Hydrol. Earth Syst. Sci., 14, 2383–2397, 2010
www.hydrol-earth-syst-sci.net/14/2383/2010/
doi:10.5194/hess-14-2383-2010

Hydrology and Earth System Sciences

Ensemble modelling of nitrogen fluxes: data fusion for a Swedish meso-scale catchment

J.-F. Exbrayat[1], N. R. Viney[2], J. Seibert[3], S. Wrede[4], H.-G. Frede[1], and L. Breuer[1]

[1]Institute for Landscape Ecology and Resources Management, Justus-Liebig-University Giessen, Heinrich-Buff-Ring 26, 35392 Giessen, Germany
[2]CSIRO Land and Water, Canberra, Australia
[3]University of Zurich, Zurich, Switzerland and Stockholm University, Stockholm, Sweden
[4]Delft University of Technology, Delft, The Netherlands and Centre de Recherche Public – Gabriel Lippmann, Belvaux, Luxembourg

Received: 5 July 2010 – Published in Hydrol. Earth Syst. Sci. Discuss.: 3 August 2010
Revised: 23 November 2010 – Accepted: 25 November 2010 – Published: 1 December 2010

Abstract. Model predictions of biogeochemical fluxes at the landscape scale are highly uncertain, both with respect to stochastic (parameter) and structural uncertainty. In this study 5 different models (LASCAM, LASCAM-S, a self-developed tool, SWAT and HBV-N-D) designed to simulate hydrological fluxes as well as mobilisation and transport of one or several nitrogen species were applied to the mesoscale River Fyris catchment in mid-eastern Sweden.

Hydrological calibration against 5 years of recorded daily discharge at two stations gave highly variable results with Nash-Sutcliffe Efficiency (NSE) ranging between 0.48 and 0.83. Using the calibrated hydrological parameter sets, the parameter uncertainty linked to the nitrogen parameters was explored in order to cover the range of possible predictions of exported loads for 3 nitrogen species: nitrate (NO_3), ammonium (NH_4) and total nitrogen (Tot-N). For each model and each nitrogen species, predictions were ranked in two different ways according to the performance indicated by two different goodness-of-fit measures: the coefficient of determination R^2 and the root mean square error RMSE. A total of 2160 deterministic Single Model Ensembles (SME) was generated using an increasing number of members (from the 2 best to the 10 best single predictions). Finally the best SME for each model, nitrogen species and discharge station were selected and merged into 330 different Multi-Model Ensembles (MME). The evolution of changes in R^2 and RMSE was used as a performance descriptor of the ensemble procedure.

In each studied case, numerous ensemble merging schemes were identified which outperformed any of their members. Improvement rates were generally higher when worse members were introduced. The highest improvements were achieved for the nitrogen SMEs compiled with multiple linear regression models with R^2 selected members, which resulted in the RMSE decreasing by up to 90%.

1 Introduction

1.1 Catchment modelling

In recent decades, anthropogenic influence on environmental systems has been demonstrated. Naturally balanced biogeochemical cycles such as the nitrogen cycles have been deeply altered (Galloway et al., 2004; Vitousek et al., 1997) since the middle of the 18th century. For about 50 years now, the increasing speed of computers allowed scientists from different fields to simulate such systems (e.g. atmosphere, hydrosphere) behaviour through different sets of mathematical equations. In hydrological sciences the catchment is considered as the basic unit and numerous different models were created from the 1960s onwards. For example, Boughton (2005) reviewed 13 different rainfall-runoff models developed in Australia alone in the second half of the 20th century. Numerical models are nowadays used as management tools from local to global scale and are able to give an approximation of the effects of different changes on a natural system (e.g. land use change, global warming).

In order to simulate both hydrology and N mobilisation and transport at the meso-scale (for catchments between

Correspondence to: J.-F. Exbrayat
(jean-francois.exbrayat@umwelt.uni-giessen.de)

100 and 100 000 km^2), multiple conceptualisations, involving different degrees of complexity, were developed (e.g. see Boughton, 2005). However, as emphasised by Breuer et al. (2008), there is no single accepted theory of catchment N cycling and models simulating the effects of nitrogen on hydrological and biogeochemical ecosystem functioning are still facing a high degree of uncertainty. Differences between models can be related to the questions they are used to address, involving different descriptions of the nitrogen balance. Some are process-based (i.e. conceptual parameters determine N turnovers rates such as in LASCAM, INCA, and HBV-N), while others are more physically-based and use parameters that are directly related to measurable quantities as for example the SWAT model. However, for the sake of simplification all models neglect some part of the well described N cycle that generally consists of ammonification, nitrification, anaerobic ammonium oxidation, denitrification, and nitrogen fixation. They sometimes totally ignore one or more N-species and corresponding turnover processes involved into this cycle, based on the assumption that models should be considered as black boxes.

Of course, due to the chaotic nature of the natural systems many simulated processes cannot be exactly described by a set of equations and this lack of knowledge involves the introduction of a certain, hardly quantifiable structural model uncertainty. Other sources of predictive uncertainty are forcing data uncertainty and parameter uncertainty, regrouped under the general term of stochastic uncertainty. It is usually difficult to assess the contribution to the total uncertainty from each of these elements. However, ensemble approaches have been proposed to investigate part of this contribution (Breuer and Huisman, 2009; Smith et al., 2004).

1.2 Ensemble modelling approach

Several global methods to assess parameter uncertainty have been described, e.g. the Monte-Carlo sampling based Generalized Likelihood Uncertainty Estimation (GLUE) approach (Beven and Binley, 1992). As parameter interactions are usually a sensitive source of uncertainty, a high number of realisations is required to cover a representative number of feasible parameter combinations and corresponding model simulations. Different combinations of parameter sets for a given model, based on a random sampling of parameter values in realistic ranges (e.g. Monte-Carlo procedures or Latin-Hypercube stratified sampling; McKay et al., 1979), are a common way to compile single-model ensembles (SME), i.e. combinations of distinct predictions obtained by perturbation of parameters, input data or initial conditions. SME built from random sampling are direct descriptions of the possible range of outcomes and illustrate part of the stochastic model uncertainty.

Multi-model ensembles (MME) are based on the combination of several deterministic model outputs. They are a state-of-the-art option for considering, or exploring, the structural model uncertainty component of the total predictive uncertainty and have been widely used in climatic and atmospheric sciences where MMEs usually outperform individual models and SMEs. However, MMEs have received little attention in hydrology even though initial MME studies of hydrological simulation were already published in the mid 1990s (Shamseldin et al., 1997).

Still, ensembles of models have been utilised in two different ways in hydrological sciences. First, some studies considered whole sets of predictions in a probabilistic way. The evaluation of these ensembles has been carried on based on skill scores which characterise the correctness of the prediction of some selected particular events, usually exceeded thresholds, in terms of correct match and false alarm rates. Good examples of such approaches were described by Renner et al. (2009) or Georgakakos et al. (2004), the latter having been realised in the frame of the Distributed Model Intercomparison Project (DMIP; Smith et al., 2004) in which calibrated and un-calibrated models were used. Probabilistic ensemble systems are typically preferred for forecasts with short lead-time and provide a direct picture of the predictive uncertainty. Recently, frameworks based on the Bayesian probabilistic theory have been developed such as the Bayesian Model Averaging technique (BMA; Raftery et al., 2005) or the hybrid Integrated Bayesian Uncertainty Estimator method (IBUNE; Ajami et al., 2007).

Some other studies combined single predictions using different statistical post-processing methods, or data-fusion schemes, in order to produce single "best" deterministic forecasts. For instance Shamseldin et al. (1997) utilised 3 combination methods to merge the output of 5 models. The philosophy behind this approach was that each model captures certain important aspects of the information available about the system and that the strengths of some may compensate weaknesses of other models, resulting in an overall better prediction. They concluded that combining outputs of rainfall-runoff models could provide better results than the best single run even with a simple averaging method. Lately, in the frame of assessing the impact of Land Use Change on Hydrology by Ensemble Modelling (LUCHEM; Breuer and Huisman, 2009) almost 30 different merging schemes were tested with 10 different model results over the same catchment as reported by Viney et al. (2009). McIntyre et al. (2005) also used model ensembles to predict discharge in ungauged or poorly gauged basin as part of the Predictions in Ungauged Basins (PUB; Sivapalan, 2003) initiative by the International Association of Hydrological Sciences. In the light of PUB, ensemble predictions are assumed to significantly increase the credibility of predictions.

We were not aware of any ensemble predictions, more particularly model combinations, in hydro-biogeochemistry to date and see this methodological approach as a first step into that direction. In this study we compiled different deterministic SMEs and MMEs by merging the outcomes of five models applied to simulate the water and nitrogen balance of a meso-scaled catchment in Sweden. Some of the fusion methods previously used in the LUCHEM project (Viney et al., 2009) were applied to the prediction of exported loads of the different N species which were considered. The reader must keep in mind that results of the different single models were not created with the aim to produce a benchmark report on the efficiency of the model structures alone. We focused our evaluation on the effect of merging results rather than on the results themselves. Differences between the models led us to make some choices (i.e. studied period, number of model realisations) that one could consider arguable (see Sect. 2). However, the main aim of this study remained to apply some data-fusion methods to different sets of nutrient predictions before comparing ensembles with single models. Results should give a primary evaluation of the applicability of the ensemble-modelling concepts to the highly uncertain N predictions with the aim to improve the global reliability of them.

The models involved were LASCAM (Sivapalan et al., 1996a,c; Viney et al., 2000) and its modified LASCAM-S version (this paper), SWAT (Arnold et al., 1998) in its 2005 version, HBV-N-D (Lindgren et al., 2007) and a new model based on the concepts proposed in INCA (Wade et al., 2002; Whitehead et al., 1998) coupled to the soil moisture equations of the HBV model (Lindström et al., 1997). This latter tool is referred as CHIMP (Combined HBV and INCA Modified in Python) throughout the text.

This article is organised as follows. Section 2 presents the catchment and the available data for model application. The models are also described as well as the methodology we adopted to create new predictions we adopted. In Sect. 3 we present the results for the single models, SMEs and MMEs N predictions. They are discussed in Sect. 4 and possible further research directions are presented in Sect. 5 along a short summary of the main conclusions that could be drawn from this study.

2 Materials and methods

2.1 The Fyris River catchment

The Fyris catchment is located in central Sweden, 90 km north of Stockholm. The Fyris River has a catchment area of 2000 km^2 and flows into Lake Ekoln, a northern part of Lake Mälaren (Sweden's third largest lake) which drains into the Baltic Sea. It is a lowland catchment whose elevation ranges between 15 and 115 m. Streams drain from the north, east and west to the outlet at Flottsund (Fig. 1). Land use is dominated by forest (mainly coniferous) which occupies about 59% of the catchment while croplands cover 33% of the area. Other minor land-use types are wetlands (4%), urban areas (2%) and lakes (2%). Forests are mainly associated with till and croplands with clay soils (Lindgren et al., 2007).

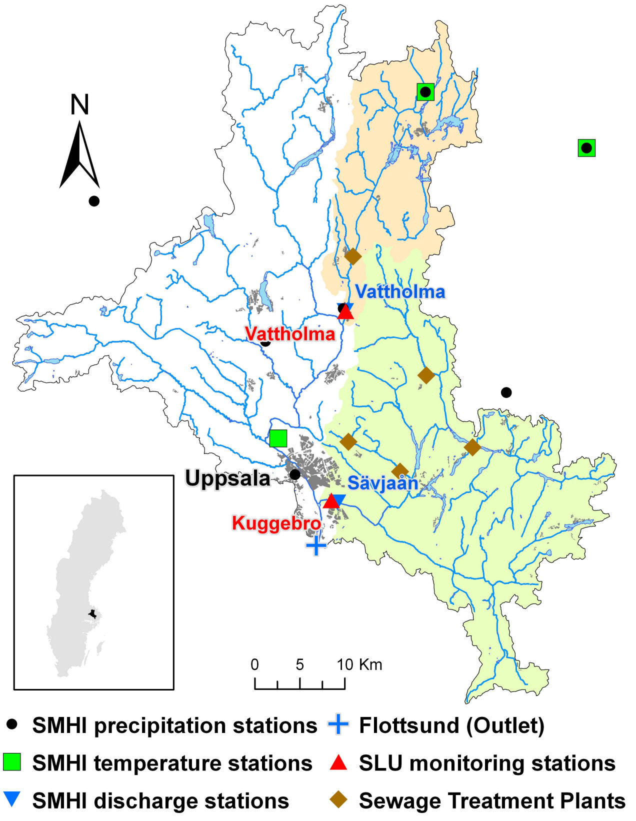
N
Vattholma
Vattholma
Uppsala
Sävjaån
Kuggebro
0 5 10 Km
● SMHI precipitation stations
🟩 SMHI temperature stations
▼ SMHI discharge stations
+ Flottsund (Outlet)
▲ SLU monitoring stations
◆ Sewage Treatment Plants

Fig. 1. The River Fyris catchment (Vattholma and Sävja subcatchments are highlighted in light-brown and light-green respectively). The names of the discharge and monitoring stations are written next to their location in blue and red, respectively.

Daily records of precipitation (8 gauges) and temperature (3 stations) collected by the Swedish Meteorological and Hydrological Institute (SMHI) were used for the 5 years study period (2000 to 2004). During this time mean annual precipitation was about 640 mm. The warmest and wettest months on average was July (>80 mm precipitation, +17 °C mean daily temperature) while the driest month was April (<40 mm precipitation) and the coldest months were December and January (−1 °C). Over the study period, two daily discharge series were available for two non-nested subcatchments: Vattholma and Sävja, with contributing areas of 281 km^2 and 699 km^2, respectively (Fig. 1). There was no gauging station available at the catchment outlet to Lake Mälaren. High flows usually occurred from late autumn to early spring. Inter-annual variability of discharge was high and thaw-refreezing events led to high temporal variability of winter discharge in some years. Mean annual runoff was 219 mm at Vattholma and 189 mm at Sävja. In-stream nitrogen input data from sewage treatment plants was also available on a daily time step for the largest plant in Uppsala, and with a biweekly or monthly resolution for four smaller ones (Fig. 1). The observed point source discharges were interpolated to a daily time step as described in Lindgren et al. (2007).

For the same period, stream chemistry data from 2 long-term measurement stations of the Swedish University of Agriculture was available for model applications. Monthly measurements of $NO_3 + NO_2$, NH_4 and Tot-N concentrations resulted in a total of 60 measurements for each station and each nitrogen species. The water quality sampling stations Vattholma and Kuggebro were located close to the gauging stations Vattholma and Sävja, respectively (Fig. 1). The Fig. 2 illustrates the monthly average concentrations of the different N-species. More treatment plants were located upstream from Sävja which was surely a key factor to explain the usually higher concentrations measured at this location (Fig. 2). Concentrations were typically higher during winter months as well. More precisely at the Vattholma station, Tot-N concentrations ranged between 0.9 and 1.4 mg/L with a contribution of 26% of $NO_3 + NO_2$ and 5% of NH_4 on average. At Sävja, the Tot-N concentrations showed a higher variability as they ranged between 1.0 and 2.9 mg/L with a contribution of 54% of $NO_3 + NO_2$ and 4% of NH_4. As shown in Fig. 2, $NO_3 + NO_2$ concentrations were the main factor explaining the Tot-N concentrations variability as a picture of their large contribution to this global measurement.

Estimates of daily exported loads were computed for these gauging stations by multiplying discharge with nitrogen concentrations measured at the sampling stations, assuming that they were representative of the mean daily concentration. This once-per-month sample was used as a single daily observation of nutrient loads. Separate sampling of the NO_2 concentration indicated that it provided a negligible contribution to the $NO_3 + NO_2$ concentration. It was therefore assumed that the NO_3 concentration is approximately equivalent to the measurements of $NO_3 + NO_2$ concentration. The combination of high concentrations and high flows during winter led to estimate large fluxes up to 700 kg and 6 tonnes of exported N per day at Vattholma and Sävja respectively. The water discharging from the Sävja sub-catchment had higher N concentrations as shown in Fig. 2, we therefore estimated higher specific N fluxes (about 13.3 g/ha d^{-1} on average) for this station than for Vattholma (6.8 g/ha d^{-1}).

2.2 Models

The five models used in the ensemble set up (LASCAM, LASCAM-S, CHIMP, SWAT and HBV-N-D) could simulate both runoff and the mobilisation and transport of different nitrogen species (see Table 1) at the landscape scale and at a daily time step. The models showed great variations in their smallest spatial units as well as the required input data, thus providing a good structural variability among the cohort (see Table 1).

For the semi-distributed models (i.e. all except HBV-N-D) we subdivided the Fyris River catchment into 70 sub-catchments. This spatial disaggregation assigned 9 and 28 upstream sub-catchments for the Vattholma and Sävja stations respectively, corresponding to mean sub-catchment areas of 31 and 25 km^2. For these models we also estimated the daily potential evapotranspiration using the Hargreaves method (Hargreaves and Samani, 1985). Daily results were then aggregated to the required temporal resolution (see Table 1). The HBV-N-D evapotranspiration input was based on monthly mean evapotranspiration estimates (Lindgren et al., 2007).

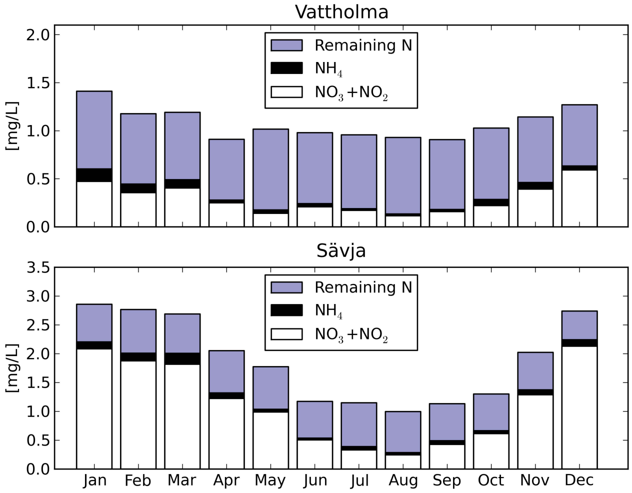

Fig. 2. Average monthly concentrations of Tot-N measured at the Vattholma and Sävja water quality stations. (Remaining N is the difference between Tot-N and the sum of the inorganic species.)

Below an overview of the various models is given, for a more detailed description of the water and nitrogen simulations the reader is referred to the original publication of the models. A short, general overview of nitrogen processes considered in these models was also provided by Breuer et al. (2008).

2.2.1 LASCAM and LASCAM-S

The semi-distributed LASCAM model was first developed for applications in arid or semi-arid regions in order to simulate water and salt balance at larger scales (Sivapalan et al., 1996a–c). Later new routines were integrated to simulate sediments (Viney and Sivapalan, 1999) and nutrients (e.g. Total-N, NO_3 and NH_4; Viney et al., 2000) mobilisation and transport. Each sub-catchment corresponds to an idealised hill slope in which 3 water and 2 nitrogen stores are interconnected. As the LASCAM model was designed to simulate dry and warm environments, the original version did not integrate any snow routine. We therefore developed the extended LASCAM-S version by implementing the degree-day approach used in the original HBV model to allow simulation of snow accumulation and melt at the sub-catchment scale. This routine is based on air temperature and a water-holding capacity of the snowpack. Depending on a threshold

Table 1. Main model characteristics. Outputs are either N loads or yields.

Model	Smallest spatial unit	Climate forcings	Vertical resolution	Outputs	N forcings
LASCAM	Sub-catchment	Daily P and annual PET	1 soil, 1 stream bank and 1 groundwater storage box	NO_3, NH_4, Tot-N	Rainfall concentration, fertilizer application
LASCAM-S	Sub-catchment	Daily P and T, annual PET	1 soil, 1 stream bank and 1 groundwater storage box	NO_3, NH_4, Tot-N	Rainfall concentration, fertilizer application
CHIMP	Land-Use	Daily P, T and PET	1 soil and 1 groundwater flow generation box	NO_3, NH_4	Wet and dry deposition, fertilizer application, STP effluents
SWAT	HRU	Daily P, maximal and minimal daily T	3 to 4 soil layers, 2 groundwater storages	NO_3, NO_2, NH_4, Organic-N	Rainfall concentration, fertilizer application, STP effluents
HBV-N-D	Grid cell	Daily P and T, monthly PET	2 linear flow generation boxes	Tot-N	Rainfall concentration, leaching coefficients, STP effluents

HRU: Hydrological Response Unit, Unique combination of a land-use with a soil type, P: Precipitation, T: Temperature, PET: Potential Evapotranspiration, STP: Sewage Treatment Plant

temperature (usually 0 °C) the snow pack melts and the water equivalent is added to the water input to the soil (Lindström et al., 1997).

The same parameter set is applied to each sub-catchment in combination with interpolated precipitation and temperature data (only for LASCAM-S). The daily potential evapotranspiration is calculated for each sub-catchment by multiplying the mean annual potential evapotranspiration by a scaling factor derived from a sinusoidal function of time. Evaporation demand is fulfilled by the 3 water stores depending on their respective levels. Nitrogen cycling at the sub-catchment scale is simulated by the following processes for both models: residue decay, plant harvest, mineralisation, volatilisation, plant uptake, nitrification, denitrification and fixation.

Water and nutrients are routed downstream. While dissolved nitrogen is not affected by any further in-stream biological or chemical reactions, water can evaporate and reinfiltrate and particulate nitrogen is affected by the erosion and sediments dynamics.

2.2.2 CHIMP

The semi-distributed INCA model requires daily effective rainfall (i.e. after canopy interception) and daily soil moisture deficit input data (Whitehead et al., 1998) which are usually difficult to assess. These variables were derived by feeding the flow generation and nitrogen routines of INCA with the output of the snow and soil moisture routines of HBV (Lindström et al., 1997). The INCA nitrogen module only outputs predictions of inorganic nitrogen species (i.e. NO_3 and NH_4) balance. All the equations were adapted from literature references (Lindström et al., 1997; Wade et al., 2002) and the two models were regrouped under the name CHIMP.

Each sub-catchment is disaggregated into up to 5 different land-use classes which all have their own parameter sets for water and nutrients balance.

The HBV snow routine is also based on the empirical degree-day approach. Evapotranspiration is calculated as a function of the input potential evapotranspiration and the HBV soil store. Below a chosen threshold of soil moisture the actual daily evaporation is computed as a linear function of the daily potential evapotranspiration. Above this threshold the total evaporation demand is fulfilled (Lindström et al., 1997). The soil routine of HBV provides the hydrological effective rainfall (e.g. water available for runoff) which is routed to the 2 INCA flow generation boxes. The soil moisture deficit required by INCA is computed as the difference between this soil storage content and the maximum value.

In both flow generation boxes different nitrogen turnover processes are simulated: plant uptake, nitrification, denitrification, fixation, mineralisation and immobilisation. The organic N store is considered as infinite so that the mineralisation rate does not depend on its magnitude. Each process is characterised by a kinetic equation which is based on turnover rates (user input) as well as a temperature and a soil moisture deficit index. Water, NO_3 and NH_4 concentrations discharge into the sub-catchment stream. As land-use classes are not spatially identified within a sub-catchment, their outputs are weighted by their respective areas to contribute to stream flow. There is no re-infiltration but nitrification and denitrification can still occur. Sewage treatment plant effluents are directly added to the stream NO_3 and NH_4 contents.

2.2.3 SWAT

The SWAT model (Arnold et al., 1998) is a semi-distributed, physically-based model (Gassman et al., 2007). It is able to simulate the long term water and nutrients balance (e.g. NO_3, NO_2, NH_4 and Organic-N, see Table 1). We used the SWAT model in its 2005 version.

Each SWAT sub-basin is divided into Hydrological Response Units (HRU). Each HRU corresponds to a single combination of a land-use class and a soil-type that can be parameterised individually. HRUs are not spatially identified within their sub-catchment. SWAT simulates snowpack and snowmelt processes at the HRU scale based on the empirical degree-day approach with a daily update of the melting rate between user defined maximum and minimum values. At the HRU scale, SWAT incorporates a simplified dynamic crop growth module. The corresponding canopy intercepts a part of the precipitation which is a function of its Leaf Area Index. Evaporation demand is first fulfilled by the canopy and eventual higher demand is partly fulfilled by the soils. In-stream discharge of each HRU is composed of several elements: surface runoff, lateral flow and baseflow. Therefore, the idealised hill slope is composed of interconnected multi-layer soil storages and a double groundwater system. Different nitrogen processes are simulated within each HRU soil layer: plant uptake, residue decay, mineralisation, nitrification, volatilisation, denitrification, fixation and leaching. Turnover rates depend on temperature and moisture, or soil water content. As HRUs are lumped at the sub-catchment scale, their contribution to in-stream water and nutrient content are summed up before being routed to the stream.

In-stream water and corresponding nutrient content routing is based on a variable storage method (Williams, 1969). Re-infiltration and biochemical nitrogen reactions are allowed: algal respiration and uptake, hydrolysis and oxidation. Turnover rates are temperature-dependent.

The SWAT setup was realised with the support of the ArcSWAT 2.1.4 for ArcGIS 9.2 extension (Olivera et al., 2006). Within 70 sub-catchments for the whole Fyris River watershed, a total of 622 HRUs were delineated by combining 5 land-use classes with 7 soil types. This corresponds to a total of 108 and 232 HRU for Vattholma and Sävja sub-basins, representing mean areas of 2.60 and 2.48 km^2, respectively. ArcSWAT automatically assigns the climatic records of the nearest station to each sub-catchment.

2.2.4 HBV-N-D

HBV-N-D is a fully distributed version of the original HBV model routines combined with the conservative solute transport model concepts of the TAC^D model (Wissmeier and Uhlenbrook, 2007). The model used a grid representation of the catchment (here $250 \times 250\,m^2$ grid cells) and is implemented in the PCRaster modelling environment (Karssenberg et al., 2001). HBV-N-D requires daily precipitation and temperature data input and weighs resulting flow and storage amounts per fractions of land-use class in each grid cell. Snow is simulated using an empirical degree-day approach at the land-use scale. Within a grid cell, HBV soil moisture and flow generation boxes can be parameterised individually for each land-use class (Lindström et al., 1997). The actual evaporation is calculated in accordance with the HBV equations following the same process described in Sect. 2.2.2. Water entering a response function is assigned a Tot-N leakage concentration. HBV-N-D is based on a single flow direction algorithm (O'Callaghan and Mark, 1984) for lateral cell to cell connection, so that water or Tot-N output from any runoff generation box is diverted into the corresponding box of the neighbouring downstream cell. When a grid cell is identified as a stream cell, a simple distribution function is applied to route the water and corresponding Tot-N content downstream. Nitrogen retention is modelled as a net effect of various biogeochemical processes such as uptake, sedimentation and denitrification. It is a function of the Tot-N concentration, the average temperature of the 10 last days and a free parameter and retention occurs in each response function box as well as in lakes or in-stream.

The HBV-N-D model application in this study is based on the identical model setup utilised in an earlier model comparison for nitrogen source apportionment (Lindgren et al., 2007). The running time of the model was a limiting factor for our study which explains the choice of a relatively short 5 years evaluation period and discrepancies in the hydrological calibration and ensemble generation procedures (Sect. 2.3).

2.3 Ensembles construction and assessment

A global overview of the methodological approach used in this study is presented in the Fig. 3. This flow chart summarises the different steps which followed to create the different types of ensembles. A more detailed chronological description of the adopted methodology is presented in the next paragraphs.

2.3.1 Hydrological calibration

Water transports particulate and dissolved chemical species through a catchment. A certain part of the stochastic uncertainty of the nutrient fluxes is then logically linked to the variations of the hydrologic parameters. In a calibration

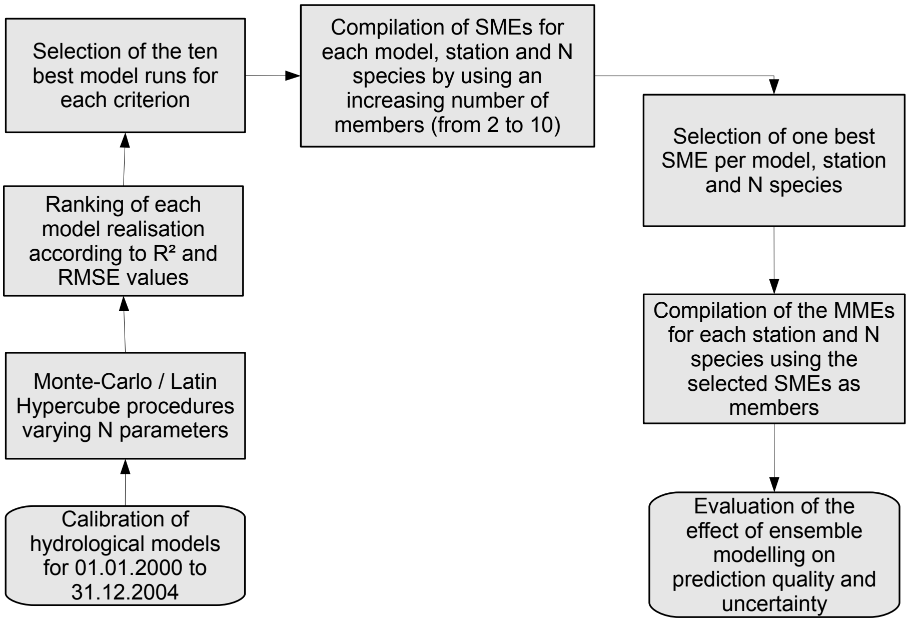
Selection of the ten best model runs for each criterion
Compilation of SMEs for each model, station and N species by using an increasing number of members (from 2 to 10)
Selection of one best SME per model, station and N species
Ranking of each model realisation according to R² and RMSE values
Compilation of the MMEs for each station and N species using the selected SMEs as members
Monte-Carlo / Latin Hypercube procedures varying N parameters
Calibration of hydrological models for 01.01.2000 to 31.12.2004
Evaluation of the effect of ensemble modelling on prediction quality and uncertainty

Fig. 3. Methodology used in this study to compile SMEs and MMEs.

context, numerous studies were based on a two-step approach. First, modellers determined the optimal parameter sets for hydrology only. Then, they calibrated the nutrient component only while keeping this optimal water balance description (e.g. Andersson et al., 2005; Viney and Sivapalan, 2001; Wade et al., 2002).

Subsequently, in this study we calibrated the hydrological components of the models against the two available discharge records (i.e. Vattholma and Sävja) in order to obtain the best water balance simulation for the whole catchment over the study period from 1 January 2000–31 December 2004. Only 60 observations of N load were available for each catchment. We therefore did not use a validation period as it would have substantially shortened the study period limiting the number of observations available for both calibration and model quality assessment. Given the methodological objective of this very first study on hydro-biogeochemical model fusion, we believe that it was acceptable to disregard this validation.

For all models except the computationally expensive HBV-N-D model, the Parameter Solution method (ParaSol; van Griensven et al., 2002) which is based on the Shuffled Complex Evolution algorithm (SCE-UA; Duan et al., n.d.) was used for parameter optimisation. The ParaSol method requires the daily sum of the squared errors (SSE, Eq. 1).

$$\mathrm{SSE} = \sum_{i=1}^{N} (O_i - S_i)^2 \tag{1}$$

In the Eq. (1), O_i and S_i are observed and simulated runoff at time step i. Their squared difference is summed for each of the N considered time steps. ParaSol automatically aggregates SSE values for each considered flux in a global objective criterion which is reduced by the SCE-UA algorithm. The Parameter Estimator (PEST; Doherty, 2004) was chosen to calibrate the computationally expensive HBV-N-D model as it usually requires fewer model realisations. The objective function was a weighted SSE with weights set as the inverse of the standard deviation of the corresponding observations.

In order to compare the goodness-of-fit resulting from the calibration efforts at each station, the results were expressed as the Nash-Sutcliffe efficiency for daily flows (NSE; Nash and Sutcliffe, 1970) which is a common standardisation of the SSE normalised by the variance of the observations (Eq. 2).

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (O_i - S_i)^2}{\sum_{i=1}^{N} \left(O_i - \overline{O}\right)^2} \tag{2}$$

In Eq. (2), O_i and S_i corresponds to notations in Eq. (1) while $\overline{O}$ is the mean observed runoff over the N considered time steps. No differences in model performance rankings are to be expected after this transformation. NSE values range between $-\infty$ and 1, the latter being achieved for a perfect fit between observations and predictions. NSE values tend to be not very sensitive to the volume error but are biased in favour of peak flows (Krause et al., 2005; Legates and McCabe Jr., 1999) which are dominant in this catchment in response to the spring snow melt. In such conditions, good NSE values could be achieved with somewhat biased predictions. To cope with this, we also evaluated our models by using the total bias to assess the performance of the models to estimate the total runoff.

2.3.2 Nitrogen ensembles construction

Different application cases of the same models might have different modelling targets depending on the aim of the study. In the case of nutrient predictions one may focus the study on concentrations if one is more interested in direct water quality estimation (e.g. Arheimer and Lidén, 2000) or on loads/yields if one is more interested in source apportionment or assessing the global contribution of fertiliser application to water bodies nutrient balances (e.g. Grizzetti et al., 2003; Viney and Sivapalan, 2001). Still, evaluating load predictions tend to be more forgiving than concentration as they are usually correlated with runoff. However, as the scope of our application was to evaluate the effect of different data-fusion methods rather than on the prediction outcomes themselves and as this analysis is computationally rather expensive, we limited out modelling target to yields in this study. Accordingly, we did not use any optimisation algorithm for N predictions in order to focus the study on the stochastic uncertainty linked to the N algorithms. A large number of model runs was realised by keeping the previously calibrated water balance parameters constant, randomly altering the parameters governing only the N mobilisation and transport. A Monte-Carlo procedure was used for LASCAM, LASCAM-S, CHIMP and SWAT, providing 40 000; 40 000; 60 000 and 20 000 realisations corresponding to 16; 16; 28 and 7 altered parameters respectively. A Latin-Hypercube stratified sampling procedure (McKay et al., 1979) was chosen for the computationally expensive HBV-N-D model, leading to overall 280 model runs for only 4 altered parameters. We are conscious that one would argue that the number of model realisations was not sufficient to explore the whole uncertainty. However, our aim was to create large sets of model realisations before testing the effect of our different data-fusion methods. Moreover optimal parameter sets may actually not exist according to the equifinality theory (Beven and Freer, 2001) or may differ for the different considered N species and stations as well. We still allocated more runs to the more parameterised models, taking into account that more parameter interactions would occur.

The large number of realisations for each model allowed us to compile several SMEs for each model, N species and measurement station independently. Each realisation was evaluated with two goodness-of-fit indicators: the coefficient of determination R^2 and the Root Mean Squared Error (RMSE; Eq. 3).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (Oi - Si)^2}{N}} \tag{3}$$

In Eq. (3) notations correspond to those used in Eqs. (1) and (2). The coefficient of determination and the RMSE were computed by comparing the estimated exported loads with the predictions of the corresponding time step. The difference between the two criteria is that R^2 requires only the dynamics, or relative differences, to be simulated correctly, while RMSE evaluates differences between observed and simulated values (Legates and McCabe Jr., 1999). While R^2 is not a suitable criterion alone, because the best achievable value of 1.0 does not imply a perfect fit. However, it provides in the case of N concentrations the useful information of whether at least the dynamics are correct. The lower the RMSE, the better the results, and by evaluating the error this last criterion is partly influenced by the bias of prediction. For each case (i.e. each N species, station and model), SMEs were compiled by using the five merging schemes in Table 2 applied to the 2 to 10 best model runs regarding each criterion respectively (i.e. R^2 and RMSE). The first three methods (ME, WM and MD) are the most simple and even the weights involved in the WM scheme do not depend on the model combination (i.e. one model will always have the same weight corresponding to either the corresponding value of R^2 or the inverse of the RMSE). On the other hand, the weight assigned to each model in a linear regression varies when combined with different models. These regression techniques have been used in several modelling contexts: meteorological forecasts (Krishnamurti et al., 2000), sea surface temperature (Fraedrich and Smith, 1989) or rainfall-runoff (Ajami et al., 2006; Shamseldin et al., 1997; Viney et al., 2009). Because normal linear regression (UR) might include a non-zero intercept leading to predict flow even if none of the members predicts flow (Viney et al., 2009), regressions with a constrained zero-intercept (CR) were adopted.

This resulted in a total of 90 SMEs per model, station and nitrogen species, and 2160 SMEs overall (CHIMP being not able to simulate Tot-N, and HBV-N-D not being able to simulate NO_3 and NH_4). The coefficients obtained by unconstrained multiple linear regression (UR) and constrained multiple linear regression (CR) with one grabbed sample per month were applied to the whole time series (i.e. the daily predictions over 5 years). Due to the occasional occurrence of negative coefficients, some negative predictions may occur and the corresponding SMEs were discarded.

For each model, N species and station, the best SME considering RMSE was selected for inclusion in the MMEs. A total of 4 SMEs was available in each case (CHIMP being not able to simulate Tot-N and HBV-N-D only predicting Tot-N and no other N solutes). Again following the 5 merging schemes outlined in Table 2 applied to each of the 11 possible combinations of 2 to 4 selected SMEs (i.e. 6 combinations of 2 models, 4 combinations of 3 models and 1 combination of 4 models), we obtained 55 different MME predictions for each N species. Regression coefficients also had to be re-calculated for each combination. Finally, we evaluated

Table 2. Overview of the adopted merging schemes for ensemble generation.

Merging scheme	Description	Abbreviation
Mean	Daily mean of the predictions	ME
Weighted mean	Daily weighted mean of the predictions[a].	WM
Median	Daily median value of the prediction	MD
Un-constrained multiple linear regression	Observations are used as dependent variables while predictions are used as independent ones.	UR
Constrained multiple linear regression	Same as above with an interception constrained through the origin	CR

[a] Weights are set at the objective function value for R^2, but its inverse for RMSE.

the evolution of both criteria for every generated MME and SME by also quantifying the improvement rate for RMSE.

3 Results

3.1 Hydrology

A summary of calibration results for the water balance components of the models is presented in Table 3. A high variability across models is observed between the different NSE and bias values. The models perform alternatively better at Vattholma (LASCAM-S, SWAT) or at Sävja (LASCAM, CHIMP, HBV-N-D). Considering the bias, which was not used for the automatic calibration procedure, results are worse for Sävja than for Vattholma, except for CHIMP. Over-predicting models at Vattholma under-predict at Sävja and vice-versa. While SWAT shows the best NSEs for each station it also presents the highest absolute biases in each case.

3.2 Nitrogen

3.2.1 Single runs overview

A summary of the best simulations for each model and each criterion (R^2 and RMSE) is given in Table 4 together with the results of the SMEs. As expected models that were presenting the best R^2 value for each N species do not necessarily have the best RMSE performance. SWAT performed the best for NO_3 simulations for both criteria. For NH_4, the best R^2 and RMSE were provided by CHIMP at Vattholma and LASCAM-S at Sävja. For Tot-N, the best R^2 values were obtained with SWAT and the best RMSE with LASCAM-S, SWAT presenting the worst RMSE values in those cases. While LASCAM-S presented better results than the original LASCAM for the hydrological predictions (Table 3), it did not always give significantly better results for nitrogen predictions, being even outperformed for Tot-N at Vattholma.

Table 3. Goodness-of-fit indicators of calibrated models runs for daily runoff prediction between 1 January 2000 and 31 December 2004 (best achieved values are highlighted in bold).

	Vattholma		Sävja	
Model	NSE	Bias (%)	NSE	Bias (%)
LASCAM	0.48	+ 10	0.53	−11
LASCAM-S	0.65	+6	0.64	−12
CHIMP	0.67	−5	0.69	**+5**
SWAT	**0.83**	−13	**0.76**	+18
HBV-N-D	0.65	−**5**	0.76	+13

3.2.2 Single-model ensembles

SMEs decreased the RMSE in all cases (Table 4 and Fig. 4). The selected SMEs presented in the Table 4 were the ones representing the best compromise between R^2 and RMSE explaining why in four cases we obtained a lower R^2 for the selected SMEs than for the best single member (i.e. Tot-N at Vattholma and for LASCAM-S, NO_3 for LASCAM-S, CHIMP and SWAT at Sävja).

Improvement of RMSE could be quantified in terms of error reduction which can also be visualised in the Fig. 4. Corresponding decreases in RMSE ranged between 9 and 92% for NO_3 at Sävja with LASCAM and NH_4 at Vattholma for SWAT respectively. The RMSE was reduced by more than 30% in 16 cases out of 24.

On the other hand the evolution of R^2 mainly showed weak improvements but still the best MMEs present a better agreement between estimated and predicted loads as illustrated in the scatter plots of the Fig. 5. More than 80% of UR ensembles had to be discarded due to the occurrence of some negative regression coefficients providing negative predictions when applied to the whole simulated time series. We used CR ensembles to circumvent this problem. The MD ensemble never increased both criteria and only rare and weak improvements were achieved by using the ME and WM ensemble models. The best SME results were always obtained

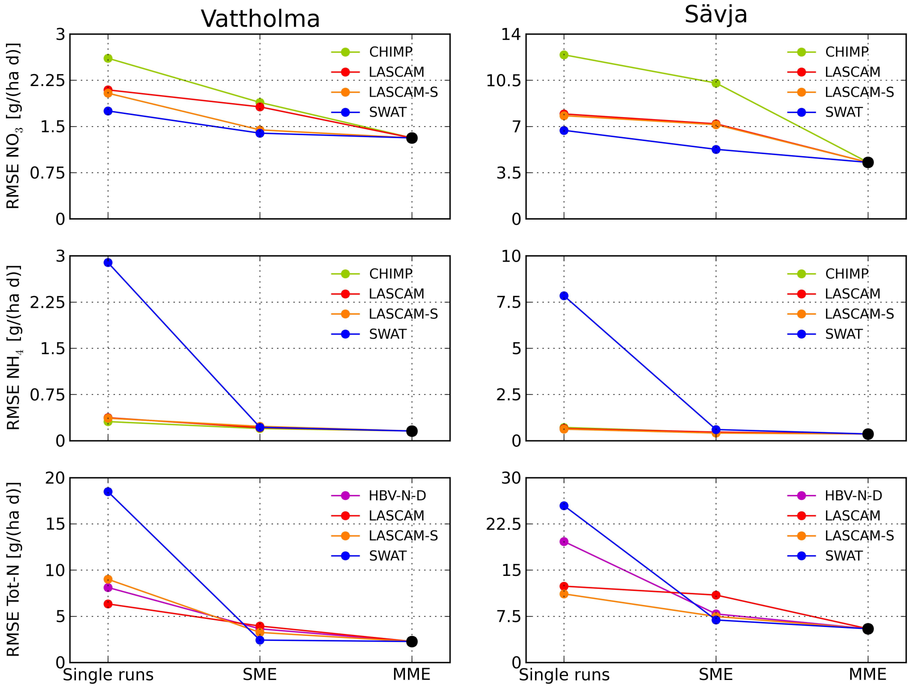

Fig. 4. Evolution of the best RMSE value between single models, SMEs and MMEs (full black circle).

by UR ensembles compiled with R^2 selected members. As was the case for single predictions LASCAM-S SMEs did not always obtain better criterion values than the original LASCAM model.

3.2.3 Multi-model ensembles

About 27% of the ME MMEs decreased RMSE values in comparison to their best member (i.e. selected SME), and 38% of the weighted average MMEs compared to the best SMEs. Similarly to SMEs, the feasibility of each MME prediction was checked prior to further model evaluation. Once again, a high number (49%) of regression schemes was discarded. However all the available UR and CR ensembles presented an improvement of both criteria compared to the predictions of the best SME. The best results were obtained by including the maximum number of ensemble members (4 for nitrogen if not dismissed due to negative unrealistic negative predictions) in UR ensembles. The best MMEs are illustrated for each N-species and station in the scatter plots of the Fig. 4 beside the best SME predictions. It showed the better agreement between observations and predictions achieved with the MME. It is interesting to notice that the improvement in RMSE of MMEs was stronger if members that were combined in a MME originally presented weaker SME results for R^2 values (e.g. an improvement in RMSE of 27% for NH_4 at Vattholma) as compared to members of MMEs that already performed well (e.g. an improvement in RMSE of 6% for Tot-N at Vattholma).

4 Discussion

At each discharge and nitrogen station, the quality of the predictions was extremely variable across models even though homogeneous input datasets were used. This behaviour has been reported by many others in hydrological modelling notably (Breuer et al., 2009; Reed et al., 2004; Refsgaard and Knudsen, 1996) but also for nutrient predictions at different scales (Diekkrüger et al., 1995; Kronvang et al., 2009a). This variability was also defined as the starting point for any ensemble prediction (Georgakakos et al., 2004; Shamseldin et al., 1997; Viney et al., 2009) with the idea to compensate weaknesses of some models with strengths of the others to improve the global prediction.

A preliminary step of this study constituted in an intercomparison between the different N models. Heterogeneous results for water and nutrient balance description were obtained. The models which provided the best hydrological predictions did not always give the best N prediction results even though loads are dominated by runoff and no global best model was to be found in our particular case study justifying in some way the need to use sets of different model conceptualisations. Corresponding nitrogen concentration predictions (not shown here) were of poorer quality as a result of combined imperfect simulations of water and nutrient balances. This also implies that the observed N concentration dynamics could not be fully represented by the chosen. However, such limitations can be typically found in solute transport modelling, when evaluating concentration

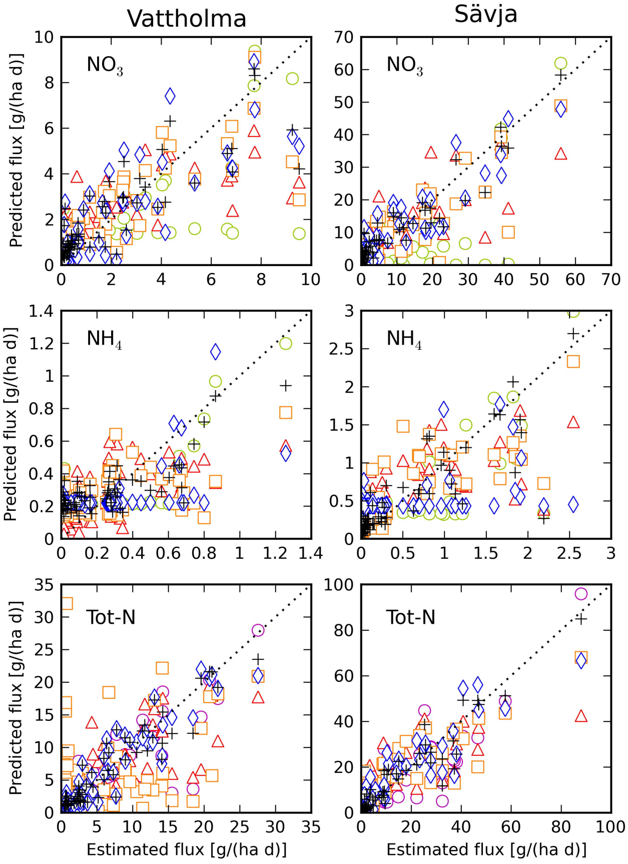

Fig. 5. Estimated daily nitrogen loads (x-axes, in g/ha d^{-1}) against different prediction from the best SMEs (y-axes, in g/ha d^{-1}): CHIMP (green circles), HBV-N-D (magenta circles), LASCAM (red triangles), LASCAM-S (orange squares), SWAT (blue diamonds). Crosses represent the prediction of the best MME for each N species at each station. Corresponding criteria are summarised in Table 4.

dynamics that were outside the calibration target of predicting nitrogen loads. A more in-depth evaluation of the effect of different calibration targets (e.g. concentration vs., loads) is certainly necessary in future applications, when the focus is on more accurate predictions,but is out of the scope of the current study. A first effort to intercompare a variety of models that have been set up to predict nitrogen flows in different agricultural systems has been published by Diekkrüger et al. (1995). Like in our study the models were set up using a common dataset, guaranteeing that eventual prediction differences depended only on the models or applied concepts themselves. Results showed a very high variability between predictions for different N turnover processes. Here we did not quantify process rates, considering our models as black boxes and only analysing the net export of the considered N species through the two outlets. This may be considered as an empirical approach but some big differences in terms of RMSE between the single models (Table 4 and Fig. 4) intrinsically imply high differences in the total N balance description and thus in the involved processes. Recently Kronvang et al. (2009a) provided the first comprehensive results of a model intercomparison project on nutrient load predictions at the mesoscale. This study included 8 nitrogen models (Kronvang et al., 2009b) and concluded that no single nutrient model could be recommended to simulate catchment scale nutrient losses. Accordingly our results (Table 4 and Fig. 4) show that the best single performers (i.e. model) vary between catchments and N species.

Considering the hydrological calibration results (Table 3) the implementation of a snow module into the LASCAM model significantly improved the water balance description, whereas very similar results for the different nitrogen species were obtained. As illustrated on Fig. 4 the SWAT model always presented the worst RMSE for N except for NO_3 while it presented the best calibration results for hydrology (Table 3). The same behaviour is observed with HBV-N-D which gave an equivalent calibration result for hydrology at Sävja while not presenting the best Tot-N prediction for this station.

These results show that improved hydrological predictions applied with the same N balance description in the case of LASCAM and LASCAM-S do not necessarily provide better nutrient export predictions. We conclude that in these models the N components behave almost independently from the water routines even though water is the driving force for the movement of any dissolved nutrients in the catchment. This theoretical statement coupled to the lack of representation of the dominating hydrological process in this region (i.e. snow) in the original version led us to modify the model. However, when considering the actual results, one could argue that these modifications may not have been necessary in the frame of only getting better N predictions as LASCAM already outperformed at least one model amongst CHIMP, SWAT and HBV-N-D in each case. We see these results as an indication that the original LASCAM model, which was not developed for such hydro-climatic conditions, might have given good nutrient predictions based on not wrong, but unknown reasons.

Moreover, high discrepancies between criteria values of each of our “best” nitrogen models (Fig. 4) indicated heterogeneous prediction qualities. The large number of runs realised previously to the SMEs data-fusion procedure guaranteed us that the mismatches of the predictions with the observations could not only be attributed to the sole parameter uncertainty. As suggested by Vrugt and Robinson (2007), uncertainty of predictions also depends on the inadequate or incomplete representation of processes which could be illustrated by the differences in nitrogen cycling conceptualisations (see Breuer et al., 2008). Moreover we demonstrated that more complex tools based on a more detailed description of processes (e.g. SWAT in comparison of LASCAM and LASCAM-S) were not necessarily better as already highlighted by Abrahart and See (2002) in the frame of another

Table 4. Nitrogen results summary between 1 January 2000 and 31 December 2004. RMSE is expressed in g/ha d^{-1}.

Models	Vattholma						Sävja					
	NO_3		NH_4		Tot-N		NO_3		NH_4		Tot-N	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LASCAM												
Best Run[a]	0.53	2.09	0.24	0.37	0.57	6.34	0.66	7.95	0.51	0.66	0.65	12.38
Selected SME[b]	0.53	1.82	0.36	0.21	0.57	3.94	0.66	7.20	0.52	0.46	0.65	10.95
LASCAM-S												
Best Run[a]	0.67	2.04	0.19	0.36	0.69	9.00	0.77	8.08	0.57	0.62	0.77	11.13
Selected SME[b]	0.67	1.45	0.23	0.23	0.08	3.25	0.69	7.13	0.63	0.41	0.77	7.47
CHIMP												
Best Run[a]	0.38	2.61	0.44	0.31			0.38	12.43	0.53	0.72		
Selected SME[b]	0.42	1.89	0.45	0.20			0.34	10.3	0.55	0.44		
SWAT												
Best Run[a]	0.68	1.75	0.32	2.89	0.84	18.49	0.84	6.70	0.16	7.84	0.82	25.45
Selected SME[b]	0.69	1.39	0.33	0.22	0.86	2.42	0.83	5.27	0.18	0.60	0.85	6.90
HBV-N-D												
Best Run[a]					0.38	8.12					0.62	19.63
Selected SME[b]					0.69	3.64					0.80	7.88
Best MME[c]	0.73	1.31	0.65	0.16	0.88	2.27	0.89	4.28	0.73	0.36	0.90	5.47

[a] Best single model runs regarding R^2 and RMSE are not necessarily obtained with the same parameter set.

[b] Selected SME R^2 and RMSE values are obtained with the same ensemble chosen to be merged in MMEs.

[c] Best MME is characterised by the both best R^2 and RMSE values.

data-fusion comparison project. Interestingly, Fig. 5 clearly shows that most of the models tend to underestimate the high loads which correspond to winter and spring months as a result of combined high flows and high concentrations (see Fig. 2). There might be a general failure in the activation of N mobilisation following snow events within all the studied models. Even though this is beyond the scope of this study, the investigation of the reasons why models behave like this has to be undertaken in the future in order to increase our process-understanding.

A good match of observed and simulated N loads is not necessarily required to achieve good R^2 values. Selecting the best single runs regarding this criterion is therefore probably source of a large predictive uncertainty. Abrahart and See (2002) demonstrated that the most efficient data fusion scheme depended on the application case. Here and as already depicted by Viney et al. (2009) and for our hydrological results, UR and CR regression ensembles created in a calibration context gave the best results for our SMEs and MMEs. However, this was only true when merging the predictions showing the highest values for the R^2 criterion even while the RMSE values were very high. Different well trended realisations were weighted in an optimal way to adjust the predicted absolute values as illustrated by some strong decrease of the RMSE values (e.g. around 90% for NH_4 and Tot-N with SWAT at Vattholma; Table 4). The risk of unrealistic values when extrapolating the coefficients obtained with monthly measurements remained very high. The ME, WM and MD schemes still gave worse predictions (usually around the RMSE value of the best single run). Viney et al. (2009) also demonstrated that that in the LUCHEM project the multiple linear regression predictions quality was significantly reduced between calibration and validation periods contrary to most of the utilised schemes (including the simplest mean ones).

Some of the single models and SMEs already achieved predictions that are almost as good as the very best MMEs (e.g. LASCAM-S and CHIMP for NH_4, SWAT for Tot-N, Table 4 and Fig. 5). However, MME results always showed the best overall model performances for both criteria. This confirmed the benefits of exploring different model structures amongst which we may not know an a priori best one. This was already suggested by Butts et al. (2004) for hydrological predictions and gave a proof of the value of standard data-fusion methods in a hydro-biogeochemical context. The different scatter plots of the Fig. 5 showed that the predictions of the best MME were always surrounded by the predictions of the introduced SMEs. For instance the predictions corresponding to the highest estimated N load export always showed a great deviation from the line symbolising the perfect fit for the different SMEs, especially for Tot-N at Sävja. In this latter case, the best MME gave a good prediction for the extreme value. It is a confirmation of the advantage of combining the different model structures and getting a part of the information about the system into each of these conceptualisations. Nevertheless the improvements were not very high compared to those of the SMEs which had already shown tremendous improvements as a result of the fusion of a large number of single model predictions. There was therefore only limited space for further improvement of the overall model performance. Moreover, as measurements and the method to estimate the loads were already sources of uncertainty, it would not be reasonable to trust a perfect fit as well. This shows that a greater quality of prediction would be achieved if the ensemble were already based on better members. Prediction users would therefore benefit from more scientifically correct model structures. This is an aspect that should not be forgotten for the sake of getting better predictions.

Constructing MMEs from the best single runs, or from calibrated runs, rather than the SME could have been another way to take into account the global prediction uncertainty linked to the full set of considered models, while introducing worse predictors along a probably higher uncertainty. Improvement would have surely been more obvious in that case as single models always showed worse performances than any SMEs (cf. improvement rates in Sect. 3.2.2 and Fig. 4). Other methods could also have been applied like Bayesian model averaging and Kalman filtering techniques which provided more accurate results and allowed a more reliable treatment of conceptual errors in some other hydrological and climatic studies (Raftery et al., 2005; Vrugt et al., 2006; Vrugt and Robinson, 2007).

5 Conclusions

A total of 2490 ensembles (SMEs and MMEs) were compiled. Data-fusion procedures have been demonstrated to greatly improve the re-prediction of different N fluxes at the mesoscale and especially to be able to produce good predictions from very poor model realisations. In every studied situation numerous combination schemes showed improvements compared to the performance of their single members. For all the studied fluxes, regression schemes were the most efficient combinations but need, as well as the adopted weighted average, comparison with observed data. This is not applicable in ungauged conditions for instance. We see ensemble predictions as a promising research direction in the domain of hydro-biogeochemical sciences. Of course, more data-fusion schemes could be tested but we should also analyse ensembles directly in a probabilistic way to assess the risk of occurrence of certain particular events (e.g. short-term concentration thresholds, long-term yield change).

However, even if lots of rainfall-runoff models exist, only few are able to simulate N mobilisation and transport. Diversity in the considered N species is also a limiting factor and we cannot make definitive statements on the effect of ensemble modelling in that case. Moreover, a lot of different averaging and probabilistic methods have been described to handle large ensembles of model predictions. Their applicability remains to be checked in the hydro-biogeochemical context. Nevertheless, the understanding of hydro-biogeochemical fluxes, and therefore the model structures, has to be improved and ensemble procedures would also benefit from better members. However, this would require matching datasets with very high temporal resolution that are most often not available and therefore concentrations would also be a better choice to use as prediction targets.

Acknowledgements. The meteorological and hydrological data was obtained from SMHI (Swedish Meteorological and Hydrological Institute) and nitrogen data was made available by SLU (Swedish University of Agricultural Sciences), Department of Aquatic Sciences and Assessment. Thanks also to the municipality of Uppsala and the Uppsala County Administration Board for providing data on wastewater treatment plants and rural households.

Edited by: A. Butturini

References

Abrahart, R. J. and See, L.: Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments, Hydrol. Earth Syst. Sci., 6, 655–670, doi:10.5194/hess-6-655-2002, 2002.

Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results, J. Hydrometeorol., 7(4), 755, doi:10.1175/JHM519.1, 2006.

Ajami, N. K., Duan, Q., and Sorooshian, S.: An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resour. Res., 43(1), W01403, doi:10.1029/2005WR004745, 2007.

Andersson, L., Rosberg, J., Pers, B. C., Olsson, J., and Arheimer, B.: Estimating catchment nutrient flow with the HBV-NP model: sensitivity to input data, Ambio, 34(7), 521–532, 2005.

Arheimer, B. and Lidén, R.: Nitrogen and phosphorus concentrations from agricultural catchments – influence of spatial and temporal variables, J. Hydrol., 227(1–4), 140–159, doi:10.1016/S0022-1694(99)00177-8, 2000.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: Model development, J. Am. Water Resour. As., 34(1), 73–89, 1998.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6(3), 279–298, doi:10.1002/hyp.3360060305, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249(1–4), 11–29, 2001.

Boughton, W.: Catchment water balance modelling in Australia 1960–2004, Agr. Water. Manage., 71(2), 91–116, doi:10.1016/j.agwat.2004.10.012, 2005.

Breuer, L. and Huisman, J.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM), Adv. Water Resour., 32(2), 127–128, doi:10.1016/j.advwatres.2008.10.010, 2009.

Breuer, L., Vaché, K. B., Julich, S., and Frede, H.-G.: Current concepts in nitrogen dynamics for mesoscale catchments/Concepts actuels relatifs à la dynamique de l'azote dans les bassins versants de méso-échelle, Hydrolog. Sci. J., 53(5), 1059, doi:10.1623/hysj.53.5.1059, 2008.

Breuer, L., Huisman, J., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM), I: Model intercomparison with current land use, Adv. Water Resour., 32, 129–146, 2009.

Butts, M., Payne, J., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, J. Hydrol., 298(1–4), 242–266, doi:10.1016/j.jhydrol.2004.03.042, 2004.

Diekkrüger, B., Söndgerath, D., Kersebaum, K. C., and McVoy, C. W.: Validity of agroecosystem models a comparison of results of different models applied to the same data set, Ecol. Model., 81(1–3), 3–29, 1995.

Doherty, J.: PEST-Model Independent Parameter Estimation User Manual: 5th edition, Watermark Numerical Computing, Brisbane, Australia, 2004.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28(4), 1015–1031, doi:199210.1029/91WR02985, 1992.

Fraedrich, K. and Smith, N. R.: Combining Predictive Schemes in Long-Range Forecasting, J. Climate, 2(3), 291–294, 1989.

Galloway, J. N., Dentener, F. J., Capone, D. G., Boyer, E. W., Howarth, R. W., Seitzinger, S. P., Asner, G. P., Cleveland, C. C., Green, P. A., Holland, E. A., Karl, D. M., Michaels, A. F., Porter, J. H., Townsend, A. R. and Vörösmarty, C. J.: Nitrogen Cycles: Past, Present, and Future, Biogeochemistry, 70(2), 153–226, doi:10.1007/s10533-004-0370-0, 2004.

Gassman, P. W., Reyes, M. R., Green, C. H., and Arnold, J. G.: The soil and water assessment tool: historical development, applications, and future research directions, T. ASAE, 2007.

Georgakakos, K. P., Seo, D., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298(1–4), 222–241, doi:10.1016/j.jhydrol.2004.03.037, 2004.

Grizzetti, B., Bouraoui, F., Granlund, K., Rekolainen, S., and Bidoglio, G.: Modelling diffuse emission and retention of nutrients in the Vantaanjoki watershed (Finland) using the SWAT model, Ecological Modelling, 169(1), 25–38, doi:10.1016/S0304-3800(03)00198-4, 2003.

Hargreaves, G. H. and Samani, Z. A.: Reference crop evapotranspiration from temperature, Appl. Eng. Agric., 1(2), 96–99, 1985.

Karssenberg, D., Burrough, P. A., Sluiter, R., and De Jong, K.: The PCRaster software and course materials for teaching numerical modelling in the environmental sciences, T. GIS, 5(2), 99–110, 2001.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, doi:10.5194/adgeo-5-89-2005, 2005.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel Ensemble Forecasts for Weather and Seasonal Climate, J. Climate, 13(23), 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2, 2000.

Kronvang, B., Behrendt, H., Andersen, H. E., Arheimer, B., Barr, A., Borgvang, S. A., Bouraoui, F., Granlund, K., Grizzetti, B., Groenendijk, P., Schwaiger, E., Hejzlar, J., Hoffmann, L., Johnsson, H., Panagopoulos, Y., Lo Porto, A., Reisser, H., Schoumans, O., Anthony, S., Silgram, M., Venohr, M., and Larsen, S. E.: Ensemble modelling of nutrient loads and nutrient load partitioning in 17 European catchments, J. Environ. Monitor., 11(3), 572–583, doi:10.1039/b900101h, 2009a.

Kronvang, B., Borgvang, S. A., and Barkved, L. J.: Towards European harmonised procedures for quantification of nutrient losses from diffuse sources – the EUROHARP project, J. Environ. Monitor., 11(3), 503–505, doi:10.1039/b902869m, 2009b.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35(1), 233–241, doi:10.1029/1998WR900018, 1999.

Lindgren, G., Wrede, S., Seibert, J., and Wallin, M.: Nitrogen source apportionment modeling and the effect of land-use class related runoff contributions, Nord. Hydrol., 38(4–5), 317, doi:10.2166/nh.2007.015, 2007.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201(1–4), 272–288, 1997.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, 14 pp., doi:200510.1029/2005WR004289, 2005.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 42(1), 55–61, 1979.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10(3), 282–290, 1970.

O'Callaghan, J. F. and Mark, D. M.: The extraction of drainage networks from digital elevation data, Comput. Vis. Graph., 28(3), 323–344, 1984.

Olivera, F., Valenzuela, M., Srinivasan, R., Choi, J., Cho, H., Koka, S., and Agrawal, A.: ARCGIS-SWAT: A geodata model and GIS interface for SWAT, J. Am. Water Resour. As., 42(2), 295–309, 2006.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133(5), 1155–1174, 2005.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D. J., and Dmip, P.: Overall distributed model intercomparison project results, J. Hydrol., 298(1–4), 27–60, 2004.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32(7), 2189–2202, 1996.

Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376(3–4), 463–475, 2009.

Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., Srinivasan, R., and Hauck, L. M.: Validation of the SWAT model on a large river basin with point and nonpoint sources, J. Am. Water Resources Ass., 37(5), 1169–1188, doi:10.1111/j.1752-1688.2001.tb03630.x, 2001.

Shamseldin, A. Y., O'Connor, K. M., and Liang, G.: Methods for combining the outputs of different rainfall-runoff models, J. Hydrol., 197(1–4), 203–229, doi:10.1016/S0022-1694(96)03259-3, 1997.

Sivapalan, M.: Prediction in ungauged basins: a grand challenge for theoretical hydrology, Hydrol. Process., 17(15), 3163–3170, 2003.

Sivapalan, M., Ruprecht, J. K., and Viney, N. R.: Water and salt balance modelling to predict the effects of land-use changes in forested catchments, 1. Small catchment water balance model, Hydrol. Process., 10(3), 393–411, 1996a.

Sivapalan, M., Viney, N. R., and Jeevaraj, C. G.: Water and salt balance modelling to predict the effects of land-use changes in forested catchments, 3. The large catchment model, Hydrol. Process., 10(3), 429–446, 1996b.

Sivapalan, M., Viney, N. R., and Ruprecht, J. K.: Water and salt balance modelling to predict the effects of land-use changes in forested catchments, 2. Coupled model of water and salt balances, Hydrol. Process., 10(3), 413–428, 1996c.

Smith, M. B., Seo, D. J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, J. Hydrol., 298(1–4), 4–26, 2004.

van Griensven, A., Francos, A., and Bauwens, W.: Sensitivity analysis and auto-calibration of an integral dynamic model for river water quality, Water Sci. Technol., 45(9), 325–332, 2002.

Viney, N. R. and Sivapalan, M.: A conceptual model of sediment transport: application to the Avon River Basin in Western Australia, Hydrol. Process., 13(5), 727–743, 1999.

Viney, N. R., Sivapalan, M., and Deeley, D.: A conceptual model of nutrient mobilisation and transport applicable at large catchment scales, J. Hydrol., 240(1–2), 23–44, doi:10.1016/S0022-1694(00)00320-6, 2000.

Viney, N. R. and Sivapalan, M.: Modelling catchment processes in the Swan-Avon river basin, Hydrol. Process., 15(13), 2671–2685, doi:10.1002/hyp.301, 2001.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32(2), 147–158, 2009.

Vitousek, P. M., Aber, J. D., Howarth, R. W., Likens, G. E., Matson, P. A., Schindler, D. W., Schlesinger, W. H., and Tilman, D. G.: Human alteration of the global nitrogen cycle: sources and consequences, Ecol. Appl., 7(3), 737–750, 1997.

Vrugt, J. A., Clark, M. P., Diks, C. G., Duan, Q., and Robinson, B. A.: Multi-objective calibration of forecast ensembles using Bayesian model averaging, Geophys. Res. Lett., 33(19), L19817, doi:10.1029/2006GL027126, 2006.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resour. Res., 43(1), W01411, doi:10.1029/2005WR004838, 2007.

Wade, A. J., Durand, P., Beaujouan, V., Wessel, W. W., Raat, K. J., Whitehead, P. G., Butterfield, D., Rankinen, K., and Lepisto, A.: A nitrogen model for European catchments: INCA, new model structure and equations, Hydrol. Earth Syst. Sci., 6, 559–582, doi:10.5194/hess-6-559-2002, 2002.

Whitehead, P. G., Wilson, E. J., and Butterfield, D.: A semi-distributed integrated nitrogen model for multiple source assessment in tchments (INCA): Part I–model structure and process equations, Sci. Total Environ., 210, 547–558, 1998.

Williams, J. R.: Flood routing with variable travel time or variable storage coefficients, T. ASAE, 12(1), 100–103, 1969.

Wissmeier, L. and Uhlenbrook, S.: Distributed, high-resolution modelling of 18O signals in a meso-scale catchment, J. Hydrol., 332(3–4), 497–510, 2007.

Zammit, C., Sivapalan, M., Kelsey, P., and Viney, N. R.: Modelling the effects of land-use modifications to control nutrient loads from an agricultural catchment in Western Australia, Ecol. Model., 187(1), 60–70, doi:10.1016/j.ecolmodel.2005.01.024, 2005.