Title:

NOVEL DATA MINING STRATEGIES FOR EXPLORING BIOGEOCHEMICAL CYCLES AND BIOSPHERE-ATMOSPHERE INTERACTIONS

June 8—10, 2009, Max Planck Institute for Biogeochemistry, Jena, Germany

Convenors:

Miguel D. Mahecha[1,2], Markus Reichstein[1], Nuno Carvalhais[3], Mirco Migliavacca[4]

(1) Max-Planck-Institut für Biogeochemie
Hans Knöll Str. 10, 07745 Jena, *Germany*
markus.reichstein@bgc.mpg.de
miguel.mahecha@bgc.mpg.de

(2) ETH Zürich
Department of Environmental Sciences
8092 Zürich, *Switzerland*
miguelm@student.ethz.ch

(3) Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
2829-516 Caparica, *Portugal*
ncarvalhais@fct.unl.pt

(4) Università Milano – Bicocca
Dipartimento Scienze dell'Ambiente e del Territorio
Piazza della Scienza 1, Milano 20126, *Italy*
m.migliavacca1@campus.unimib.it

Key words:

Ecosystem-atmosphere interactions, Green house gases, Biogeochemical cycle data, Data mining, Information theory, Nonlinear machine learning

Outline: exploratory workshop

Abstract:

Regional and global observations of biogeochemical cycles, such as ecosystem-atmosphere fluxes or biospheric responses to climate, lead to high dimensional spatiotemporal data matrices. Environmental diagnostics and prognostics under non-stationary climate conditions require insight in the coupling of multiple data streams, posing novel problems to statistics and bioinformatics. Merging the expertise from relevant backgrounds is expected to allow the identification of common goals and open new methodological perspectives.

Background

One major challenge in terrestrial biogeochemical cycle research is the characterization of the spatiotemporal coupling of various biogeochemical cycles in ecosystem-atmosphere interactions. This holds especially true for fluxes of green house gases (GHG), energy and information. The underlying processes of these exchange processes are mutually interlinked, setting up nonlinear feedback systems with the global climate, terrestrial biosphere and human activities (IPCC, 2007; Heimann and Reichstein, 2008). Understanding the observed patterns and relating them to underlying mechanisms is by no means trivial: observed temporal features, e.g., in the carbon and water fluxes and their mutual relationships are non-stationary (Reichstein et al., 2005), act differently on different time scales (Katul et al., 2001), are shaped by ecosystem memory (Seneviratne et al., 2006) and consequently contain respective symptoms, e.g., hysteresis effects (Richardson et al., 2006; Mahecha et al., 2007).

The development of new ground based and remotely sensed measurement techniques provided a dense grid of spatiotemporal observations. Fluxes and concentrations of GHGs, climate patterns and biophysical signatures of the terrestrial biosphere are observable at different temporal and spatial scales. The respective measurements provide insight into the temporal development of ecosystem-atmosphere interactions along broad geographical gradients; nowadays over reasonable sampling periods, e.g., many sites level records of GHG fluxes are more than a decade in length (Baldocchi, 2008). Under the variable climate conditions these data bases are becoming an invaluable asset for integrative assessments of spatiotemporal ecosystem-atmosphere interactions.

The diversity of spatiotemporal data sets collected by European projects and initiatives on continental or global scale is becoming increasingly available and expected to broaden the basis of biospheric assessments (e.g. CarboEurope-IP, NEESPI, or http://www.spot-vegetation.com, JRC-fAPAR and the ICP-programs). The respective projects are working along particular questions of high actual societal relevance, e.g., for obtaining continental scale carbon balances.

Objectives of the Workshop

The available and prospect data collections are compelling for exploration in the near future beyond the currently ongoing projects, research questions and analysis techniques. The overall goal of this workshop is thus to explore suitable methodological approaches for, and conceptual challenges posed by, observations of ecosystem-atmosphere interactions. Consequently, we expect to gain insight on methodological aspects on, as well as novel process understanding of multidimensional spatiotemporal relations within and between the high diversity of biosphere-atmosphere fluxes. These exchange processes may be interpreted beyond matter and energy fluxes, for example in terms of information transfer and the signatures of complexity (Hauhs and Lange, 2008). We follow a multidisciplinary perspective by bringing together experts on biogeochemical data and ecological interpretation along with researchers actively involved in the development of multidimensional time series analysis, data mining, pattern recognition, and machine learning methods. By bringing together applied and theoretical expertise

in the handling of spatiotemporal data cubes, extreme value statistics, and information science we promote an exchange of ideas contributing both to methodical and epistemological aspects of the particular problem of how to learn the patterns behind ecosystem-atmosphere-interactions as monitored from different sources.

From the perspective of the environmental and atmospheric sciences, we are interested in developing novel diagnostics of biosphere-atmosphere interactions. In particular, the hidden underlying features of the data arrays, their mutual relationships, weak couplings and signatures of interactivity require the consideration of alternative viewpoints and analysis strategies. We aim at linking actual biogeochemical research questions to different classes of (empirical) data analysis strategies; and the types of interactions that can be described by bio-cybernetic approaches. All of this is tied to localizing and initiating an active transfer and adaptation of knowledge on information content exploration methodologies.

From the perspective of time series and spatial statistics, bioinformatics and biological cybernetics, this workshop should explore where the particularities of biosphere-atmosphere fluxes (e.g., their high degrees of uncertainty, their temporal fragmentation, and the particular extreme value characteristics) require further adaptations of existing techniques, or pose new methodical research problems. We expect the unique constellation of scientists to unravel potentials in the interdisciplinary space between bioinformatics and (non-linear) statistics, and the research of biosphere-atmosphere interactions.

Potential outcomes

The workshop promotes a novel view on the assessment of biogeochemical cycle's interrelations. Over the course of the last decades the predominant approaches to understanding the coupling of the terrestrial biosphere and the atmosphere were either process based or used very simple empirical relationships. Methodologies from other fields were only or timidly transferred. In the particular context of biogeochemistry, the ongoing scientific integration of biology and information sciences has been widely ignored. However, the highly advanced fields of spatiotemporal statistics, bioinformatics, and biological cybernetics have identified common problems and build the basis for modern analysis strategies, especially well received, e.g., in medical sciences (e.g., in neurosciences or genetics). We expect that comparable interdisciplinary advances would significantly broaden the horizon of biogeochemistry. This view is supported by first singular studies, which have illustrated how the different biospheric monitoring data could be analyzed on similar methodological ground than common to bioinformatics (e.g., Yang et al. 2007).

The potential scientific impact of the proposed exploratory workshop would consist on: (i) identifying and synthesising where the transfer or development of techniques could actively support and promote the investigation of biosphere-atmosphere interactions; as well as (ii) exploring where biogeochemistry and the related field of climatology and ecology poses different exigencies to "conventional" bioinformatics or advanced statistics. Thus, the short term goal is to increase awareness and identify common key points between scientific communities. We expect that within months following the workshop this "mutual information"-transfer could lead to a variety of cooperation and the formulation of common goals and to perform the first practical analyses in the track of the workshop.

The synergistic effects of the workshop embody the potential to gain empirical insight on the signatures of climate change where conventional (statistical or semi-empirical modelling) methods fail. Indeed, the recognition and description of anomalies of biogeochemical variables (and especially their interactions) is often ahead from fully understanding causalities (Ciais et al., 2005). In this context also the efforts for empirical prognostic modelling could benefit from a more accurate empirical description of the nonlinear internal feedback system of the biogeochemical cycles.

Organization of event

The participation of environmental scientists from different European and non European countries is mandatory since the workshop needs expertise from all major ecosystem types, and varying geo-ecological constraints (ranging from semi-arid to northern temperate areas, Reichstein et al., 2007). In this regard, we aim at inviting people with experience in synthesis analysis of biogeochemical cycle data, in conjunction with climate and/or model-data analysis and integration perspectives.

A similarly wide horizon is expected from the community representing the methodological development. Their expertise is widely spread over Europe, but somehow "clustered" according to scientific schools. In particular, we hope to generate the interest of people with a strong background in machine learning methods (ranging from regression trees, and artificial neural networks to kernel-based methods). We also want to reach the community of time series analysis. Here we expect contributions regarding the explicit analysis of multidimensional time series in terms of trend assessments, extreme value statistics and the investigation of long range correlations. Although the scientific achievements of the time series statisticians have already found broad application in climate science we are facing novel challenges arising from the instationarity and nonlinearity of ecosystem-atmosphere interactions. Furthermore, we expect this group to have a significant role in bridging pure environmental and pure methodological perspectives.

Expected benefit

As outlined above, the workshop has a strong interdisciplinary character. Bringing together disparate communities in a focused workshop generates the expectation for synergies and future collaborations between participants, yielding a considerable outreach. New perspectives will be offered to research in the particular context of biogeochemical cycles assessments. These are highly required for sustaining the leading role of European inthe investigation of ecosystem-atmosphere interactions, at a global level.

Initiatives, emerging from the workshop are expected to provide a basis for improved future assessments of European GHG fluxes, including uncertainty estimates and the elimination of redundancies. We also expect that interesting findings might have considerable feedback to other European scale projects, e.g., within the 7th Framework Programme new observational networks have been established (e.g., ICOS) which will lead to a high demand of novel techniques in data mining and the analysis of complex spatiotemporal data cubes. We expect that the participants will stay in contact after the workshop where small interdisciplinary working groups will provide the basis for future European scale collaborations.

References

Baldocchi, D. (2008) *Turner review no. 15. Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems.* Australian Journal of Botany, 56:1–26.

Ciais, P., Reichstein**,** M.**,** Viovy, N., Granier, A., Ogée, J., Allard, V., Buchmann, N., Aubinet, M., Bernhofer, C., Carrara, A., Chevallier, F., Noblet, N.D., Friend, A., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J.M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J.F., Sanz, M.J., Schulze, E.D., Vesala, T. and Valentini, R. (2005) *Europe-wide reduction in primary productivity caused by the heat and drought in 2003.* Nature 437: 529-533.

Hauhs, M. and Lange, H. (2008) Classification of runoff in headwater catchments: A physical problem?, Geography Compass, 2: 235-254.

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

IPCC (2007) Fourth Assessment Report of the IPCC Intergovernmental Panel on Climate Change.

Katul, G., Lai, C.-T., Schäfer, K., Vidakovic, B., J., A., Ellsworth, D., and Oren, R. (2001) *Multiscape analysis of vegetation surface fluxes: from seconds to years.* Advances in Water Resources, 24: 1119–1132.

Mahecha, M. D., Reichstein, M., Lange, H., Carvalhais, N., Bernhofer, C., Grnwald, T., Papale, D., and Seufert, G. (2007) *Characterizing ecosystem-atmosphere interactions from short to interannual time scales.* Biogeosciences, 4:743–758.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Valentini, R., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Gr¨unwald, T., Havrnkov, K., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Rambal, S., Rotenberg, E., Sanz, M., Seufert, G., Vaccari, F., Vesala, T., and Yakir, D. (2005) *On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm.* Global Change Biology, 11:1–16.

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.* Geophysical Research Letters, 34: L01402.262.

Richardson, A., Braswell, B., Hollinger, D., Burman, P., Davidson, E., Evans, R., Flanagan, L., Munger, J. W., Savage, K., Urbanski, S. P., and Wofsy, S. C. (2006) *Comparing simple respiration models for eddy flux and dynamic chamber data.* Agricultural and Forest Meteorology, 141: 219–234.

Seneviratne, S., Koster, R., Guo, Z., Dirmeyer, P., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C.-H., Mocko, D., Oleson, K., and Verseghy, D. (2006) *Soil moisture memory in agcm simulations: Analysis of global landatmosphere coupling experiment (glace) data.* Journal of Hydrometeorology, 7: 1090–1112.

Yang, F., Ichii, K., White, M.A., Hirofumi, H., Michaelis, A., Votava, P., Zhu, A.-X., Huete, A., Running, S.W., and Nemani, R.R. (2007) *Developing a continental scale measure of gross primary production by combining MODIS and AmeriFlux data through support vector machine approach. Remote Sensing of Environment.* 110: 109–122.