetation mixture, CL), and non-vegetation (Water, urban and barren or sparsely vegetated areas) regions. The 36 Ameriflux sites are shown as plus symbols.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

Table 1
Name, latitude, longitude, vegetation structure and years of data available for each flux site in this study

Name	Latitude (°)	Longitude (°)	Vegetation structure [a]	Year
Forest				
Indiana MMSF, IN (IN) [b, c]	39.3232	−86.4134	Mixed hardwood deciduous forest dominated by sugar maple, tulip poplar, sassafras, white oak, and black oak	2000–2002
Blodgett, CA (BL)	38.8953	−120.6328	Mixed evergreen coniferous forest dominated by ponderosa pine (>70%)	2000–2004
University of Michigan, MI (UM) [b]	45.5598	−84.7138	Mid-aged conifer and deciduous, northern hardwood with mostly deciduous, old growth hemlock	2000–2003
Niwot Ridge Forest, CO (NI) [b]	40.0329	−105.5464	Subalpine coniferous forest dominated by subalpine fir, engelmann spruce, and lodgepole pine	2000–2004
Howland Forest, ME (HO)	45.2041	−68.7403	Boreal — northern hardwood transitional forest with 41% red spruce, 25% eastern hemlock, 23% other conifers and 11% hardwoods	2000–2004
Harvard Forest, MA (HA) [b, c]	42.5378	−72.1715	Temperate deciduous forest dominated by red oak, red maple, black birch, white pine, and hemlock	2000–2004
Lost Creek, WI (LC) [b]	46.0827	−89.9792	Alder-willow deciduous wetland	2000–2004
Willow Creek, WI (WC) [b]	45.9059	−90.0799	Temperate/boreal forest, lowland and wetland forest, upland hardwoods dominated by white ash, sugar maple, basswood, green ash, and red oak	2000–2004
Park Falls, WI (PF)	45.9459	−90.2723	Deciduous and coniferous forest with 70% deciduous (aspen, birch, maple, basswood, alder) and 30% conifer (balsam fir, jack pine, black spruce, white cedar)	2000–2004
Metolius Intermediate, OR (MM) [b, c]	44.4524	−121.5572	Temperate coniferous forest dominated by pinus ponderosa, purshia tridentate, and arctostaphylos patula	2002–2004
Metolius Old Young, OR (MY) [b, c]	44.4372	−121.5668	Temperate coniferous forest dominated by ponderosa pine	2000–2002
Metolius Old, OR (MO) [b, c]	44.4992	−121.6224	Temperate coniferous forest dominated by pinus ponderosa, purshia tridentate	2000–2000
Sylvania Wilderness Area, MI (SW) [b]	46.2420	−89.3477	Old-growth eastern hemlock/sugar maple/basswood/yellow birch	2001–2004
Duke Forest Pine, NC (DP)	35.9782	−79.0942	Even-aged loblolly pine plantation, deciduous, oak-hickory type, mixed hardwood species, evergreen coniferous. Tree density = 3700/ha	2000–2003
Black Hills, SD (BH)	44.1580	−103.6500	Conifer forest dominated by Ponderosa pine	2001–2004
Donaldson, FL (DN) [b]	29.7548	−82.1633	Pine plantation dominated by pinus elliottii	2000–2002
KSC Scrub Oak, FL (KO) [b]	28.6086	−80.6715	Scrub-oak palmetto dominated by schlerophyllous evergreen oaks and saw palmetto serenoa repens	2000–2003
KSC Slash Pine, FL (KP)	28.4583	−80.6709	Old pine flatwoods ecosystem	2002–2002
Mize, FL (MI) [b]	29.7648	−82.2448	Pine plantation dominated by pinus elliottii	2000–2003
Wind River, WA (WR) [b]	45.8205	−121.9519	Old coniferous, temperate rainforest, evergreen forest dominated by Douglas fir, western hemlock	2004–2004
Walker Branch, TN (WB)	35.9588	−84.2874	Mixed-species, broad-leaved forest, deciduous forest, oak/hickory dominated by oak, hickory, maple, tulip poplar and loblolly pine	2002–2004
Ozark, MO (OZ) [b]	38.7441	−92.2001	Oak hickory forest located in the transitional zone between central hardwood region and grassland region of USA	2004–2004
Non-forest				
Walnut River, KS (WA)	37.5208	−96.8550	C3/C4 mixed grassland, tallgrass prairie	2001–2004
Bondville, IL (BO)	40.0061	−88.2919	Annual rotation between corn and soybeans	2000–2004
Mead Rainfed, NE (MF) [b]	41.1797	−96.4396	Maize–soybean rotation	2001–2004
Mead Irrigated, NE (MI) [b]	41.1651	−96.4766	Continuous maize	2001–2004
Mead Rotation, NE (MR) [b]	41.1649	−96.4701	Maize–soybean rotation	2001–2004
Vaira Ranch, CA (VA)	38.4067	−120.9507	Grazed C3 grassland opening in a region of oak/grass savanna	2000–2004
Tonzi Ranch, CA (TO)	38.4316	−120.9660	Blue oak trees in the overstory and annual grasses in the understory	2001–2004
Fort Peck, MT (FO)	48.3079	−105.1005	Temperate grassland	2000–2004
Sky Oaks Old Chaparral, CA (SO)	33.3739	−116.6230	Chaparral on hilly terrain	2000–2004
Sky Oaks Young Chaparral, CA (SO)	33.3772	−116.6230	Chaparral on hilly terrain	2000–2004
Canaan Valley, WV (CV)	39.0633	−79.4208	Temperate grassland	2004–2004
Goodwin Creek, MS (GO)	34.2500	−89.9700	Temperate grassland	2002–2004
Audubon Grasslands, AZ (AU)	31.6000	−110.5104	Desert grassland	2002–2004
Arm Oklahoma, OK (AO)	36.6050	−97.4850	Winter wheat, some pasture and summer crops	2003–2004

The names in the parentheses are abbreviations of flux sites.
[a] Descriptions on vegetation structures come from site information available at http://public.ornl.gov/ameriflux/.
[b] NEE provided.
[c] GPP provided.

temperature. Eq. (1) describes a respiration model of two carbon pools with the first term on the right hand side representing the above-ground biomass respiration while the second term representing soil respiration (Gu et al., 2002). We removed all nighttime fluxes with u^* below u_L^* or above u_H^* and filled the resulting R_e gaps using Eq. (1). We did not conduct R_e filtering if PAR, T_a, T_s, or u^* was missing or if fewer than five nighttime measurements were present.

Third, we computed daytime GPP by subtracting NEE from R_e (upward flux positive by convention). If the result was smaller than zero, we set GPP to zero. We marked GPP as missing if daytime PAR, T_a, or T_s was missing and set GPP to zero during nighttime.

We then processed the tower-based GPP to 8-day averages to correspond with satellite compositing intervals. Because the tower-based GPP had only 65% coverage due to system failures or data rejection, we gap-filled the flux observations and treated missing values as follows (Falge et al., 2001): (1) if more than 70% of data were missing in an 8-day period, we marked the period as missing; (2) if a particular time of day was missing in all 8 days, i.e. all eight 2 a.m. values were missing, we marked the period as missing; (3) if neither condition 1 nor 2 were met, we filled missing values with the mean from the non-missing days, i.e. if a single 2 a.m. value was missing from the 8-day period, we filled it with the mean of the remaining seven 2 a.m. values.

2.3.2. LST, EVI, SWR and land cover

For LST, we used the MODIS 8-day average 1 km daytime subsets (Cook et al., 2004) consisting of 7 km × 7 km regions centered on flux towers for each AmeriFlux site. At each time step, we computed flux site LST as the average of the pixels marked as good quality (mandatory quality assurance (QA) flag being zero in the QA data) (Wan et al., 2002). If none of the 49 values was of good quality, we treated the period as missing. For the 2004 conterminous U.S. extrapolation, we obtained the MODIS 8-day average 1 km daytime LST product (MOD11A2; Wan et al., 2002), which has a deviation of ±1 °C compared to ground measurements (Wan et al., 2004; Wan et al., 2002).

For EVI, we again used the subsets for the AmeriFlux sites and the standard MODIS product (MOD13A2; Huete et al., 2002) for continental application. Based on RMSE from ground validations, EVI RMSE is about 0.03 (Gao et al., 2003). EVI is composited on a 16-day basis; for both the AmeriFlux and continental applications, we therefore assigned each 16-day composite EVI to the corresponding two 8-day periods.

We used satellite- and ground-based inputs for SWR. For continental application, we obtained daily 0.5° resolution SWR from the Surface Radiation Budget project (SRB), derived from the Geostationary Operational Environmental Satellite (Pinker et al., 2002) and processed 8-day averages. The daily SWR from GOES-SRB has an RMSE of 24 W/m^2 (~2.07 MJ/m^2/day) (Pinker et al., 2003). Due to the coarse resolution of remotely sensed SWR, we opted to use ground-base SWR (converted from PAR by assuming 45% incident shortwave radiation is PAR) for SVM tuning, training, and testing over the AmeriFlux sites.

We obtained AmeriFlux land cover from the site descriptions. Due to data availability, we regrouped the land cover classes into two categories (Fig. 1): forest (evergreen needleleaf forest, deciduous broadleaf forest, and mixed forest) and non-forest (savanna, shrubland, grassland, and cropland). For the conterminous U.S., we obtained land cover from the MODIS land cover product (MOD12Q1) (Friedl et al., 2002) and again regrouped classes to forest and non-forest (Fig. 1). Based on ground comparisons at the continental scale, the accuracy of MODIS land cover is 70–85% (Friedl, 2003).

2.3.3. MOD17 GPP

We compared the performance of the SVM GPP against the standard MODIS GPP product MOD17A2 (Heinsch et al., 2006; Running et al., 2004). As for LST and EVI, we used GPP subsets for the AmeriFlux sites. Although MOD17 GPP performs well in forest ecosystems, overestimating tower-based GPP by only 20–30% (Heinsch et al., 2006; Running et al., 2004), other researches have found that MOD17 overestimates low GPP and underestimates high GPP (Rahman et al., 2005; Turner et al., 2006), suggesting that alternate methods, such as the SVM approach presented here, may provide useful alternatives.

2.3.4. NDVI and FPAR

Although we based our primary SVM vegetation characterization on EVI, we also explored the use of two additional remote sensing products: NDVI and FPAR. As for LST and EVI, we used subsets for the AmeriFlux sites. NDVI was derived from the standard MODIS product MOD13A2 (Huete et al., 2002) while FPAR was derived from MOD15A2 (Myneni et al., 2002).

2.3.5. Inputs for e_{max} estimation

For e_{max} estimation, we used four types of data: (1) 8-day composite MODIS FPAR products (MOD15A2; Myneni et al., 2002); (2) daily 0.5° resolution PAR converted from GOES-SRB SWR described in Section 2.3.2; (3) daily minimum temperature (T_{min}), and (4) daily vapor pressure deficit (VPD). To generate T_{min} and VPD, we first created 8 km surfaces of daily maximum temperature (T_{max}) and T_{min} using Ordinary Kriging (Jolly et al., 2005) with two independent point observation datasets from National Climatic Data Center (NCDC) and Climate Prediction Center (CPC). An average of 1650 stations was used for the spatial interpolation over the conterminous U.S. in 2004. We then calculated daytime VPD as the residuals of saturated and actual vapor pressures using the DAYMET algorithm (Thornton et al., 1997). The saturated vapor pressure was calculated from daytime temperature derived from T_{max} and T_{min} (Thornton et al., 1997). The actual vapor pressure was calculated by assigning T_{min} as the dew point temperature (Campbell and Norman, 1998).

2.4. Experiments

2.4.1. SVM tuning, training, and testing

Using AmeriFlux tower-based GPP observations, we tuned and trained the SVM with 2000–2003 data and tested the SVM with 2004 data. Our input variables were LST, EVI, SWR, and land cover, and our target concept was GPP. After removing 8-day periods in which one or more of these variables were missing, we had a total of 2603 data examples from 33 flux sites in the GPP training set and a total of 736 data examples in the GPP test set. We scaled all the input variables to the range of −1 to +1 based on the minimum and maximum values of the

2000–2003 data, as per standard SVM techniques to eliminate the influence of variables with different absolute magnitudes.

The configuration of SVM regression requires three types of parameters: C for the cost of errors, ε for the width of an insensitive error band (ε-insensitive band), and kernel parameters. The parameter C determines the tradeoff between model complexity and the training error, with higher values of C decreasing the impact of the model complexity on the optimization formulation. The parameter ε controls the tolerance for training errors. Data examples that have training errors smaller than ε are ignored in the optimization formulation. Finally, the kernel parameters vary with the selection of the kernel function.

In this study, we configured SVM as follows. First, we selected the radial basis function (RBF) kernel, as opposed to linear, polynomial, or sigmoid kernels, because it is highly flexible and requires only one parameter, σ (Chang & Lin, 2005). Second, we tuned C, ε, and σ using a grid search with a three-fold cross validation training process (Chang & Lin, 2005). In this approach, the training examples are randomly divided into three non-overlapping subsets; training is performed three times on two of the subsets with the remaining subset reserved for testing; parameters yielding the lowest cross validation errors are selected. We initially conducted a coarse grid search (Chang & Lin, 2005) for C (2^{-1}, 2^{0}, 2^{1}, …, 2^{4}), ε (2^{-4}, 2^{-3}, 2^{-2}, …, 2^{2}), and σ (2^{-3}, 2^{-2}, 2^{-1}, …, 2^{4}) and identified the C, ε, and σ combination producing the lowest mean cross validation RMSE. We then used a progressively finer grid search until the variance of the RMSE was smaller than 0.01. Third, using the selected (C, ε, σ), we conducted a final training of the SVM with the 2000–2003 AmeriFlux data.

Lastly, we tested the trained model on the test set for three groups: forest, non-forest, and forest and non-forest combined. The GPP testing dataset (2004) was from 24 flux sites (12 forest and 12 non-forest sites) based on data availability. We evaluated SVM performance using RMSE, R^2, scatterplots of predictions versus observations, seasonal variations between the predictions and observations, and residual analysis.

2.4.2. Contribution of input variables on GPP variations

We examined the contribution of each input variable on SVM GPP predictions by sequentially removing one of the input variables (LST, EVI, SWR, and land cover) and replicating the cross validation training process. We assessed the contribution of each input variable with the mean cross validation RMSE and R^2 from the cross validation training process.

2.4.3. Comparison of tower-based GPP with SVM GPP and MOD17 GPP

We compared the performance of SVMs to that of MOD17 GPP at the 24 Ameriflux sites in 2004 using RMSE, R^2, and scatterplots of predictions versus observations. Further, we compared annual SVM GPP against MOD17 GPP. We used only those sites with at least 30 non-missing 8-day averages (out of a possible 45) and gap filled missing periods with linear interpolation. Extrapolation was performed if the missing 8-day averages were at the beginning or end of the year and was set to zero if the extrapolated value was smaller than zero.

2.4.4. Comparisons of the predictive performance of EVI with NDVI and FPAR

We compared the predictive performance of EVI with NDVI and FPAR by replacing EVI with NDVI and FPAR to form two new sets of input variables: (LST, NDVI, SWR, and land cover) and (LST, FPAR, SWR, and land cover) and replicating the tuning, training and testing processes described in Section 2.4.1. We assessed the performance with the RMSE and R^2 on the test set.

2.4.5. Generalization from AmeriFlux sites to the conterminous U.S.

Based on research showing that the AmeriFlux network is representative of conterminous U.S. ecoregions (Hargrove et al., 2003) and that the 36 flux sites in this study included most of the active flux sites in the AmeriFlux network (Fig. 1), we reasoned that the knowledge learned at flux sites can be extrapolated to the conterminous U.S. In order to generalize the model learned from flux sites to the conterminous U.S., we first conducted a new training of the SVM with the entire 2000–2004 dataset using the selected (C, ε, σ) from Section 2.4.1. The trained models were then used to investigate the spatial and temporal distribution of GPP over the conterminous U.S. for 2004.

2.4.6. Estimation of e_{max} for the conterminous U.S.

As an illustration of the SVM GPP estimations for ecological applications, we examined the spatial distribution of $e_{\max}$ over the conterminous U.S. by coupling SVM GPP estimation with the MOD17 GPP algorithm (Heinsch et al., 2006; Running et al., 2004). MOD17 GPP is calculated as follows:

$$GPP = e_{\max} \times m(T_{\min}) \times m(\mathrm{VPD}) \times \mathrm{PAR} \times \mathrm{FPAR} \qquad (2)$$

where $m(T_{\min})$ and $m(\mathrm{VPD})$ are multipliers for $e_{\max}$ used to reduce the conversion efficiency in the condition of cold temperature and high vapor pressure deficit, both of which reduce photosynthesis. The multipliers range linearly from 0 (total inhibition) to 1 (no inhibition). Eq. (2) includes five unknown parameters: (1) $e_{\max}$, (2 and 3) daily $T_{\min}$ at which $m(T_{\min})=1$ ($\mathrm{TMIN}_{\mathrm{start}}$) and $m(T_{\min})=0$ ($\mathrm{TMIN}_{\mathrm{full}}$), and (4 and 5) daily VPD at which $m(\mathrm{VPD})=1$ ($\mathrm{VPD}_{\mathrm{start}}$) and $m(\mathrm{VPD})=0$ ($\mathrm{VPD}_{\mathrm{full}}$).

For the parameter estimation, we used Levenberg–Marquardt nonlinear least squares algorithms with box-constraints (Kanzow et al., 2004) to minimize the RMSE between monthly variations in SVM GPP and MOD17 GPP for each pixel in the conterminous U.S. in 2004. Because the covariance between parameters in the relationships to GPP was high when all five parameters were left unconstrained and simultaneously optimized, we used MOD17 GPP approach to set the parameter values for $\mathrm{VPD}_{\mathrm{start}}$, $\mathrm{VPD}_{\mathrm{full}}$, and $\mathrm{TMIN}_{\mathrm{full}}$ (Running et al., 1999), and only optimized $e_{\max}$ and $\mathrm{TMIN}_{\mathrm{start}}$. We set the box constraints for $e_{\max}$ to between 0 and 3 gC/MJ and for $\mathrm{TMIN}_{\mathrm{start}}$ to between 5 and 15 °C. We further compared the optimized $e_{\max}$ with the $e_{\max}$ used in MOD17 by averaging the optimized $e_{\max}$ for six land covers: (1) evergreen needleleaf forest; (2) deciduous broadleaf forest, (3) mixed forest, (4) closed/open

shrubland plus woody savannas and savannas; (5) grassland, and (6) cropland plus cropland/natural vegetation mixture.

3. Results and discussion

3.1. AmeriFlux sites

3.1.1. SVM performance

Using the full input training set of LST, EVI, SWR and land cover, the parameter combination of $C=14.929$, $\varepsilon=0.063$ and $\sigma=3.732$ produced the smallest mean cross validation RMSE of 1.83 gC/m^2/day and an R^2 of 0.76. Using these parameter values and the trained SVM, the testing on the 2004 AmeriFlux data produced a RMSE for the forest and non-forest combined sites of 1.87 gC/m^2/day with an R^2 of 0.71. Non-forest sites had 2.05 gC/m^2/day RMSE and 0.63 R^2 while forest sites had 1.63 gC/m^2/day RMSE and 0.79 R^2 (Fig. 2), indicating that the SVM performed better in forest than in non-forest sites.

The SVM represented most features of the tower-based GPP seasonality in the 2004 AmeriFlux test data (Fig. 3). For some specific sites, episodes of under- or over-prediction occurred. Four examples of and possible explanations for site-specific SVM GPP over-prediction are presented below. The SVM GPP at Blodgett (Fig. 3 — 1, ponderosa pine) typically agreed well with tower-based GPP until the end of July. We speculate that at this point our R_e calculation method failed to capture the large measured summer soil respiration variations (Misson et al., 2006); errors in this case may therefore be due to errors in our calculation of tower GPP, not predictions from the SVM GPP. The over-prediction at Walnut River (Fig. 3 — 16, grassland) was influenced by unusually high EVI in the summer of 2003 and 2004 (~0.1 higher than in 2001 and 2002). The tower-based GPP at Tonzi Ranch (Fig. 3 — 19, savanna) was 2–3 gC/m^2/day lower during the early summer of 2004 than those in 2001–2003 yet the seasonal variations of LST, EVI and SWR during 2004 were similar to those in 2000–2003, indicating that GPP at Tonzi Ranch in the early summer of 2004 was influenced by other factors such as unusual wind and/or soil moisture patterns. Although vegetation LUE declines during plant senescence (Escudero & Mediavilla, 2003), EVI may remain high due to favorable water and nutrient conditions, possibly accounting for over-prediction at ARM Oklahoma (Fig. 3 — 24, winter wheat) in May 2004.

Under-prediction errors occurred in the summer of 2004 at Niwot Ridge Forest (Fig. 3 — 2), Metolius Intermediate (Fig. 3 — 7), Wind River (Fig. 3 — 11), Mead Irrigated (Fig. 3 — 14) and Bondville (Fig. 3 — 17). Although factors not explicitly included in the SVM, such as soil moisture and nutrition uptake, may account for some under-prediction errors, we speculate that use of 8-day mean conditions may inadequately represent the effects of non-linear GPP processes, i.e. the SVM drivers of LST, EVI, and SWR may be incapable of producing extremely high GPP values occurring during short periods of super-optimal physiological and boundary layer conditions. Further, unusually large fluxes might come from certain atmospheric turbulent events rather than real ecological or physiological processes (Gu et al., 2002), i.e. processes beyond the SVM prediction capability.

Based on examination of site-specific RMSE and R^2, SVM performance varied by flux site and land cover (Table 2). For forest sites, Black Hills had the lowest RMSE of 1.00 gC/m^2/day and Wind River had the highest RMSE of 3.08 gC/m^2/day. In terms of R^2, the lowest R^2 of 0.39 occurred at Blodgett and the highest R^2 of 0.93 occurred at Sylvania Wilderness Area. In comparison to forest sites, non-forest sites had greater variation of RMSE and R^2: RMSE varied from 0.62 gC/m^2/day at Aububon Research Ranch to 3.31 gC/m^2/day at Mead Irrigated, and R^2 varied from 0.00 at Aububon Research Ranch to 0.95 at Walnut River.

Residual analysis showed that RMSE averaged across all AmeriFlux sites showed a strong seasonality (Fig. 4). In absolute magnitudes, winter had low prediction errors (<1.0 gC/m^2/day) while warm season errors often exceeded 2.0 gC/m^2/day. Yet when expressed as a percentage of RMSE to the mean tower-based GPP, the pattern was reversed: winter errors were often above 80% but summer errors rarely exceeded 60%. This is not surprising because the rapid changes of ecosystem productivity in spring are likely to introduce great uncertainty in tower-based GPP measurement and remotely sensed input variables (LST, EVI and SWR). Further, we suspect that the

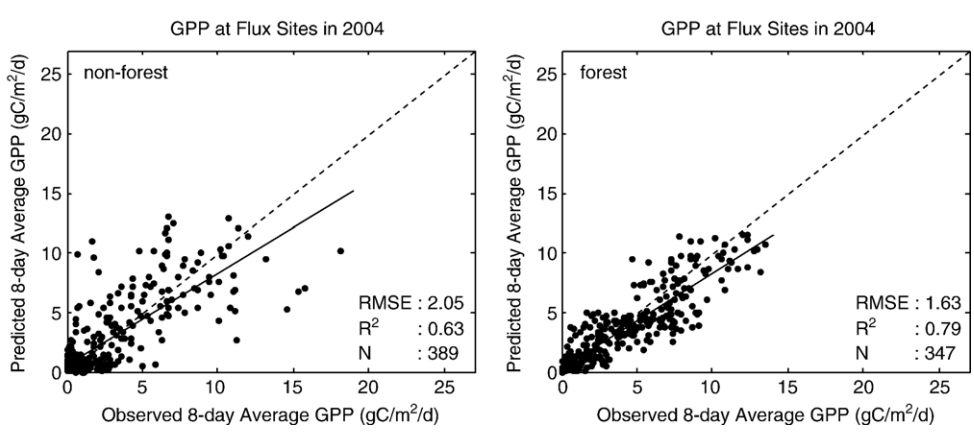

Fig. 2. Scatterplots of the observed AmeriFlux GPP versus predicted SVM GPP in 2004 for non-forest and forest sites. The SVM was trained with non-forest and forest combined, but tested with non-forest and forest separately. Dashed lines show a 1:1 relationship; solid lines show a least squares regression line.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

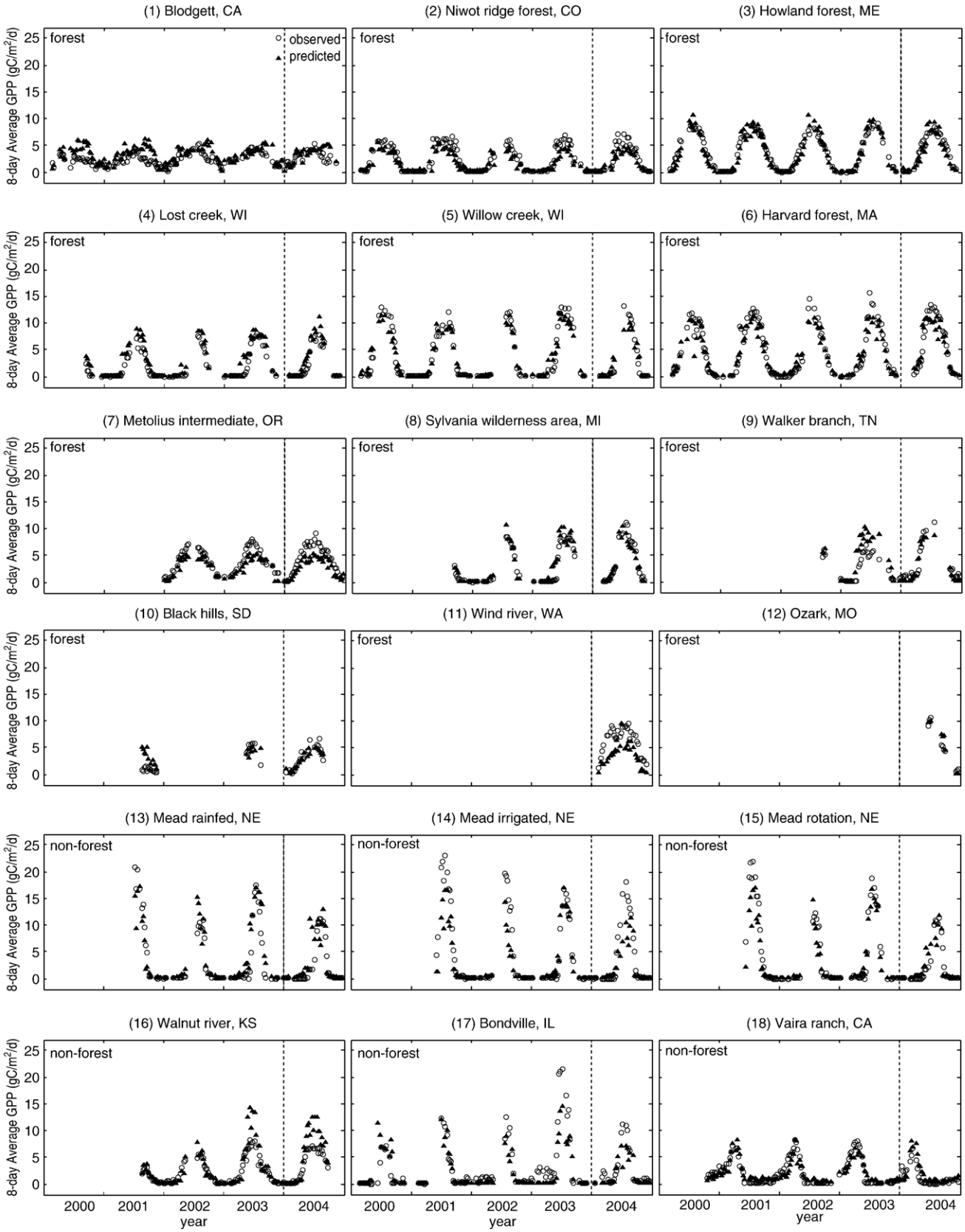

Fig. 3. Comparison of seasonal variations of the observed AmeriFlux (open circle) and predicted SVM (closed triangle) 8-day average GPP at flux sites for 2000–2004 (sites that do not have data available in 2004 testing set are not shown). The predicted GPP for 2000–2003 are training results; the predicted GPP for 2004 are testing results. The vertical dashed lines separate testing results from training results.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

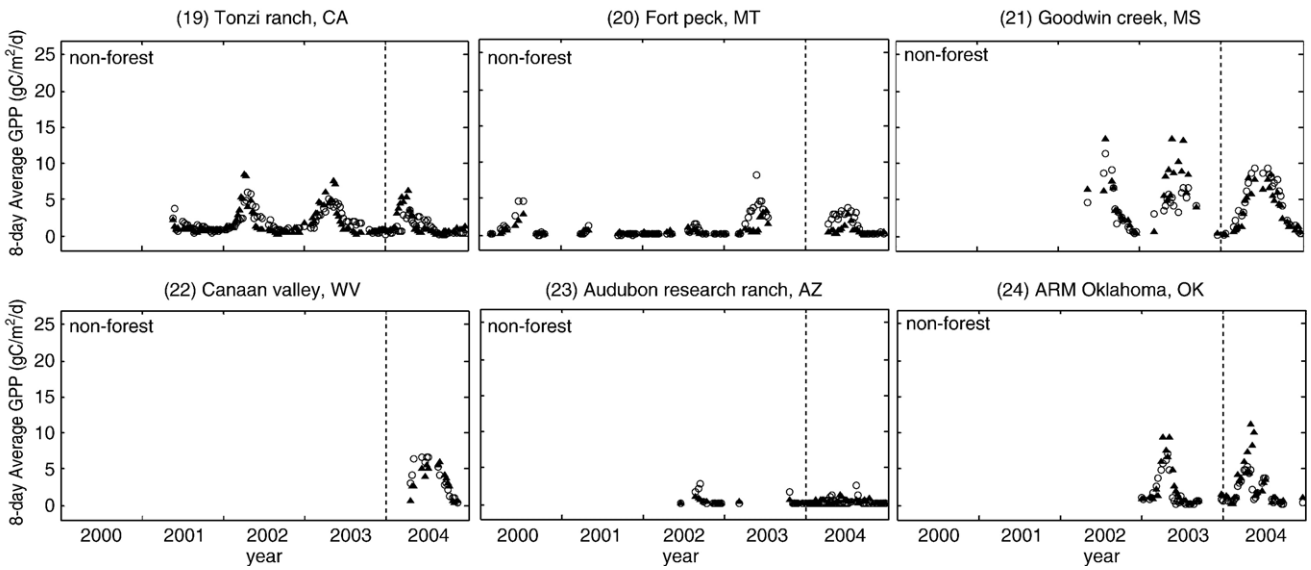

Fig. 3 (*continued*).

measurement error from the explanatory variables had additive or multiplicative influence on GPP prediction. Thus, potential users should carefully consider the sensitivity of the system under study to the seasonality of absolute and relative GPP prediction errors.

Analysis of the mean GPP from AmeriFlux and SVM also showed that SVM performance varied by flux sites and land cover (Fig. 5). The predicted and tower-based mean GPP was within 40% of the 1:1 line for all sites except Vaira Ranch (VA), Walnut River, Fort Peck (FO), ARM Oklahoma (AO) and Wind River (WR). High over-prediction occurred at Vaira Ranch (VA, 52.1%), Walnut River (WA, 49.8%), ARM Oklahoma (AO, 47.7%) whereas high under-prediction occurred at Fort Peck (FO, 66.3%) and Wind River (WR, 40.8%). In terms of land cover, the predicted mean GPP was within 20% of the 1:1 line for all forest sites except Niwot Ridge (NI), Metolius Intermediate (MM), Wind River (WR) and Lost Creek (LC); but of the non-forest sites, only Audubon Grassland (AU), Goodwin Creek (GO), Mead Rotation (MR) and Canaan Valley (CV) were within 20% of the 1:1 line. The overall prediction error was 31.8% for non-forest and 17.7% for forest. Given the heterogeneity of the AmeriFlux data and the simplicity of the model inputs, model performance was promising.

Table 2
Site name and statistics (RMSE and R^2) for GPP at flux sites in 2004 for the number (N) of available MODIS and tower 8-day average datasets

Name	RMSE (gC/m^2/day)	R^2	N
Forest			
Blodgett, CA	1.25	0.39	33
Niwot Ridge Forest, CO	1.24	0.89	38
Howland Forest, ME	1.07	0.89	34
Harvard Forest, MA	1.82	0.87	31
Lost Creek, WI	1.43	0.92	29
Willow Creek, WI	1.59	0.90	25
Metolius Intermediate, OR	1.90	0.87	41
Sylvania Wilderness Area, MI	1.08	0.93	26
Black Hills, SD	1.00	0.80	23
Wind River, WA	3.08	0.59	31
Walker Branch, TN	1.18	0.88	26
Ozark, MO	1.32	0.90	10
Non-forest			
Walnut River, KS	2.74	0.95	32
Bondville, IL	2.17	0.64	30
Mead Rainfed, NE	2.59	0.63	37
Mead Irrigated, NE	3.31	0.81	36
Mead Rotation, NE	1.86	0.76	36
Vaira Ranch, CA	1.42	0.66	37
Tonzi Ranch, CA	1.53	0.23	42
Fort Peck, MT	1.46	0.46	25
Canaan Valley, WV	1.68	0.51	16
Goodwin Creek, MS	1.27	0.91	30
Audubon Grasslands, AZ	0.62	0.00	39
Arm Oklahoma, OK	2.36	0.43	29

3.1.2. Contribution of input variables on GPP variations

As measured by changes in cross validation error statistics, the removal of EVI caused the largest reduction in performance of the SVM GPP (Table 3): RMSE increased from 1.83 gC/m^2/day to 2.68 gC/m^2/day and R^2 decreased from 0.76 to 0.49.

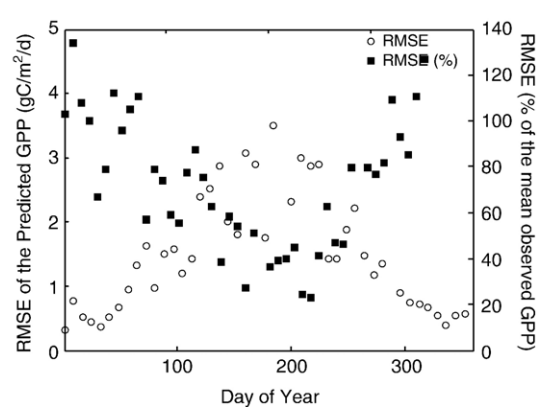

Fig. 4. Seasonal variations of GPP prediction errors averaged across all flux sites. Left axis shows RMSE error expressed as gC/m^2/day; right axis shows RMSE error expressed as a percent of average tower-based GPP.

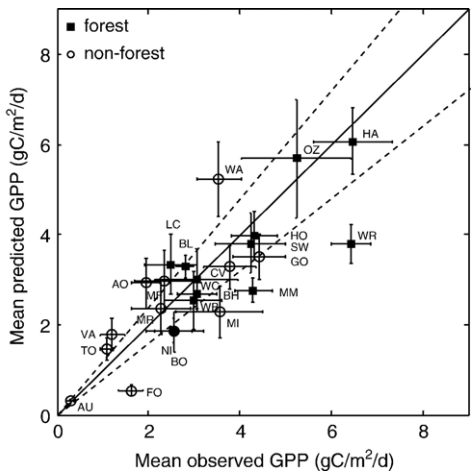

Fig. 5. Mean observed AmeriFlux GPP versus mean predicted SVM GPP in 2004 for forest and non-forest sites. Error bars are standard errors (defined as the standard deviation divided by the square root of the number of observations) of the observed and predicted 8-day average GPP. Solid line shows a 1:1 relationship. Dotted lines are the confidence limits at 20% range defined by mean predicted GPP = 0.8 × mean observed GPP for the lower limit and mean predicted GPP = 1.2 × mean observed GPP for the upper limit. Abbreviations of flux sites refer to Table 1.

Removal of LST produced comparatively minor changes: RMSE rose to 1.97 gC/m^2/day and R^2 fell to 0.72. The removal of land cover led to an increased RMSE of 1.91 gC/m^2/day and a decreased R^2 of 0.74. The removal of SWR had the smallest impact: the RMSE only increased by 0.06 gC/m^2/day and the R^2 decreased by 0.02. This is consistent with other studies where positive and statistically significant relationships have been reported between vegetation index and photosynthesis (Paruelo et al., 1997; Rahman et al., 2005; Sellers et al., 1992). However, our input variable ranking was based on the 8-day averages within 7 km × 7 km regions. Thus, potential users should be cautious on the relative importance of the input variables reported in this study because the importance may change with different spatial and temporal resolutions.

3.1.3. Comparison with MOD17 GPP

In terms of R^2 and RMSE calculated from 8-day values, SVM GPP outperformed MOD17 GPP in 2004 for both forest and non-forest sites. SVM GPP had 2.05 gC/m^2/day RMSE with an R^2 of 0.63 for non-forest sites and 1.63 gC/m^2/day RMSE with an R^2 of 0.79 for forest sites while MOD17 GPP had 2.71 gC/m^2/day RMSE with an R^2 of 0.39 for non-forest sites and 2.08 gC/m^2/day RMSE with an R^2 of 0.67 for forest sites. While both SVM GPP and MOD17 GPP performed better in forest sites than in non-forest sites, MOD17 GPP showed a systematic underestimation of the tower-based GPP for non-forest sites.

Based on annual data, both SVM GPP and MOD17 GPP again performed better in forest sites and SVM again outperformed MOD17 for non-forest sites but had similar performance for forest sites (Table 4). For SVM GPP in non-forest sites, the highest over-prediction occurred at Walnut River (77.6%) and the highest under-prediction occurred at Mead Irrigated (40.7%), while for forest sites, the highest over-prediction occurred at Blodgett (14.9%) and the highest under-prediction occurred at Wind River (41.9%). MOD17 GPP underestimated all sites except Blodgett, Audubon Grassland, Vaira Ranch and Tonzi Ranch with the highest underestimation of 66.1% at Mead Irrigated for non-forest sites and 31.7% at Harvard Forest for forest sites. Overall, annual SVM GPP had an error of 32.1% for non-forest sites and 22.2% for forest sites, while annual MOD17 GPP had an error of 50.3% for non-forest sites and 21.5% for forest sites.

Heinsch et al. (2006) reported a relative overestimation of 20.0–30.0% from MOD17 based on a comparison with 15 flux sites (mostly forest ecosystem). We speculate the difference may be due to different tower-based GPP calculations and different spatiotemporal coverage. For example, based on the

Table 3
The impact of removing one of the four input variables on the predicting performance of SVM on GPP

Variable removed	RMSE (gC/m^2/day)	R^2	C	ε	σ
None	1.83	0.76	14.929	0.063	3.732
LST	1.97	0.72	12.126	0.125	5.278
EVI	2.68	0.49	10.556	0.095	8.000
SWR	1.89	0.74	10.556	0.072	6.964
Land cover	1.91	0.74	12.126	0.065	6.964

The results shown are the average from three-way cross validation on the training set.

Table 4
Comparison of annual total GPP between tower-based calculation, SVM, and MOD17 for sites which had no less than thirty 8-day averages available in 2004

Name	Tower-based (gC/m^2/year)	SVM (gC/m^2/year)	Relative error (%) (SVM) [1]	MOD17 (gC/m^2/year)	Relative error (%) (MOD17) [1]
Forest					
Blodgett, CA	947	1088	14.9 (+)	1058	11.7 (+)
Niwot Ridge Forest, CO	869	650	25.2 (−)	742	14.6 (−)
Howland Forest, ME	1297	1208	6.9 (−)	1155	10.9 (−)
Harvard Forest, MA	1737	1900	9.4 (+)	1187	31.7 (−)
Metolius Intermediate Site, OR	1548	1013	34.6 (−)	1081	30.2 (−)
Wind River, WA	1909	1109	41.9 (−)	1340	29.8 (−)
Non-forest					
Walnut River, KS	990	1759	77.6 (+)	683	31.1 (−)
Bondville, IL	926	663	28.4 (−)	458	50.5 (−)
Mead Rainfed, NE	773	1017	31.6 (+)	514	33.5 (−)
Mead Irrigated, NE	1403	832	40.7 (−)	476	66.1 (−)
Mead Rotation, NE	786	851	8.2 (+)	481	38.9 (−)
Vaira Ranch, CA	474	686	44.9 (+)	602	27.1 (+)
Tonzi Ranch, CA	386	521	34.7 (+)	595	53.9 (+)
Goodwin Creek, MS	1650	1315	20.3 (−)	780	52.7 (−)
Audubon Grasslands, AZ	117	114	2.3 (−)	233	98.8 (+)

[1] Relative error is defined as (estimation − observation)/observation × 100.
(+) Indicate over-estimation.
(−) Indicate under-estimation.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

Table 5
Comparison of the predicted performance of EVI with NDVI and FPAR

	Non-forest		Forest	
	RMSE (gC/m^2/day)	R^2	RMSE (gC/m^2/day)	R^2
EVI	2.05	0.63	1.63	0.79
NDVI	2.27	0.57	1.77	0.74
FPAR	2.48	0.56	1.81	0.59

site-specific GPP provided by Harvard Forest, we calculated the annual tower-based GPP at Harvard Forest to be 1737 gC/m^2/year in 2004 (Table 4). However, Heinsch et al. (2006) reported an estimation of 1600 gC/m^2/year for the tower-based GPP at Harvard Forest in 2001–2002. We thus concluded the overestimation of MOD17 in the study of Heinsch et al. (2006) might be the result of their different calculation of the tower-based GPP and/or different spatiotemporal coverage from our study.

3.1.4. Comparison of the predictive performance of EVI with NDVI and FPAR

SVM GPP with EVI had 2.05 gC/m^2/day RMSE with an R^2 of 0.63 for non-forest sites, and 1.63 gC/m^2/day RMSE with an R^2 of 0.79 for forest sites on the 2004 AmeriFlux data (Table 5). Replacing EVI with NDVI yielded an RMSE 2.27 gC/m^2/day with an R^2 of 0.57 for non-forest sites, and an RMSE of 1.77 gC/m^2/day with an R^2 of 0.74 for forest sites. Replacing EVI with FPAR produced an RMSE of 2.48 gC/m^2/day with an R^2 of 0.56 for non-forest sites, and an RMSE of 1.81 gC/m^2/day with an R^2 of 0.59 for forest sites. Therefore we concluded that EVI had better predictive performance on GPP variations than NDVI or FPAR.

3.2. Generalization from AmeriFlux sites to the conterminous U.S.

The purpose of our conterminous U.S. GPP application was to assess whether or not the SVM technique produced spatio-temporal GPP estimates consistent with expected patterns. In the four 8-day periods in 2004 representing spring, summer, fall and winter over the conterminous U.S., the SVM model trained at the AmeriFlux sites generally captured the expected GPP features (Fig. 6). Temporally, April GPP was low because of low temperature and low radiation in the onset of growing seasons; July GPP was high because of high precipitation, peak vegetation, and intensive radiation; September GPP dropped as vegetation senesced; and December GPP was lowest with

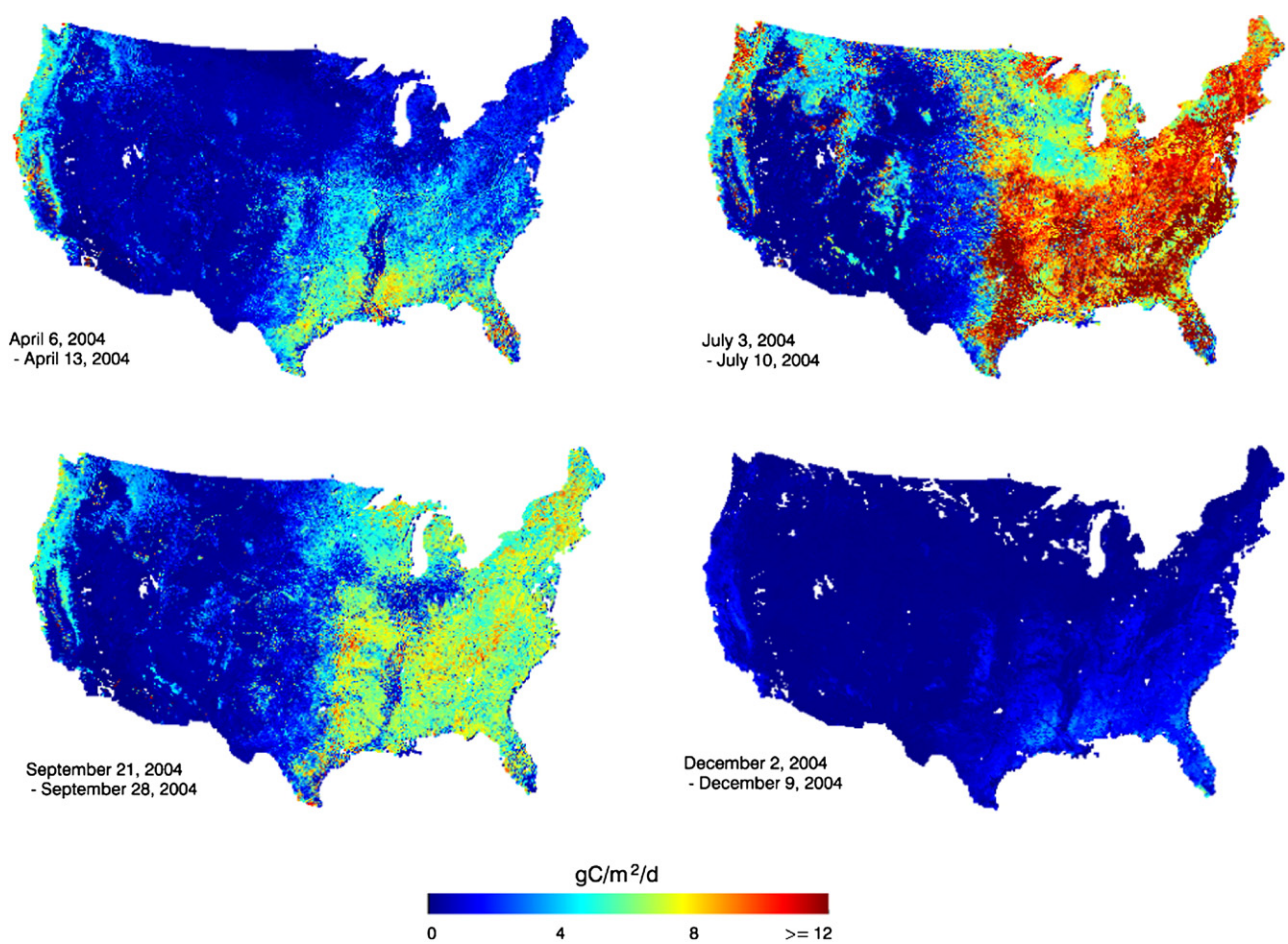

Fig. 6. Eight-day average SVM GPP for the conterminous U.S. for April 6–April 13, July 3–July 10, September 21–September 28, and December 2–December 9 of 2004. GPP greater than 12 gC/m^2/day was truncated to 12 gC/m^2/day for display purposes. Areas shown as white are either barren or contained missing data in one or more input variables (LST, EVI, SWR).

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

coldest temperatures and lowest radiation. Spatially, April GPP was high in California and Florida due to high temperature; July GPP was highest in the southeastern U.S., due to high moisture and energy availability; and September GPP showed patterns consistent with early phenological decline of agricultural regions in the Mississippi River valley and Midwest.

Examination of annual total GPP in 2004 showed that eastern deciduous forest, and eastern and southern mixed evergreen forest areas were the most productive regions in the U.S. (Fig. 7). The Great Lakes, New England mixed forest, Rocky Mountain evergreen forest and Pacific coastal evergreen forest were modest in productivity. Low GPP regions included the Great Plains (extending from eastern Montana and North Dakota to Texas) and the Great Basin, Sonoran, Mojave, and Chihuahua Deserts.

3.3. Estimation of e_{max} for the conterminous U.S.

The spatial patterns of $e_{\max}$ were strongly heterogeneous across the conterminous U.S. (Fig. 8). Most semi-arid regions had very low $e_{\max}$ (<0.5 gC/MJ) but arid regions of south-western U.S. had extremely high $e_{\max}$ exceeding 2.0 gC/MJ. While this finding may be due to difficulties in estimating FPAR and/or SVM GPP, $e_{\max}$ may also be adapted to optimize assimilation during short periods of optimal environmental conditions. Cropland regions and eastern forests had high $e_{\max}$ (>1.5 gC/MJ) (Fig. 8).

Estimated $e_{\max}$ in each land cover class (Table 6) was qualitatively consistent with ground-based observations (e.g. Gower et al., 1999). First, except for cropland, $e_{\max}$ was greater for forest than for non-forest. Second, $e_{\max}$ was greater for deciduous than for evergreen forests. Third, croplands had high $e_{\max}$. This agreement with ground observations suggests the overall credibility of our $e_{\max}$ estimation.

Comparison of $e_{\max}$ in the MOD17 GPP algorithm with our estimation strengthens the need for $e_{\max}$ refinement in the MOD17 model, especially in cropland regions (Table 6). The

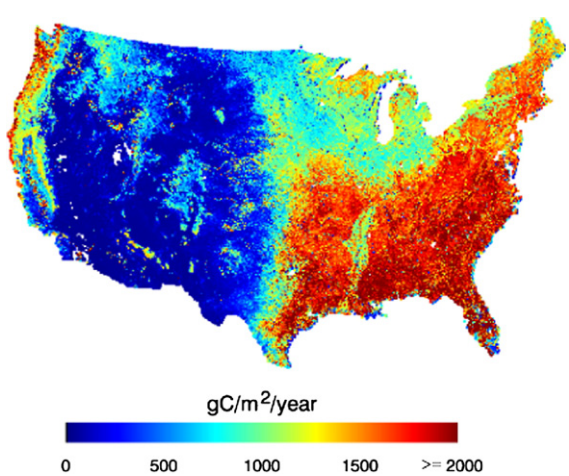

Fig. 7. Annual total SVM GPP for the conterminous U.S. in 2004. GPP greater than 2000 gC/m^2/year was truncated to 2000 gC/m^2/year for display purposes. Areas shown as white are either barren or contained missing data in one or more input variables (LST, EVI, SWR).

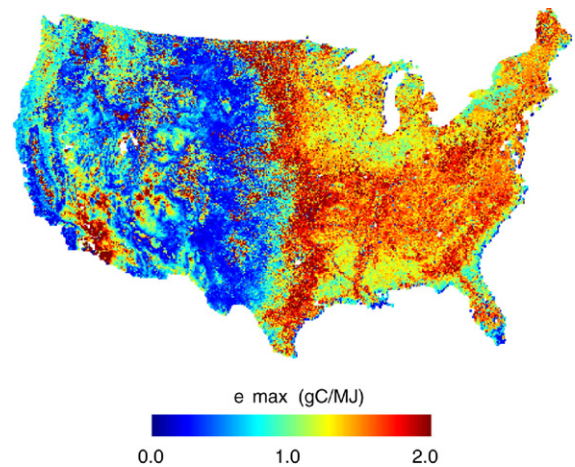

Fig. 8. Estimated maximum light use efficiency ($e_{\max}$) for the conterminous U.S. in 2004.

underestimation of MOD17 GPP products in non-forest regions could be explained by its low $e_{\max}$ values in grassland and cropland. Along with several studies which reported large uncertainties of $e_{\max}$ in the MOD17 GPP algorithm (e.g. Heinsch et al., 2006; Turner et al., 2005), our analysis has the potential for improving the ecophysiological parameter estimations in the MOD17 GPP algorithm.

3.4. Limitations and advantages of the proposed method

At least six factors may have limited the continental GPP generalization: (1) use of 0.5° resolution GOES-SRB SWR for the conterminous U.S. tended to smooth GPP variations; (2) use of 16-day composite EVI for 8-day periods also tended to smooth GPP; (3) the 36 flux sites included in this study may not fully represent the spatiotemporal variation of actual GPP; (4) the gap-filling method adopted in this study might not be appropriate for all sites; (5) canopy storage flux uncorrected may introduce errors in tower GPP; and (6) the tower-based GPP may have measurement errors jeopardizing the model generalization ability. Additional flux data would mitigate some of these issues, as would finer resolution EVI (temporally) and SWR (spatially).

The 36 AmeriFlux sites (Table 1) included in this study represent a wide variety of vegetation types and climate regimes. Depending on tower heights and site topography, the footprints

Table 6
Estimated and MODIS GPP algorithm based $e_{\max}$ (gC/MJ) in each vegetation class

Land cover	This study	MOD17 GPP
ENF (Evergreen needleleaf forest)	1.02 (0.40)	1.01
DBF (Deciduous broadleaf forest)	1.56 (0.36)	1.16
MF (Mixed forest)	1.31 (0.36)	1.12
SH, SV (Closed/open shrubland and woody savannas and savannas)	0.79 (0.59)	0.81
GL (Grassland)	0.86 (0.73)	0.68
CL (Cropland and cropland/natural vegetation mixture)	1.47 (0.53)	0.68

() shows its standard deviations.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016

of flux sites can vary from a hundred meters to several thousand meters (Kim et al., 2006). In our SVM GPP analysis, we chose a 7 km × 7 km region as the spatial representativeness for each flux site. Such simplification ignored landscape heterogeneity and could lead to inadequate representation of the nonlinearity of vegetation photosynthesis and respiration processes. The influence of vegetation structure and site topography on the selection of appropriate scales for SVM GPP analysis deserves more research.

Our GPP estimation method provides a potentially powerful and independent tool for terrestrial carbon cycle studies. First, because our method is driven by satellite data, it can possibly monitor more spatially detailed patterns in surface heterogeneity than can process-based models. Second, we can use the SVM GPP for independent tests of ecosystem models because it does not require surface meteorological data. Third, the spatiotemporal variations in data-driven GPP can potentially give insights for ecosystem sciences and processes, as demonstrated here with the spatial inversion of e_{max} for optimization of LUE models.

In summary, although there are potentially confounding factors and it is difficult to validate GPP distribution over the conterminous U.S., our results show that the SVM trained at AmeriFlux sites generally captured the expected spatiotemporal variations of GPP over the conterminous U.S. This result, although not quantitatively conclusive, strongly suggests that SVM models trained at flux sites can be generalized to larger regions.

4. Conclusions

Using a combination of tower-based GPP, machine learning techniques, and remotely sensed inputs, we developed a technique to predict GPP over the conterminous U.S. with an average test error of 1.87 gC/m^2/day and a method to estimate e_{max} for the conterminous U.S. We found that EVI was the most important factor for GPP prediction, suggesting that finer temporal resolution for EVI (possibly by including Terra and Aqua satellites) could substantially enhance regional and/or continental GPP estimates. The method can also be improved by increasing the number of GPP ground observations. A central limitation of the SVM technique is that the knowledge learned is encoded as weights that are not directly comprehensible to humans. However, methods exist with which to convert the structure learned by SVMs to a more understandable format, such as rules, thus enhancing our understanding of GPP processes at different scales. With these improvements, the combination of satellite data with ground observation of GPP through machine learning should be able to provide prediction of GPP with sufficient accuracy and timeliness for application in regional to continental natural resource management services and may serve as a supplemental tool for existing GPP retrieval methods. Furthermore, based on the product accuracy and the spatiotemporal scale considered in this study, the SVM based GPP prediction can also be useful for documenting hydro-ecological models and for improving terrestrial biosphere models at regional to continental scales.

Acknowledgement

The SVM software from LIBSVM (Chang & Lin, 2005) facilitated this study. M.A. White was supported by the NASA New Investigator Program. NASA's Science Mission Directorate funded part of this research through EOS and REASon grants to RRN. The funding from the Chinese Academy of Sciences through the One-Hundred Talents Program and the Chinese Academy of Sciences International Partnership Project (Human Activities and Ecosystem Changes (Project Number: CXTD-Z2005-1)) to A-Xing Zhu was also appreciated. AmeriFlux was funded by the Department of Energy, the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF). Special thanks to all scientists and supporting staffs at AmeriFlux sites. Finally, we acknowledged three anonymous reviewers for their comments.

References

Baldocchi, D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, *9*, 479.

Baldocchi, D., Falge, E., Gu, L. H., Olson, R., Hollinger, D., Running, S., et al. (2001). FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, *82*, 2415−2434.

Campbell, G. S., & Norman, J. M. (1998). *An introduction to environmental biophysics.* Springer, New York 286 pp.

Chang, C. -C., & Lin, C. -J. (2005). *LIBSVM—A library for Support Vector Machines.* http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chapin, F. S., Mooney, H. A., Chapin, M. C., & Matson, P. (2004). *Principles of terrestrial ecosystem ecology.* Springer.

Cook, R. B., Margle, S. M., Holladay, S. K., Heinsch, F. A., & Schaaf, C. B. (2004). Subsets of remote sensing products for AmeriFlux sites: MODIS ASCII Subsets, Ameriflux Annual Meeting, Boulder, CO.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press 189 pp.

Drolet, G. G., Huemmrich, K. F., Hall, F. G., Middleton, E. M., Black, T. A., Barr, A. G., et al. (2005). A MODIS-derived photochemical reflectance index to detect inter-annual variations in the photosynthetic light-use efficiency of a boreal deciduous forest. *Remote Sensing of Environment*, *198*, 212−224.

Escudero, A., & Mediavilla, S. (2003). Decline in photosynthetic nitrogen use efficiency with leaf age and nitrogen resorption as determinants of leaf life span. *Journal of Ecology*, *91*(5), 880.

Falge, E., Baldocchia, D., Olsonb, R., Anthonic, P., Aubinetd, M., Bernhofere, C., et al. (2001). Gap filling strategies for long term energy flux data sets. *Agricultural and Forest Meteorology*, *107*(1), 71−77.

Friedl, M. A. (2003). *Validation of the consistent-year V003 MODIS land cover product.* Internal PI document. http://www-modis.bu.edu/landcover/userguidelc/consistent.htm

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., et al. (2002). Global land cover mapping from MODIS: Algorithms and early results. *Remote Sensing of Environment*, *83*(1–2), 287−302.

Gao, X., Huete, A. R., & Didan, K. (2003). Multisensor comparisons and validation of MODIS vegetation indices at the semiarid Jornada Experimental Range. *IEEE Transactions on Geoscience and Remote Sensing*, *41* (10), 2368−2381.

Gilmanov, T. G., Tieszen, L. L., Wylie, B. K., Flanagan, L. B., Frank, A. B., Haferkamp, M. R., et al. (2005). Integration of CO_2 flux and remotely-sensed data for primary production and ecosystem respiration analyses in the

Northern Great Plains: Potential for quantitative spatial extrapolation. *Global Ecology and Biogeography*, *14*(3), 271−292.

Goulden, M. L., Munger, J. W., Fan, S., Daube, B. C., & Wofsy, S. C. (1996). Exchange of carbon dioxide by a deciduous forest: Response to interannual climate variability. *Science*, *271*(5255), 1576−1578.

Gower, S. T., Kucharik, C. J., & Norman, M. (1999). Direct and indirect estimation of leaf area index, fapar, and net primary production of terrestrial ecosystems. *Remote Sensing of Environment*, *70*, 29−51.

Gu, L., Baldocchi, D., Verma, S. B., Black, T. A., Vesala, T., Falge, E. M., et al. (2002). Advantages of diffuse radiation for terrestrial ecosystem productivity. *Journal of Geophysical Research (Atmospheres)*, *107*(D6), 4050. doi:10.1029/2001JD001242

Gu, L., Falge, E. M., Boden, T., Baldocchi, D. D., Black, T. A., Saleska, S. R., et al. (2005). Objective threshold determination for nighttime eddy flux filtering. *Agricultural and Forest Meteorology*, *128*, 179−197.

Hargrove, W. W., Hoffman, F. M., & Law, B. E. (2003). New analysis reveals representativeness of the AmeriFlux network. *EOS Transactions AGU*, *84* (48), 529.

Haykin, S. (1998). *Neural networks: A comprehensive foundation.* Prentice Hall 842 pp.

Heinsch, F. A., Zhao, M., Running, S. W., Kimball, J. S., Nemani, R. R., Davis, K. J., et al. (2006). Evaluation of remote sensing based terrestrial productivity from MODIS using regional tower eddy flux network observations. *IEEE Transactions on Geoscience and Remote Sensing*, *44* (7), 1908−1925.

Huete, A., Didan, K., Miura, T., & Rodriguez, E. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1–2), 195−213.

Jolly, W. M., Graham, J. M., Michaelis, A., Nemani, R. R., & Running, S. W. (2005). A flexible, integrated system for generating meteorological surfaces derived from point sources across multiple geographic scales. *Environmental Modeling & Software*, *20*, 873−882.

Kanzow, C., Yamashita, M., & Fukushima, M. (2004). Levenberg–Marquardt methods for constrained nonlinear equations with strong local convergence properties. *Journal of Computational and Applied Mathematics*, *172*, 375−397.

Kim, J., Guo, Q., Baldocchi, D. D., Leclerc, M. Y., Xu, L., & Schmid, H. P. (2006). Upscaling fluxes from tower to landscape: Overlaying flux footprints on high-resolution (IKONOS) images of vegetation cover. *Agricultural and Forest Meteorology*, *136*(3–4), 132−146.

Metz, B., Davidson, O., Coninck, H. d., Loos, M., & Meyer, L. (2006). *Special report of the intergovernmental panel on climate change, Intergovernmental Panel on Climate Change (IPCC).*

Misson, L., Gershenson, A., Tang, J., Boniello, R., McKay, M., Cheng, W., et al. (2006). Influences of canopy photosynthesis and summer rain pulses on root dynamics and soil respiration in a young ponderosa pine forest. *Tree Physiology*, *26*, 833−844.

Monteith, J. (1972). Solar radiation and productivity in tropical ecosystems. *Journal of Applied Ecology*, *9*, 747−766.

Myneni, R. B., Knyazikhin, Y., Privette, J. L., & Glassy, J. (2002). Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, *83*(1–2), 214−231.

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., et al. (2003). Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *Science*, *300*(5625), 1560−1563.

Paruelo, J. M., Epstein, H. E., Lauenroth, W. K., & Burke, I. C. (1997). ANPP estimates from NDVI for the central grassland region of the United States. *Ecology*, *78*, 953−959.

Pinker, R. T., Laszlo, O., Tarpley, J. D., & Mitchell, K. (2002). Geostationary satellite parameters for surface energy balance. *Advances in Space Research*, *30*(11), 2427−2432.

Pinker, R. T., Tarpley, J. D., Laszlo, I., Mitchell, K. E., Houser, P. R., Wood, E. F., et al. (2003). Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research*, *108*, 8798. doi:10.1029/2002JD003301

Rahman, A. F., Sims, D. A., Cordova, V. D., & El-Masri, B. Z. (2005). Potential of MODIS EVI and surface temperature for directly estimating per-pixel ecosystem C fluxes. *Geophysical Research Letters*, *32*, L19404. doi:10.1029/2005GL024127

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., & Hashimoto, H. (2004). A continuous satellite-derived measure of global terrestrial primary productivity. *BioScience*, *54*(6), 547−560.

Running, S. W., R.R., N., Glassy, J. M., & Thornton, P. E. (1999). *MODIS daily photosynthesis (PSN) and annual net primary production (NPP) product (MOD17).* Algorithm theoretical basis documents. http://www.ntsg.umt.edu/modis/ATBD/ATBD_MOD17_v21.pdf

Sasai, T., Ichii, K., Nemani, R. R., & Yamaguchi, Y. (2005). Simulating terrestrial carbon fluxes using the new biosphere model "biosphere model integrating eco-physiological and mechanistic approaches using satellite data" (BEAMS). *Journal of Geophysical Research*, *110*, G02014. doi:10.1029/2005JG000045

Sellers, P. J., Berry, J. A., Collatz, G. J., Field, C. B., & Hall, F. G. (1992). Canopy reflectance, photosynthesis, and transpiration III. A reanalysis using improved leaf models and a new canopy integration scheme. *Remote Sensing of Environment*, *42*, 187−216.

Thornton, P. E. (1998). *Regional ecosystem simulation: Combining surface- and satellite-based observations to study linkages between terrestrial energy and mass budgets.* Missoula, MT: The University of Montana 280 pp.

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, *190*, 214−251.

Turner, D. P., Ritts, W. D., Cohen, W. B., Gower, S. T., Running, S. W., Zhao, M., et al. (2006). Evaluation of MODIS NPP and GPP products across multiple biomes. *Remote Sensing of Environment*, *102*, 282−292.

Turner, D. P., Ritts, W. D., Cohen, W. B., Gower, S. T., Zhao, M., Running, S. W., et al. (2003). Scaling gross primary production (GPP) over boreal and deciduous forest landscapes in support of MODIS GPP product validation. *Remote Sensing of Environment*, *88*, 256−270.

Turner, D. P., Ritts, W. D., Cohen, W. B., Maeirsperger, T. K., Gower, S. T., Kirschbaum, A. A., et al. (2005). Site-level evaluation of satellite-based global terrestrial GPP and NPP monitoring. *Global Change Biology*, *11*(4), 666−684.

Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley 736 pp.

Vapnik, V., & Chervonenkis, A. (1991). The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, *1*(3), 283−305.

Wan, Z., Zhang, Y., Zhang, Q., & Li, Z. (2002). Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. *Remote Sensing of Environment*, *83*(1–2), 163−180.

Wan, Z., Zhang, Y., Zhang, Q., & Li, Z. (2004). Quality assessment and validation of the MODIS global land surface temperature. *International Journal of Remote Sensing*, *25*(1), 261−274.

White, M. A., Thornton, P. E., Running, S. W., & Nemani, R. R. (2000). Parameterization and sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: net primary production controls. *Earth Interactions*, *4*, 1−85.

Wylie, B., Johnson, D., Laca, E., Saliendra, N., Gilmanov, T., & Reed, B. (2003). Calibration of remotely sensed, coarse resolution NDVI to CO_2 fluxes in a sage–brush–steppe ecosystem. *Remote Sensing of Environment*, *85*, 243−255.

Xiao, X., Zhang, Q., Hollinger, D., Aber, J., & Moore, B. (2005). Modeling gross primary productivity of an evergreen needleleaf forest using MODIS and climate data. *Ecological Applications*, *15*(3), 954−969.

Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., et al. (2006). Prediction of continental scale evapotranspiration by combining MODIS and AmeriFlux data through Support Vector Machine. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(11), 3452−3461.

Zhan, H., Shi, P., & Chen, C. (2003). Retrieval of oceanic chlorophyll concentration using Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(12), 2947−2951.

Zhao, M., Heinsch, F. A., Nemani, R. R., & Running, S. W. (2005). Improvement of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment*, *95*, 164−176.

Zhu, A. X., & Scott, M. D. (2001). Effects of spatial detail of soil information on watershed modeling. *Journal of Hydrology*, *248*(1–4), 54−77.

Please cite this article as: Yang, F. et al. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* (2007), doi:10.1016/j.rse.2007.02.016