Exploiting simulation model results in parameterising a Bayesian network – A case study of dissolved organic carbon in catchment runoff

[1]Koivusalo, H., [1]T. Kokkonen, [1]H. Laine, [2]A. Jolma and [1]O. Varis

[1] Laboratory of Water Resources, Helsinki University of Technology, [2] Laboratory of Cartography and Geoinformatics, Helsinki University of Technology, E-Mail: Harri.Koivusalo@tkk.fi

Keywords: *Bayesian network; Decision support, Dissolved organic carbon, Simulation model.*

EXTENDED ABSTRACT

This work is part of CLIME project (Climate and Lake Impacts in Europe), which assesses climate change effects on lake dynamics. In CLIME, a decision support system (CLIME-DSS) is based on a causal Bayesian network that summarises the most important relationships between climate variables and lake characteristics. A Bayesian network is a probabilistic graphical model, where nodes represent random variables and arcs between the nodes represent conditional dependencies. In a Bayesian network, relationship between the dependent variable and its explanatory variables is described for discrete variables as a conditional probability table (CPT). The aim of this study is to demonstrate how expert knowledge provided by researchers, and results of an environmental simulation model, are exploited in constructing a Bayesian network. A case study addresses the impact of climate change on concentrations of dissolved organic carbon (DOC) in catchment runoff.

The environmental simulation model is a DOC model (Jennings and Naden, 2004), which is coupled with the hydrological routine of the Generalized Watershed Loading Function (GWLF) model. The output of the model is the daily stream water DOC concentration and the daily load of DOC entering a lake. The DOC/GWLF model has been calibrated and validated against historical data from three catchments in Europe.

A Bayesian network for describing interrelations between the climate and DOC concentrations is constructed on the basis of expert opinions and the structure of the DOC/GWLF simulation model. Those variables that are present both in the DOC model, and in the network structure based on the expert opinions, are included in the final structure of the Bayesian network. One Bayesian network is constructed for each of the three study sites.

The GWLF/DOC model was run under a variety of climatic conditions using one set of calibrated parameter values at a time. The meteorological input variables were compiled from results of Regional Climate Models (RCM). Subsequently, the RCM and GWLF/DOC model results were analysed to compute conditional frequency tables for the links between each dependent node and its explanatory variables in the Bayesian network. The procedure of Kokkonen et al. (2005) was utilised to estimate link strength values from the conditional frequency information. The link strength values were optimised against the conditional frequencies determined from the model simulations. Finally, all values in the CPTs were generated using the optimised link strength values.

In order to apply the Bayesian networks within the study region, RCM results are utilised for creating distributions of explanatory variables for all computation grid cells in Europe. The distributions are constructed for different scenarios characterising current and future climatic conditions. These spatial data on the climatic variables together with the Bayesian networks allow the CLIME-DSS users to study the predicted climate change effects across Europe.

Three Bayesian networks were applied to predict how decomposition and summertime DOC concentrations change in the future in Lough Leane in Ireland. The application revealed that variability of the predicted annual decompositions and summer DOC concentrations was very different between the three Bayesian networks. The predicted direction of change in DOC concentrations, however, from the control scenario to the future climate scenario was same for all three Bayesian network parameterisations. The model application demonstrates how Bayesian networks can be used as diagnostics for assessing the conformity of model regionalisation.

1. INTRODUCTION

The predicted climate change along with its potential impacts on physical, chemical, and biological lake processes sets a challenge to lake management in the future. Meteorological forcing, such as precipitation, evaporation, and air temperature, exert a significant control on catchment hydrological processes, lake flushing rates, lake residence times, thermal stratification, loadings of dissolved and suspended material, and productivity of a lake. According to the predictions of the Intergovernmental Panel on Climate Change (IPCC), in case of an increasing population and an insufficient extent of emission control technologies, greenhouse gas and aerosol emissions will lead to a large increase in air temperature and to altered precipitation patterns. Computational methods are needed for quantifying how lake variables respond to these changing conditions.

The CLIME project (Climate and Lake Impacts in Europe) aims at developing a suite of methods and models that can be used to manage lakes and catchments under future as well as current climatic conditions. In CLIME, regional climate scenarios, and existing catchment and lake models are combined to support lake management in the light of the water quality criteria prescribed in the European Union Water Framework Directive. CLIME aims at integrating expert knowledge and simulation model results in a form of a decision support system (CLIME-DSS) that illustrates and summarises the main results of the project to interest groups outside the research community. The CLIME-DSS is based on probabilistic Bayesian networks that characterise causal dependencies between climate, catchment, and lake variables.

In Europe, one of the key water quality issues that is likely to become increasingly important in the future is the leaching of coloured water from catchments having a large percentage of organic soils. The brown colour of water, which suggests a high concentration of dissolved organic carbon (DOC), results mainly from humic acids produced by the decomposition of organic matter within the catchment. Variation of climatic and hydrological conditions is a major factor controlling the decomposition of organic matter and transport of DOC to water bodies (Boyer et al., 1996; Dawson et al., 2002). Increases in water colour have been reported in several catchments in Europe, where these increases have been attributed to changes in climate (e.g. Freeman et al., 2001; Hongve et al., 2004, Jennings and Naden, 2004). Climatic variables control the drying and wetting patterns of soil, which have a major effect on the decomposition rate of organic material. Furthermore, changes in the climate have an impact on runoff volumes, which affect the DOC load transported to a lake (Holmberg, 2003). In CLIME, a DOC model (Jennings and Naden, 2004) is coupled with the hydrological computation scheme of the Generalised Watershed Loading Functions (GWLF, Haith et al., 1996) to simulate production, leaching and transport of DOC. The model is calibrated against historical data and applied to future climate scenarios.

The aim of this study is to demonstrate how expert knowledge provided by CLIME researchers, and results of the DOC/GWLF model, are exploited in constructing a Bayesian network for the CLIME-DSS. The first task is to conduct an expert survey to determine the variables characterising DOC processes, and to identify the causal dependencies between the variables. The experts' choice of the most important variables is compared with the variables included in the DOC model. The second task is to parameterise the Bayesian network characterising the response of DOC concentrations to changes in climate. The parameterisation is based on DOC model results and the methodology presented in Kokkonen et al. (2005). Finally, it is explored how the Bayesian network can be applied to regionalizing the model results from one location in Europe to another.

2. MATERIALS AND METHODS

2.1. Bayesian networks in CLIME-DSS

The CLIME-DSS is an expert system that is used for 1) visualising how hydrological and meteorological variables are predicted to change across Europe in future climate, and 2) illustrating how these changes are reflected in variables characterising lake water quality. The system allows the user to select a location in Europe, and explore how the variability of a water quality indicator, such as DOC level, responses to the climate change. Lake water quality assessments are based on Bayesian networks that are solved using the BNJ software (Bayesian Network Tools in Java, http://bndev.sourceforge.net). A Bayesian network is a probabilistic graphical model, where nodes represent random variables and arcs between the nodes represent conditional dependencies. In a Bayesian network a relationship between a child variable and its parent variables is described for discrete variables as a conditional probability table (CPT) (Pearl, 1988; Neapolitan, 2004). In this study, the method of Kokkonen et al. (2005) is applied to describe the CPTs with aid of link strength values.

2.2. Expert elicitation

Construction of a Bayesian network for DOC starts from elicitation of expert information. Figure 1 shows a graphical illustration of the question areas included in the expert elicitation. A questionnaire was sent to 40 experts; 12 of them returned it. Based on the answers, the most important DOC variables and their dependencies were identified. In addition, the variable characterising the level of DOC in lake water was selected, and the predicted changes in meteorological variables were ranked in a descending order of importance.

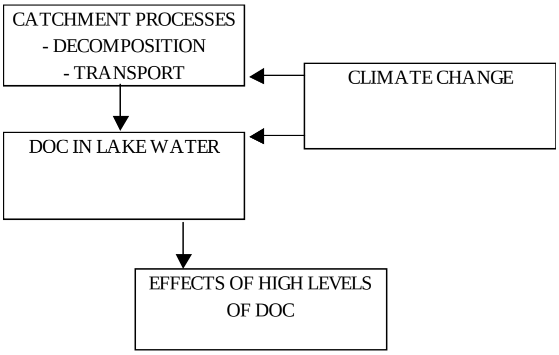

Figure 1. Question areas in the elicitation of expert knowledge about DOC processes.

2.3. Climate scenarios

The climate data for the CLIME-DSS have been compiled from simulation results of Regional Climate Models (RCMs). RCMs have been applied to generate scenarios for a control period from 1960 to 1990, and for a future period from 2070 to 2100. The future simulations include two different greenhouse gas and aerosol emission scenarios, A2 and B2 (Nakicenovic and Swart, 2000). In A2, population growth is assumed to be low during the 21st century. However, energy use and gross domestic product (GDP) growth are assumed to be high. B2 represents an intermediate alternative of all the IPCC scenarios: population growth, energy use, and GDP growth are all considered to be moderate in the future decades.

Climate simulations are carried out with two RCMs. These are the RCAO (Rossby Centre, SMHI, Sweden) and the HadRM3p (Hadley Centre, U.K. Met Office) models, which operate on grids with a cell size of ca. 50x50 km^2. Details on the RCM simulations are reported in Samuelsson (2004). In addition to meteorological variables, each climate scenario produces predictions of hydrological variables, such as evaporation, soil moisture, snow cover, and runoff.

2.4. GWLF/DOC model applications

In addition to the expert elicitation results, GWLF/DOC model simulations provide information to formulation of Bayesian networks in CLIME-DSS. The construction of the DOC model in CLIME rests on the work by Naden (1991) and Naden and Watts (1998), who developed models for assessing response of water colour to changing weather patterns. Jennings and Naden (2004) provide a description of the coupled DOC and GWLF model, where the hydrological fluxes are simulated using the curve number approach of GWLF, and decomposition and transport of organic matter are described in the DOC model. The model takes as the meteorological input daily time series of air temperature and precipitation. Decomposition of organic matter, which is primarily controlled by soil moisture conditions, produces dissolved carbon that is subject to washing out with runoff. The output of the model is the daily stream water DOC concentration and the daily load of DOC entering a lake.

The GWLF model has been calibrated and validated against historical streamflow data, and the DOC model against DOC load data from three catchments in Europe. The study catchments are Trout Beck (11.4 km^2, 54.1° N, 2.1° W) in the UK, Lough Leane (130 km^2, 52.1° N, 9.4° W) in Ireland, and Mustajoki (76.8 km^2, 61.0° N, 25.1° E) in Finland. Trout Beck is typical of peaty upland areas in the northern Pennines underlain by a clayey soliflucted till deposit. *Eriophorum*, *Calluna* and *Sphagnum* are the dominant vegetation species in the catchment, and more than 90% of the area is blanket bog. Lough Leane is located in an upland mountain peat area overlying Old Red Sandstone. About 90% of the catchment is covered with moor/bog/scrub grassland, with heather and grass dominating. Mustajoki lies in the southern boreal vegetation zone with surficial deposits typically characterized by moraine, with some highly permeable sand and gravel deposits, and organic peat layers. The proportion of forest land is 67%, peat land 20%, and agricultural land 13% from the total catchment area. Norway spruce, Scots pine, birch, and European aspen are the typical tree species in the catchment. Jennings and Naden (2005) assessed qualitatively the model fit against measured DOC load data from the three sites. Visual inspection of the results revealed that the model performance was better in Lough Leane than in the other two catchments. In this study, one Bayesian network is constructed for each of the three study sites.

3. RESULTS

3.1. Identification of the structure of the Bayesian network for DOC

Figure 2a illustrates the structure of the network that was constructed based on the expert elicitations. The ellipses are those explanatory and dependent variables that were mentioned most frequently by the experts, and the arcs indicate dependencies between the variables as seen by most experts. The annual average rate of decomposition of organic matter is dependent on the annual average soil moisture, the annual average air temperature, and the fraction of organic soils within a catchment. The seasonal DOC concentration depends on the annual decomposition rate, the runoff volumes in the same and the previous season, and in-lake DOC processes. Comparison of the expert network against the DOC model structure (Figure 2b) reveals that in the model there is no direct dependency between air temperature and decomposition. The control of the organic soil on the decomposition rate is embedded in the calibrated values of model parameters and the initial carbon store. In-lake DOC processes are not included in the model.

In this study, those variables that are present both in the DOC model and in the network structure based on the expert opinions are included in the Bayesian network for DOC. It is worth noting that expert elicitation results represent the opinion of a large group of scientists, and therefore it would have been desirable to formulate the Bayesian network based on the elicitation results alone. However, the selection of the variables was conditioned on the model structure, because the parameterisation of the Bayesian network was carried out on the basis of model results. The variables included in the Bayesian network of the CLIME-DSS are underlined in Figure 2a. The parent variables (soil moisture and runoff nodes) attain their distributions from the RCM results as explained in Section 3.3. Identification of the CPTs for child variables (decomposition and DOC nodes) is explained in Section 3.2.

Instead of using absolute values, the present study deals with deviations from prescribed reference levels. In other words, all variables in the Bayesian network for DOC describe how annual/seasonal values deviate from the reference level. The reference level is the long-term mean value of a variable over the control period from 1960 to 1990. Deviations are used for two reasons. Firstly, the CLIME project concentrates on describing changes in the lake water quality that arise from the predicted climate change, and deviations from a reference level are illustrative in presenting such water quality changes. And secondly, regionalisation of results on deviations is assumed to be more robust than regionalisation of results on absolute values.

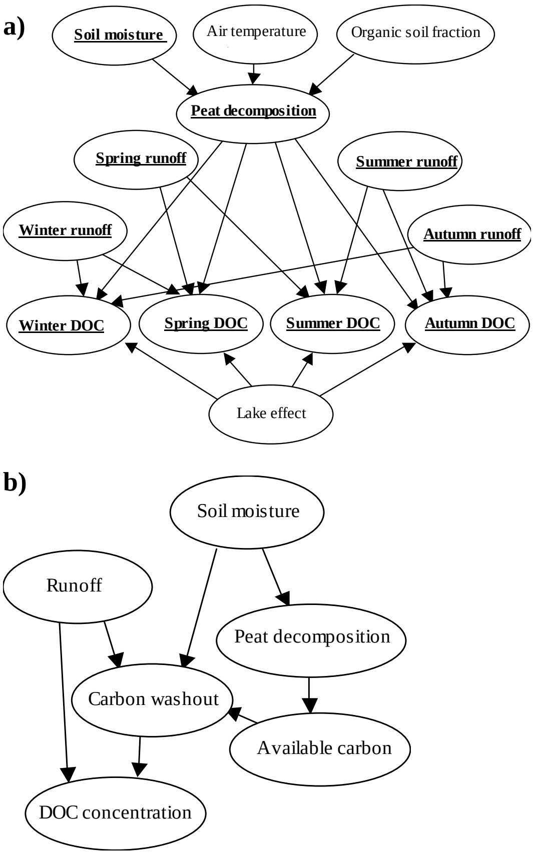

Figure 2. Structure of the DOC network identified by the experts (a) and structure of the DOC simulation model (b). The underlined variables in (a) are included in the Bayesian network for DOC.

3.2. Parameterisation of CPTs

In order to set up the Bayesian network for DOC, the GWLF/DOC model was run under a variety of climatic conditions using the calibrated parameter values from the three catchments (see Section 2.4). The meteorological input was compiled from RCM results for 63 grid cells located within the CLIME region. The meteorological input comprised RCM simulations for the control, A2, and B2 scenarios. A cumulative probability distribution of deviations was computed using the model results from all 63 grid cells. Based on this cumulative probability distribution, a range of variability was determined for every variable, and all variables were discretised into 5 classes, where the middle class represented a “no deviation” state (Figure 3).

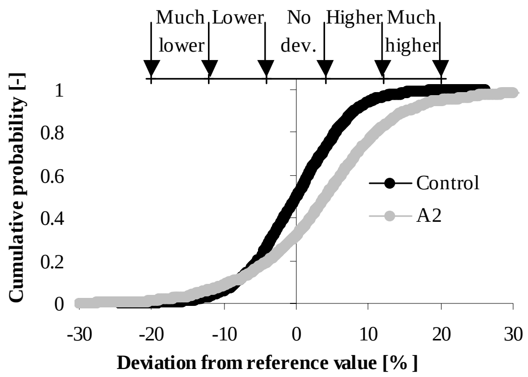

Figure 3. A cumulative probability distribution for the deviation in annual decomposition for control and A2 scenarios. The results were computed with the GWLF/DOC model parameterisation for Mustajoki. Discretisation to five classes is also shown.

Subsequently, the RCM and GWLF/DOC model results from the 63 grid cells were analysed to compute conditional frequency tables for the Bayesian network shown in Figure 2a. Table 1 shows the conditional frequency table computed for the link between the deviation in annual decomposition and soil moisture.

Table 1. The conditional frequency table for the link between the deviation in annual decomposition and soil moisture (Mustajoki DOC/GWLF parameterisation).

	Decomposition				
Soil moisture	*Much lower*	*Lower*	*No deviation*	*Higher*	*Much higher*
Much lower	0	0	0	6	1056
Lower	0	0	227	2190	748
No deviation	0	303	1698	292	0
Higher	164	540	102	0	0
Much higher	286	55	0	0	0

The procedure of Kokkonen et al. (2005) was utilised to estimate link strength values from the conditional frequency information. The link strength values were optimised against the conditional frequencies determined from the model simulations. The value of an optimised link strength parameter is an indicator of the dependency between the child node and one of its parent nodes. Table 2 shows optimised link strength values between the deviation in the annual soil moisture and the deviation in the annual decomposition for the three CLIME sites. All link strengths are negative, indicating that drier conditions lead to a more efficient decomposition.

Table 2. Optimised link strength values between the deviation in the annual soil moisture and the deviation in the annual decomposition for three CLIME sites (Lough Leane, Mustajoki, Trout Beck).

	Decomposition		
	Lough Leane	*Mustajoki*	*Trout Beck*
Annual moisture	-0.432	-0.598	-0.374

When studying the link strengths between deviations of seasonal DOC concentrations and their parents (Table 3), increase in runoff is identified to lead to a decrease in the DOC concentrations, and an increase in decomposition is found to result in an increase in DOC. Also, most of the time deviation in decomposition is more important than deviation in runoff in explaining the deviations in seasonal DOC concentrations.

Table 3. Optimised link strength values between the deviations in the seasonal (a - winter, b - spring, c - summer, and d - autumn) DOC concentrations and their parents for three CLIME sites (Lough Leane, Mustajoki, Trout Beck).

a) ***Winter DOC concentration***

	Lough Leane	*Mustajoki*	*Trout Beck*
Autumn runoff	-0.41	-0.28	-0.32
Winter runoff	-0.30	-0.28	-0.52
Decomposition	0.63	0.28	0.52

b) ***Spring DOC concentration***

	Lough Leane	*Mustajoki*	*Trout Beck*
Winter runoff	-0.50	-0.22	-0.55
Spring runoff	-0.43	0	-0.12
Decomposition	0.67	0.22	0.35

c) ***Summer DOC concentration***

	Lough Leane	*Mustajoki*	*Trout Beck*
Spring runoff	-0.23	-0.24	-0.07
Summer runoff	0.00	0.00	-0.14
Decomposition	0.54	0.24	0.41

d) ***Autumn DOC concentration***

	Lough Leane	*Mustajoki*	*Trout Beck*
Summer runoff	-0.21	-0.29	-0.21
Autumn runoff	-0.42	0.00	-0.12
Decomposition	0.61	0.49	0.58

3.3. Application of the Bayesian network

Climate model simulation results are utilised for creating a distribution of the deviations for each explanatory (parent) variable for all RCM grid cells in Europe, and for all three scenarios (control, A2, and B2). These input data allow the users to study the predicted climate change effects across the European continent. To demonstrate the capabilities of the CLIME-DSS, predicted changes in decomposition and DOC concentrations are studied here as an example for one location, Lough Leane.

Figure 4 plots distributions of deviations in annual decomposition and Figure 5 shows deviations of summer DOC concentrations using the three parameterisations from Lough Leane, Trout Beck, and Mustajoki. The results are presented for control and A2 scenarios. It is noteworthy that also for the control period some of the years have higher (or lower) values for the modelled variables compared with the long-term average over the control period.

The differences between the bars for control and A2 scenarios indicate that both decomposition and summer DOC concentrations are predicted to increase in the future for each Bayesian network parameterisation. While decomposition shows a clear increase for Lough Leane and Trout Beck parameterisations, the seasonal DOC concentrations are less affected. This is explained by the parameterisation and structure of the Bayesian network. In a Bayesian network, where the link strength values are not perfect (absolute value < 1), the uncertainty of the state distribution of the variables increases when moving down the hierarchy of the network (see Figure 2a).

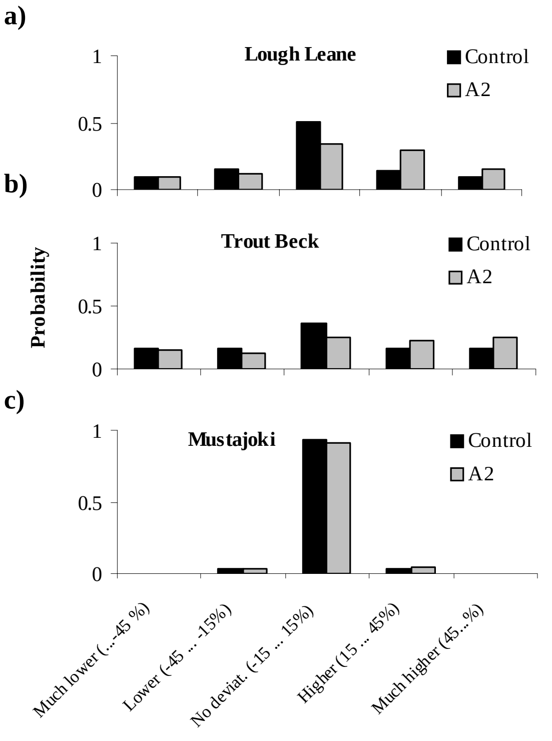

Figure 4 Deviation in annual decomposition for the control and A2 scenarios using three different Bayesian network parameterisations: Lough Leane (a), Trout Beck (b), and Mustajoki (c).

It is evident that the different parameterisations yield different distributions for deviations both in the control and A2 periods. The GWLF/DOC model parameterised for Trout Beck yields a much greater variability in DOC concentrations in response to different climatic conditions than the model calibrated for Mustajoki. The difference in the variability of DOC concentrations is a reflection of the variability in decomposition. And variability of decomposition is in turn explained by soil moisture distribution, which according to the GWLF/DOC model appears to be the primary factor controlling the DOC response to climate change. Since the results from the three model parameterisations are clearly different, regionalisation of the model results to catchments having different soil moisture response to climate change is not warranted.

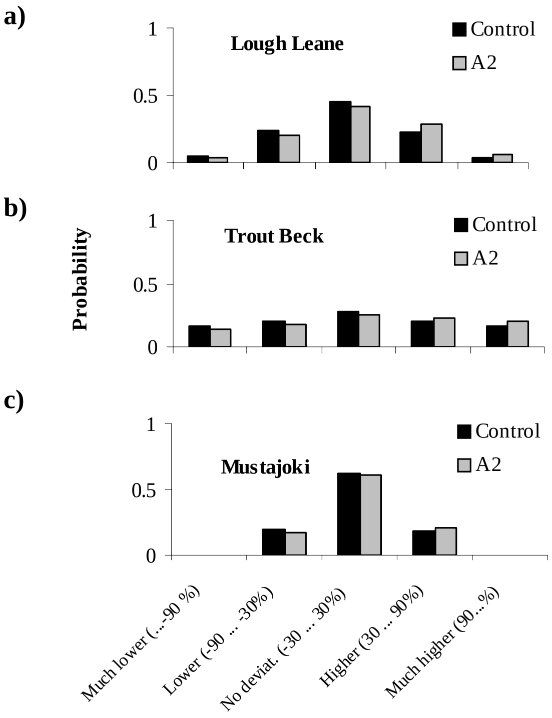

Figure 5. Deviation in summer DOC concentration for the control and A2 scenarios using three different Bayesian network parameterisations: Lough Leane (a), Trout Beck (b), and Mustajoki (c).

4. CONCLUSIONS

A methodology for combining expert knowledge and results of a DOC simulation model into a Bayesian network was presented. Use of simulation model results in estimation of conditional probability tables is straightforward, when the variables included in the Bayesian network are in accordance with the structure of the simulation model.

In a case study three Bayesian networks were parameterised utilising results of a DOC model that was calibrated to data available from three different sites. Application of the three Bayesian networks revealed that variability in the predicted deviations of both annual decomposition and summer DOC concentration was clearly different between the three parameterisations. The predicted direction of change in DOC concentrations, however, from the control scenario to the future climate scenario was same for all three Bayesian network parameterisations. The model application demonstrates how Bayesian networks can be used as a diagnostic tool for assessing the conformity of model regionalisation.

5. ACKNOWLEDGMENTS

This research was mainly funded by the European Commission within the Environment and Sustainable Development Programme under contract EVK1-CT-2002-00121 (CLIME). The first author is thankful for a travel grant awarded by the Maj and Tor Nessling Foundation. The second author acknowledges funding available from an Academy of Finland research position. Contributions from the climate and DOC experts of CLIME, as well as the software from the BNJ development team are greatly acknowledged. Thanks are due to prof. Pertti Vakkilainen for his continuing support.

6. REFERENCES

Boyer, E.W., Hornberger, G.M., Bencala, K.E. and McKnigh, D. (1996), Overview of a simple model describing variation of dissolved organic carbon in an upland catchment, *Ecological Modelling*, 86, 183-188.

Dawson, J.J.C., Billett, M.F., Neal, C., and Hill, S. (2002), A comparison of particulate, dissolved and gaseous carbon in two contrasting upland streams in the UK, *Journal of Hydrology*, 257, 226-246.

Haith, D.A., Mander, R. and Wu, R.S. (1996), Generalized watershed loading functions, user's manual, Department of Agricultural and Biological Engineering, Cornell University, Ithaca, New York.

Holmberg, M. (2003), Modelling Studies on Soil-Mediated Response to Acid Deposition and Climate Variability, Systems Analysis Laboratory Research Reports A87, Helsinki University of Technology.

Freeman, C., Evans, C.D., Montieth, D.T., Reynolds, B., and Fenner, N. (2001), Export of organic carbon from peat soils, *Nature*, 412, 785.

Jennings, E. and Naden, P. (2004), CLIME WP 5 Report on Historical Modelling of DOC, pp. 42

Kokkonen, T., Koivusalo, H., Laine, H., Jolma, A. and Varis, O. (2005), Method for defining conditional probability tables with link strength parameters for a Bayesian network, In: Proceedings of the International Congress on Modelling and Simulation, MODSIM 2005, 12-15 December, Melbourne, Australia.

Naden, P.S. (1991), Modelling water colour in upland catchments, Report to Yorkshire Water, Institute of Hydrology, Wallingford, UK.

Naden, P.S. and Watts, C.D. (1998), Development of the EPIC colour prediction model: Phase II. Report for Yorkshire Water, Institute of Hydrology, Wallingford, UK,

Nakicenovic, N., and Swart, R., (Eds), (2000), IPCC Special Report on Emissions Scenarios, United Kingdom, Cambridge University Press.

Neapolitan, R.E. (2004), Learning Bayesian networks, Prentice Hall, Upper Saddle River, NJ, USA.

Pearl, J. (1988), Probabilistic Reasoning in Intelligent Systems, Palo Alto, CA: Morgan Kaufmann, 1988.

Samuelsson, P. (2004), RCAO and HadRM3p simulation results with focus on CLIME lake sites, WP2 Deliverable, SMHI Rossby Centre, SE-602 36 Norrköping, Sweden.