Agricultural and Forest Meteorology 151 (2011) 508–520

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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

# Classification and assessment of turbulent fluxes above ecosystems in North-America with self-organizing feature map networks

Andres Schmidt*, Chad Hanson, James Kathilankal, Beverly E. Law

*Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, OR 97331, USA*

ARTICLE INFO

*Article history:*
Received 2 June 2010
Received in revised form
20 December 2010
Accepted 21 December 2010

*Keywords:*
AmeriFlux
Carbon uptake
Pattern recognition
SOFM neural networks
Turbulent exchange process

ABSTRACT

We determined key environmental and meteorological drivers that influence the magnitude of the fluxes of carbon dioxide, latent heat, and sensible heat using data from different vegetation types at 56 sites of the AmeriFlux network in combination with a self-organizing feature map neural network. The technique was combined with a subsequent *k*-means clustering procedure to classify the turbulent fluxes. This method is a direct approach to assess the importance of several environmental parameters for the turbulent exchange rates above vegetated areas, based on eddy covariance measurements. The synthesis of the available dataset was achieved by merging 225 pre-clusters of fluxes that were found by the SOFM into 9 final clusters exhibiting mean $CO_2$ fluxes from $-5.8\ \mu mol\ m^{-2}\ s^{-1}$ to $3.8\ \mu mol\ m^{-2}\ s^{-1}$.

The spatial and temporal comprehensive dataset covering 305 site years, combined with the independence from mechanistic ecological assumptions of the SOFM network approach provides an opportunity to assess the strength of the influence of various environmental parameters on the turbulent fluxes over vegetated areas. After ranking the environmental and meteorological variables according to their cluster separation strength, the pattern of turbulent fluxes turned out to be three times more sensitive to photosynthetic photon flux density and vegetation type than air temperature. Furthermore, the results indicate a strong influence of the water vapor deficit on the magnitude of the turbulent exchange.



## 1. Introduction

The turbulent vertical exchange between the land surface and the atmosphere covers a significant portion of the global budget of energy, water, and carbon dioxide. These turbulent fluxes are a driving factor of climate on regional as well as global scales (Schimel et al., 2001). Because ecosystems and their respective exchange rates are influenced by varying environmental conditions, the atmospheric fluxes of carbon dioxide, water and energy themselves are considered as indicators for climatic changes. The complex interdependency of ecosystem processes and environmental conditions is the focus of much current research (Cox et al., 2000; Saleska et al., 2003; Thompson et al., 2004; Desai et al., 2008; Beer et al., 2010). Consequences of climate changes on the carbon budget on large spatial scales are challenging to predict due to inherent assumptions and uncertainties in various approaches. Meteorological parameters such as incoming radiation, temperature, and the availability of water are predicted to change globally due to ongoing massive anthropogenic emission of greenhouse gases (Solomon et al., 2007). Recent studies also indicate an acceleration of the global terrestrial carbon cycle (Bond-Lamberty and Thomson, 2010). Nevertheless, the sensitivity of different ecosystems to climate is still far from being understood (Petoukhov et al., 2000; Friedlingstein et al., 2006; Heimann and Reichstein, 2008). Several approaches have been taken to determine the magnitude of ecosystem response to climate variability, including process modeling (e.g. Dufresne et al., 2002; Friedlingstein et al., 2003; Gourdji et al., 2008; Heimann and Reichstein, 2008), statistical analysis of measurements (e.g. Law et al., 2002; Risch and Frank, 2010; Yadav et al., 2010), or time series analysis (e.g. Stoy et al., 2009). In this study we present a novel combination of methods to assess the importance of different meteorological and geographical conditions for the ability of ecosystems to exchange carbon dioxide, heat, and water vapor. The approach is based on the clustering of long-term flux data of the *AmeriFlux* network measured above different biomes across America (e.g. Hargrove et al., 2003). For this purpose artificial neural *Kohonen* networks or self-organizing feature maps (SOFM) were used to recognize similarities in the turbulent fluxes of $CO_2$, sensible heat, and latent heat. An SOFM network can be used as an advanced cluster analysis tool because of their ability to recognize similarities of presented input vectors (e.g. Kohonen, 1982, 1997; Patterson, 1996; Haykin, 1999). In recent years SOFM networks have been successfully applied in various scientific disciplines to find hidden structures in large or high-dimensional datasets (e.g. Yang et al., 1998; Wang et al., 2001; Tigrine-Kordjani et al., 2007; Ghasemi et al., 2009; Murty et al., 2009). The ability

* Corresponding author. Tel.: +1 541 737 6175; fax: +1 541 737 1393.
*E-mail address:* andres.schmidt@oregonstate.edu (A. Schmidt).

0168-1923/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.agrformet.2010.12.009

**Table 1**
AmeriFlux sites analyzed in this study.

| Site name | Ameriflux ID | State | Latitude [°] | Longitude [°] | Vegetation type | Incorporated years |
|---|---|---|---|---|---|---|
| ARM SGP Burn | US-ARb | OK | 35.550 | −98.040 | Grassland | 2005, 2006 |
| ARM SGP Control | US-ARc | OK | 35.547 | −98.040 | Grassland | 2005, 2006 |
| ARM SGP Main | US-ARM | OK | 36.606 | −97.489 | Cropland | 2002–2009 |
| Audubon Research Ranch | US-Aud | AZ | 31.591 | −110.509 | Grassland | 2002–2009 |
| Barrow | US-Brw | AK | 71.323 | −156.626 | Tundra | 1998–2002 |
| Bartlett Experimental Forest | US-Bar | NH | 44.065 | −71.288 | Deciduous broadleaf forest | 2004–2007 |
| Blodgett Forest | US-Blo | CA | 38.895 | −120.633 | Evergreen needleleaf forest | 1997–2007 |
| Bondville | US-Bo1 | IL | 40.006 | −88.290 | Cropland | 1996–2008 |
| Brookings | US-Bkg | SD | 44.345 | −96.836 | Grassland | 2004–2009 |
| Canaan Valley | US-CaV | WV | 39.063 | −79.421 | Grassland | 2004–2007 |
| Chestnut Ridge | US-ChR | TN | 35.931 | −84.332 | Deciduous broadleaf forest | 2006, 2007 |
| Cottonwood | US-Ctn | SD | 43.950 | −101.847 | Grassland | 2006–2009 |
| Donaldson | US-SP3 | FL | 29.755 | −82.163 | Evergreen needleleaf forest | 2000–2004 |
| Duke Forest Loblolly Pine | US-Dk3 | NC | 35.978 | −79.094 | Evergreen needleleaf forest | 1998–2005 |
| Fermi Agricultural | US-IB1 | IL | 41.859 | −88.223 | Cropland | 2005–2008 |
| Fermi Prairie | US-IB2 | IL | 41.841 | −88.241 | Grassland | 2004–2008 |
| Flagstaff Managed Forest | US-Fmf | AZ | 35.143 | −111.727 | Evergreen needleleaf forest | 2005–2008 |
| Flagstaff Unmanaged Forest | US-Fuf | AZ | 35.089 | −111.762 | Evergreen needleleaf forest | 2005–2008 |
| Flagstaff Wildfire | US-Fwf | AZ | 35.445 | −111.772 | Woody savanna | 2005–2008 |
| Fort Peck | US-FPe | MT | 48.308 | −105.102 | Grassland | 2000–2008 |
| Freeman Ranch Mesquite Juniper | US-FR2 | TX | 29.950 | −97.996 | Woody savanna | 2004–2006 |
| Goodwin Creek | US-Goo | MS | 34.255 | −89.874 | Grasslands | 2002–2006 |
| Great Mountain Forest | US-GMF | CT | 41.967 | −73.233 | Mixed Forest | 1999–2000 |
| Harvard Forest | US-Ha1 | MA | 42.538 | −72.172 | Deciduous broadleaf forest | 1992–2006 |
| Kendall Grassland | US-Wkg | AZ | 31.737 | −109.942 | Grassland | 2004–2007 |
| Kennedy Space Center Scrub Oak | US-KS2 | FL | 28.609 | −80.672 | Closed shrubland | 2003–2006 |
| Kennedy Space Center Slash Pine Flatwoods | US-KS1 | FL | 28.458 | −80.671 | Evergreen needleleaf forest | 2002,2003 |
| Little Washita | US-LWW | OK | 34.960 | −97.979 | Grassland | 1997,1998 |
| Marys River Fir Site | US-Fir | OR | 44.647 | −123.552 | Evergreen needleleaf forest | 2006–2009 |
| Mead Irrigated | US-Ne1 | NE | 41.165 | −96.477 | Cropland | 2001–2008 |
| Mead Irrigated Rotation | US-Ne2 | NE | 41.165 | −96.470 | Cropland | 2001–2008 |
| Mead Rainfed | US-Ne3 | NE | 41.180 | −96.440 | Cropland | 2001–2008 |
| Metolius First Young Pine | US-Me5 | OR | 44.437 | −121.567 | Evergreen needleleaf forest | 2000–2002 |
| Metolius Intermediate Pine | US-Me2 | OR | 44.452 | −121.557 | Evergreen needleleaf forest | 2002–2007 |
| Metolius Old Pine | US-Me4 | OR | 44.499 | −121.622 | Evergreen needleleaf forest | 2000 |
| Missouri Ozark | US-MOz | MO | 38.744 | −92.200 | Deciduous broadleaf forest | 2004–2008 |
| Mize | US-SP2 | FL | 29.765 | −82.245 | Evergreen Needleleaf Forest | 2001–2004 |
| Morgan Monroe State Forest | US-MMS | IN | 39.323 | −86.413 | Deciduous broadleaf forest | 2000–2002, 2006–2007 |
| Niwot Ridge | US-NR1 | CO | 40.033 | −105.546 | Evergreen needleleaf forest | 1998–2007 |
| North Carolina Clearcut | US-NC1 | NC | 35.811 | −76.712 | Evergreen needleleaf forest | 2004–2006 |
| North Carolina Loblolly Pine | US-NC2 | NC | 35.803 | −76.668 | Evergreen needleleaf forest | 2005–2007 |
| Park Falls | US-PFa | WI | 45.946 | −90.272 | Mixed Forest | 1996–2001 |
| Ponca Winter Wheat | US-Pon | OK | 36.767 | −97.133 | Cropland | 1997–2000 |
| Santa Rita Mesquite Savanna | US-SRM | AZ | 31.821 | −110.866 | Woody savanna | 2004–2007 |
| Shidler Tallgrass Prairie | US-Shd | OK | 36.933 | −96.683 | Grassland | 1997–2000 |
| Sky Oaks New | US-SO4 | CA | 33.384 | −116.640 | Closed shrubland | 2004–2006 |
| Sky Oaks Old | US-SO2 | CA | 33.374 | −116.623 | Woody savanna | 1997–2006 |
| Sky Oaks Young | US-SO3 | CA | 33.377 | −116.623 | Closed shrubland | 1997–1998, 2004–2006 |
| Sylvania Wilderness | US-Syv | MI | 46.242 | −89.348 | Mixed Forest | 2001–2006 |
| Tonzi Ranch | US-Ton | CA | 38.432 | −120.966 | Woody savanna | 2001–2008 |
| UMBS | US-UMB | MI | 45.560 | −84.714 | Deciduous broadleaf forest | 2004–2006 |
| Vaira Ranch | US-Var | CA | 38.407 | −120.951 | Grassland | 2000–2008 |
| Walker Branch | US-WBW | TN | 35.959 | −84.287 | Deciduous broadleaf forest | 2000–2007 |
| Walnut River | US-Wlr | KS | 37.521 | −96.855 | Grassland | 2001–2004 |
| Willow Creek | US-WCr | WI | 45.806 | −90.080 | Deciduous broadleaf forest | 1999–2006 |
| Wind River Crane Site | US-Wrc | WA | 45.821 | −121.952 | Evergreen needleleaf forest | 1998–2006 |

to recognize complex patterns along multiple dimensions makes SOFM networks an ideal tool to analyze large datasets of turbulent fluxes and find similarities in flux patterns and cluster the dataset accordingly. After that the respective flux patterns can be examined with respect to the influencing meteorological variables. When clustering the flux vectors, the data are indirectly sorted according to the meteorological and environmental parameters which are known to drive the magnitudes of turbulent exchanges.

The present approach allows this inherent information about the drivers to be extracted and evaluated. Through statistical analysis of the cluster differences we were able to rank the available environmental parameters according to their importance on the magnitude of the turbulent exchange rates between the surface and the atmosphere within the boundary layer. The advantage of the presented inductive method is that the results are derived directly in response to the relations and magnitudes of observed net $CO_2$ fluxes, sensible heat fluxes, and latent heat fluxes without any model uncertainties, previously made assumptions or approximations.

## 2. Data and methods

### 2.1. The AmeriFlux source data

We used observation data collected within the past 17 years from 56 *AmeriFlux* sites in the USA (Table 1), totaling 305 site

**Table 2**
Available variables used for the clustering and the further analyses of the clusters.

| Active data (used for clustering) Flux data | Passive data (used for interpretation) | | |
|---|---|---|---|
| | Meteorological data | Geographical data and site information | Time-related information |
| | Air temperature | | |
| | Wind velocity | | |
| | Wind direction | | |
| | Soil temperature | Latitude | |
| Net $CO_2$ flux | Precipitation | Longitude | Year |
| Latent heat flux | | | Day of year |
| Sensible heat flux | | | |
| | Relative humidity | Elevation of site | |
| | Vapor pressure deficit | Vegetation type | |
| | Net radiation | | |
| | Incoming PPFD | | |
| | Friction velocity | | |

years of data. To evaluate the effect of various environmental parameters on exchange rates of different biomes included in the *AmeriFlux* database (Baldocchi et al., 2001; Hargrove et al., 2003), the net $CO_2$ flux (*FC*), sensible heat flux (*H*), and latent heat flux (*LE*) were used as input variables to build clusters. The *AmeriFlux* Level 2 database analyzed in this study (ftp://cdiac.ornl.gov/pub/ameriflux/data/Level2, April 2010) contains datasets with a temporal resolution of 30 min or 60 min. For the purpose of standardization and to remove the effect of the diurnal course of the turbulent fluxes on the clustering procedure, daily values were used by averaging or summarizing, respectively. This also reduced the effects of sampling random errors in the measuring data (Baldocchi, 2003). Furthermore, by using daily data records the computational cost decreased to a maintainable amount.

The exchange of carbon dioxide, water vapor, and energy are the variables of interest to describe biological exchange activity of ecosystems (e.g. Baldocchi et al., 2001; Haupt-Herting et al., 2001). Thus, only those *AmeriFlux* L2 data records with available values of net $CO_2$ flux, latent heat flux and sensible heat flux were extracted. Since the procedure is based on the clustering of flux data and the comparison of the respective meteorological variables the algorithm is not affected by discontinuous time series. Thus, as every gap filling implies an additional error for the data which might influence the SOFM procedure we used only the available, measured Level 2 data of the *AmeriFlux* database.

To remove outliers and peaks in the data series statistically we applied a 5-$\sigma$ criterion, i.e. if a value deviated from the average of the series by more than 5 standard deviations it was considered an outlier and was excluded from further analyzes. This chosen factor of 5 is in accordance with the *Chebyshev's Theorem* and implies that at least 96% of the statistical acceptable data are within these boundaries without restriction related to the data distribution (e.g. Gnedenko, 1988; Amidan et al., 2005).

Several incomplete data records were excluded in order to make sure that the used data are sufficient to represent the patterns that occur in the turbulent ecosystem exchange of the 56 sites to allow a meaningful interpretation. In particular, the data records needed to include all available variables that potentially influence the fluxes as well as the spatially and temporally corresponding flux values. Therefore, only variables with an availability of at least 80% within the analyzed period from 1992 through 2009 were selected for further analyses to achieve a complete and comprehensive dataset. After quality control and the removing of outliers, 78,691 daily data records each comprising 19 variables (Table 2) remained for further analyses. Hence, in addition to the fluxes of *LE*, *H*, and $CO_2$, the final dataset also included the corresponding environmental conditions such as vegetation type, meteorology, geographical information

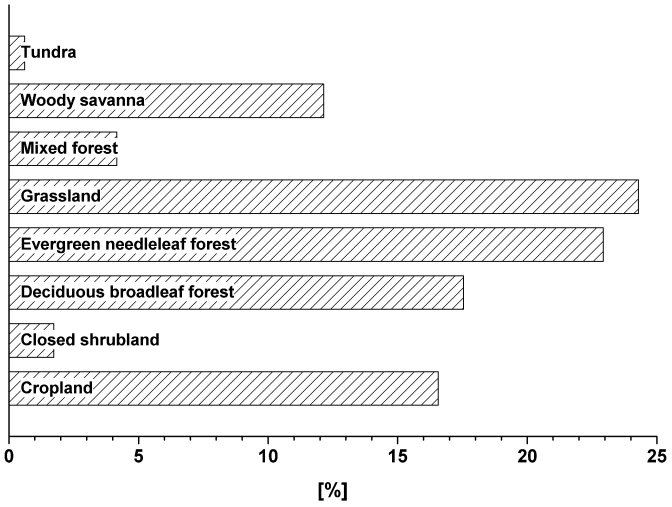


**Fig. 1.** The 8 vegetation types represented in the analyzed long-term *AmeriFlux* dataset and the percentages of the respective origins of the available flux data records.

(i.e. latitude, longitude, and elevation), and the date (i.e. year and DOY).

The active part of the dataset consists of the values of the three fluxes that were clustered, whereas the passive data variables were only used for the interpretation of the clustering results but were not included in the clustering procedure itself (Table 2).

The vegetation types of the different measurement sites are classified based on the 17 IGBP land cover types (International Geosphere-Biosphere Programme).

The used dataset covers 8 different vegetation types above which the turbulent fluxes where measured. Their percentages are given in Fig. 1.

The available dataset is dominated by fluxes measured above grasslands and evergreen needleleaf forests. However, due to the large spatial and temporal range of the comprehensive dataset (Table 1), the other vegetation types are also represented to examine the effect of the vegetation type and further environmental parameters on the magnitude of the estimated turbulent fluxes.

### 2.2. Self-organizing feature map neural networks

SOFM neural networks consist of two layers and exhibit only one layer that contains active and adaptable neurons. The functions in these active neurons are used to map the higher-dimensional input data onto a lower-dimensional, mostly two-dimensional topologi-

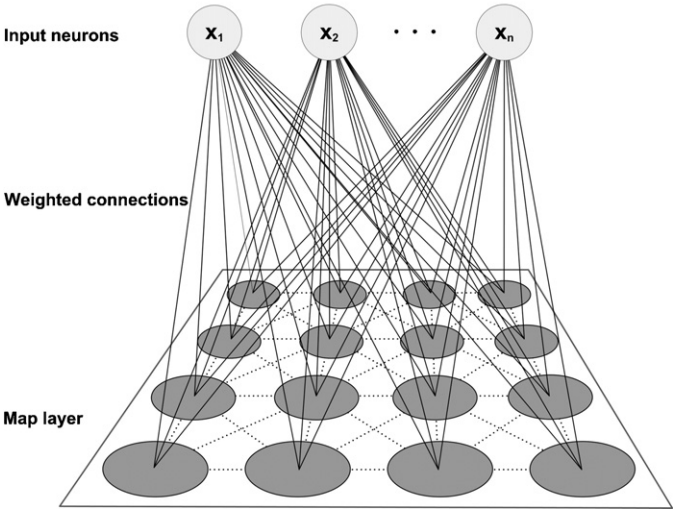


**Fig. 2.** Schematic layout of a self-organizing feature map network. The $n$ input neurons are connected to each neuron in the map lattice. The neurons in the map layer are also connected to each other. The actual network used in this study consists of 225 neurons.

cal grid (i.e. map) whereas the neurons in the input layer are only used to transfer the input vectors to the active nodes arranged in the SOFM plane (Fig. 2).

The general advantage of neural networks, i.e. that no previous information about the data are needed because of their ability to learn, also applies to SOFM networks. SOFM networks learn to recognize groups of similar input vectors so that neurons that are spatially near to each other in the map layer respond to similar input vectors. Thus, this topology preservation ensures that similar input vectors that are close together in the $n$-dimensional input space will be assigned to neurons that are close together on the flat feature map (or even to the same neuron). All computations for this contribution including development of the SOFM were performed with MATLAB® (Rev. 2009b, The MathWorks, Inc., Natick, Massachusetts, USA).

In contrast to common feed-forward neural networks, that are used e.g. for regression purposes or data gap filling in flux time series (Van Wijk and Bouten, 1999; Papale and Valentini, 2003; Schmidt et al., 2008) there is no comparison between a target value of a training dataset and the output of the network to adopt the internal network parameters by a back propagation from the output layer. An SOFM adapts its spatial structure during an unsupervised learning procedure and learns to recognize the distribution of a given input dataset as well as its topology in the input space (Kohonen, 1997; Du, 2009).

Every active neuron in the map grid layer is also connected to a weight vector that can be adapted to the input data. During each epoch (learning step), all input vectors are transferred in a random order to all neurons in the map grid. For each input vector $x$ presented to the network, the competitive learning procedure determines the neuron with a weight vector that is nearest to the current input vector. Based on this principle, the winning neuron can be determined by finding the minimum of the Euclidian distance $d$ between the input $x$ and all weights $w$ in the $n$-dimensional input space (Eq. (1)), with

$$d_{\min} = \min \left|\left| x - w_j \right|\right| = \min \left[ \sum_n (x_n - w_{jn}) \right]^{1/2}. \quad (1)$$

After that, only the weight vector of this winning neuron and the weights of its assigned neighbors are updated and shifted even closer towards the respective input vector. As a consequence, the activation of the adapted neurons will be even higher if a similar input is presented to the SOFM during the remaining learning epochs. To obtain the same range between 0 and 1 for all used input vector components the raw values $x_i$ of all variables were linearly transformed using the *min–max normalization* before being transferred to the SOFM neurons (Kohonen, 1990; Priddy and Keller, 2005) as given in Eq. (2),

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \quad (2)$$

Here, $x_{\max}$ is the maximum and $x_{\min}$ the minimum of the respective variable in consideration of the complete used dataset. For the interpretation of the results these values can be reconverted to the real physical units and original range.

The SOFM network approach is among vector quantization algorithms and can be considered an enhancement of the $k$-means clustering method (McQueen, 1965; Broomhead and Lowe, 1988; Likas et al., 2003). During a $k$-means clustering procedure the input vectors are partitioned into to $k$ clusters, and the neuron centers are set to represent the best center – 'best centroid' in a multi-dimensional space for the respective cluster. The algorithm iteratively determines the minimum of the sum-of-squares clustering-function $J$ given as,

$$J = \sum_k \sum_{i \in S_k} \left|\left| w_{ik} - c_k \right|\right|^2, \quad (3)$$

with

$$c_k = \frac{1}{N_k} \sum_{i \in S_k} w_i, \quad (4)$$

Here $c_k$ is the mean of the data subset $S_k$ with $N_k$ elements, building the cluster $k$ (Bishop, 1995). At the end of this clustering procedure, each weight vector $w$ is assigned to the cluster center $c_k$ that is nearest. A radial basis distance function (e.g. a Gaussian function) with center $c$ and a defined width of its basis can be used as representation of the cluster. A more detailed description of the $k$-means clustering method using radial basis functions is given in e.g. McQueen (1965), Likas et al. (2003), Ripley (2005), or Hair et al. (2009). In contrast, during the SOFM learning procedure not only the winning neuron is updated but also its neighbors in a defined range. This neural interconnectivity gives the SOFM a better ability to recognize similarities of datasets that show a large variety and range compared to the $k$-means clustering algorithm (Clare and Cohen, 2001).

The strength of the final neighbor connections between the neurons in the map layer are defined by distances between their weight vectors in the input space or *feature space* (Kohonen, 1990; Tavan et al., 1990). The update of a the weight vectors $w_j$ after epoch $l$ is given by,

$$w_j(l+1) = w_j(l) + \eta(l) \cdot \phi(l) \cdot (x(l) - w_j(l)). \quad (5)$$

Where the learning rate $\eta$ is a value between 0 and 1 that describes the steps, with which the weight vector values are iteratively adjusted. Both, the learning rate $\eta$ and the value of the neighborhood range function $\phi(l)$ decrease with increasing $l$ (Kohonen, 1990; Mulier and Cherkassky, 1994). In this study the exponentially decreasing learning rate was calculated according to,

$$\eta(l) = \eta_{\text{start}} \left( \frac{\eta_{\text{end}}}{\eta_{\text{start}}} \right)^{l/l_{\max}}. \quad (6)$$

For the network used in this study the settings $\eta_{\text{start}} = 0.9$ as the initial learning rate and $\eta_{\text{end}} = 0.01$ as the minimum learning rate were applied. The total number of reiterations of the input vector presentations to the network was set to $l_{\max} = 20,000$. The chosen initial and ending learning rates are commonly used values to achieve

a coarse and fast adoption of the weight vectors in the beginning and ending with a small learning rate for the final accurate weight vector optimization. As long as the learning rate exhibits a clear gradient the results are not highly sensitive to the choices of $\eta_{start}$ and $\eta_{end}$. The optimal value for $l_{max}$ was found by iteratively increasing the number of reiterations until the SOFM clustering results (i.e. node allocations) remained stable.

A Gaussian function (Eq. (7)) was chosen as neighborhood range function for the SOFM training in this study,

$$\phi(c_j, l) = \exp\left[\frac{-\left\|c_{win} - c_j\right\|^2}{(\eta(l) \times m)^2}\right]. \quad (7)$$

Here $c_j$ gives the center position of neuron number $j$ in the map lattice, $m$ is the number of nodes per dimension in the topological map (in our case 15) and $c_{win}$ is the vector representing the cell position of the winning neuron in the flat map grid.

The expression $(\eta(l) \cdot m)^2$ in the two-dimensional Gaussian function (Eq. (7)) defines the radius of the winning neuron's neighborhood. Hence, the closer the neighbors are to the center of the wining neuron, the higher the value of the respective bell-shaped RBF surface point and the bigger the modification of the respective adjacent weight vector since $\phi \to 0$ if $\left\|c_{win} - c_j\right\|$ increases.

This gradual decrease of the neighborhood and the learning rate leads to a coarse sorting of the input data at the beginning of the learning process and a fine-tuning of the center positions during the final epochs. Moreover, because of the monotonically decreasing functions $\eta(l)$ and $\phi(l)$ the system converges to a stable state when $l = l_{max}$.

Therefore, the final state of a neurons weight vector is the result of $l_{max}$ functional combinations of its weight vector position in the input space (Eq. (1)) and its position in the flat map lattice (Eqs. (5)–(7)). According to this, the SOFM algorithm implies the reduction of dimensions from the higher-dimensional input space to the two-dimensions of the map plane.

A reduction of the dimensionality is always attended by a loss of information (Roweis and Saul, 2000; Hastie et al., 2001). It has been proven that by projecting the spatial arrangement of the input values on the flat output map, the strength of the neighbor connections between the neurons in the SOFM are also capable of representing the distances of the assigned input vectors in the $n$-dimensional input space (Kohonen, 1990, 1997; Haykin, 1999). For this reason SOFM networks are an appropriate method to keep the loss of information that originally can be found in the input data as small as possible to gain meaningful data accumulations that represent the physical similarity of the fluxes in the respective clusters.

### 2.3. Combination of SOFM pre-clusters

At the end of the learning procedure, areas of neurons with small distances between their weight vectors will have developed on the map. These agglomerations of neurons are then combined to the final clusters. This was done to get a manageable amount of clusters used for the further analyses and interpretations. Without knowing the actual number of clusters in the dataset, the number of initial reference vectors (i.e. the number of SOFM neurons) can be chosen irrespective of the number of real clusters. The network arranges itself around the present clusters while recognizing the distribution and the topology of the data in the input space iteratively (Kohonen, 1984; Rojas, 1996; Haykin, 1999).

In order to find the best network topology, we applied quantitative criteria by slightly increasing the number of neurons until empty pre-clusters appeared in the resulting map. This was done to set the number of nodes to an appropriate amount for the current dataset. Secondly, by finding an acceptable low value for the summarized distances of each input vector to its respective nearest weight vector, the clustering performance of the tested networks was quantified (Vesanto and Alhoniemi, 2000; Liu et al., 2007).

Several SOFM networks with various numbers of nodes and topologies were tested to find the most appropriate SOFM for the given problem. An SOFM network with the topology 3–225 (three input neurons and a $15 \times 15$ SOFM lattice) turned out to exhibit a map with no empty SOFM pre-clusters and the lowest sum of distances. As a radial basis function was used to calculate the neighborhood activation, a hexagonal layout of the neurons was chosen so that, with exception of the neurons at the borders of the lattice, every neuron is equidistantly surrounded by 6 direct neighbor neurons. Furthermore, a hexagonal layout has been proven to provide a good topology preservation of the input data (Arsuaga-Uriarte and Diaz-Martin, 2005).

Due to its high performance in finding the centroids of spherical clusters (McQueen, 1965, 1967) the $k$-means algorithm was applied in a following step after the first 225 pre-clusters were built by the SOFM. In contrast to the application of an SOFM network the number of clusters should be chosen prior to the analysis according to the expected number of clusters when applying the $k$-means clustering algorithm (Duda et al., 2001, Hair et al., 2009). To find the appropriate number of final clusters in the second step (Fig. 3), the Davies–Bouldin validity index $I_{DB}$ (Davies and Bouldin, 1979; Jain and Dubes, 1988; Maulik and Bandyopadhyay, 2002) was used to merge the SOFM pre-clusters into the final clusters. The Davies–Bouldin validity index incorporates the within-cluster distances between each cluster member and the center of the respective cluster as well as the between-cluster distances. Using this method based on the *average-linkage* values (Eq. (8)) the appropriate number of final clusters for the used dataset was found by determining the minimum of $I_{DB}$, with,

$$I_{DB}(k) = \frac{1}{k}\sum_{q=1}^{k} \max_{l \neq q} \left[\frac{(1/N_q)\sum_m \left\|w_m - c_q\right\| + (1/N_l)\sum_n \left\|w_n - c_l\right\|}{\left\|c_q - c_l\right\|}\right]. \quad (8)$$

Here, $k$ is the number of final clusters, $c_l$ and $c_q$ give the centers positions of two final clusters at a time, $N$ is the number of elements in the clusters, and $w$ gives the SOFM weight vector positions assigned to the SOFM pre-clusters (i.e. neurons) $q$ and $l$, respectively. Consequently, the Davies–Bouldin index has a small value for a good clustering that exhibits well-separated and compact spherical clusters. Thus, within the clustering procedure presented in this work the sorting of the turbulent exchange expressed by *FC*, *H*, and *LE* acts on two levels. First, similar input vectors are merged into several groups that are close together on the flat output map by application of an SOFM network. Afterwards, these pre-clusters, composed of agglomerations of the weight vectors of neurons responding to similar inputs, are grouped together in the final clusters by using a $k$-means clustering algorithm. Since the pre-clusters built by the SOFM are local averages of the fluxes the clustering of the pre-clusters is less sensitive to random variations than the clustering of the original data (Vesanto and Alhoniemi, 2000).

### 2.4. Cluster analysis and interpretation

After the active variables (i.e. the flux vectors) consisting of the *FC*, *H* and *LE* values were arranged in the final clusters, the corresponding passive variables (i.e. the environmental parameters) were extracted from the time series data records assigned to the different clusters. By referring back to the complete 19-dimensional daily data records to which the clustered flux vectors belong, the influences of the simultaneously measured environmental param-

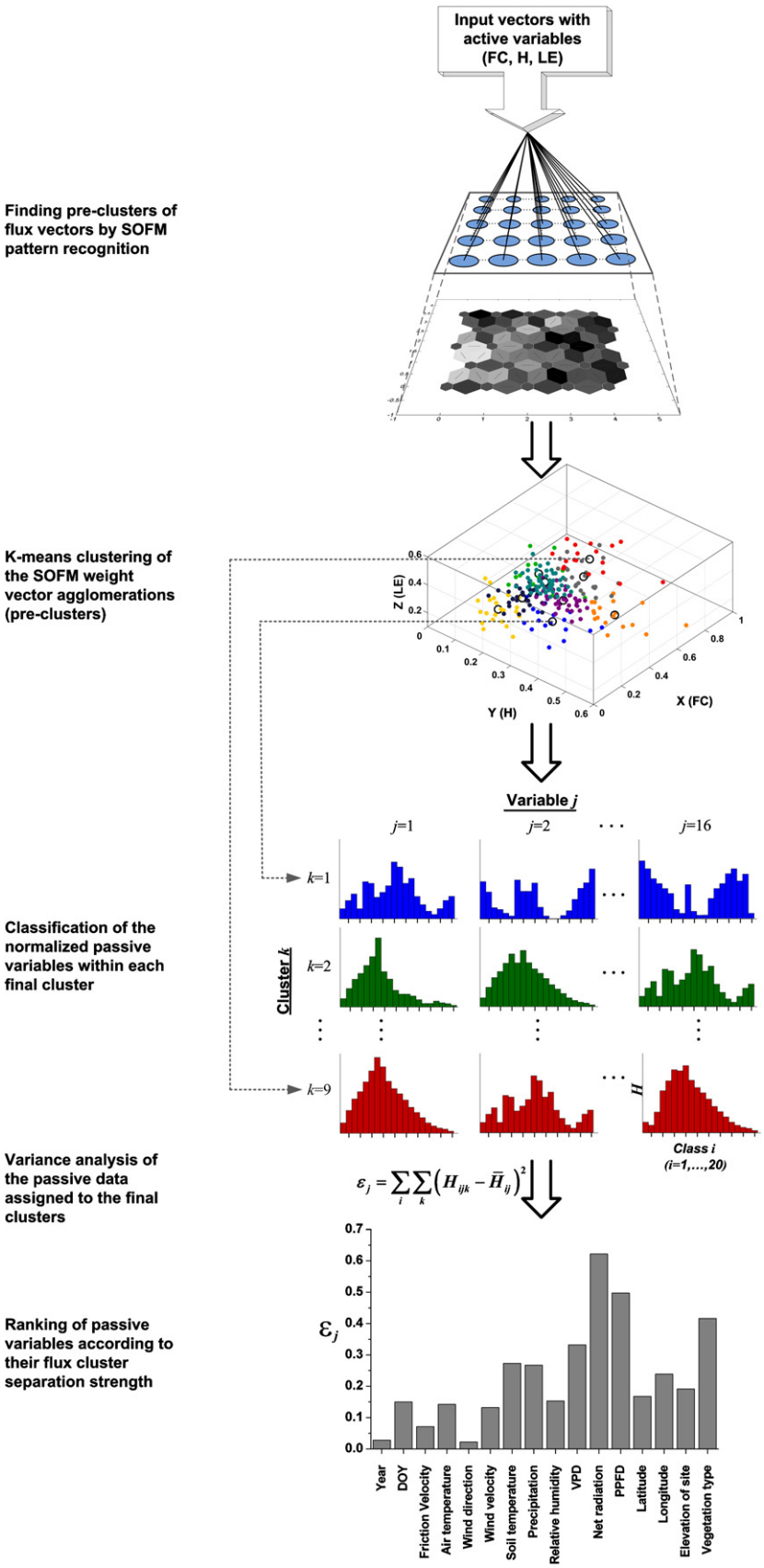


**Fig. 3.** Schematic overview of the applied procedure for the flux data clustering and analysis. The clusters have only been built based on the active variables (i.e. flux values) while the environmental conditions that affected the fluxes are quantitatively described by the corresponding passive variables.

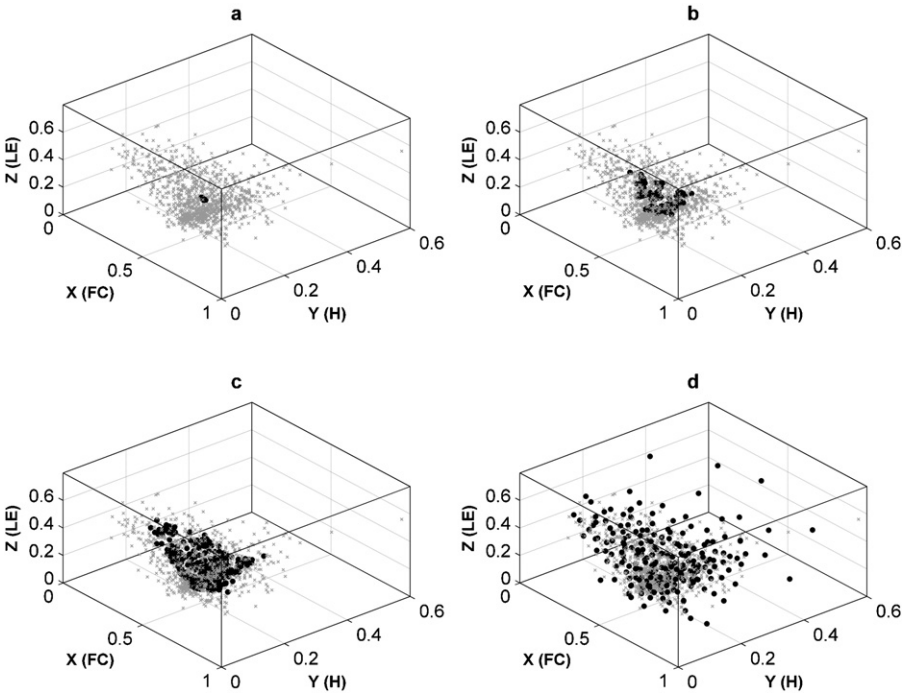


**Fig. 4.** The unsupervised adaption of the weight vector positions (black points) to the topology of the linearly normalized inputs (gray crosses) after 1 (a), 50 (b), 500 (c), and 10,000 (d) epochs. For the purpose of visibility only every 70th input is shown.

eters on the magnitude of the flux vectors were assessed. For this purpose, the variation of the frequency distributions of each passive variable among the different clusters has to be determined. The dissimilarities of the distributions of the passive variables have to be examined along two dimensions of which one is the cluster number $k$ the respective data value $x_{ijk}$ of variable $j$ is assigned to. The second dimension is the class $i$ within the distribution of variable $j$ in cluster $k$.

The more different the distributions of a variable among the clusters the higher its influence on the fluxes as the clusters were mathematically only built based on the flux input vectors. The fluxes *FC*, *H*, and *LE* in the final clusters are assorted (clustered) to be as different as possible among the clusters but as similar as possible within the single clusters. Hence, if a passive variable such as the air temperature would be very similar among all clusters the influence on the fluxes obviously is not high as the corresponding flux vectors differ as much as possible among the clustered groups.

The Assessment of dissimilarity of the passive variables in the clusters was achieved by calculating the error $\varepsilon$ as a measure for the dissimilarity of a variable among the different clusters. This inter-cluster variation of each passive variable $j$ was calculated according to,

$$\varepsilon_j = \sum_i \sum_k (H_{ijk} - \bar{H}_{ij})^2, \tag{9}$$

where $H_{ijk}$ gives the relative frequency of a class $i$ of variable $j$ found in the values allocated to the cluster $k$. $\bar{H}_{ij}$ denotes the average over all clusters of the frequencies of the respective class. Within some distributions of variables in the clusters not all classes are occupied. Using the $\varepsilon$ value as given in Eq. (9) provides the advantage that unoccupied classes of discontinuous variables such as elevation or vegetation type do not influence the Epsilon sums which therefore allows an unbiased assessment of the dissimilarity between the frequency distributions. By comparison of the $\varepsilon_j$ values the relative strength of the influence of the passive variables on the observed fluxes can be derived and compared.

Since all classes for all normalized variables in all clusters have the same range, the ratio between the 16 resulting $\varepsilon$-values of the passive variables is not sensitive to the absolute number of classes as long as the classification is not chosen too coarse. Otherwise, merging values which differ significantly would lead to a loss of statistical information between those values and the corresponding fluxes. This effect did not occur as a subdivision into 20 classes was conducted. The number of classes was chosen according to the respective dataset in order to assure a fine enough classification that represents the raw data as exactly as possible on the one hand and without exhibiting too many empty classes in the various distributions on the other hand.

This method enabled us to get the required information about the flux cluster separation strength of the passive variables giving the environmental conditions during the recording of the corresponding turbulent fluxes. Here, the information about the cluster separation strength is represented by the $\varepsilon_j$ values that give the accumulated inter-cluster differences of the relative frequencies (Eq. (9)).

A schematic diagram of the applied multi-step procedure is presented in (Fig. 3).

## 3. Results and discussion

As the used input vectors consist of the net $CO_2$ flux, the sensible heat flux, and the latent heat flux, the weight vectors of the SOFM neurons are also three-dimensional $w_j(w_{j1}, w_{j2}, w_{j3})$. Therefore, the adaption of the neurons' weight vectors during the learning phase can be directly visualized to show the performance of the network. The iterative adaption of the weight vectors to the flux vectors is clearly cognizable (Fig. 4).

The unified distance matrix (U-matrix) was used to visualize the SOFM accumulations on the resulting two-dimensional feature map (Fig. 5). The U-matrix shows the distances between the higher-dimensional weight vectors of the neurons that are arranged in the map lattice (Ultsch and Siemon, 1990; Vesanto, 1999; Nikkila et al., 2002; Taşdemir, 2008). Here, a large distance between the neurons

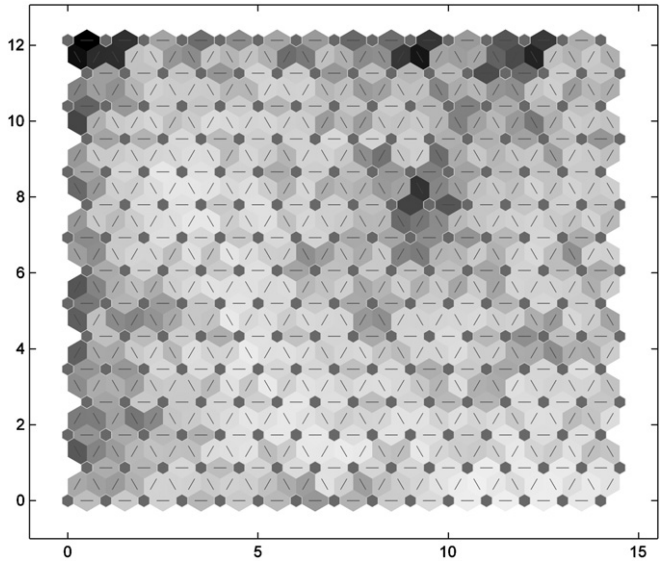


**Fig. 5.** Visualization of the neighbor weight vector distances on the two-dimensional SOFM by the U-matrix. The hexagons mark the positions of the single feature map neurons.

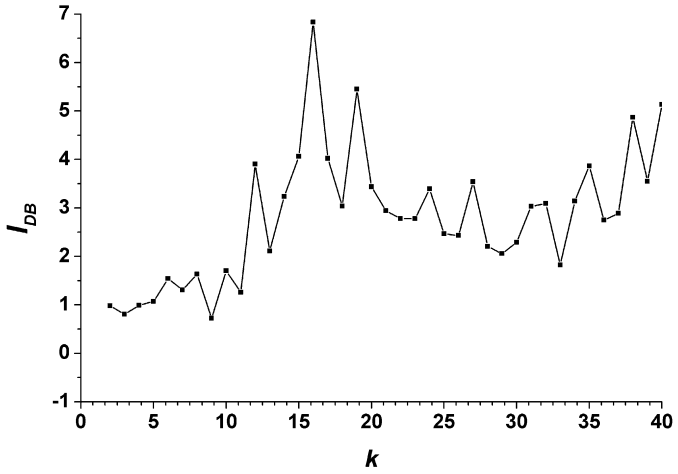


**Fig. 6.** The Davies–Bouldin validity index vs. the number of final clusters $k$ composed of the 225 SOFM pre-clusters.

is indicated by a darker shade of the connection between the neurons, which can be interpreted as a class border. The lighter areas indicate a small distance between the neurons (i.e. their weight vectors), indicating closer proximity to the weight vectors in the input space that is spanned by the flux vectors $x$(*FC*, *H*, *LE*). Hence, light shaded areas can be interpreted as clusters. Fig. 5 shows the visualized U-matrix of the network after 20,000 learning epochs. Several agglomerations of weight vectors associated with the map neurons can be recognized.

In the lower right area of the map a bright area is shown for instance. Also the separated pre-cluster that consist of a single neuron in column 10 and row 8 is visible. This neuron obviously has values assigned that differ significantly from the other flux vectors. Also neurons in the first row at columns 2, 9, and 12, exhibit high distances to all other neighbors.

Fig. 5 shows the ability of the SOFM to recognize data peaks and to detect outliers when used for data mining procedures (e.g. Kohonen, 1997; Munoz and Muruzabal, 1998). Although outliers in a statistical sense were removed previous to the SOFM training, these values are still rare extreme values. The respective 5 extreme value pre-clusters that altogether contain 237 assigned flux values were excluded for the following $k$-means algorithm to improve the final clustering and therewith the reliability of the results. The plotted U-matrix gives only a first impression of the success of the pre-clustering and helps to ensure that similar data groups actually have been recognized.

It has been proven that the applied two-step clustering procedure offers the advantage of a high computational efficiency and more reliable clustering results compared to a direct reduction of the number of SOFM nodes to the number of final clusters (Vesanto and Alhoniemi, 2000). The SOFM is superior in finding similar multidimensional values compared to a direct $k$-means clustering especially if the values cover a wide range and contain random noise (Liu and Gader, 2002; Mohebi and Sap, 2009). Nevertheless, when the values are statistically represented in the spherical input space by the SOFM weight vectors, the $k$-means algorithm is an appropriate and common method to find the final clusters quickly (Duda et al., 2001). Hence, the data can be reduced to a clearly arranged number of clusters with simultaneous consideration of the variability of the fluxes. In addition, the noise reduction is another benefit of the two-step clustering approach.

To find the number of final clusters in the feature map, a more accurate and distinct method using the Davies–Bouldin validity index (Eq. (8)) was applied.

The validity index $I_{DB}$ was calculated for a number of 2 through 40 clusters. As shown in Fig. 6, $I_{DB}$ has its minimum when the 225 SOFM pre-clusters are combined to 9 final clusters. Thus, the 225 SOFM pre-clusters are arranged around the 9 centers of the final clusters so that each SOFM pre-cluster is assigned to its nearest center. The different flux vector properties of the clusters can already be derived from the three-dimensional arrangement of the cluster centers and the cluster members. The values merged in cluster 8 exhibit the highest positive $CO_2$ fluxes whereas the fluxes indicating the highest uptake of $CO_2$ are assigned to cluster 2 (Fig. 7).

Furthermore, the spatial separation of the pre-clusters shown in Fig. 7 also proves the ability of the SOFM to recognize groups of flux values that differ significantly as well as to merge data that show similar fluxes. A summary of the properties of the flux vectors in the resulting clusters is given in Fig. 8. Low mean daily exchange rates of $CO_2$, sensible heat, and latent heat are mostly assigned to cluster 1. This is explained by the fact that this cluster contains a high portion of low wintertime data of the incorporated *AmeriFlux* sites with low plant activity and more stable atmospheric conditions. This explains the small average of $H$ (Fig. 8) in cluster 1 and also the low incoming shortwave radiation and PPFD values which average $51.08 \pm 40.92\,\mathrm{W\,m^{-2}}$ and $245.33 \pm 163.25\,\mu\mathrm{mol\,m^{-2}\,s^{-1}}$, respectively. By contrast the distribution of months associated with the flux measurements in the other clusters exhibit higher frequencies in the periods with increased plant activity in terms of high photosynthesis and respiration rates in spring and summer.

Even though the cluster averages have been calculated from various independent fluxes including measurements of different sites and years, the final cluster data shows clear patterns. Irrespective of the spatial and temporal distribution of the original fluxes this allows an interpretation of the functional relations between driving environmental and meteorological parameters and the corresponding clustered fluxes.

The high $CO_2$ uptake in cluster 2 in connection with the high mean fluxes of *LE* and *H* shows cases with a well developed turbulence regime where the respective ecosystems act as remarkable $CO_2$ sinks. The high standard deviations in relation to the mean values of *LE* (39%) and *H* (99%) in cluster 2 indicate that the high $CO_2$ uptake takes place under variable meteorological conditions. Thus, even if the heat fluxes are low, the average $CO_2$ uptake, which shows a lower standard deviation (±19%), is still high in most of the cases. Cluster 3 and 4 also exhibit a noticeable $CO_2$ uptake and signifi-

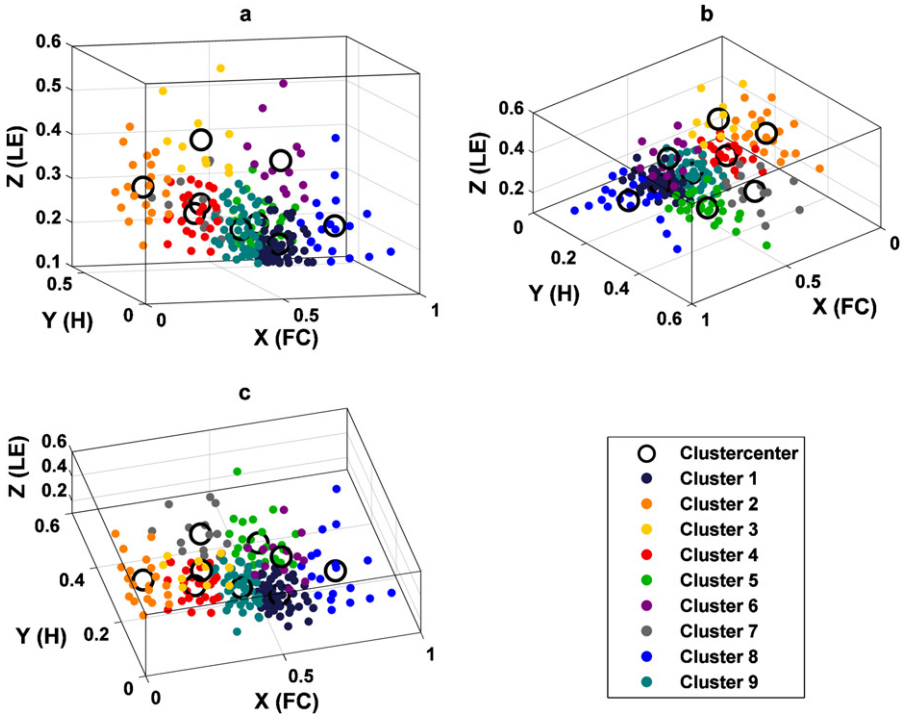


**Fig. 7.** The SOFM weight vectors marked according to their cluster affiliations and the position of the centers of the final clusters shown from various view angles using a right-handed coordinate system (azimuth:−15° elevation:10° (a), azimuth:140° elevation:50° (b), and azimuth:−15° elevation:70° (c)).

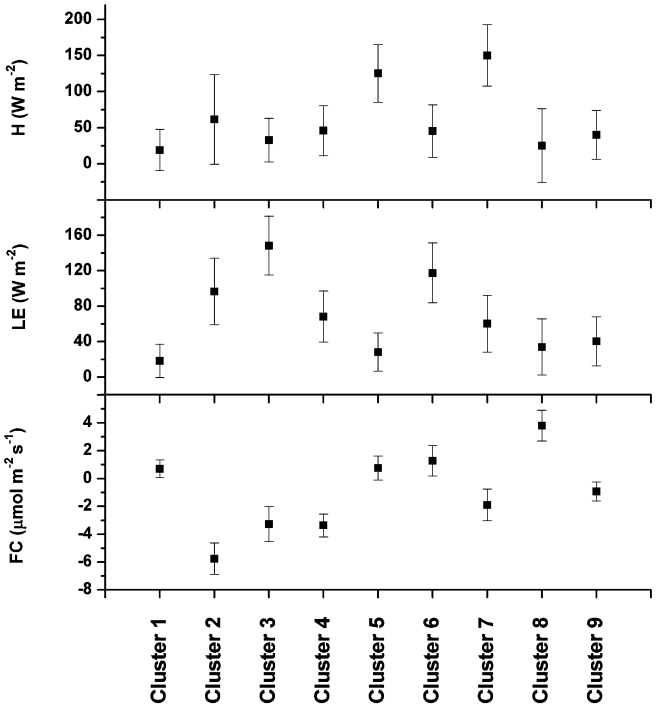


**Fig. 8.** The mean values of the measured fluxes in the final clusters. The error bars show the $\pm 1\sigma$ standard deviations to indicate the variability within the clusters.

cant energy exchange. In addition, cluster 3 contains fluxes under meteorological conditions with a higher amount of available water and air moisture causing the increased transport of latent heat. By contrast, the fluxes that are assigned to cluster 8 show the vegetated areas acting as sources of $CO_2$. Because the corresponding mean latent heat fluxes are small, and we know that physiologically, gross photosynthesis and transpiration are coupled through stomatal activity, it can be assumed that situations with low gross photosynthesis rates are numerously represented in this cluster. Respiration is also likely higher when sensible heat flux is high, as long as water is not limiting.

A more detailed and sites specific analysis can be achieved by setting the clustering results in a spatial relation to the incorporated *AmeriFlux* tower sites.

### 3.1. Spatial flux patterns and land-cover

On the map of the USA shown in Fig. 9, the cluster affiliations of the different sites that were analyzed in this study are shown. Due to the high variability of turbulent fluxes over the years and the variability of the relations between *FC*, *H*, and *LE* the fluxes that were measured at single sites are consequently not all assigned to the same cluster. The bars at the measurement tower positions on the map of mainland USA (Fig. 9) show the percentage cluster affiliations of the flux vectors measured at the respective site. The IGBP ecosystem related information is extracted from MODIS (Moderate Resolution Imaging Spectroradiometer) data of 2007. The pixels were aggregated to adopt the spatial resolution (0.02°) to the scale of the mapped area presented in Fig. 9.

The map generally shows that sites in the Northeast and the Midwest areas have a higher percentage of fluxes affected by the cold temperatures and low ecosystem exchange rates of $CO_2$ and energy during wintertime (cluster 1). Furthermore, large areas of cropland in the Midwest are partially not tilled during this time. In contrast, the sites in the Pacific Northwest show more fluxes from the clusters 4, 5, and 7. As shown in Fig. 9, the northwest sites are dominated by evergreen needleleaf forests (ENF) that are still able to take up $CO_2$ during some periods of the relatively mild winters. This explains the ongoing moderate uptake of $CO_2$ that is represented in cluster 4. Higher NEE values of ENF systems in comparison to DBF ecosystems, both exposed to similar temperature and precipitation regimes, were also observed by Desai et al. (2008). However, in contrast to this work, the previous study was

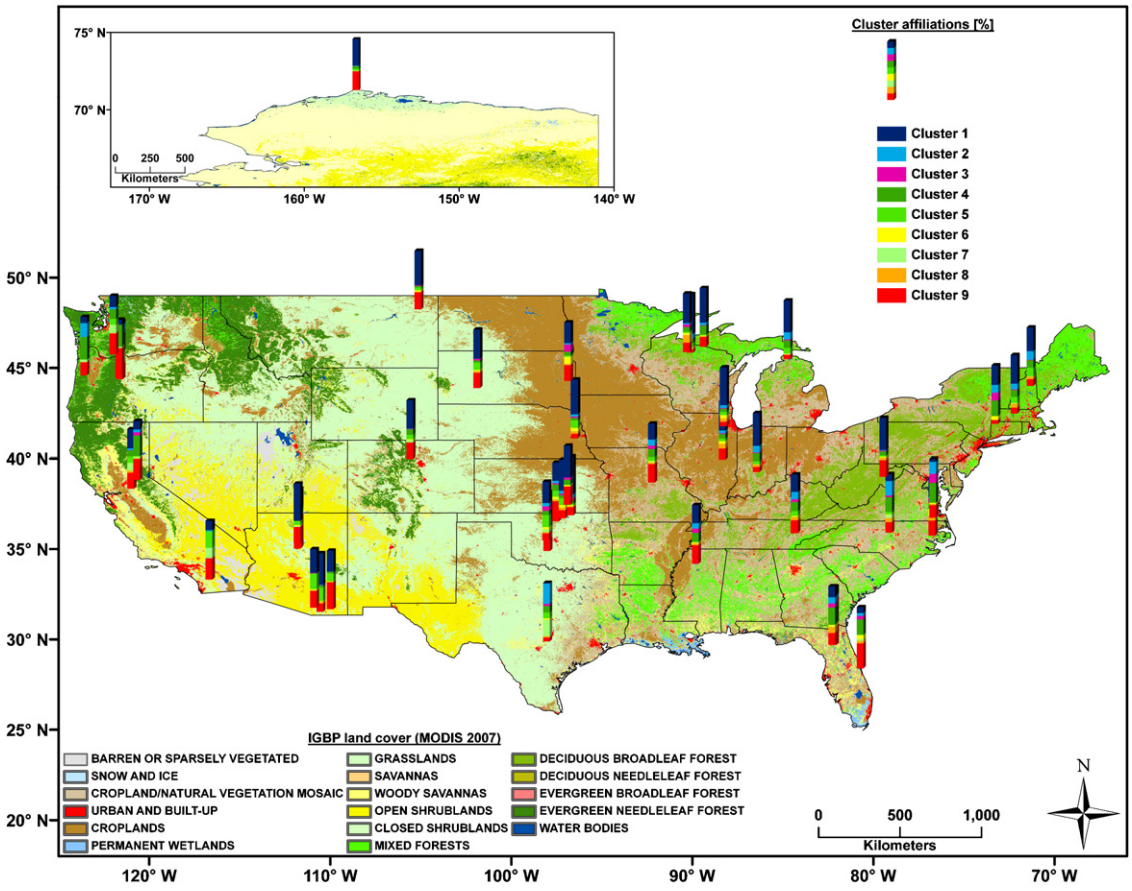


**Fig. 9.** Map of mainland USA showing the land cover, the position of the *AmeriFlux* sites, and the percentage cluster affiliations of the corresponding flux data.

restricted to data measured during the growing season. Nevertheless, the affiliations to cluster 5 of the Pacific Northwest sites indicate that the rate of net $CO_2$ exchange (release) is small even if the sensible heat flux occasionally increases. This indicates transient situations with turbulence induced by incoming radiation but no significant effect on ecosystem respiration.

The woody savanna site in Texas (*Freeman Ranch Mesquite Juniper*, Table 1) shows a high percentage of fluxes assigned cluster 2 which is characterized by the highest $CO_2$ uptake values and also of cluster 7 which indicates high fluxes of sensible heat during the years 2004–2006. By contrast, the tundra site in northern Alaska is dominated by cluster 1 and 9 and contributes portions to clusters 4 (5.3%) and 5 (3.7%) and only negligible percentages (<1%) to clusters 2, 6, and 7. This shows that the tundra vegetation in this cold and dry area exhibits significant $CO_2$ exchanges only during the short warmer periods in summer. Thus, it can be concluded, that this Alaskan tundra site acted as a weak $CO_2$ sink over the years, which was also found through direct analysis of the flux measurements by Lund et al. (2009).

A similar pattern of flux activities was recognized by the SOFM for the *Fort Peck* site Montana (Fig. 9). This grassland site located on 643 m a.s.l. exhibits a mean annual precipitation sum of about 500 mm and a mean temperature of 5.13 °C, based on data from 2000 to 2005. The low average values of *FC*, *LE*, and *H* in cluster 1 caused by the values measured during the harsh and cold winters are clearly shown (Figs. 7 and 8). Cluster 7 however is characterized by high average fluxes of sensible heat and with an average of 846.4 $\mu mol\,m^{-2}\,s^{-1}$ cluster 7 also exhibits the highest PPFD values. The respective mean $CO_2$ uptake and the mean transport of water vapor (i.e. *LE*) are moderate (Fig. 8). Such conditions are frequently observed above semiarid shrubland and grassland communities (e.g. Spano et al., 2006; Serrano-Ortiz et al., 2007) which probably also explains the noticeable percentage of the fluxes of the shrubland sites (*Sky Oaks* sites, Table 1) in the southwest that were assigned to this cluster as shown in Fig. 9.

### 3.2. Assessment of flux drivers

Since the passive variables of each of the 9 clusters were classified into 20 classes exhibiting the same width, we yield a $9 \times 16 \times 20$ data array containing 2880 frequency values. This comprises the $k$ (9) clusters each containing $j$ (16) passive variables that are subdivided into $i$ (20) classes. To determine the driving parameters that significantly influence ecosystem exchange rates in this dataset, the accumulated inter-cluster differences for each of the 20 classes of every passive variable were calculated through application of the algorithm given in Eq. (9). Based on the linearly normalized values used for the SOFM and $k$-means procedures the passive variables could be ranked relatively according to their cluster segregation strength (i.e. their influence on the flux vectors, Fig. 3).

The final results show, that net radiation and incident PPFD (photosynthetic photon flux density measured as photosynthetically active radiation in the 400–700 nm waveband) are the most important variables that explain variances in the atmosphere–ecosystem exchange of carbon and energy (Fig. 10).

Furthermore, vegetation type and vapor pressure deficit (VPD) are parameters that significantly determine the cluster affiliation of the vectors representing the magnitude of the fluxes, followed by soil temperature and precipitation. Vegetation type can be considered as a long-term integrated response to environmental variables, particularly temperature and moisture, so it must be assumed that there is some co-variation that is not discernible.

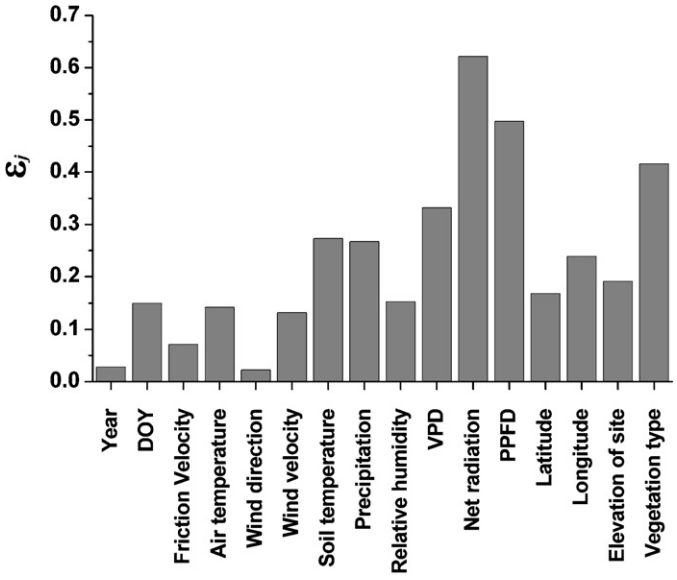


**Fig. 10.** Accumulated inter-cluster differences $\varepsilon_j$ of the class allocations for each passive variable.

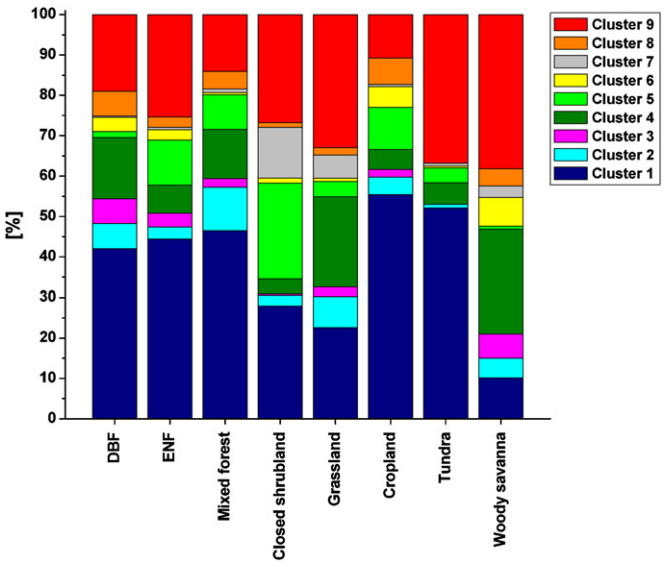


**Fig. 11.** Percentage cluster apportionment of the fluxes related to the 8 vegetation types.

Although VPD in the *AmeriFlux* database is calculated from relative humidity (RH) and air temperature, VPD turned out to have an influence on the fluxes that is about twice as high in, terms of the cluster separation strength, as the influence of RH or air temperature, respectively (Fig. 10). With respect to the dependency of fluxes on the VPD (e.g. Turner et al., 1985; Mahrt and Vickers, 2002; Lund et al., 2009), this result affirms the important role of the vegetation cover and physiological response via stomatal activity.

Soil temperature exhibits a noticeable effect on clustering of flux vectors (Fig. 10), in contrast to air temperature. Because soil respiration accounts for 60–70% of ecosystem respiration (e.g. Law et al., 1999; Ryan and Law, 2005; Borken et al., 2006; Euskirchen et al., 2006) and is highly correlated with soil temperature (Raich and Tufekciogul, 2000; Martin and Bolstad, 2005), it may strongly influence the sensitivity of $CO_2$ fluxes to soil temperature in this analysis. Furthermore this supports the assumption that a remarkable portion of the measured net $CO_2$ fluxes has to be attributed to soil respiration rather than to plant respiration.

The geographical parameters latitude, longitude and the elevation of the sites show a moderate effect on the fluxes according to their inter-cluster variances (Fig. 10). Their importance is ranked similar to that of the day of year (DOY). It is likely that the value of the DOY is increased through its correlation with the amount of incoming radiation that of course exhibits a strong seasonal run.

With respect to the vegetation types the cluster affiliations show that the flux values of the different forest types exhibit a similar distribution of cluster percentages (Figs. 9 and 11). With exception of the group of deciduous broadleaf forests (DBF) which has less of the flux vectors assigned to cluster 5 which is also characterized by a noticeable average amount of $CO_2$ uptake. The high percentage of the fluxes above the shrubland biomes assigned to cluster 5 (Fig. 11) is, with respect to the cluster-specific averages (Fig. 8), obviously related to the frequently high values of $H$ above the arid and semi-arid areas where the shrubland sites are located (Fig. 9).

The tundra vegetation type in our dataset was only represented by the *Barrow* site in northern Alaska (Table 1). The clear assignment of the respective long-term flux patterns of this tundra site shown in Fig. 11 additionally validates the ability of the presented method to recognize and arrange the fluxes correctly. The cluster affiliations of the respective fluxes reflect the plant activities in connection with the general meteorological conditions at cold and dry tundra sites (Figs. 8 and 11).

Overall, the results found in this synthesis of the *AmeriFlux* data support findings of other studies that comprised data from single sites or various ecosystems. The relatively small influence of the air temperature on ecosystem respiration (Fig. 10) matches the findings of other studies (e.g. Law et al., 2002; Reichstein et al., 2007).

This also accounts for the remarkable influence of the net radiation and PPFD (Garbulsky et al., 2010; Pieruschka et al., 2010) and the type of vegetation (Valentini et al., 2000; Desai et al., 2008), respectively. In an analysis of 38 sites, Law et al. (2002) found a strong correlation between GPP and the product of temperature and site water balance, which was represented by the sum of monthly precipitation minus evapotranspiration. Although soil water availability can have a major effect on $CO_2$ and water vapor exchange (Irvine and Law, 2002; Thomas et al., 2009; Baldocchi et al., 2004) our dataset contains only a small number of sites that regularly experience drought during the growing season and continuous soil moisture was not available for this analysis. In addition, our analysis does not decompose daytime fluxes into photosynthesis and respiration, so we were not able to confirm such previous findings.

The presented SOFM method implies that turbulent exchange rates assigned to the same cluster will show a similar response to changing magnitudes of variables that were found to be important in terms of their strength to passively segregate the different SOFM weight vector agglomerations. Hence, the change of the underlying vegetation type by deforestation or converting a grassland area into cropland for agricultural purposes, for instance, will show a massive effect on the budget of $CO_2$, water, and energy budget of the region. In comparison to fluxes above croplands, the fluxes above mixed forests and deciduous broadleaf forests exhibit a higher percentage assigned to cluster 2 and cluster 4, respectively (Figs. 9 and 11) that both are characterized by high means of net $CO_2$ uptake (Fig. 8).

According to the ranked passive variables, the large response of fluxes to the vegetation type indicates that a change of vegetation cover leads to variations in the $CO_2$ fluxes that are even larger than the known variations of turbulent exchanges associated only with the changing of seasons, represented by the classified DOY (Fig. 10). Moreover, on a larger scale the expected increase of $CO_2$ uptake through increasing air temperatures caused by global warming is

likely to be superposed by the expected increase of air moisture and cloud formation and the corresponding, much higher effect of decreased incoming radiation and particularly PPFD.

## 4. Conclusions

The results show that the presented method is sufficient to detect the patterns of turbulent fluxes across a large dataset with variations in space and time as well as in terms of meteorological and environmental conditions.

This spatial and temporal comprehensive dataset, combined with the independence from mechanistic ecological assumptions of the SOFM network approach provides a unique opportunity to validate and assess modeling efforts. The relative ranking of driving variables shows that radiation (net radiation and the PPFD) and type of vegetation are the most important parameters that determine the amount of turbulent exchange above vegetated areas. The quantitative results show that vegetation type and PPFD turned out to have an effect on the pattern of the turbulent flux magnitudes that is much higher than the effect of air temperature.

With respect to the ongoing discussion about the effects of the global change this study shows that rising air temperature itself will have a smaller direct effect on the turbulent exchange over ecosystems than that of most other variables accounted for in this study. This is especially interesting in consideration of the fact that the amount of water vapor and therewith the formation of clouds is expected to increase due to the global warming process (Solomon et al., 2007) as, with respect to the findings in this study, for the continental USA the effect of the increased temperature on the fluxes is likely be overcompensated at many sites by the counteractive effect of reduced incoming radiation caused by the generally increased cloud cover.

Strong effects of soil temperature and precipitation on the fluxes are clearly shown. In particular the conspicuous inter-cluster variance of soil temperature indicates that net carbon uptake is strongly influenced by the temperature response of soil processes contributing to soil respiration.

Furthermore, the large response of fluxes to vegetation type in connection with the average *FC* values and vegetation types in the clusters indicate that deforestation and transformation into agriculturally used areas will probably lead to a decreased carbon sequestration in most cases.

Due to the findings of this study, the application of the presented method on a global dataset using the participating databases of the *FLUXNET* organization in connection with high performance computers is encouraged as a very interesting and important goal for future studies.

## Acknowledgements

This research was supported by the Office of Science (BER), U.S. Department of Energy (DOE, Grant no. DE-FG02-06ER64307). The authors thank Manuela P. Huso for her statistical advice and support. The authors also would like to thank the *AmeriFlux* network organization for providing the data and making the public access easy. Furthermore, we would like to thank the principal investigators of the participating sites and their coworkers for their outstanding work and the *AmeriFlux QA/QC laboratory* for helping to assure the high quality of data of the *AmeriFlux* database.

## References

Amidan, B.G., Ferryman, T.A., Cooley, S.K., 2005. Data outlier detection using the Chebyshev theorem. In: IEEE Aerospace Conference , pp. 3814–3819.

Arsuaga-Uriarte, E., Diaz-Martin, F., 2005. Topology preservation in SOM. Int. J. Appl. Math. Comp. Sci. 1, 19–22.

Baldocchi, D.D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. Global Change Biol. 9, 479–492.

Baldocchi, D., Falge, E., Gu, L.H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X.H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82, 2415–2434.

Baldocchi, D.D., Xu, L., Kiang, N.Y., 2004. How plant functional-type, weather, seasonal drought, and soil physical properties alter water and energy fluxes of an oak-grass savanna and an annual grassland. Agric. For. Meteorol. 123, 13–39.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Altaf Arain, M., Baldocchi, D., Bonan, G.B., Bondeau, A., Cescatti, A., Lasslop, A., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K.W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F.I., Papale, D., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science 329, 834–838.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Bond-Lamberty, B., Thomson, A., 2010. Temperature-associated increases in the global soil respiration record. Nature 464, 579–582.

Borken, W., Savage, K., Davidson, E.A., Trumbore, S.E., 2006. Effects of experimental drought on soil respiration and radiocarbon efflux from a temperate forest soil. Global Change Biol. 12, 177–193.

Broomhead, D.S., Lowe, D., 1988. Multivariable functional interpolation and adaptive networks. Complex Syst. 2, 321–355.

Clare, A.P., Cohen, D.R., 2001. A comparison of unsupervised neural networks and k-means clustering in the analysis of multi-element stream sediment data. Geochem. Explor. Environ. Anal. 2, 119–134.

Cox, P.M., Betts, R.A., Jones, C.D., Spal, A.S., Totterdell, I.J., 2000. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature 408, 184–187.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1, 224–227.

Desai, A.R., Noormets, A., Bolstad, P.V., Chen, J., Cook, B.D., Davis, K.J., Euskirchen, E.S., Gough, C., Martin, J.G., Ricciuto, D.M., Schmid, H.P., Tang, J., Wang, W., 2008. Influence of vegetation and seasonal forcing on carbon dioxide fluxes across the Upper Midwest, USA: implications for regional scaling. Agric. For. Meteorol. 148, 288–308.

Du, K.-L., 2009. Clustering: a neural network approach. Neural Networks 23, 89–107.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley & Sons, New York, NY, USA.

Dufresne, J.-L., Fairhead, L., Le Treut, H., Berthelot, M., Bopp, L., Ciais, P., Friedlingstein, P., Monfray, P., 2002. On the magnitude of positive feedback between future climate change and the carbon cycle. Geophys. Res. Lett. 29, 1405, doi:10.1029/2001GL013777.

Euskirchen, E.S., Pregitzer, K.S., Chen, J., 2006. Carbon fluxes in a young, naturally regenerating jack pine ecosystem. J. Geophys. Res. 111, D01101, doi:10.1029/2005JD005793.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H.D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.G., Schnur, R., Strassmann, K., Weaver, A.J., Yoshikawa, C., Zeng, N., 2006. Climate-carbon cycle feedback analysis: results from the (CMIP)-M-4 model intercomparison. J. Climate 19, 3337–3353.

Friedlingstein, P., Dufresne, J.-L., Cox, P.M., Rayner, P., 2003. How positive is the feedback between climate change and the carbon cycle? Tellus 55B, 692–700.

Garbulsky, M.F., Peñuelas, J., Papale, D., Ardö, J., Goulden, M.L., Kiely, G., Richardson, A.D., Rotenberg, E., Veenendaal, E.M., Filella, I., 2010. Patterns and controls of the variability of radiation use efficiency and primary productivity across terrestrial ecosystems. Global Ecol. Biogeogr. 19, 253–267.

Ghasemi, J.B., Ahmadi, S., Brown, S.D., 2009. A quantitative structure–retention relationship study for prediction of chromatographic relative retention time of chlorinated monoterpenes. Environ. Chem. Lett., doi:10.1007/s10311-009r-r0251-9.

Gnedenko, B.V., 1988. Theory of Probability, 6th ed. Mir Publishers, Moscow, Russia.

Gourdji, S.M., Mueller, K.L., Schaefer, K., Michalak, A.M., 2008. Global monthly-averaged $CO_2$ fluxes recovered using a geostatistical inverse modeling approach: 2. Results including auxiliary environmental data. J. Geophys. Res. 113, D21115, doi:10.1029/2007JD009733.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2009. Multivariate Data Analysis, seventh ed. Prentice Hall, Upper Saddle River, NJ, USA.

Hargrove, W.W., Hoffman, F.M., Law, B.E., 2003. New analysis reveals representativeness of AmeriFlux network. EOS Trans. AGU 84, 529–544.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York, USA.

Haupt-Herting, S., Klug, K., Fock, H.P., 2001. A new approach to measure gross $CO_2$ fluxes in leaves. gross $CO_2$ assimilation, photorespiration, and mitochondrial respiration in the light in tomato under drought stress. Plant Physiol. 126, 388–396.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, second ed. Prentice Hall, Upper Saddle River, NJ, USA.

Heimann, M., Reichstein, M., 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. Nature 451, 289–292.

Irvine, J., Law, B.E., 2002. Seasonal soil $CO_2$ effluxes in young and old ponderosa pine forests. Global Change Biol. 8, 1183–1194.
Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ, USA.
Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.
Kohonen, T., 1984. Self-organization and Associative Memory. Springer, Heidelberg, Germany.
Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78, 1464–1480.
Kohonen, T., 1997. Self-organizing Maps, second ed. Springer, Heidelberg, Germany.
Law, B.E., Falge, E., Gu, L., Baldocchi, D.D., Bakwin, P., Berbigier, P., Davis, K., Dolman, A.J., Falk, M., Fuentes, J.D., Goldstein, A., Granier, A., Grelle, A., Hollinger, D., Janssens, I.A., Jarvis, P., Jensen, N.O., Katul, G., Mahli, Y., Matteucci, G., Meyers, T., Monson, R., Munger, W., Oechel, W., Olson, R., Pilegaard, K., Paw, K.T., Thorgeirsson, H., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2002. Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. Agric. For. Meteorol. 113, 97–120.
Law, B.E., Ryan, M.G., Anthoni, P.M., 1999. Seasonal and annual respiration of a ponderosa pine ecosystem. Global Change Biol. 5, 169–182.
Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recogn. 36, 451–461.
Liu, Y., Cukic, B., Schumann, J., Jiang, M., 2007. Performance analysis of dynamic cell structures. In: Chen, K., Wang, L. (Eds.), Trends in Neural Computation. Springer, Berlin, Heidelberg, Germany, pp. 367–389.
Liu, J., Gader, P., 2002. Neural networks with enhanced outlier rejection ability for off-line handwritten word recognition. Pattern Recogn. 35, 2061–2071.
Lund, M., Lafleur, P.M., Roulet, N.T., Lindroth, A., Christensen, T.R., Aurela, M., Chojnicki, B.H., Flanagan, L.B., Humphreys, E.R., Laurila, T., Oechel, W.C., Olejnik, J., Rinne, J., Schubert, P., Nilsson, M.B., 2009. Variability in exchange of $CO_2$ across 12 northern peatland and tundra sites. Global Change Biol., doi:10.1111/j.1365-2486.2009.02104.x.
Mahrt, L., Vickers, D., 2002. Relationship of area-averaged carbon dioxide and water vapour fluxes to atmospheric variables. Agric. For. Meteorol. 112, 195–202.
Martin, J.G., Bolstad, P.V., 2005. Annual soil respiration in broadleaf forests of northern Wisconsin: influence of moisture and site biological, chemical, and physical characteristics. Biogeochemistry 73, 149–182.
Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. 24, 1650–1654.
McQueen, J.B., 1965. On convergence of k-means and partitions with minimum average variance. Ann. Math. Stat. 36, 1084.
McQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability ,. University of California Press, Berkley, pp. 281–297.
Mohebi, E., Sap, N.M.N., 2009. Rough Kohonen neural network for overlapping data detection. In: Sap, N.M.N., Kim, T., Fang, W.C., Lee, C., Arnett, K.P. (Eds.), Advances in Software Engineering. Springer, Berlin, Heidelberg, Germany, pp. 183–196.
Mulier, F., Cherkassky, V., 1994. Learning rate schedules for self-organizing maps. Pattern Recogn. 2, 224–228.
Munoz, A., Muruzabal, J., 1998. Self-organizing maps for outlier detection. Neurocomputing 18, 33–60.
Murty, U.S.N., Banerjee, A.K., Arora, N., 2009. Application of Kohonen maps for solving the classification puzzle in AGC kinase protein sequences. Interdiscip. Sci. Comput. Life Sci. 1, 173–178.
Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., Wong, G., 2002. Analysis and visualization of gene expression data using self-organizing maps. Neural Networks 15, 953–966.
Papale, D., Valentini, R., 2003. A new assessment of European forest carbon exchanges by eddy fluxes and artificial neural network spatialization. Global Change Biol. 9, 525–535.
Patterson, D.W., 1996. Artificial Neural Networks. Prentice Hall, Singapore, Singapore.
Pieruschka, R., Huber, G., Berry, J.A., 2010. Control of transpiration by radiation. Proc. Natl. Acad. Sci. U.S.A. 107, 13372–13377.
Petoukhov, V., Ganopolski, A., Brovkin, V., Claussen, M., Eliseev, A., Kubatzki, C., Rahmstorf, S., 2000. CLIMBER-2: a climate system model of intermediate complexity. Part I: Model description and performance for present climate. Climate Dyn. 16, 1–17.
Priddy, L.K., Keller, E.P., 2005. Artificial Neural Networks, an Introduction. SPIE Press, Bellingham, Washington, USA.
Raich, J.W., Tufekciogul, A., 2000. Vegetation and soil respiration: correlations and controls. Biogeochemistry 48, 71–90.
Reichstein, M., Papale, D., Valentini, R., Aubinet, M., Bernhofer, C., Knohl, A., Laurila, T., Lindroth, A., Moors, E., Pilegaard, K., Seufert, G., 2007. Determinants of terrestrial ecosystem carbon balance inferred from European eddy covariance flux sites. Geophys. Res. Lett. 34, L01402, doi:10.1029/2006GL027880.
Ripley, B.D., 2005. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK.
Risch, A.C., Frank, D.A., 2010. Diurnal and seasonal patterns in ecosystem $CO_2$ fluxes and their Controls in a temperate grassland. Rangeland Ecol. Manage. 63, 62–71.
Rojas, R., 1996. Neural Networks—a Systematic Introduction. Springer, Berlin, Germany/New York, NY, USA.
Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.
Ryan, M.G., Law, B.E., 2005. Interpreting, measuring and modeling soil respiration. Biogeochemistry 73, 3–27.
Saleska, S.R., Miller, S.D., Matross, D.M., Goulden, M.L., Wofsy, S.C., da Rocha, H.R., de Camargo, P.B., Crill, P., Daube, B.C., de Freitas, H.C., Hutyra, L., Keller, M., Kirchhoff, V., Menton, M., Munger, J.W., Hammond Pyle, E., Rice, A.H., Hudson, S., 2003. Carbon in Amazon forests: unexpected seasonal fluxes and disturbance induced losses. Science 302, 1554–1557.
Schimel, D.S., House, J.I., Hibbard, K.A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B.H., Apps, M.J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A.S., Field, C.B., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R.A., Melillo, J.M., Moore, B., Murdiyarso, D., Noble, I., Pacala, S.W., Prentice, I.C., Raupach, M.R., Rayner, P.J., Scholes, R.J., Steffen, W.L., Wirth, C., 2001. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. Nature 414, 169–172.
Schmidt, A., Wrzesinsky, T., Klemm, O., 2008. Gap filling and quality assessment of $CO_2$ and water vapour fluxes above an urban area with radial basis function neural networks. Boundary-Layer Meteorol. 126, 389–413.
Serrano-Ortiz, P., Kowalski, A.S., Domingo, F., Rey, A., Pegoraro, E., Villagarcía, L., Alados-Arboledas, L., 2007. Variations in daytime net carbon and water exchange in a montane shrubland ecosystem in southeast Spain. Photosynthetica 45, 30–35.
Solomon, S., Qin, D., Manning, M., Alley, R.B., Berntsen, T., Bindoff, N.L., Chen, Z., Chidthaisong, A., Gregory, J.M., Hegerl, G.C., Heimann, M., Hewitson, B., Hoskins, B.J., Joos, F., Jouzel, J., Kattsov, V., Lohmann, U., Matsuno, T., Molina, M., Nicholls, N., Overpeck, J., Raga, G., Ramaswamy, V., Ren, J., Rusticucci, M., Somerville, R., Stocker, T.F., Whetton, P., Wood, R.A., Wratt, D., 2007. Technical summary. In: Qin, S.D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK/New York, NY, USA.
Spano, D., Duce, P., Sirca, C., Zara, P., Marras, S., Pisanu, S., Arca, A., Ventura, A., 2006. Energy and $CO_2$ exchanges of a Mediterranean shrubland ecosystem. In: Proceedings of the 27th Conference on Agricultural and Forest Meteorology , San Diego, CA, USA, May.
Stoy, P.C., Richardson, A.D., Baldocchi, D.D., Katul, G.G., Stanovick, J., Mahecha, M.D., Reichstein, M., Detto, M., Law, B.E., Wohlfahrt, G., Arriga, N., Campos, J., McCaughey, J.H., Montagnani, L., Paw, U., Sevanto, K.T., Williams, S.M., 2009. Biosphere-atmosphere exchange of $CO_2$ in relation to climate: a cross-biome analysis across multiple time scales. Biogeosciences 6, 2297–2312.
Taşdemir, K., 2008. Exploring topology preservation of SOMs with a graph based visualization. Lecture Notes in Computer Science 5326, 180–187.
Tavan, P., Grubmüller, H., Kühnel, H., 1990. Self-organization of associative memory and pattern classification: recurrent signal processing on topological feature maps. Biol. Cybern. 64, 95–105.
Tigrine-Kordjani, N., Chemat, F., Meklati, B.Y., Tuduri, L., Giraudel, J.L., Montury, M., 2007. Relative characterization of rosemary samples according to their geographical origins using microwave-accelerated distillation, solid-phase microextraction and Kohonen self-organizing maps. Anal. Bioanal. Chem. 389, 631–641.
Thomas, C.K., Law, B.E., Irvine, J., Martin, J.G., Pettijohn, J.C., Davis, K.J., 2009. Seasonal hydrology explains interannual and seasonal variation in carbon and water exchange in a semi-arid mature ponderosa pine forest in Central Oregon. J. Geophys. Res. 114, G04006, doi:10.1029/2009JG001010.
Thompson, S.L., Govindasamy, B., Mirin, A., Caldeira, K., Delire, C., Milovich, J., Wickett, M., Erickson, D., 2004. Quantifying the effects of $CO_2$-fertilized vegetation on future global climate and carbon dynamics. Geophys. Res. Lett. 31, L23211, doi:10.1029/2004GL021239.
Turner, N.C., Schulze, E.-D., Gollan, T., 1985. The responses of stomata and leaf gas exchange to vapour pressure deficits and soil water content. Oecologia 65, 348–355.
Ultsch, A., Siemon, H.P., 1990. Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the International Neural Network Conference ,. Kluwer, Dordrecht, The Netherlands, pp. 305–308.
Valentini, R., Matteucci, G., Dolman, A.J., Schulze, E.-D., Rebmann, C., Moors, E.J., Granier, A., Gross, P., Jensen, N.O., Pilegaard, K., Lindroth, A., Grelle, A., Bernhofer, C., Grünwald, T., Aubinet, M., Ceulemans, R., Kowalski, A.S., Vesala, T., Rannik, Ü., Berbigier, P., Loustau, D., Gumundsson, J., Thorgeirsson, H., Ibrom, A., Morgenstern, K., Clement, R., Moncrieff, J., Montagnani, L., Minerbi, S., Jarvis, P.G., 2000. Respiration as the main determinant of carbon balance in European forests. Nature 404, 861–865.
Van Wijk, M.T., Bouten, W., 1999. Water and carbon fluxes above European coniferous forests modelled with artificial neural networks. Ecol. Model. 120, 181–197.
Vesanto, J., 1999. SOM-based data visualization methods. Intell. Data Anal. 2, 111–126.
Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Networks 11, 586–600.
Wang, X.-Z., Yoshizawa, M., Tanaka, A., Abe, K., Imachi, K., Yambe, T., Nitta, S., 2001. An automatic monitoring system for artificial hearts using a hierarchical self-organizing map. Int. J. Artif. Organs 4, 198–204.
Yadav, V., Mueller, K.L., Dragoni, D., Michalak, A.M., 2010. A geostatistical synthesis study of factors affecting gross primary productivity in various ecosystems of North America. Biogeosciences 7, 2655–2671.
Yang, H., Jegla, J.D., Griffiths, P.R., 1998. Classification and recognition of compounds in low-resolution open-path FT-IR spectrometry by Kohonen self-organizing maps. Fresen. J. Anal. Chem. 362, 25–33.