ROSENZWEIG, M. L. (1971) Paradox of enrichment: Destabilization of exploitation ecosystems in ecological time. Science, **171**, 385.
ROVINSKY, A., MENZINGER, M. (1994) Differential flow instability in dynamical systems without an unstable (activator) subsystem. Phys. Rev. Lett., **72**, 2017-2020.
SATNOIANU, R. A., MERKIN, J. H. & SCOTT, S. K. (1998) Spatio-temporal structures in a differential flow reactor with cubic auto catalator kinetics. Physica D, **124**, 345-367.
SATNOIANU, R. A., MAINI, P. K. & MENZINGER, M. (2001) Parameter space analysis, pattern sensitivity and model comparison for Turing and stationary flow distributed waves (fds). Physica D, **160**, 79-102.
SAUNDERS, P. T. & BAZIN, M. J. (1974) On the stability of food chains. J. theor. Biol., **52**, 121.
TILMAN, D. & WEDIN, D. (1991) Oscillations and chaos in the dynamics of perennial grasses. Nature, **353**, 653.
VANCAPPELLEN, PH. & WANG, Y. (1996) Cycling of iron and manganese in surface sediments: A general theory for the coupled transport and reaction of carbon oxygen, nitrogen, sulfur, iron and manganese. Amer. J. Sci., **296**, 197-243.
VOLTERRA, V. (1926) Fluctuations in the abundance of a species considered mathematically. Nature, **118**, 558.

M. Baurmann, T. Gross & U. Feudel, Institute of Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University of Oldenburg, P.O. Box 2503, D-26111 Oldenburg, Germany; e-mail: m.baurmann@icbm.de.

REDUCTION OF A COMPLEX BIOGEOCHEMICAL MODEL WITH NEURAL NETWORK AND CLUSTERING TECHNIQUES

K. Bernhardt, T. A. Sperr & K. W. Wirtz

Introduction

In the context of global change, mathematical models of biogeochemical cycles can support the understanding of complex processes controlling the emission of greenhouse gases. In this work a special focus is laid on nitrous oxide, estimated to contribute to global warming to an amount of about 6% (IPCC, 2001).Even though N_2O outflux from coastal and shelf sediments is found to be about one third of the total global emissions (SEITZINGER & KROEZE, 1998), the formation of N_2O is often not incorporated in standard model approaches (e. g. WANG & VAN CAPPELLEN, 1996, SOETAERT *et al*., 1996; HUNTER *et al*., 1998; RINN *et al*., 1999). In part, this gap derives from the scarcity of available datasets preventing an adequate parameterisation of process-oriented modelling frames.

In this study we nonetheless adopt a newly built integrated and mechanistic modelling framework as proposed by (WIRTZ, 2003). Albeit using a relatively large data-set for validation, many process parameters of this model cannot be directly evaluated. The model furthermore requires a set of boundary conditions which characterize a local site and are therefore hardly applicable on a global scale. Both characteristics lead to two key problems in environmental modelling:

1. Can meaningful predictions be made at large model uncertainty?
2. Does a complex and intrinsically uncertain model allow for being up-scaled to larger scales?

In this study we present an approach which simultaneously addresses the issue of up-scaling and model uncertainty using data mining techniques such as neural network-based techniques like the Self-Organizing Map (SOM) (KOHONEN, 2001).

One of our guiding hypothesis is that a reduced-form representation can be proposed if the full account of the uncertainty ranges of all model processes is known. The existence of such a reduced representation is supported by nonlinear data analysis where, for example, a neural network trained with an empirical, high-dimensional dataset of sediment biochemistry (KROPP & KLENKE, 1997) revealed a limited number of distinct system states. We suggest that a similar reducibility should be inherent to the high-dimensional output of a complex biogeochemistry model.

Yet, such a behaviour has never been expressed in strict quantitative terms. Whereas this can be addressed with the SOM algorithm, the low dimensional state maps representing a well trained SOM often do not allow direct interpretation. A further clustering of the SOM results as suggested by VESANTO & ALHONIEMI (2000) improve this aspect, but it does not provide a suitable verbal, graphical or quantitative representation. Thus, the discussion of key mechanisms ruling the system behaviour as well as the coupling to a larger scale modelling frame require further transformation of clustering results for which we suggest the form of a simple transition graph.

Model description

The biogeochemical model used in this work is also described by Wirtz (this volume). It builds upon a synthesis of standard approaches such as those of HUNTER (1998), WANG & VAN CAPPELLEN (1996) and SOETAERT *et al*. (1996), extended into various directions.

Four arrays of biogeochemical reactions are accounted for: degradation of several classes of particulate organic carbon (POC), oxidation of dissolved organic carbon (DOC) organised in different groups of higher and lower molecular weight, re-oxidation of reduced substances and mineral precipitation. Major chemical products are CO_2, mineralised nutrients and methane.

Many improvements were made particularly to enhance the applicability of the model to near-shore sediments characterised by temporally and spatially variable conditions. Besides an overall temperature dependence these are the microbial control of nearly the entire kinetics, the adaptation and competition processes of microorganisms and an array of additional transport mechanisms for chemical or biological species either in the dissolved or particulate phase.

The transport part combines, e. g., non-local exchange between the water column and deeper sediment, trapping of particulate species in the upper sediment matrix or the adhesion behaviour of bacteria. To keep computational loads low, we used a one-dimensional setup in our analysis. N_2O formation is calculated on the basis of a detailed module for the cycling of different nitrogen species.

Data generation under uncertainty

The variation of model-parameter values in physically meaningful ranges constitutes a simple but effective means to reflect parameter uncertainty. Until now, the data amount produced by parameter variations has mainly been used to assess model sensitivities which, in turn, may improve the understanding of the simulated system in general or prepare subsequent stages of model reduction in future studies (KÖHLER & WIRTZ, 2002; WIRTZ, 2001).

In this study, parameter variations do not only reflect underlying uncertainties but also enable the generation of an arbitrary large dataset needed by the following stages of nonlinear analysis. Since we simultaneously varied 62 parameters, a Monte-Carlo algorithm based on two different distribution functions was used. If minimal and maximal range values differed by more than one order of magnitude, a log-normal distribution was applied, whereas random parameter values for smaller ranges were distributed equally.

To compromise between the availability of computer power and the demand resulting from a high-dimensional parameter space, 1000 variations were performed. The varied parameters can be categorized by the following basic groups:

1) Temperature dependence,
2) rate and half-saturation of microbial kinetics for all pathways,
3) formation rates of inorganic precipitates,
4) bacterial growth parameters,
5) physical properties of the sediment,
6) bacterial parameters concerning nitrous oxide formation and
7) a selection of ambient seawater concentrations like those of O_2 or reducible iron.

For each parameter variation we employ the same driving forces, i. e. a set of time series (temperature, nitrate, ammonium and POC) measured in the Spiekeroog backbarrier tidal flat system.

To sum up, we generate realisations of an annual cycle with fixed outer conditions and different inner parameterisation. Besides the lack of quantitative data for process parameters, these different realisations represent spatial and organismic heterogeneity (WIRTZ, 2001).

The system state was characterized by 12 aggregated variables, which were calculated using the model variables. We used five summed quantities to capture the microbial activity and the chemical state and three flux variables of nitrogen components as these are of special interest concerning nitrous oxide dynamics. In addition, we included the daily value of all four boundary time series. The complete list of aggregated state variables can be found in Table 1.

Table 1. Aggregated state variables of the sediment model.

Symbol	Comment
TfCO2	Total CO_2 production - Sum over all main reactions and all boxes
TOC	Total organic carbon (DOC + POC) - Sum over all boxes
TO2	Total oxygen – Sum over all boxes
TNO3	Total nitrate – Sum over all boxes
TRM	Total reduced matter – H_2S, Mn(II), Fe(II), FeS, MnS and pyrite. Sum over all boxes
FNO3	Flux of nitrate – Water to □sediment.
FNH4	Flux of ammonium – Water to □sediment.
FN2O	Flux of nitrous oxide – Water to sediment.
pelPOC	Pelagic particulate organic carbon - Annual cycle measured 1995
pelTemp	Pelagic temperature – Annual cycle measured 1995
pelNO3	Pelagic nitrate - Annual cycle measured 1995
pelNH4	Pelagic ammonium - Annual cycle measured 1995

The Self-Organizing Map algorithm

The SOM-algorithm (KOHONEN, 2001) provides the mapping of a multidimensional data set onto a set of topologically ordered so-called prototype vectors. These vectors w_i are arranged in a regular hexagonal grid, resembling neighbourhood relations between adjacent nodes. The prototype vectors are iteratively updated according to the Euclidean distance between the training data and the nearest prototype according to this norm. The update procedure of prototype $w_i(t+1)$ is then calculated as

$$w_i(t+1) = w_i(t) + \alpha_i(t) \cdot (X(t) - w_i(t)) \,, \qquad (1)$$

with learning rate factor $\alpha_i(t)$ and input vector $X(t)$. The neighborhood relation of $\alpha_i(t)$ takes a Gaussian form:

$$\alpha_i(t) = \varepsilon(t) \cdot \exp\left(- \frac{\| r_i - r_{w1(t)} \|^2}{\sigma(t)^2} \right), \qquad (2)$$

with r_i denoting spatial coordinates and $w1(t)$ being the best matching prototype at time t.

Before training each component of the data was normalised by mean and standard deviation. We used a two-dimensional map of 28 x 42 nodes and an initial neighbourhood parameter of $\sigma = 21$ decreasing linearly to 1, preventing the distortion of the ordered map by subsequent fine-tuning steps. The map was initialised using the first two principial components of the data to shorten the training process.

Clustering and construction of the state-transition network

The prototype vectors of the trained SOM were clustered with the k-means clustering algorithm (MACQUEEN, 1967; VESANTO & ALHONIEMI, 2000) for cluster numbers ranging from 2 to 100. The Davies-Bouldin Index (DBI) (DAVIES & BOULDIN, 1979) is used in this work to measure cluster quality.

Any of the 1000 times 365 simulation days calculated during the variation corresponds to a defined cluster providing a distribution of transitions between prototypes or clusters, respectively. To evaluate the first-order dynamics between aggregated states defined by clusters we determined simple means of transition-probabilities.

Principal component analysis

Parallel to the combined SOM-clustering method a principal component analysis on the variation data was also performed. It revealed that the first seven Eigenvectors of the covariance matrix are needed to explain 90 percent of the variance. Thus, it may not be possible to reduce the data to a small linear subspace.

Results and discussion

The result of the principal component analysis demonstrates the necessity of a nonlinear method for data analysis. Components for each reference vector of the trained map are shown in Fig. 1, using a grayscale coding. Beside a global ordering, e. g. into a northern summer and a southern winter part (see variable pelTemp), local structures emerge, indicating around 20 to 100 homogeneous regions/clusters. In the context of this work the south-eastern corner of the map showing high negative values of N_20 flux (FN2O) is of special interest.

An integral but at the same time much more puzzled picture emerges after superposing the 12 individual component planes. Local patches with similar contribution of each component could be robustly extracted by the simple

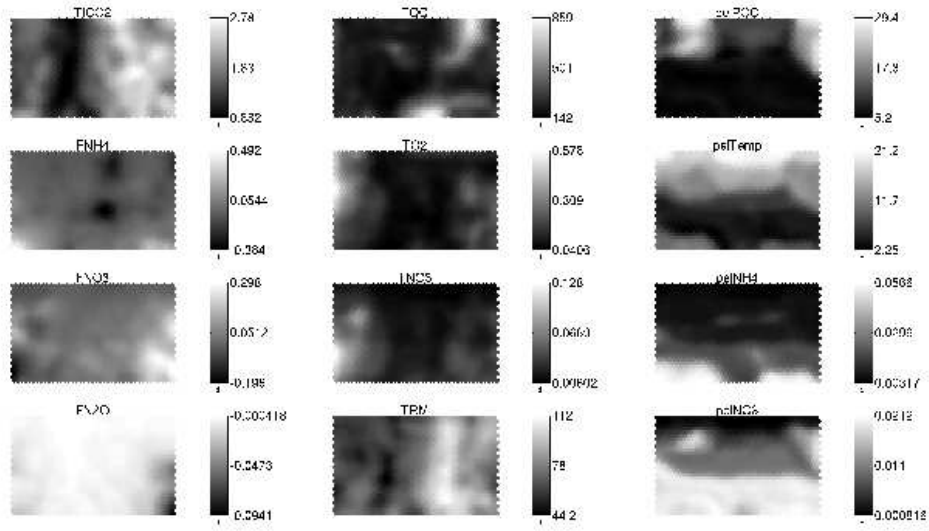

Fig. 1. SOM component planes for reference vectors (see Table 1). Each grid point represents a member of the trained map. Relative importance of the specific component is visualised by shading intensity.

clustering algorithm used. Calculations of the DBI (not shown) yielded a local minimum at 49 clusters in accordance with the impression obtained by looking at the maps.

The set of clusters obtained (see Fig. 2) is considered to represent typical states the system is able to reach. Temporal dynamics can be re-inserted by mapping each day of the input data onto the nearest cluster of the network and by determining empirical transition-probabilities between the clusters.

Fig. 3 shows an example for the clusters characterized by high N_2O emissions (compare with Figs. 1 and 2 for cluster numbering). This transition graph indicates that only 8 input clusters show relevant probabilities to reach one of the 5 clusters with high N_2O efflux and a feedback into a subset of these input states.

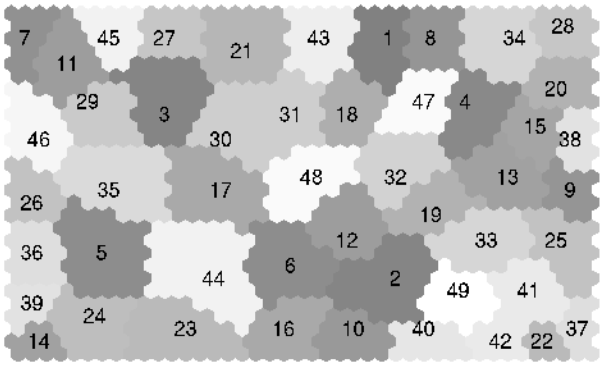

Fig. 2. The trained 28 x 42 SOM with 49 clusters. A cluster can be interpreted as an effectively aggregated state of the simulated biogeochemical system.

Drawbacks of SOM and future work

Considering the modelling aims presented in the previous sections, some theoretical as well as technical limitations of Self-Organizing Maps must be outlined (BARALDI & BLONDA, 1999):

- SOM does not preserve topology of the underlying data if the dimension of the input space is larger than three (the maximum dimension of the SOM network for visualisation).
- Prototype parameter estimates may be severely affected by noise points and outliers, as learning rates in SOM are independent of the actual distance between the input pattern and its nearest template vector.

Especially the lack of topology-preservation of higher-dimensional input spaces can result in a poor representation of the data. This discrepancy is further amplified by the impact of outliers and noise on the general structure of SOM.

A possible alternative to SOM designed to overcome this kind of problems is FOSART (fully self-organizing simplified adaptive resonance theory; BARALDI & BLONDA, 1999; BARALDI & ALPAYDIN, 2002). FOSART generates prototype vectors directly from the data and is able to dynamically create and remove links between these templates. It therefore accounts for the inherent complexity of the data and is an example for topology-preserving mappings.

Thus, a first step of further studies should be the analysis of the effective dimensionality of the data and the implementation of adequate methodologies to preserve its topology. Further improvements concerning, e. g., the fitness functions used to quantify the state of organization of the map (BAUER *et al*., 1999; POLANI, 1997) or the goodness of clustering (VESANTO & ALHONIEMI, 2000; ROTH *et al*., 2002) should be made. In this way, a stable algorithm should be found which can be used to help the user/modeller to reduce complicated process-based models. Potential applications of such a generic modelling tool range from more theoretical investigations of system stability as envisaged by FEUDEL *et al*. (this volume) to the practical use of model-based knowledge for decision support (WIRTZ, 2001).

Up to a certain degree it is possible to inversely map each cluster to approximate values of typical monitoring observables or of standard model variables. After doing so, the transition matrix obtained in this study could, e. g., be implemented as a lookup-model usable in global change assessments.

Nevertheless, the major challenge of the presented transformation approach remains its complete reformulation in terms of measurable variables. In order to link the highly abstracted transition graph to empirical research a more intense assimilation as well as common interpretation of field data is needed.

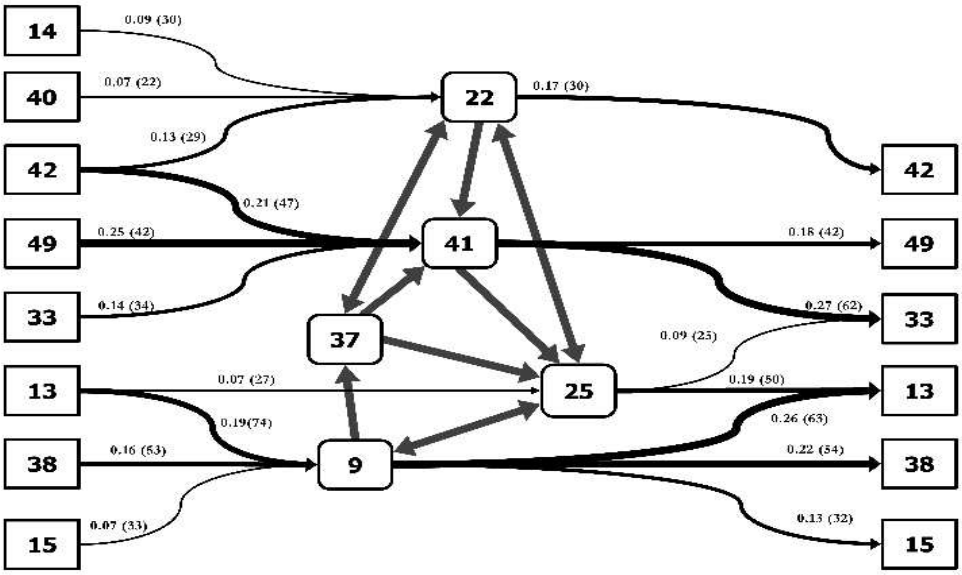

Fig. 3. Transition graph for clusters indicating high N_2O emissions. On the left side, typical predecessor states of high N_2O events are collected, on the right side are the successors. Transition probabilities are attached as small numbers.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft through the Research Group *BioGeoChemistry of Tidal Flats* (FOR 432/1). We thank Wolfgang Ebenhöh and Jürgen Kropp for their support.

References

BARALDI, A. & ALPAYDIN, E. (2002) Constructive Feedforward ART Clustering Networks - Part II. IEEE Trans. Neural Networks, **13**, 645-661.

BARALDI, A. & BLONDA, P. (1999) A survey of Fuzzy clustering algorithms for pattern recognition - Part II. IEEE Trans. Syst. Man, Cybern., **29**, 778-785.

BAUER, H.-U., HERRMANN, M. & VILLMANN, T. (1999) Neural maps and topographic vector quatization. Neural Networks, **12**, 659-676.

DAVIES, D. L. & BOULDIN, D. W. (1979) A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. (PAMI), **1**, 224-227.

HUNTER, K. S., WANG, Y. & VAN CAPPELLEN, P. (1998) Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. J. Hydrol., **209**, 53-80.

IPCC (2001). Working Group I: The scientific basis. Technical Summary, Intergovernmental Panel on Climate Change, Geneva, Switzerland.

KOHONEN, T. (2001) Self-organizing maps, 3rd edition. Springer, Berlin.

KÖHLER, P. & WIRTZ, K. W. (2002) Linear understanding of a huge aquatic ecosystem model using a group-collecting sensitivity study. Environ. Softw. Managem., **17**, 613-635.

KROPP, J. & KLENKE, T. (1997) Phenomenological pattern recognition in the dynamical structures of tidal sediments from the German Wadden Sea. Ecol. Mod., **103**, 151-170.

MACQUEEN, J. (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam, L. M. & Neyman, J. (eds.), Proc. Fifth Berkeley Symp. Math. Stat. Prob., **1**, 281-297. University of California Press, Berkeley, CA.

POLANI, D. (1997) Fitness functions for the optimization of self-organizing maps. In; BÄCK, T. (ed.), Proc. Seventh Int. Conf. Genetic Algorithms (ICGA97), 776-783, San Francisco.

PRINN, R., JACOBY, H., SOKOLOV, A., WANG, C., XIAO, X., YANG, Z., ECKHAUS, R., STONE, P., ELLERMAN, D., MELILLO, J., FITZMAURICE, J., KICKLIGHTER, D., HOLIAN, G., & LIU, Y. (1999) Integrated global system model for climate policy assesment: Feedbacks and sensitivity studies. Clim. Change, **41**, 469-546.

ROTH, V., LANGE, T., BRAUN, M. & BUHMANN, J. M. (2002) A resampling approach to cluster validation. In: HÄRDLE, W. & RÖNZ, B. (eds.), Proc. Comp. Stat.: 15th Symp. (Berlin, 2002) *(COMPSTAT 2002)*, 123-128. Physica-Verlag, Heidelberg.

SEITZINGER, S. P. & KROEZE, C. (1998) Global distribution of nitrous oxide production and N inputs in freshwater and coastal marine ecosystems. Glob. Biogeochem. Cycles, **12**, 93-113.

SOETAERT, K., HERMAN, P. M. J. & MIDDELBURG, J. J. (1996) Dynamic response of deep-sea sediments to seasonal variation: a model. Limnol. Oceanogr., **41**, 1651-1668.

VESANTO, J. & ALHONIEMI, E. (2000) Clustering of the Self-Organizing Map. IEEE Trans. Neural Networks, **11**, 586-600.

WANG, Y. & VAN CAPPELLEN, P. (1996) A multicomponent reactive transport model of early diagenesis: Application to redox cycling in coastal marine sediments. Geochim. Cosmochim. Acta, **60**, 2993-3014.

WIRTZ, K. W. (2001) Strategies for transforming fine scale knowledge to management usability. Mar. Poll. Bull., **43**, 209-214.

WIRTZ, K. W. (2003) Control of biogechemical cycling by mobility and metabolic strategies of microbes in the sediments: an integrated model study. FEMS Microbiol. Ecol., accepted.

K. Bernhardt, T. A. Sperr & K. W. Wirtz, Institute of Chemistry and Biology of the Marine Environment, Carl von Ossietzky University of Oldenburg, Carl-von-Ossietzky-Str. 9-11, P.O. Box 25 03, D-26111 Oldenburg, Germany; phone: +49 441 7985230; e-mail: k.bernhardt@icbm.de.