Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment 106 (2007) 399–413
Remote Sensing of Environment
www.elsevier.com/locate/rse

Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains

Bruce K. Wylie [a], Eugene A. Fosnight [a,*], Tagir G. Gilmanov [b], Albert B. Frank [c], Jack A. Morgan [d], Marshall R. Haferkamp [e], Tilden P. Meyers [f]

[a] *Science Applications International Corporation (SAIC), U.S. Geological Survey (USGS) Center for Earth Resources Observation and Science (EROS), Sioux Falls, SD 57198, USA*
[b] *South Dakota State University, Brookings, SD 57007, USA*
[c] *U.S. Department of Agriculture, Agricultural Research Service, Northern Great Plains Research Laboratory, Mandan, ND 58554, USA*
[d] *U.S. Department of Agriculture, Agricultural Research Service, Fort Collins, CO 80526, USA*
[e] *U.S. Department of Agriculture, Agricultural Research Service, Miles City, MT 59301, USA*
[f] *National Oceanic and Atmospheric Administration, Air Resources Laboratory, Oak Ridge, TN 37830, USA*

Received 8 November 2005; received in revised form 5 September 2006; accepted 9 September 2006

Abstract

Rangeland carbon fluxes are highly variable in both space and time. Given the expansive areas of rangelands, how rangelands respond to climatic variation, management, and soil potential is important to understanding carbon dynamics. Rangeland carbon fluxes associated with Net Ecosystem Exchange (NEE) were measured from multiple year data sets at five flux tower locations in the Northern Great Plains. These flux tower measurements were combined with 1-km^2 spatial data sets of Photosynthetically Active Radiation (PAR), Normalized Difference Vegetation Index (NDVI), temperature, precipitation, seasonal NDVI metrics, and soil characteristics. Flux tower measurements were used to train and select variables for a rule-based piece-wise regression model. The accuracy and stability of the model were assessed through random cross-validation and cross-validation by site and year.

Estimates of NEE were produced for each 10-day period during each growing season from 1998 to 2001. Growing season carbon flux estimates were combined with winter flux estimates to derive and map annual estimates of NEE. The rule-based piece-wise regression model is a dynamic, adaptive model that captures the relationships of the spatial data to NEE as conditions evolve throughout the growing season. The carbon dynamics in the Northern Great Plains proved to be in near equilibrium, serving as a small carbon sink in 1999 and as a small carbon source in 1998, 2000, and 2001. Patterns of carbon sinks and sources are very complex, with the carbon dynamics tilting toward sources in the drier west and toward sinks in the east and near the mountains in the extreme west. Significant local variability exists, which initial investigations suggest are likely related to local climate variability, soil properties, and management.
Published by Elsevier Inc.

Keywords: Carbon flux; Grassland ecosystems; Northern Great Plains; Data-driven models; Piece-wise regression models; Net Ecosystem Exchange

1. Introduction

Grassland systems, faced with large-scale agricultural conversions, are some of the most altered systems in the world (Butcher, 2004; White et al., 2000; WRI, 2000). Rangelands make up 40% of the Earth's surface (WRI, 2000) within which temperate grasslands contain about 18% of global carbon reserves (Burke et al., 1997). As future demands on ecosystems increase, the value of ecosystem services, including carbon mitigation and ecosystem health will also continue to increase (Costanza et al., 1997).

This study describes an adaptive data-driven piece-wise regression methodology to estimate Net Ecosystem Exchange (NEE) at 10-day time steps during the growing season. These estimates are summed and added to winter flux estimates to create 1-km resolution maps of annual carbon fluxes for the ecoregion. An analysis of the spatial patterns and responses

* Corresponding author.
E-mail address: fosnight@usgs.gov (E.A. Fosnight).

0034-4257/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2006.09.017

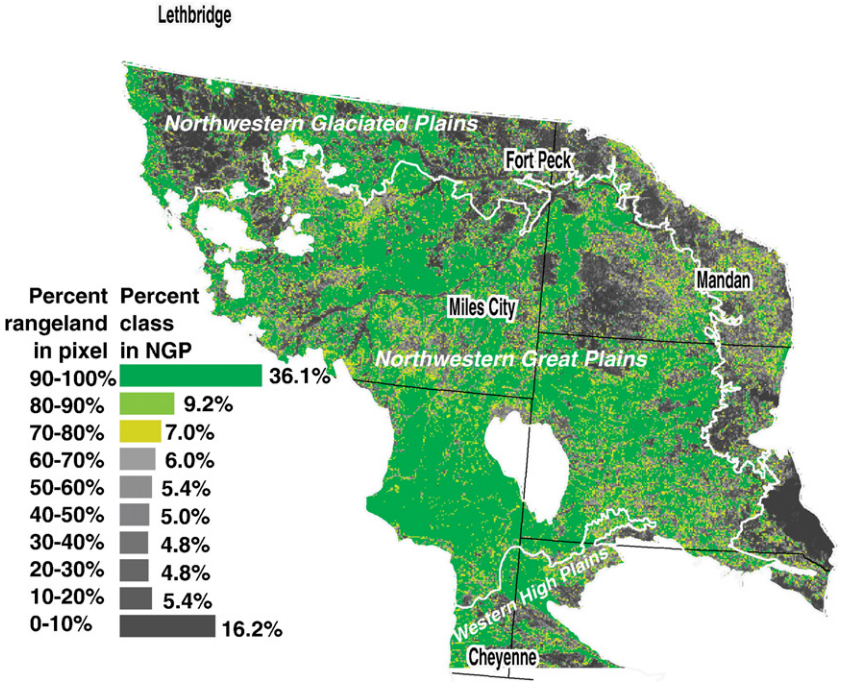

Fig. 1. Distribution of grass and shrub lands in the Northern Great Plains. The boundaries of the ecosystems are in white. The locations of the five flux towers are shown.

through time will lead to a better understanding of climatic variability and land management practices on the Net Ecosystem Exchange of carbon.

The Northern Great Plains grassland ecosystem for this study (see Fig. 1) includes the Northwestern Glaciated Plains, Northwestern Great Plains and the Western High Plains north of 41° N latitude (CEC, 1997; McMahon et al., 2001; Omernik, 1987). The Northern Great Plains comprises a transition from moister and more intensive agricultural regions to the east to dryer lands dominated in the south and west by native grasslands, where agriculture is controlled by access to irrigation. The boundary between the Northwestern Great Plains and the Western High Plains marks the transition between blue grama-buffalo grass (*Bouteloua gracilis*) and winter wheat (*Triticum aestivum*) to the south and mostly wheatgrass–needlegrass (*Pascopyrum smithii–Stipa spp.*) and spring wheat to the north (EPA, 2005). How will Northern Great Plains rangelands respond to predicted increases in winter precipitation and drier summer conditions and increased weather variability (Wigley, 1999) or the introduction of new varieties of drought resistant crops (Higgins et al., 2002) within this highly variable and responsive ecological system? Methodologies described in this paper will assist in identifying how these causal variables are reflected in changing carbon dynamics.

Micro-meteorological flux towers improve our understanding of ecosystem responses to climate and quantify carbon dynamics locally at great detail. Continental (Wofsy & Harriss, 2002) and international programs (Cihlar et al., 2003) have prioritized the scaling up of localized flux tower measurements to identify, monitor, and understand carbon sink and source areas. The relationships between grassland CO_2 and spectral vegetation indices (e.g., Bartlett et al., 1990; Churkina et al., 2005; Gilmanov et al., 2005; Wylie et al., 2004) provide opportunities for scaling up localized tower measurements to larger geographical areas.

Carbon absorbed and released as a result of biological activity needs to be summed throughout the ecoregion to quantify biological carbon sinks and sources. However, carbon fluxes can only be directly measured for approximately 1-km fetch areas, and the cost of direct measurement limits the number of locations that can be measured. The key to understanding ecosystem carbon dynamics lies in discovering robust relationships between detailed knowledge collected at representative local sites and spatial data that describe the entire ecoregion.

Two complementary approaches are possible to quantify the relationships between flux tower measurements and spatial data. The first approach is to define theoretical biophysical models of carbon dynamics, to adapt these models to available spatial data, and to calibrate and validate the models using flux tower measurements. The second approach, described in this paper, is to develop data-driven models at the flux towers using tower measurements and spatial data measurements at the tower. These data-driven models are evaluated in regard to known vegetation physiology and are then applied across the ecosystem.

This paper describes (1) the spatial and tower data, (2) the development of a data-driven model, (3) techniques to assess the robustness of the model, and (4) the results of applying the model to estimate NEE across the entire ecoregion. The spatial variables must be able to quantify carbon fluxes in a manner that can be justified given known physical characteristics of carbon dynamics. To achieve robust estimates, the spatial distribution of the flux towers and the years sampled at the towers must adequately sample the variability of the environmental extremes in the

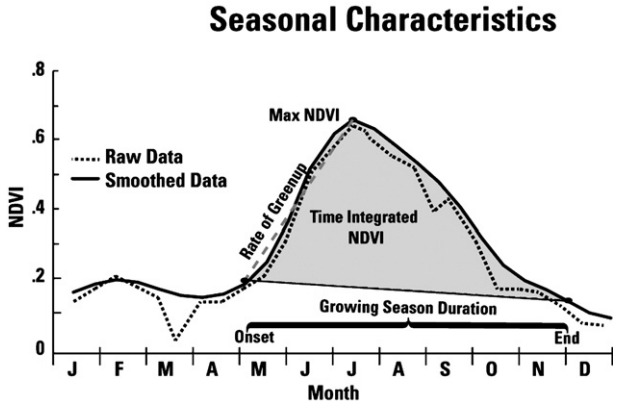

Fig. 2. Seasonal parameterization of the NDVI annual time series (after Reed et al., 1994).

ecoregion. These conditions must be satisfied to confidently scale measurements at flux towers to estimate annual carbon fluxes for entire ecoregions.

The approach defined below creates a regional database of Net Ecosystem Exchange that can be used to investigate within and among year patterns of carbon flux.

2. Data

2.1. Spatial data sets

Carbon fluxes can be described as a function of vegetation type and condition, soil characteristics, and climate. To scale from flux towers to regions requires that appropriate regional spatial data representing this information be available to model the variation in carbon flux. These spatial data must have both explanatory and predictive power to create regional carbon flux maps for use in identifying, explaining, and quantifying carbon sinks and sources. The resulting regional NEE database is in an Albers map projection with 1-km pixels and has an NEE layer for every 10 days during the growing season.

The National Land Cover Database (NLCD), derived from the Landsat Thematic Mapper (TM) (Homer et al., 2004), was used to identify the 1-km pixels for which at least 70% of the 30-m NLCD pixels were classified as grass or shrub lands (Fig. 1). 52.3% of the pixels in the Northern Great Plains are classed as grass or shrub lands. The NLCD has well documented accuracy characteristics (Wickham et al., 2004).

Vegetation productivity was estimated from derivative products created from the SPOT VEGETATION sensor (JRC, 2003). The Normalized Difference Vegetation Index was selected to monitor change in vegetation productivity, including monitoring the effect of water and temperature stress and changes in land management or land use (Tucker, 1979).

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{nir}} - \rho_{\mathrm{red}}}{\rho_{\mathrm{nir}} + \rho_{\mathrm{red}}}, \text{ where } \begin{cases} \rho_{\mathrm{red}} = \text{red band} \\ \rho_{\mathrm{nir}} = \text{near infrared band} \end{cases}$$

NDVI correlations with grassland biophysical parameters such as green leaf area index, green fPAR, and green biomass have been demonstrated (Cayrol et al., 2000; Wylie et al., 2002). The 1-km SPOT VEGETATION NDVI daily data are aggregated to 10-day composites using maximum NDVI compositing techniques to minimize the effect of off-nadir pixels and atmospheric attenuation. The 10-day composites were filtered temporally with a weighted least-squares approach to further minimize the effect of atmospheric attenuation and to allow the derivation of seasonal characteristics (Swets et al., 1999).

Seasonal characteristics were derived from annual smoothed NDVI time series (Reed, 2006; Reed et al., 1994). Phenologic parameters selected for use in the model were the NDVI value at onset of season (SOSN), the number of days from onset of season (SSOST), the date of onset of season (SOST), and total integrated NDVI (TIN) (Fig. 2). Day of year (DOY) and days from summer solstice (SOLS) were considered in addition to SSOST and SOST as means to incorporate time of year into the model. The parameters extracted from the NDVI time series, as do all of the variables, introduce uncertainties into the model. Of particular concern is the sensitivity of start of the season to snow cover in northern regions and sparse vegetation with associated aerosols in arid and semi-arid regions. Nonetheless the phenology extracted from remotely sensed data has the advantage characterizing biologically significant parameters at relatively high spatial resolution.

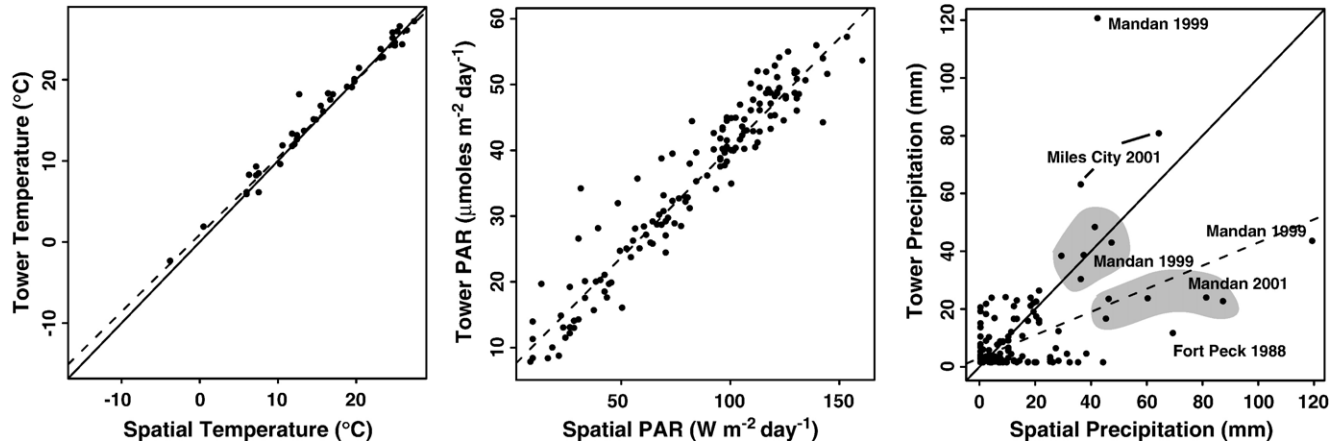

Fig. 3. Comparison of temperature (a), PAR (b), and precipitation (c) measured at the tower versus estimates extracted from the spatial database for the tower.

Table 1
Flux tower descriptions (derived from Gilmanov et al., 2005)

Site years	Ecosystem	Elevation (m)	Precipitation (mm)	Temperature mean Jan (°C) Mean Jul (°C)	Network	Sensor
Lethbridge 2000–2001	Northern mixed-short grass prairie	960	378	−8.6 18.0	Ameriflux	Eddy-covariance
Fort Peck 2000	Northern mixed Prairie	634	310	−11.9 18.0	Ameriflux	Eddy-covariance
Miles City 2000–2001	Northern mixed Prairie	719	343	−8.7 23.5	Agriflux	BREB
Mandan 1999–2001	Mixed prairie	518	404	−12.2 21.2	Agriflux	BREB
Cheyenne 1998	Mixed prairie	1910	397	−2.5 17.5	Agriflux	BREB

Climate variables are derived from satellite imagery calibrated with ground weather information. Daily Photosynthetically Active Radiation (PAR) (Frouin & Pinker, 1995), daily total precipitation, and daily average temperature (CDC, 2005; Xie & Arkin, 1996) were aggregated to 10-day summaries. These climate variables are very influential with respect to carbon dynamics. Regional estimates of temperature and PAR agree closely with measurements at the flux towers (Fig. 3a and b). These variables tend to vary smoothly through space. Precipitation, which is among the most critical variables for NEE estimates, measured at the tower and from the regional data are not highly correlated (Fig. 3c). Precipitation, frequently a localized event (Frank, 2003), is not well represented in the regional precipitation spatial data set. As a consequence, variation in precipitation is, in effect, captured indirectly by the NDVI data.

The mean temperature varies from a low of 4 °C in the north to a high of 13 °C in the south. The rainfall ranges from annual minimums of 26.9 cm in the northwest to annual maximums of 47.5 cm in the southeast. Snowfall ranges from annual minimums of 71.9 cm in the northeast and annual minimums of 133.6 cm in the west. The climate variables are modeled at 0.25° or nominally 30 km. The coarse resolution of these data introduces significant uncertainties into a 1-km resolution model.

Four variables-percent grass (PCTGRASS), percent C_4 (C4PCT), percent surface clay content (PCTCLAYSURF), and percent surface CO_3 (CO3SURF)—from the State Soil Geographic (STATSGO) Data Base were evaluated for inclusion in the NEE model (USDA, 1995). Percent grass and percent C_4 capture soil related vegetation potential. Percent surface clay content provides a measure of water holding capacity, and percent surface CO_3 quantifies soil inorganic carbon. The soils database and its attributes are a one-time estimate mapped for STATSGO at 1:250,000 with a minimum mapping unit of 6.25 km^2 and only 100–200 delineations per quad (USDA, 1995). Many inclusions exist and most delineations are large and generalized. The percent grass and percent C_4 represent a single snapshot in time.

2.2. Flux towers

Data collected and research conducted at flux towers make significant contributions to understanding and quantifying carbon dynamics. Among the networks of flux towers that monitor carbon dynamics are the Ameriflux (Running et al., 1999) and Agriflux (Svejcar et al., 1997) networks. The U.S. Department of Agriculture (USDA) Agriflux network of 12 locations, some with multiple sites, collects information quantifying the effects of environmental conditions and agricultural management decisions on carbon exchange between the land and atmosphere. AmeriFlux is a research network used in collecting, synthesizing, and disseminating long-term measurements of CO_2, water, and energy exchange for a variety of terrestrial landscapes across the United States and throughout the Americas. There are about 130 AmeriFlux sites with data available through the FLUXNET "network of regional networks" (http://www.fluxnet.ornl.gov/fluxnet/index.cfm), and about half of them have been in operation for 5 years or longer; a few sites have data records of 10 years or longer. The AmeriFlux network is led by the Department of Energy (DOE) with joint support from the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), USDA, National Science Foundation (NSF), and U.S. Geological Survey (USGS) (DOE, 2003).

The five flux towers in the region provide data and research for use in model development and validation (Table 1). Both the Bowen ratio energy balance (BREB) approach (Raupach, 1988) and the eddy covariance (EC) approach (Baldocchi, 2003; Moncrieff et al., 1997) are used to quantify water and carbon fluxes at the Agriflux towers, while the eddy covariance approach is used at all Ameriflux towers. The Bowen ratio technique is limited to low stature vegetation (Angell et al., 2001). Both the Bowen ratio (Angell et al., 2001) and the eddy covariance (Flanagan & Johnson, 2005) techniques have been shown to agree with chamber measurements. Some uncertainty is introduced through the pooling of the EC and BREB measurements. Ongoing research by Morgan and others is designed to quantify this uncertainty (Morgan, 2006).

These rangeland flux towers are well distributed throughout the Northern Great Plains (Fig. 1). The Northern Great Plains extends from Colorado to Alberta. The Fort Peck, Miles City, and Mandan towers lie within the heart of the ecoregion. The Lethbridge and Cheyenne towers provided important samples to represent and constrain northern and southern limits. The Lethbridge data were used to construct and assess the model, but the model could not be

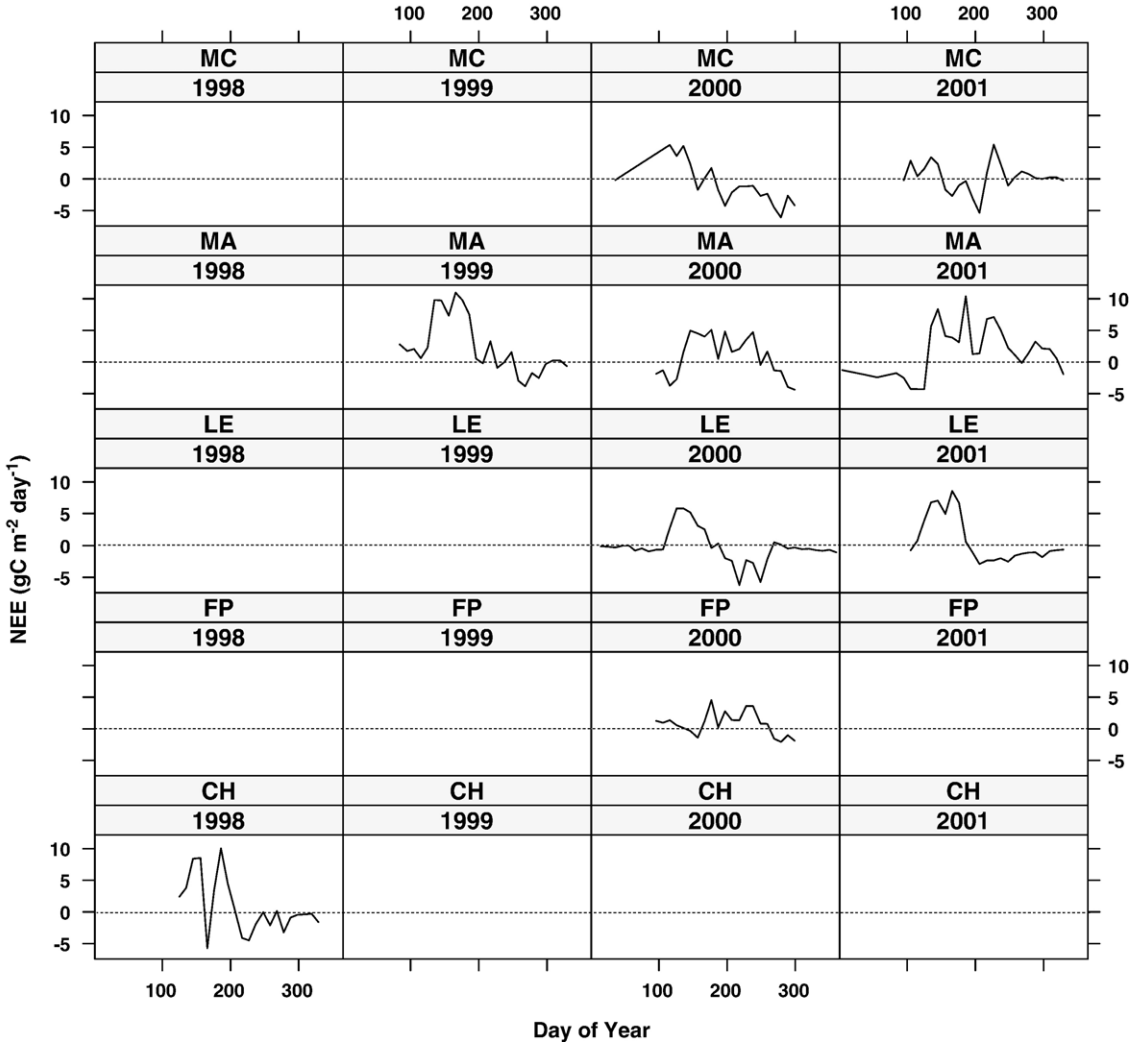

Fig. 4. Net Ecosystem Exchange measured at flux towers: Miles City (MC), Mandan (MA), Lethbridge (LE), Fort Peck (FP), and Cheyenne (CH).

scaled up for the Canadian segment of the ecoregion since some of the spatial data were not available for Canada.

Fig. 4 shows the years and locations for which NEE measurements are available for 1998 through 2001. Most of the flux tower data were for the 2000 and 2001 growing seasons—7 of the 9 site years. Since data for 1998 were only available at Cheyenne, and Cheyenne had data for 1998 only, the Cheyenne tower represents both a spatial and temporal extreme.

An inspection of the NEE time series (Fig. 4) shows that substantial variability exists among years and sites. As these measurements are used to train the regional model, we hypothesize that the variation among the tower-years used sufficiently bound the variation in the climatic conditions for the ecoregion. Meyers (2001) stressed the importance of including drought and wet years in carbon models to account for climate variability extremes.

2.3. Flux tower data analysis

Flux towers provide detailed, but localized measurements, of CO_2 fluxes between the atmosphere and the land surface, summarized at a 20- or 30-min time steps. The carbon flux estimates have a growing season and a dormant season component. Whether an area is a source or a sink is dictated by the balance between growing season fluxes dominated by photosynthesis and dormant season fluxes dominated by respiration (Frank et al., 2002; Haferkamp & MacNeil, 2004). Rangeland systems tend toward equilibrium (Baron et al., 2006; Dugas et al., 1999; Frank, 2002; Frank et al., 2002; Sims & Singh, 1978), though water availability resulting from climatic variation and management of the land can drive this balance one way or the other (Gilmanov et al., 2005).

Gilmanov et al. (2005) described the methodology to fill data gaps and integrate the flux measurements to derive 10-day average estimates of NEE (Fig. 5a). At each flux tower, many environmental variables are collected including temperature, Photosynthetically Active Radiation (PAR), and precipitation (Fig. 5c d and e). Precipitation at the Mandan flux tower typify the temperate continental grassland regime dominated by summer production and winter respiration.

Net Ecosystem Exchange (NEE) is the difference between gross primary production, the carbon absorbed through photosynthesis, and respiration, the carbon respired by plants or released

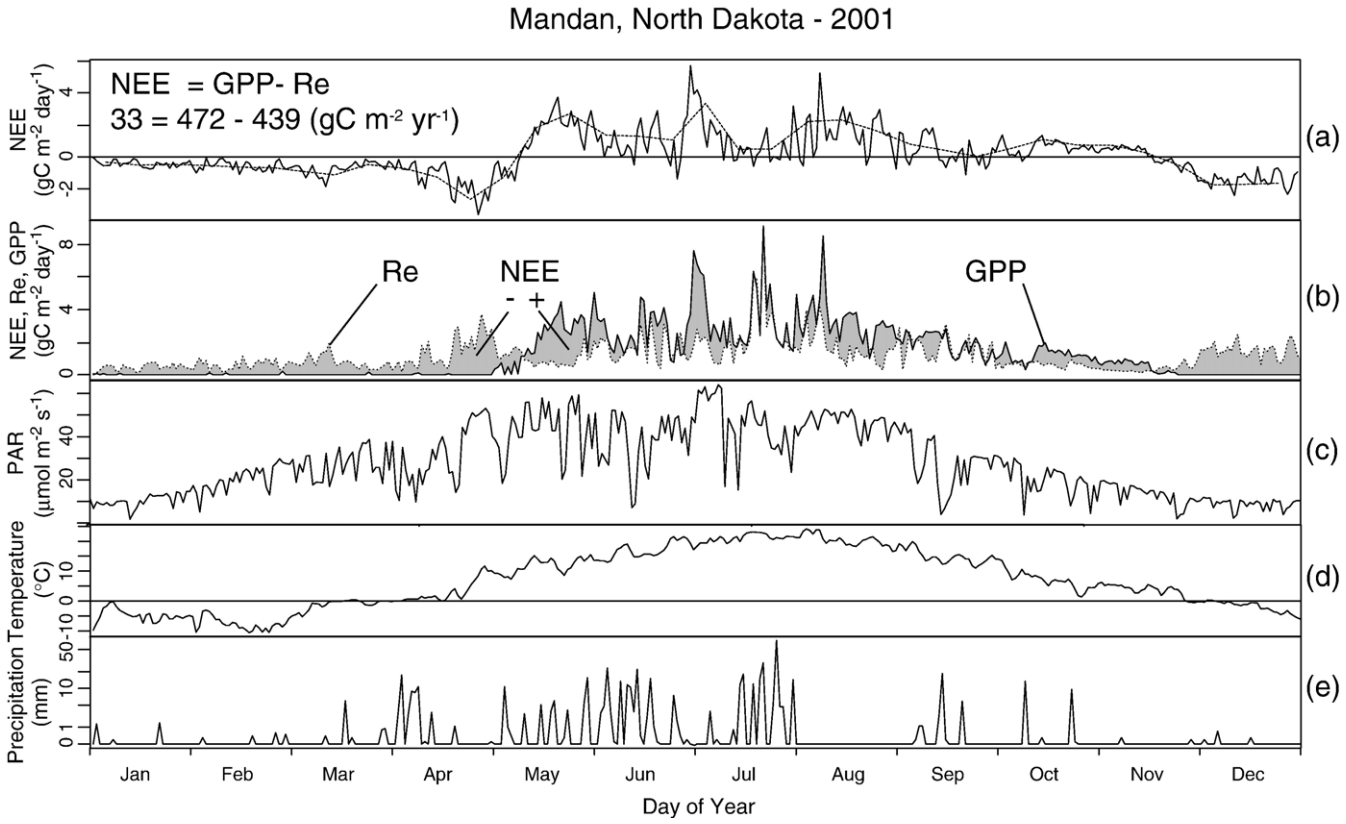

Fig. 5. (a) Daily Net Ecosystem Exchange—solid line, ten-day mean Net Ecosystem Exchange—dotted line; (b) Gross Primary Production—solid line, Respiration—dotted line, grey areas above the GPP line are net respiration, and grey areas below the GPP line are net production; (c) Photosynthetically Active Radiation; (d) Temperature; and (e) precipitation.

through decomposition of organic debris (Fig. 5b). If production exceeds respiration, then a net carbon sink exists for the period measured. In 2001 the rangeland at the Mandan, North Dakota, flux tower (Fig. 5a) was a carbon sink with a net assimilation of 121 gC m^{-2} $year^{-1}$.

Winter fluxes were measured at the Mandan tower for 2001 and the Miles City tower for 2000. The winter fluxes were estimated for the ecoregion based on research conducted by Gilmanov (2002), Frank et al. (2002), and Baron et al. (2006). The regional mean winter flux was estimated to be 0.53 gC $m^{-2}day^{-1}$ (Table 2).

Table 2
Winter flux estimates

Flux tower	Winter seasons	Mean *C* flux (gC $m^{-2}day^{-1}$)	Reference
Mandan, ND	Nov 16, 2001–April 15, 2002	0.69	Gilmanov (2002)
Mandan, ND	Oct 24, 1999–Apr 24, 2000 Oct 24, 2000–Apr 24, 2001	0.50	Frank et al. (2002)
Miles City, MT	Jan 1–May 16 and Nov 16–Dec 31, 2000	0.50	Gilmanov (2002)
Lacombe, AB	Oct 1, 2003–Mar 31, 2004	0.44	Baron et al. (2006)
Average		0.53	

This mean value was added to the C flux estimates integrated over the growing season.

3. Methods

3.1. Piece-wise regression models

A class of algorithms and methodologies called data mining (Witten & Frank, 2000), machine learning (Fielding, 1999), data-driven modeling (Solomatine, 2002), or data-dredging (Burnham & Anderson, 2002), depending individual biases or perspectives, were evaluated for scaling up the flux tower information to regions. These data-driven models, with their highly desirable characteristic of combining a variable selection and data mining tool with an interpretable model result, are increasingly used in the environmental community (Bell, 1999; De'ath & Fabricius, 2000; Fielding, 1999; O'Connor & Wagner, 2004; Zhang et al., 2005a).

Piece-wise regression models were selected as the most appropriate approach for scaling the flux tower data to ecoregions. Piece-wise regressions have been used for many years to handle biological problems that are inherently discrete and nonlinear (Toms & Lesperance, 2003). Burnham and Anderson (2002) stress the importance of parsimony—the importance of avoiding both over- and under-fitting of models. Data-driven models, if appropriately applied, can effectively capture the causal relationships

Table 3
Variables selected for use in cubist model (temperature and precipitation spatial variables are filled with temperature and precipitation measurements at the Lethbridge tower, since no spatial estimates were available)

Theme	Utilization for stratification		Importance for prediction	
	Count	Pct	Weight	Pct
PAR	486	43.5	1.727	14.4
NDVI	350	31.3	3.375	28.1
Days since start of season (SSOST)	231	20.7	0.347	2.9
Temperature (TEMP)	40	3.6	3.464	28.9
Precipitation (PPT)	11	1.0	1.228	10.2
Time integrated NDVI (TIN)	0	0.0	0.849	7.1
Value at start of season (SOSN)	0	0.0	1.012	8.4

Excluded variables were SOLSTICE, DOY, SOST, C4PCT, CO3SURF, PCTCLAYSURF, PCTGRASS, and OMERNICK.

that exist in the data. If data-driven models do not make biological sense, then the models may be over-fit; that is, they may be capturing noise, rather than meaningful relationships, in the data. To state the obvious, over-fit models can be dangerous if used to scale point measurements to form regional estimates. In this project, Cubist[1] rule-based piece-wise regression models (www.rulequest.com) were defined to scale flux tower measurements up to regions (Quinlan, 1992; Rulequest, 2004).

3.2. Northern Great Plains rangeland NEE model

The Cubist statistical software was used to implement a rule-based piece-wise regression model to estimate the carbon flux at 1-km pixels for each 10-day period throughout the growing season. NEE, NDVI, and TEMP are scaled to a byte data range to better accommodate storage, analysis, and display of the variables. The underlying assumption is that environmental information spaces created using phenologically sensitive variables can be fitted by linear regression equations. These piece-wise equations are adaptive, allowing NEE to be estimated for any given 10-day period for any year for any rangeland pixel within the ecoregion.

The variables contribute within Cubist to (1) partition the information space into sub-spaces that minimize within space and maximize between space variation and (2) within each information space, a multiple regression equation is fit. Cubist produces an unordered set of rules to define these environmental information spaces. A pixel may satisfy more than one rule. When more than one rule applies, the result is the mean value of the predictions provided by the regressions (Rulequest, 2004). The variables in each regression equation are sorted in decreasing order based on their relative explanatory power within the individual information spaces.

The model analysis seeks to ensure that (1) the most effective variables are selected, (2) meaningful error estimates are established, and (3) the best possible predictive model is identified for estimating NEE throughout the ecoregion.

[1] Any use of trade, product, or firm names is for description purposes only and does not imply endorsement by the U.S. Government.

4. Results

4.1. Variable selection

Many Cubist models were tested and evaluated before selecting the final model that was used for prediction. An inspection of these models suggest that seven variables (Table 3) selected from the 15 spatial data sets described above would best explain the variation in NEE. Our goal is to create a meaningful, parsimonious model that maximizes cross-validated R^2 and minimizes the number of variables.

The number of times a spatial variable was used in the rules to stratify the training data can be quantified as the frequency or percent utilization. This count gives an indication of the importance of each spatial variable for stratification (Table 3). Similarly, a nonlinear "Importance for Prediction" weight is determined for each variable for each regression equation and summed across the regression equations. This gives an indication of the importance of each spatial variable in predicting NEE. Variables that did not contribute to stratification and explained little of the variation in the information spaces were excluded from the final model.

Because NEE is the confounded effect of both Re and GPP, functional relationships may not always be clear, but this model's dependence on PAR, NDVI, and days since start of season (SSOST) for stratification (rule criteria) make ecological sense. These variables quantify the amount of light available for photosynthesis, photosynthetic potential, and phenological development. The model equations within the "rules" rely heavily on NDVI and temperature and to a lesser degree on PAR and precipitation. Moisture and temperature are often considered the primary drivers of respiration. PAR is an important input in light efficiency models. Yearly lagged effects, such as increased nutrient availability after a drought, may be partially captured by time integrated NDVI (TIN). Residual herbage levels from the previous year affect soil water and temperature. This effect is crudely accounted for in the NDVI value at the start of season (SOSN).

Regression tree models were evaluated in the *R* statistical software package to increase the basic understanding of the relationships among the variables (R Development Core Team, 2005; Venables & Ripley, 2002). The *R* regression tree analysis identifies competing variables at each split (Fig. 6). Variables were evaluated based on how well they explained the expected variation in NEE. Competing variables were selected over primary variables if they better explained NEE dynamics and plant phenology. An inspection of the Cubist variable rankings in light of competing variables identified by *R* helps verify Cubist rules that better explain NEE with fewer and more biologically relevant variables. The relative magnitude of the improvement among competing variables provides a measure of the comparative value of each competing variable (Therneau & Atkinson, 1997, 2005).

The first split is at an NDVI value of 140 (or 0.40 in original NDVI units). This split is very strong in all models in Cubist and in *R*. In both the left and right splits, the most important variable is a date variable. For low values of NDVI, the most important

```
Competing variables
Node number 1:
    NDVI < 140.5    to the left,  improve=0.260
    PAR < 81.5      to the left,  improve=0.171
    SSOST < 114.5   to the right, improve=0.165
    DOY < 186.5     to the right, improve=0.160
    SOLS < 19       to the right, improve=0.151
Node number 2:
    SSOST < 88.5    to the right, improve=0.178
    DOY < 186.5     to the right, improve=0.163
    SOLS < 9        to the right, improve=0.161
    NDVI < 127.5    to the left,  improve=0.072
    PAR < 81.5      to the left,  improve=0.0698
Node number 3:
    DOY < 186.5     to the right, improve=0.371
    DOLS < 19       to the right, improve=0.299
    PAR < 100.5     to the left,  improve=0.250
    SSOST < 116.5   to the right, improve=0.237
    TEMP < 136.5    to the left,  improve=0.056
Node number 4:
    PAR < 95.5      to the left,  improve=0.279
    SSOST < 33.5    to the left,  improve=0.267
    TEMP < 112.6 to the left,  improve=0.265
    NDVI < 121.5    to the left,  improve=0.263
    TIN < 60.5      to the right, improve=0.238
```

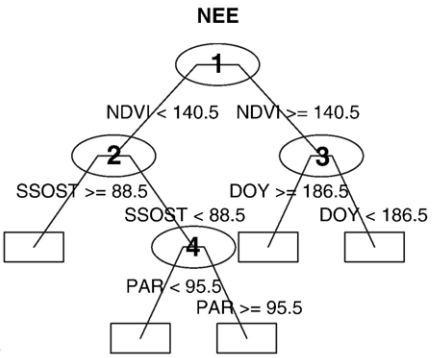

Fig. 6. Regression tree from *R* showing competing variables for each split. NDVI and temperature are rescaled to 0–255.

date variable is "since start of season (SSOST)." For high values of NDVI, the most important date variable is "day of year (DOY)" (Fig. 6). At most splits, the three date variables, SSOST, DOY, and "since summer solstice (SOLS)," are competing variables (Fig. 6). The three variables are highly co-linear. While DOY and SOLS are directly related (Fig. 7c), SSOST introduces new information directly related to phenology. In Fig. 7a and b, SSOST can be seen to contain site-specific information that does not exist in DOY, that may provide an improved ability to characterize sites and years. The use of SSOST alone provides both parsimony and phenologic relevance.

4.2. Error estimates through cross-validation

Care must be taken to ensure the implementation of robust biologically reasonable models given the proclivity to over-fit models developed from training data with a small sample size and significant noise (Quinlan, 1996). Cubist models compensate for small sample size by using cross-validation to assess models for robustness and provide a realistic estimation error. In random cross-validation, data are divided into *n* random subsets. A model is developed or trained using *n*-1 of the subsets and tested on the subset that was withheld. A model is created for each of the *n* subsets, where each iteration is a fold. Through a drop one site or year approach, we further determine the model stability if individual years or sites are not available.

Cross-validation is used to determine the stability of the models, to calculate realistic error estimates, and to identify influential samples. We need to accept the reality that the training data only meet minimal size and distribution assumptions. This is true whether the flux tower data are used for empirical scaling up or for validation/calibration of a physical model. Flux towers are

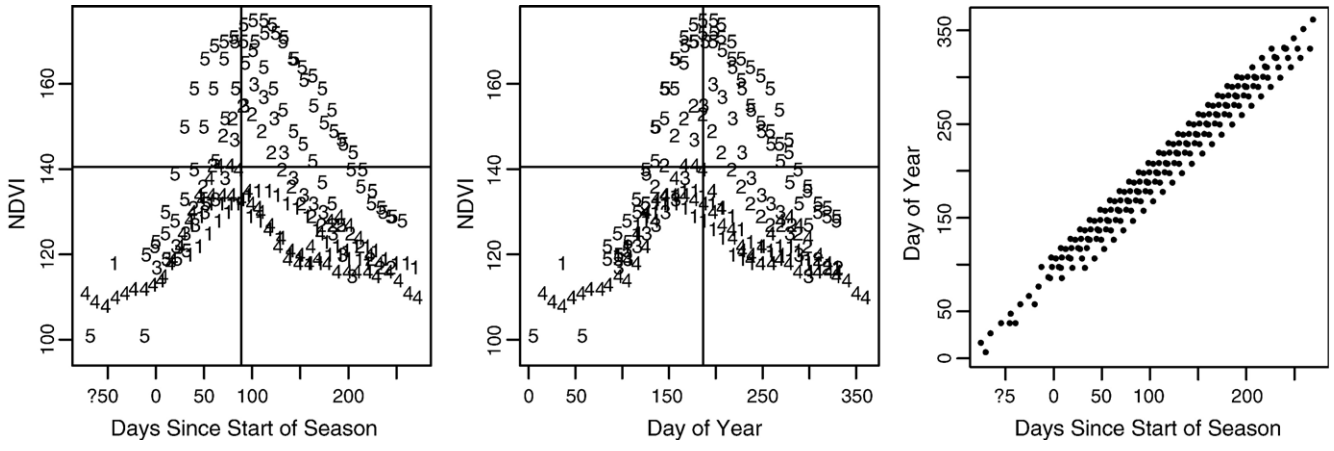

Fig. 7. The time variables day of year (DOY), and days since start of season (SSOST) are highly correlated and have similar, but not identical, relationships with NDVI. 1 is Miles City; 2 is Cheyenne; 3 is Fort Peck; 4 is Lethbridge; and 5 is Mandan.

sited with independent research objectives rather than to support a comprehensive mapping objective. Cross-validated estimates of error are important conservative measures of the robustness of the model. These three cross-validation methodologies, random, by year, and by site, were used to evaluate the stability and robustness of models derived from the available towers and selected spatial variables.

4.3. Random cross-validation

The Cubist model is a five-fold cross-validated model, where the number of folds is selected to be approximately the same as the number of sites or years. In the random cross-validated model, the database is randomly divided into five equal subsets. By randomly selecting the withheld samples and testing against samples, which are not used in the model construction, unbiased and realistic error estimates are determined. An inspection of the rules provides a subjective estimate of the stability of the rules generated and the variables used in the models.

With the exception of Fold 5, splits at NDVI ~140 and SSOST ~85 control the stratification. After NDVI and SSOST, the next most important variable is PAR. Fold 5, which only uses PAR as a splitting variable, has the largest mean absolute error of the five-folds. In the *R* regression tree model (Fig. 6), PAR is the first competing variable below NDVI for the first split, but has considerably less explanatory power. An inspection of PAR and NEE measurements at the Mandan Flux tower (Fig. 5) shows that NEE is responsive to changes in PAR.

The Mean Absolute Difference (MAD) is used by Cubist to compare the model results, where MAD is in data units (gC $m^{-2}day^{-1}$).

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

The flux tower measurements of NEE range from −6.2 to 11.0 gC $m^{-2}day^{-1}$. The interquartile range, which bounds 50% of the data, is 3.8 gC $m^{-2}day^{-1}$ with a lower quartile at −1.4 gC $m^{-2}day^{-1}$ and an upper quartile at 2.4 gC $m^{-2}day^{-1}$. The mean value is 0.7 gC $m^{-2}day^{-1}$, while the median value is −0.01 gC $m^{-2}day^{-1}$, suggesting a small negatively skewed distribution.

MAD for the random cross-validations range from 0.38 to 0.60 (Table 4), which are small, as are the relative errors (RE), in proportion to the range of the estimates of NEE. The model formed in Fold 5 controlled by PAR is atypical, and this is reflected in the high MAD for Fold 5. However, when compared to the interquartile range, MAD for the cross-validations is acceptably small for even the largest instance.

Table 4
Mean Absolute Difference (MAD) values for five-fold random cross-validation

Random sample as test	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Sample size (test)	43	43	43	44	44
Mean Absolute Difference (test)	0.49	0.43	0.48	0.38	0.60

Table 5
Cross-validation by site: LE is Lethbridge; MA is Mandan; PK is Fort Peck; CN is Cheyenne; and MC is Miles City

Site withheld as test	LE	MA	FP	CN	MC	None
Sample size (train)	159	144	196	196	173	217
Mean Abs. Difference (train)	0.45	0.29	0.36	0.37	0.38	0.35
Relative Error (train)	0.14	0.09	0.11	0.12	0.12	0.11
Sample size (test)	58	73	21	21	44	0
Mean Abs. Difference (test)	0.46	0.66	0.31	0.55	0.59	
Relative Error (test)	0.14	0.21	0.10	0.17	0.18	

4.4. Cross-validation by site and year

De'ath and Fabricius (2000) suggested that cross-validation by site is an effective method to compensate for the effect of a limited number of sites with multiple samples selected from each site. Likewise, cross-validation by year provides a means to evaluate the influence of individual years. Cross-validation by year or site has a simple and practical objective: do the models continue to be effective even if one of the sites or years of training data is missing.

To test for the effect of sites, five site models were created with each one trained using four of the sites, while the withheld site is used for testing. The results of site models are compared to the results from the full model to determine whether any of the sites are influential. The objective is to better understand the impact of the individual sites on the model. If an influential site is withheld, then the model should not perform as well as a model that includes the samples from that site. Therefore, influential sites provide important contributions to the model. The five models with sites withheld are compared to the model with all sites (Table 5). The MADs for the by site training data (MADs from 0.29 to 0.45) are in the same order of magnitude as are the MADs for the test data (MADs from 0.31 to 0.66).

Mandan, the site with the largest test MAD value and a substantial difference between training and test values, has the largest sample size (73 of the 217 samples) and the greatest annual precipitation. Total precipitation at Mandan for the 1999, 2000, and 2001 growing seasons (April through October) was 496, 406, and 437 mm, respectively, while the non-Mandan site with the highest total growing season precipitation was only 342 mm. However, even the exclusion of this most influential site does not produce an excessive MAD value. The inclusion of

Table 6
Cross-validation by year

Year withheld as test	1998	1999	2000	2001	None
Sample size (train)	196	192	120	143	217
Mean Absolute Difference (train)	0.37	0.33	0.44	0.39	0.35
Relative Error (train)	0.12	0.10	0.14	0.12	0.11
Sample size (test)	21	25	97	74	0
Mean Absolute Difference (test)	0.55	0.91	0.49	0.48	
Relative Error (test)	0.17	0.29	0.16	0.15	

Mandan provides a model that is robust across a wider range of conditions.

Likewise, the MADs for the by year training data are in the same order of magnitude as are the MADs for the test data (Table 6). The largest MAD value for all of the cross-validations was for 1999. Spring rainfall was extremely high for the 1999 growing season at Mandan, and Mandan was the only site that had data for 1999 (0.91). This was the extreme year for the extreme site, so the high value of MAD for this cross-validation was not unexpected.

The methodology is robust under cross-validation. Conditions across the region vary considerably. Rulequest, in its Cubist tutorial, emphasizes the importance of having extreme values represented to minimize the need for extrapolation (Rulequest, 2004). The available training data include the influential years and sites needed to capture extreme environmental conditions.

4.5. Model description

After a thorough study of how the models respond to cross-validation and variable selection, a final model was trained using the seven selected variables and all of the flux tower data. The model was applied to the spatial database to estimate NEE for each 10-day period. The rules stratifying the piece-wise regression are listed on Figs. 8 and 9.

The interrelationships among the variables by rule can be seen in Fig. 8. The scatterplot of NDVI versus SSOST shows the clean partition formed by Rule 1 (red) and 3 (blue) for low values of NDVI. The split point for SSOST is very close to maximum NDVI. PAR and TEMP come into play for high values of NDVI as Cubist minimizes variation as vegetation approaches and passes peak vegetation greenness.

Fig. 9 shows the spatial distribution (May, July, and September for 1998 through 2001) of the rules. Recall that

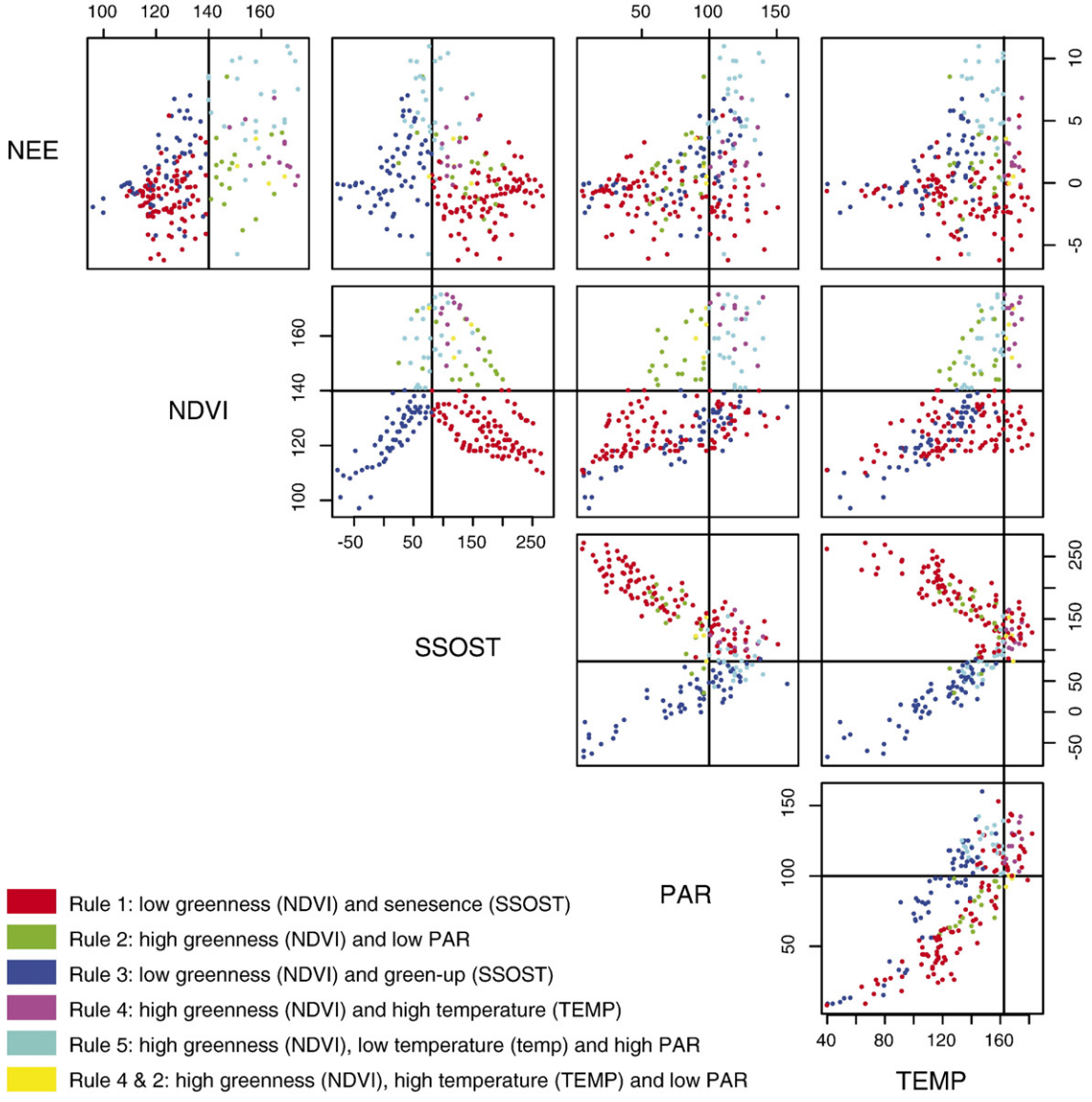

Fig. 8. Scatterplots of the four stratifying variables and NEE at the towers. The splits used are overlain on the scatterplots. The colors on the scatterplots correspond to the colors used in the rule maps in Fig. 9.

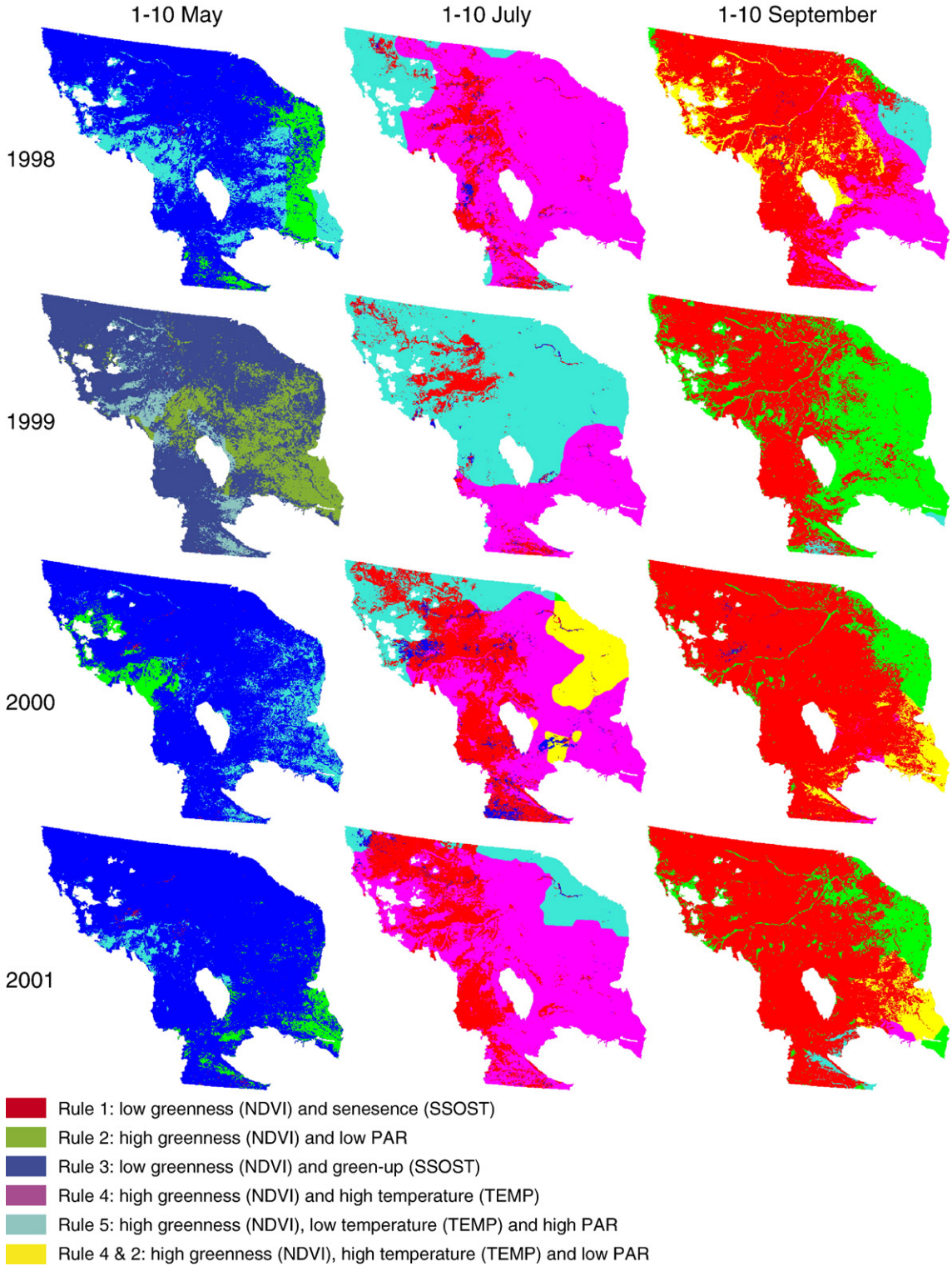

Fig. 9. Rule maps showing the distribution of the rule use through time and space.

the Cubist model is a dynamic rule-base parameterization of the carbon flux regression model. As the growing season progresses different multiple regression models become active and move across the landscape in response to changing environmental conditions.

In May, Rule 3, an early green-up environmental information space, dominates the ecoregion, although Rules 2 and 5 identify environmental spaces that are already very green in early May. By early July most of the ecoregion is dominated by Rules 4 and 5 delineating regions of high biomass with the two rules differentiated by higher temperatures in the south and lower temperatures in the North. Some areas controlled by Rule 1 have already advanced into senescence. By early September, most of the ecoregion is dominated by Rule 1. Rules 2, 4, and 5 bound

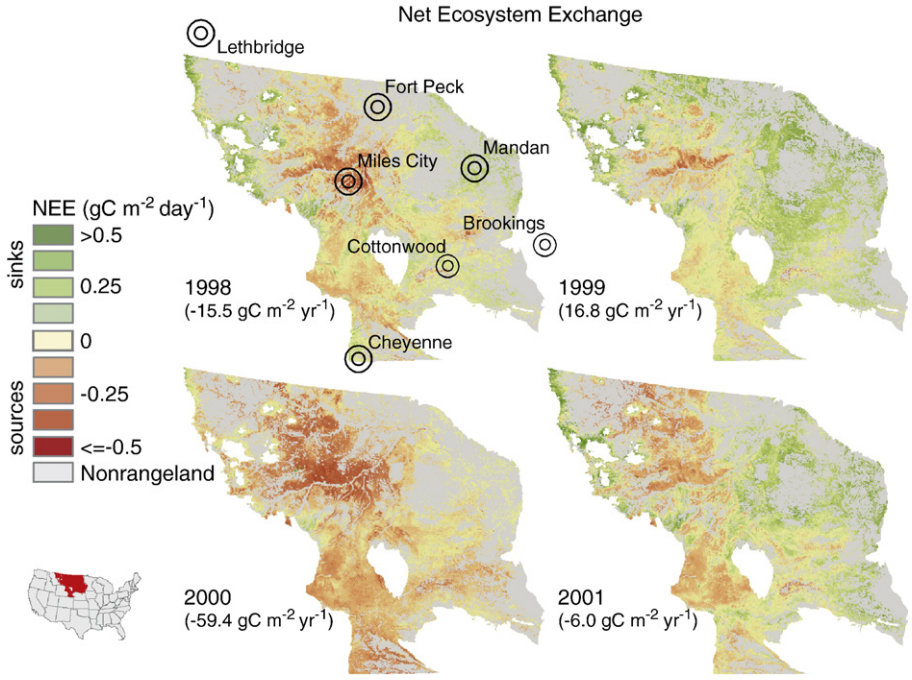

Fig. 10. Annual NEE estimates for 1998 through 2001. Greens are carbon sinks, and reds are carbon sources. The new Cottonwood and Brookings flux tower sites will help improve future carbon flux estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

areas of high vegetation vigor in the east where more rainfall is available. Rainfall is not directly included in the model, but NDVI acts as a surrogate for rainfall. The model adapts dynamically to different environmental conditions through the year by applying appropriate rules as changing environmental conditions dictate.

4.6. Annual estimates

We used the rule-based piece-wise regression model to estimate NEE for each 1-km pixel every ten days during the growing season. By summing these 10-day estimates, we obtained

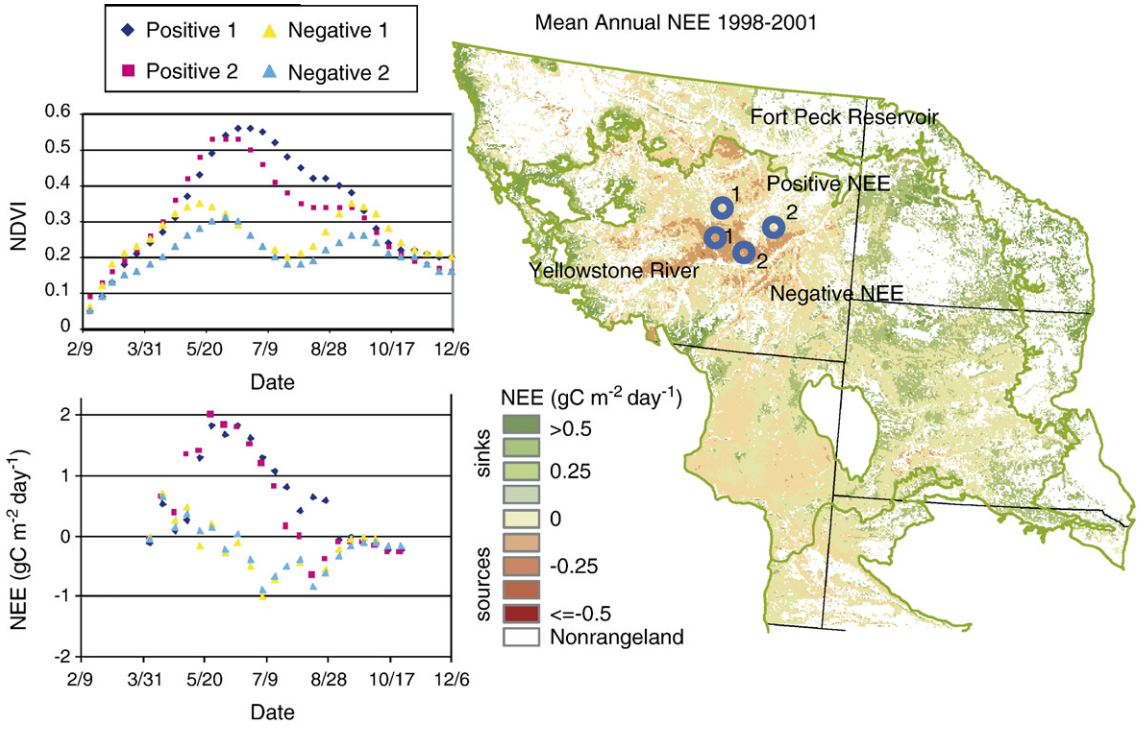

Fig. 11. A stable carbon source shows a localized early decrease in both NDVI and NEE. The seasonal graphs of NEE and NDVI are for 2001. Positive NEE circles 1 and 2 on map can be seen in the graphs as high peaks in June 2001, while the negative NEE circles 1 and 2 on the map can be seen as lows in July 2001.

a growing season carbon flux estimate. Finally, by adding winter flux estimates to the growing season estimates, a total carbon estimate is determined for the rangelands within the Northern Great Plains ecoregion (Fig. 10).

For the four years estimated, the carbon fluxes for the Northern Great Plains rangelands are in approximate equilibrium. The ecoregion was found to be a small sink in 1999 and a small source in 1998, 2000, and 2001, although the time series is too short to provide a more definitive estimate of the state or trends in carbon fluxes in the region.

Current research is directed to understanding the driving forces that cause patterns such as the stable carbon source along the Yellowstone River and near the Fort Peck Reservoir, shown in Fig. 11. The seasonal graphs clearly show a sharp downturn in NDVI and NEE at two source locations near the Yellowstone River when compared to two sink locations north of the river. The carbon source near the Fort Peck Reservoir is associated with a clay soil, which is not true along the Yellowstone River, where high growing-season temperatures are associated with the source along the Yellowstone River, but not with the source near the Fort Peck Reservoir.

Future research calls for field studies to explore these consistent source and sink areas and to establish causal relationships for patterns in the NEE data. This research may identify patterns of carbon flux that represent local persistent weather patterns, management practices, or soil characteristics or may uncover false patterns that are artifacts of the spatial data sets used. The maps of annual NEE in Fig. 10 document annual trends in carbon flux. The 10-day database used to determine annual estimates of NEE will be investigated further, particularly in those anomalous areas to establish causal relationships and to confirm patterns.

5. Discussion and conclusions

The rule-based piece-wise model permits an estimation of NEE that holds up well under cross-validation and existing knowledge of NEE process models. The essence of a data-driven model is the constant search for new and better data that can help explain and predict NEE within a known and accepted theoretical framework. New sources of data continue to be investigated for inclusion in the model.

Three fundamental data assumptions must be met to establish confidence and limits for the regional flux estimates. One, the spatial distribution of the flux towers must adequately sample the spatial variability of the environmental extremes in the ecoregion. Two, the years sampled must adequately bound the temporal variability of the environmental extremes in the ecoregion. Three, the spatial variables selected for use in the model must adequately and robustly quantify carbon fluxes in a manner that can be justified given known biophysical characteristics of carbon dynamics.

A corollary to these assumptions is the need to identify gaps in available data to improve the quality of future models. A data-driven approach has inherent strengths. The available spatial data are imperfect surrogates for the information needed to satisfy process models. Empirical, data-driven models adapt to the idiosyncrasies of the available data and help build an understanding of the relationships between NEE and the available data. An outcome of this analysis is an improved understanding of gaps in the knowledge base, which is needed to accurately and precisely estimate carbon fluxes.

Of particular interest are more effective measures of water availability and temperature, particularly in water limited ecosystems (Austin & Vivanco, 2006). Haferkamp and MacNeil (2004) stressed the importance of April–June precipitation when more than 90% of the biomass of cool-season grasses and 75% of the biomass of warm-season grasses are produced by the end of June. NDVI presently acts as a de facto surrogate for water availability, but spatial data quantifying soil moisture are under investigation by NASA, NOAA, USDA, and others (Griffiths & Wooding, 1996; Lakshimi, 2004; Lu & Meyer, 2002). Other aggregations of precipitation, such as precipitation lags or a moving window of precipitation accumulation, will be investigated.

A "growing degree days" variable is expected to be important in future grassland ecoregion models (Frank & Hofmann, 1989). Other spatial variables under consideration are remotely sensed measures of vegetation residue and snow cover (Daughtry et al., 2004; Hall et al., 1998; Nagler et al., 2003; Riggs et al., 2003). The construction of new flux towers is being supported to improve the spatial distribution of rangeland flux towers. It is hoped that towers at Cottonwood and Brookings, South Dakota (Fig. 10), will become available to help anchor the south and central Northern Great Plains estimates.

The flux tower operators have conducted significant rangeland climatic and management research upon which interpretation of regional models of carbon flux depend (Flanagan et al., 2002; Frank, 2002; Frank et al., 2002; Haferkamp & MacNeil, 2004; Heitschmidt et al., 2005; Meyers, 2001; Morgan et al., 2004, Wever et al., 2002; Zhang et al., 2005b). Detailed studies and data collected at sites provide the understanding of the carbon dynamics needed to lend credence to regional carbon studies.

A substantial amount of the uncertainty in the annual flux estimates lies in the estimate of winter fluxes (Frank, 2002). Methodologies will be investigated to reduce the uncertainty associated with winter fluxes. The model needs to be extended not only to new ecoregions but also to additional years to better bound the inherent variability of these rangeland ecosystems.

Our intent is to extend the methodology into other rangeland ecoregions to capture the full variability of the environmental conditions represented in rangeland ecosystems. A coherent set of rule-based piece-wise regression models will evolve to explain and predict NEE throughout North American rangelands.

Acknowledgements

This project would not have been possible without the strong collaboration and support of the following: USGS Earth Surface Dynamics, Land Remote Sensing, and Geographic Analysis and Monitoring Programs, NOAA Atmospheric Turbulence and Diffusion Division, the collaborative CO_2 flux scaling project (University California, Davis) funded through the US Agency for International Development Global Livestock Collaborative

Research Programs (USAID GL-CRSP) and USDA Agricultural Research Service, USDA Agriflux, and USGS National Center, EROS. Without the vital contributions of data and science by the flux tower operators, A.B. Frank, L. B. Flanagan, J.A. Morgan, M. R. Haferkamp, and T. P. Meyers, the project would not be possible. The authors thank Bradley Reed (USGS Center for Earth Resources Observation and Science (EROS)) for assistance with NDVI data, Norman Bliss, Zhengxi Tan, Chris Wright (EROS) and two anonymous reviewers for judicious and insightful critiques and edits, and Ruth Anne Doyle and Li Zhang (EROS) for help with GIS and mapping.

References

Angell, R. F., Svejcar, T. J., Bates, J., Saliendra, N. Z., & Johnson, D. A. (2001). Bowen ratio and closed chamber carbon dioxide flux measurements over sagebrush steppe. *Agricultural and Forest Meteorology*, *108*, 153−161.

Austin, A. T., & Vivanco, L. (2006). Plant Litter decomposition in a semi-arid ecosystem controlled by photodegradation. *Nature*, *442*, 555−558.

Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystem: Past, present and future. *Global Change Biology*, *9*, 479−492.

Baron, V. S., Young, D. G., Dugas, W. A., Mielnick, P. C., La Bine, C., Skinner, R. H., et al. (2006). Net Ecosystem Carbon Dioxide over a temperate, short-season grassland: Transition from cereal to perennial forage. In J. S. Bhatti, R. Lal, M. J. Apps, & M. A. Price (Eds.), *Climate change and managed ecosystems* New York: Taylor and Francis.

Bartlett, D. S., Whiting, G. J., & Hartman, J. M. (1990). Use of vegetation indices to estimate intercepted solar radiation and net carbon dioxide exchange of a grass canopy. *Remote Sensing of Environment*, *30*, 115−128.

Bell, J. F. (1999). Tree-based methods. The use of classification trees to predict species distributions. In A. Fielding (Ed.), *Machine learning methods for ecological applications* (pp. 89−105). Norwell, Massachusetts: Kluwer Academic Publishers.

Burke, I. C., Lauenroth, W. K., & Milchunas, D. G. (1997). Biogeochemistry of managed grassland in central North America. In E. A. Paul, E. T. Elliot, K. Paustian, & C. V. Cole (Eds.), *Soil organic matter in temperate agroecosystems: Long term experiments in North America* (pp. 85−101). Boca Raton: CRC Press Inc.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information—theoretic approach.* New York: Springer-Verlag.

Butcher, G. (2004). State of the birds USA. *Audubon*, *106*, 45−51.

Cayrol, P., Dedieu, G., Mordelet, P., Nouvellon, Y., Chehbouni, A., & Kergoat, L. (2000). Grassland modeling and monitoring with SPOT-4 VEGETATION instrument during the 1997–1999 SALSA experiment. *Agricultural and Forest Meteorology*, *105*(1–3), 91−115.

CDC (2005). CPC US Unified precipitation Date, NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA. Available at http://www.cdc.noaa.gov/ on 4 February 2004.

CEC (1997). *Ecological regions of North America: Toward a common perspective.* Montreal: Commission for Environmental Cooperation.

Churkina, G., Schimel, D., Braswell, B. H., & Xiao, X. (2005). Spatial Analysis of growing season length control over net ecosystem exchange. *Global Change Biology*, *11*, 1777−1787.

Cihlar, J., Denning, S. Francey, R., Gommes, R., Heimann, M., Kabat, P., Olsen, R., Scholes, R., Townshend, J., Tschirley, J., Valentini, R., & Wickland, D. (2003). Implementation plan for the terrestrial and atmospheric carbon observation initiative (TCO), Integrate Global Observing Strategy (IGOS) (Available at http://www.fao.org/gtos/doc/TCO_Implan_finedit.doc on 30 October 2006).

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., et al. (1997). The value of the world's ecosystem services and natural capital. *Nature*, *387*, 253−260.

Daughtry, C. S. T., Hunt Jr., J. E., & McMurtrey III, J. E. (2004). Assessing crop residue cover using shortwave infrared reflectance. *Remote Sensing of Environment*, *90*, 126−134.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178−3192.

DOE (2003). U.S. climate change technology program: Research and current activities. Available at http://www.climatetechnology.gov/library/2003/currentactivities/measure-monitor.htm on 4 February 2005.

Dugas, W. A., Heuer, M. L., & Mayeux, H. S. (1999). Carbon dioxide fluxes over Bermuda grass, native prairie, and sorghum. *Agricultural and Forest Meteorology*, *93*, 121−139.

EPA, (2005). Level III Ecoregions. Available at http://www.epa.gov/wed/pages/ecoregions/level_iii.htm on 5 November 2005.

Fielding, A. H. (Ed.). (1999). *Machine learning methods for ecological applications.* Norwell, Massachusetts: Kluwer Academic Publishers.

Flanagan, L. B., & Johnson, B. G. (2005). Interacting effects of temperature, soil moisture and plant biomass production on ecosystem respiration in a northern temperate grassland. *Agricultural and Forest Meteorology*, *130*, 237−253.

Flanagan, L. B., Wever, L. A., & Carlson, P. J. (2002). Seasonal and interannual variation in carbon dioxide exchange and carbon balance in a northern temperate grassland. *Global Change Biology*, *8*, 599−615.

Frank, A. B. (2002). Carbon dioxide fluxes over a grazed prairie and seeded pasture in the Northern Great Plains. *Environmental Pollution*, *116*, 397−403.

Frank, A. B. (2003). Six years of CO2 flux measurements for a moderately graxed mixed-grass prairie. *Environmental Management*, *3*(Supplement 1), S426−S431.

Frank, A. B., & Hofmann, L. (1989). Relationship among grazing management, growing degree-days, and morphological development for native grasses on the Northern Great Plains. *Journal of Range Management*, *42*(3), 199−202.

Frank, A. B., Liebig, M. A., & Hanson, J. D. (2002). Soil carbon dioxide fluxes in northern semiarid grassland. *Soil Biology & Biochemistry*, *34*, 1235−1241.

Frouin, R., & Pinker, R. T. (1995). Estimating Photosynthetically Active Radiation (PAR) of the Earth's surface from satellite observations. *Remote Sensing of Environment*, *51*, 98−107.

Gilmanov, T. G. (2002, Sept. 30). *Wintertime CO_2 flux in four ecosystems of the USDA/ARS rangeland CO_2 flux network in relation to environmental factors, internal report to the Raytheon Company.* USGS/EROS.

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.

Griffiths, G. H., & Wooding, M. G. (1996). Temporal monitoring of soil moisture using ERS-1 SAR data. *Hydrological Processes*, *10*(9), 1127−1138.

Haferkamp, M. R., & MacNeil, M. D. (2004). Grazing effects on carbon dynamics in the northern mixed-grass prairie. *Environmental Management*, *33*(Suppl. 1), S462−S479.

Hall, D. K., Foster, J. L., Verbyla, D. L., Klein, A. G., & Benson, C. S. (1998). Assessment of snow cover mapping accuracy in a variety of vegetation cover densities in central Alaska. *Remote Sensing of Environment*, *66*, 129−137.

Heitschmidt, R., Klement, K., & Haferkamp, M. R. (2005). Interactive effects of drought and grazing on Northern Great Plains Rangelands. *Rangeland Ecology and Management*, *58*, 11−19.

Higgins, K. F., Naugle, D. E., & Forman, K. J. (2002). A case study of changing land use practices in the Northern Great Plains, U.S.A. An uncertain future for waterbird conservation. *Waterbirds*, *25*(2), 42−50.

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing*, *70*(7), 829−840.

JRC (2003). Global land cover 2000 database. European commission, joint research centre, 2003. Available at http://www-gvm.jrc.it/glc2000 on 4 February 2005.

Lakshimi, V. (2004). *Use of satellite remote sensing in hydrological predictions in ungaged basins*. Paper presented at the Geo-Imagery Bridging Continents XXth ISPRS Congress, 12–23 July 2004, Istanbul, Turkey.

Lu, Z., & Meyer, D. J. (2002). Study of high SAR backscattering caused by an increase of soil moisture over a sparsely vegetated area: Implications for characteristics of backscattering. *International Journal of Remote Sensing*, *23*(6), 1063−1074.

McMahon, G., Gregonis, S. M., Waltman, S. W., Omernik, J. M., Thorson, T. D., Freeouf, J. A., et al. (2001). Developing a spatial framework of common ecological regions for the conterminous United States. *Environmental Management*, *28*(3), 293−316.

Meyers, T. P. (2001). A comparison of summertime water and CO_2 fluxes over rangeland for well watered and drought conditions. *Agricultural and Forest Meteorology*, *106*, 205−214.

Moncrieff, J. B., Massheder, J. M., Verhoeh, A., Elbers, J. A., Heusinkveld, B., Scott, S., et al. (1997). A system to measure surface fluxes of momentum, sensible heat, water vapor and carbon dioxide. *Journal of Hydrology*, *188–189*, 589−611.

Morgan, J. A. (2006). Personal communication.

Morgan, J. A., Lecain, D. R., Reeder, J. D., Schuman, G. D., Derner, J. D., Lauenroth, W. K., Parton, W. J., & Burke, I. C. (2004). *Drought and Grazing Impacts on CO_2 Fluxes in the Colorado Shortgrass Steppe*. Paper presented at the Ecological Society of America.

Nagler, P. L., Inoue, Y., Glenn, E. P., Russ, A. L., & Daughtry, C. S. T. (2003). Cellulose absorption index (CAI) to quantify mixed soil-plant litter scenes. *Remote Sensing of Environment*, *87*, 310−325.

O'Connor, R. J., & Wagner, T. L. (2004). A test of a regression-tree model of species distribution. *The Auk*, *12*(2), 604−609.

Omernik, J. M. (1987). Ecoregions of the conterminous United States: Map supplement. *Annals of the Association of American Geographers*, *77*(1), 118−125.

Quinlan, J. R. (1992). *Learning with Continuous Classes*. Paper presented at the Proceedings Fifth Australian Joint Conference on Artificial Intelligence, Hobart, Tasmania, Singapore.

Quinlan, J. R. (1996). *Boosting first-order learning, Algorithmic Learning Theory, 7th International Workshop, ALT '96, Sydney, Australia, October 1996, Proceedings*.

R Development Core Team (2005). R: A language and environment for statistical computing, from http://www.R-project.org

Raupach, M. R. (1988). Canopy transport processes. In W. L. Steffen & O.T. Denmead (Eds.), *Flow and transport in the natural environment: Advances and applications* (pp. 95−127). New York: Springer-Verlag.

Reed, B. C. (2006). Trend analysis of time-series phenology of North America derived from satellite data. *GIScience & Remote Sensing*, *43*(1), 1−15.

Reed, B. C., Brown, J. F., VanderZee, D., Loveland, T. L., Merchant, J. W., & Ohlen, D. O. (1994). Measuring phenological variability from satellite imagery. *Journal of Vegetation Science*, *5*, 703−714.

Riggs, G. A., Hall, D. K., Salomonson, V. V. (2003). MODIS snow products user guide for collection 4 data products. Available at http://modis-snow-ice.gsfc.nasa.gov/sug_main.html on 5 February 2005.

Rulequest (2004). An overview of Cubist, from http://www.rulequest.com/cubist-unix.html on 11 May 2005.

Running, S. W., Baldocchi, D. D., Turner, D. P., Gower, S. T., Bakwin, P. S., & Hibbard, K. A. (1999). A global terrestrial monitoring network integrating tower fluxes, flask sampling, ecosystem modeling and EOS satellite data. *Remote Sensing of Environment*, *70*, 108−127.

Sims, D. A., & Singh, J. S. (1978). The structure and function of ten western North American grasslands. III. Net primary production, turnover, and efficiencies of energy capture and water use. *Journal of Ecology*, *66*, 573−597.

Solomatine, D. P. (2002). Data-driven modelling: Paradigm, methods, experiences. Paper presented at the Proceedings 5th International Conference on Hydroinformatics, Cardiff, United Kingdom, July 1–5, 2002 (pp. 757–763). Available at http://www.ihe.nl/hi/sol/papers/ on 5 November 2005.

Svejcar, T. J., Mayeux, H. S., & Angell, R. F. (1997). The rangeland carbon dioxide flux project. *Rangelands*, *19*, 16−18.

Swets, D. L., Reed, B. C., Rowland, J. R., Marko, S. E. (1999). A weighted least-squares approach to temporal smoothing of NDVI 1999, ASPRS Annual Conference, From Image to Information, Portland, Oregon, May 17–21, 1999. Proceedings: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing, CD-ROM, 1 disc.

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines.* : Mayo Foundation.

Therneau, T.M., Atkinson, E.J. (2005). RPART: Recursive Partitioning, from S-PLUS 6.x. Available at http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm on 5 November 2005.

Toms, J. D., & Lesperance, M. L. (2003). Piecewise regression: A tool for identifying ecological thresholds. *Ecology*, *84*(8), 2034−2041.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*, 127−150.

USDA (1995). State Soil Geographic (STATSGO) Data Base: Data use information. Available at http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo/ on 5 November 2005.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S.* New York: Springer.

Wever, L. A., Flanagan, L. B., & Carlson, P. J. (2002). Seasonal and interannual variation in evapotranspiration, energy balance and surface conductance in a northern temperate grassland. *Agricultural and Forest Meteorology*, *112*, 31−49.

White, R. P., Murray, S., & Rohweder, M. (2000). *Pilot analysis of global ecosystems: Grassland ecosystems.* Washington, D.C.: World Resources Institute.

Wickham, J. D., Stehman, S. V., Smith, J. H., & Yang, L. (2004). Thematic accuracy of the 1992 National Land-Cover Data for the western United States. *Remote Sensing of Environment*, *91*, 452−468.

Wigley, T. M. L. (1999). *The science of climate of climate change: Global and U.S. perspectives.*Arlington, Virginia: Pew Center on Global Climate Change. 48 pp.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations.* San Francisco, California: Morgan Kaufmann Publishers.

Wofsy, S. C., & Harriss, R. C. (2002). *The North American Carbon Program (NACP): A report of the NACP Committee of the U.S. carbon cycle science steering group.* Washington, D.C.: U.S. Global Change Research Program.

WRI (2000). Chapter 2—Taking stock of ecosystems-grassland ecosystems. *World Resources 2000–2001: People and ecosystems—the fraying web of life* (pp. 119−131). Available at http://pubs.wri.org on 5 November 2005.

Wylie, B. K., Gilmanov, T. G., Johnson, D. A., Saliendra, N. Z., Akshalov, K., Tieszen, L. L., et al. (2004). Intra-Seasonal mapping of CO2 Flux in Rangelands of northern Kazakhstan at one-kilometer resolution. *Journal of Environmental Management*, *33*(Suppl. 1), S482−S491.

Wylie, B. K., Meyer, D. J., Tieszen, L. L., & Mannel, S. (2002). Satellite mapping of biophysical parameters at the biome scale over the North American grasslands: A case study. *Remote Sensing of Environment*, *79*, 266−278.

Xie, P., & Arkin, P. A. (1996). Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, *78*, 2539−2558.

Zhang, B., Valentine, I., & Kemp, P. D. (2005). A decision tree approach modelling functional group abundance in a pasture ecosystem. *Agriculture, Ecosystems and Environment*, *110*(3–4), 279−288.

Zhang, Y., Grant, R. F., Flanagan, L. B., Wang, S., & Verseghy, D. L. (2005). Modelling CO_2 and energy exchanges in a northern semiarid grassland using the carbon- and nitrogen-coupled Canadian Land Surface Scheme (C-CLASS). *Ecological Modelling*, *181*, 591−614.