year to another. Three different studies of three different sites linked this inter-annual variation to three different drivers: spring temperature, precipitation and VPD. The La Thuile data set provides the required data to re-visit this hypothesis and assign (conditional) probabilities to the potential drivers. The braves among the ecologists who accepted the La Thuile challenge soon had to admit their recklessness: the data set contains at least one dimension too much to allow for the usage of established approaches such multiple-regressions, wavelets or self-organizing maps.

Depending on the selected approach and question at hand, the diurnal variability may be removed from the data set by using variables that are representative at the daily scale. The seasonal variability can be accounted for by de-trending the data which implies that the residuals are then analyzed. Although pre-processing reduces the information content of the data set, it may be acceptable from an ecological point of view. The analysis then proceeds and yields for every site in the analysis an individual time series of regression coefficients, U-matrix or wavelet coefficients for respectively multiple regressions, self-organizing maps and wavelets. Although these outputs look similar, quantitative comparison across sites is hampered by the lack of formal tools to post-process these outputs across sites. This spatial aspect is my so-called too-many-th dimension.

If developed, these tools should as well account for the difference between the calendar year and the timing of the processes under study. For example: the period between the time that NEE is dominated by CO₂-uptake and the maximum CO₂ uptake might be of particular interest for an analysis. Although this period is easy to identify, it occurs earlier in the calendar year and last longer in temperate compared to boreal regions. Hence, post-processing techniques should be capable of comparing relationships (described by means of regression coefficients, U-matrix or wavelet coefficients) with a different time and length of occurrence.

Related: Luyssaert et al. (2007); In this study a self-organizing map (SOM) is used to describe the relationships between a set of variables describing the weather condition and the CO₂ exchange between forest and the atmosphere. Although the SOM results in interesting relationships, the analysis comes short in synthesizing these relationships across sites.

### 3.19 RGB imagery as a useful data source in ecological research

**Hella E. Ahrends**

Recent studies demonstrate the suitability of standard digital cameras or webcams for phenological observations in forests. This data mining strategy is user-friendly, cost-effective and highly objective. RGB cameras provide archivable documentation, data that can be used for quantitative and qualitative evaluations and the applicability of a technique independent from observer skills. They provide information on e.g. meteorological conditions (such as visibility, fog), changes in canopy structure, visible plant responses to drought stress or snow cover and, thus, are most useful in ecosystem monitoring and ecological research. We suggest that this technique has high potential to bridge the gap between spatially integrated species-averaged information from satellites and point observations at the species level. Some applications, mainly with respect to remote sensing phenology, are shown.

Related: Ahrends et al. (2008)

### 3.20 Environmental informatics and data mining in analysis of CO₂ data

**Mika Sulkava**

There are many frameworks for learning from data. Commonly used frameworks include statistical learning, exploratory and confirmatory data-analysis, machine learning, knowledge discovery in databases, data mining, and pattern recognition. Environmental statistics and environmental informatics are fields related to learning from environmental data (Sulkava, 2008). The most significant difference between these fields is the importance of the different frameworks of learning. Data analysis in statistics is often confirmatory, i.e., hypotheses are tested and
either confirmed or rejected. In addition, data is usually collected from carefully designed experiments, which involve sampling, randomization, replication, controlling for confounding variables, etc. In many cases environmental data from, e.g., monitoring networks cannot be treated the same way as data from controlled experiments. The characteristics of these frameworks and fields and their suitability to solving scientific problems related to CO₂ fluxes and concentrations are discussed.

Three examples of how methods of environmental informatics—especially data mining and exploratory data analysis—have been successfully applied in analysis of CO₂ data are presented. First, it is shown how parametric curve fitting, time series segmentations, cross-correlations, sparse regression, and neural networks, such as the Self-Organizing Map (SOM, Kohonen, 2001) have been used in finding drivers of anomalies in CO₂ fluxes of pine forests at different latitudes. These analyses revealed that anomalies in net ecosystem exchange (NEE) are dominated by anomalies in gross photosynthesis (GPP, Luyssaert et al., 2007). Second, it is presented how time series segmentation and trend detection (e.g., Sulkava et al., 2007) with permutation tests and methods for combining p-values have been used for studying changes in CO₂ concentrations between different parts of Europe. This methodology revealed many statistically significant trends in the differences between a coastal station and inland stations. The time-series were analyzed at different resolutions, which exposed some relatively fast changes in addition to the longer-term trends. Third, the use of clustering of the SOM (Vesanto and Sulkava, 2002) in a sampling problem for studying the representativeness of the European CO₂ flux tower network is demonstrated (Canfora et al., 2009). The analysis showed that the network mostly represents the European domain rather well in respect to the variables considered. However, North-Eastern and South-Eastern climatic and ecophysiological conditions were found to be poorly sampled for several plant functional types. The results of all cases above are encouraging for using data-driven methods of environmental informatics in learning from CO₂ data.

Related: Vesanto and Sulkava (2002); Luyssaert et al. (2007); Sulkava et al. (2007); Sulkava (2008); Canfora et al. (2009)

3.21 The evolving patterns of monsoonal precipitation over India

Nishant Malik*, Norbert Marwan & Jürgen Kurths

We present an analysis of a high resolution daily rainfall gridded data set from 1951 to 2007 for India. In the light of changing characteristics of precipitation at the global scale due to atmospheric warming it is of extreme importance to understand how the spatio-temporal behaviour of monsoonal precipitation has evolved over the last few decades. We employ some standard linear trend detection methods and principal component analysis to study the evolution of monsoonal precipitation patterns. We use recurrence plots to calculate spatial complexities and there evolution over the time scales concerned in the data set and correlate them to some new features that has emerged in the rainfall events. Monsoonal precipitation usually occurs in the form of large scale spatial activity and here we show a method to find these scales and divide India into different monsoonal regions based on a measure of event synchronization and then compare it using other measures like cross-correlation and mutual information.

3.22 Information driven ecohydrologic self–organization

Praveen Kumar

Interaction between processes in nature leads to self-organization. For example, the dynamics of vegetation growth is affected by prevailing above and below ground conditions, but the energy, water and carbon fluxes in and out of the vegetation modify this environment itself. This self-organized feedback, facilitated by the variability of the component processes that interact, is continually evolving. Yet, ascertaining the precise role of the variability in this feedback dynamics remains an open question. Towards this goal, using observations of moisture, energy, and carbon fluxes from Fluxnet towers, we measure the predictive information provided by one variable to the
5 Bibliography


