Performance assessment of nitrate leaching models for highly vulnerable soils used in low input farming based on lysimeter data

Piet Groenendijk[a,*], Marius Heinen[a], Gernot Klammler[b], Johann Fank[b], Hans Kupfersberger[b], Vassilios Pisinaras[c], Alexandra Gemitzi[c], Salvador Peña-Haro[d], Alberto García-Prats[e], Manuel Pulido-Velazquez[f], Alessia Perego[g], Marco Acutis[g], Marco Trevisan[h]

a Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands

b Joanneum Research, Forschungsgesellschaft mbH, Leonhardstraße 59, 8010 Graz, Austria

c Democritus University of Thrace, Department of Environmental Engineering, Vas. Sofias 12, Xanthi, 67100, Greece

d Institute of Environmental Engineering, ETH Zurich, Wolfgang-Pauli-Str. 15, CH-8093 Zurich, Switzerland

e Universitat Politècnica de València, Department of Hydraulic Engineering and Environment, Camino de Vera, 46022 Valencia, Valencia, Spain

f Universitat Politècnica de València, Research Institute of Water and Environmental Engineering (IIAMA), Camino de Vera, 46022 Valencia, Valencia, Spain

g University of Milan, Department of Agricultural and Environmental Science, Via G. Celoria 2 20133, Milan, Italy

h Università Cattolica del Sacro Cuore, sede di Piacenza, Via Emilia Parmense, 84 29100, Piacenza, Italy

*Corresponding author: Piet Groenendijk, Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands, Email: piet.groenendijk@wur.nl Tel.: +31 317 486434

Abstract

The agricultural sector faces the challenge of ensuring food security without an excessive burden on the environment. Simulation models provide excellent instruments for researchers to gain more insight into relevant processes and best agricultural practices and provide tools for planners for decision making support. The extent to which models are capable of reliable extrapolation and prediction is important for exploring new farming systems or assessing the impacts of future land and climate changes.

A performance assessment was conducted by testing six detailed state-of-the-art models for simulation of nitrate leaching (**ARMOSA, COUPMODEL, DAISY, EPIC, SIMWASER/STOTRASIM, SWAP/ANIMO**) for lysimeter data of the Wagna experimental field station in Eastern Austria, where the soil is highly vulnerable to nitrate leaching.

Three consecutive phases were distinguished to gain insight in the predictive power of the models: 1) a blind test for 2005 – 2008 in which only soil hydraulic characteristics, meteorological data and information about the agricultural management were accessible; 2) a calibration for the same period in which essential information on field observations was additionally available to the modellers; and 3) a validation for 2009 – 2011 with the corresponding type of data available as for the blind test. A set of statistical metrics (mean absolute error, root mean squared error, index of agreement, model efficiency, root relative squared error, Pearson's linear correlation coefficient) was applied for testing the results and comparing the models.

None of the models performed good for all of the statistical metrics. Models designed for nitrate leaching in high input farming systems had difficulties in accurate predicting leaching in low input farming systems that are strongly influenced by the retention of nitrogen in catch crops and nitrogen fixation by legumes. An accurate calibration does not guarantee a good predictive power of the model. Nevertheless all models were able to identify years and crops with high and low leaching rates.

Keywords

Lysimeter, model comparison, nitrate leaching, performance assessment, predictive power, simulation model

1. Introduction

Agriculture is the major land use in Europe (ca. 50% of overall land area) and has strongly increased its use of external inputs (fertiliser, pesticides and water) over the last 50 years. The environmental effects of intensive agriculture include a decline in biodiversity, eutrophication of ecosystems and surface waters, acidification, global warming, air pollution and diffuse nitrate pollution of groundwater. A global challenge is to produce enough food for the ever-growing population and at the same time minimizing the loss of reactive nitrogen (N) to the environment. Since the 1980s, agriculture in Western Europe has managed to reduce its N surpluses, owing to stringent national and European community policies (Vitousek et al., 2009; Grizzetti et al., 2011).

The main aim of the Nitrates Directive (EU, 1991: Directive 91/676/EEC) is to reduce water pollution caused or induced by nitrates and phosphorus from agricultural sources. The Nitrates Directive legally restricts farm application of manure to 170 kg ha^{-1} of nitrogen, or in case of derogation to inputs up to 250 kg ha^{-1} (Oenema, 2004). An implementation measure of the Nitrates Directive is the establishment of codes of Good Agricultural Practice. Recommended measures include, among others, the application of crop rotations, the cultivation of a soil winter cover and catch crops to prevent nitrate leaching and run-off during wet seasons. Catch crops create a new challenge in the assessment of environmental effects of crop rotations. In theory, catch crops take up N that would otherwise be lost, and, after incorporation of the crop residues into the soil, make this N available to the succeeding crop via mineralization. However, the influence of a catch crop on the nitrogen supply to the succeeding crop can vary greatly and range from a positive to a negative effect (Nett et al., 2011). The effect is determined by the N uptake capacity, the rooting depth of a catch crop, the weather and soil conditions as well as the rooting depth of the succeeding crop (Thorup-Kristensen, 2006).

Models are an important tool for assessment of environmental impacts of a certain agricultural practice and are also an instrument for increasing the understanding of the biological, pedological and hydrological factors that affect productivity and the risk of nitrate leaching. For this reason, for more than 30 years simulation models have been developed and applied in the research on nitrate leaching. The different model descriptions are a reflection of the intended purpose, the physical conditions and the available data for model application and the knowledge and skill of the model developer. Technical implementations have evolved from stand-alone model codes to modelling platforms comprising modular models able to include and compare different process descriptions.

Calibration and validation of models contributes to their reliability. In addition also an analysis of the implemented process descriptions and the mutual comparison of models provides information on the predictive power. Several model comparison studies have been conducted in which nitrate leaching models were compared (De Willigen and Neeteson, 1985; Vereecken et al., 1991; De Willigen, 1991; Diekkrüger et al., 1995; Moreels et al., 2003; Kersebaum et al., 2007; Jabro et al., 2012). Most of them were related to ordinary agricultural conditions with a single crop on a typical agricultural soil. Thus, there is no information (comparison) available for situations in soils that are highly vulnerable to nitrate leaching in combination with low-input conditions and the use of catch crops.

It is widely recognised that despite the deterministic nature of process oriented models they often have a limited validity range for certain climatic, pedological, hydrological and agronomic circumstances characterised by high inputs. It is not clear whether the models are able to produce relatively reliable predictions for low input conditions. A better insight into the model performance for such uncommon circumstances underpins conclusions about the predictive power.

In this study a number of models were inter-compared for low input conditions of one of the lysimeters of the Wagna experimental research station, Austria (Klammler and Fank, 2014; this issue) for three typical conditions for which they were not designed: 1) the crop rotation which included an uncommon crop (oil pumpkin), 2) catch crops for which the N-uptake was not measured, and 3) the soil consisted of a shallow soil vulnerable to nitrate leaching on top of a high conductive gravel layer. The objectives of this study were: 1) to assess the performance of state-of-the-art nitrate leaching models as they are used in the scientific research community, for the above mentioned conditions, 2) to inter-compare the models for analysing their predictive power, and 3) to identify strengths and weaknesses of bio-physically based models.

2. Materials and Methods

2.1 Description of the lysimeter

Observations were used of a lysimeter located in the agricultural experimental field station in Wagna in Eastern Austria (46° 46.113'N, 15° 33.140'E; altitude 265 m; Klammler and Fank, 2014 (this issue)). Since 1987 different cultivation strategies are investigated concerning nitrogen-fertilizer input, nitrate leaching and crop yields. In 2004, the cultivation changed into comparing low-input farming and organic farming, each covering 50% of the test site. Since then, two of the test plots have been equipped with two weighable, monolithic, high-precision lysimeters (2 m depth, 1 m^2 surface). The lysimeter in the conventional tillage test plot (KON-system) is subject for this study. Cultivation practices including crop species, sowing and harvest dates, and fertilizer applications in the test plot are presented in Table 1.

<<Table 1 >>

The lysimeters are equipped with soil water samplers, soil moisture probes, matrix sensors/tensiometer and soil temperature probes at four measuring depths (0.35, 0.6, 0.9, 1.8 m). An accompanied measuring profile for soil moisture, matrix potential and soil temperature is also installed outside the lysimeters (same depths as inside the lysimeter) to determine if the conditions inside the lysimeter are representative for the rest of the field. At the bottom of the lysimeter (depth 1.8 m) a suction cups rake was installed which kept the pressure head at this depth equal to that outside the lysimeter. The water sucked off was collected, weighted and sampled for the determination of the nitrate concentration. While quantity of seepage water was recorded automatically in 0.1 mm resolution by a tipping bucket, nitrogen concentration in the accumulated leachate was analysed in an approximately weekly interval. Furthermore, a weather station is installed at agricultural test site in Wagna for the recording of air temperature, relative humidity, shortwave solar radiation, wind speed, wind direction, precipitation, sunshine duration and atmospheric pressure at high temporal resolution (Klammler and Fank, 2014; this issue). Annual precipitation rates and cumulative probabilities of the rates relative to the values of the period 1961 – 2011 are presented in Table 2.

<<Table 2>>

Annual rainfall amounts during the calibration years can be considered as moderate, the first year of the validation period is characterised by an extreme high rainfall and during the last year of the validation a low precipitation amount was recorded.

2.2 Description of models

This performance assessment study was conducted as part of the EU-FP7 GENESIS project (2009 – 2014) by six partners. Six well-known detailed models for European research on field-scale crop and soil water and soil nitrogen dynamics were chosen: **ARMOSA**, **CoupModel (COUP)**, **DAISY**, **EPIC**, **SIMWASER-STOTRASIM** and **SWAP-ANIMO**. It goes beyond the scope of this paper to give full details on the process descriptions of the six models used. Brief descriptions will be given in text and inter-comparison of processes and various other characteristics can be found in Supplemental Materials. All models are one-dimensional.

- **ARMOSA** has recently been developed specifically for the Lombardy region in Italy to assess the regional soil vulnerability to nitrate leaching (Perego et al., 2013). The model allows the simulation at field and multi-field level. The model is based on the **SWAP** (version 2.07) approach for simulating the water flow (Van Dam, 2000), on **STAMINA** for simulating the crop development and growth (Ferrara et al., 2011; Richter et al., 2010) and on **SOILN** for simulation of the soil organic matter and nitrogen cycle and nitrate leaching (Bergström et al., 1991).

- **CoupModel (COUP),** a coupled heat and mass transfer model for soil plant-atmosphere systems, was initially developed to simulate conditions in forest soils, but it has been further developed to simulate conditions in any type of soil, independent of plant cover (Jansson and Karlberg, 2004). COUP applicability is very wide as it includes water, heat, tracer, chloride, nitrogen and carbon modules that can be incorporated in the modelling process. COUP development, calibration procedures and applications are presented by Jansson (2012).
- **DAISY** is a soil-plant-atmosphere system model designed to simulate crop production, soil water dynamics, and nitrogen dynamics in crop production at various agricultural management practices and strategies (Hansen et al., 1990). The agricultural management model allows for building complex management scenarios (Hansen, 2002). The model has been validated in a number of major comparative tests (Diekkrüger et al., 1995; Hansen et al., 1991a,b; Jensen et al., 1997; Smith et al., 1997; Svendsen et al., 1995; Vereecken et al., 1991; De Willigen, 1991).
- **EPIC** (Williams et al., 1984; 1989) is a cropping systems simulation model, which was developed to estimate soil productivity as affected by erosion throughout the United States during the 1980's. **EPIC** is a field scale model, but linked to a GIS it has been applied in several regional model applications (Burkart et al., 1999; Sohier et al., 2009). Furthermore the **EPIC** model has been applied to study the effect of agricultural practices and biofuels cultivation on N leaching at the European scale (Bouraoui and Aloe, 2007; Van der Velde et al., 2009).
- **SIMWASER** (Stenitzer, 1988) simulates the water flow in soil. A unique feature of the model is the description of actual rooting depths based on both root biomass simulated for a crop and on the penetration resistance of the soil. **STOTRASIM** (Feichtinger, 1998) is fully coupled to **SIMWASER** and simulates nitrogen and basic carbon dynamics of agriculturally used soils. The model has already been applied to the region of southeast Styria (Fank et al., 2006). The name of these coupled models is abbreviated as **SIM-STO**.
- The **SWAP** model, version 3.2 (Van Dam et al., 2008) simulates water flow in the soil – plant – atmosphere domain in an integrated manner. The **ANIMO** model (Groenendijk et al., 2005) is sequentially coupled to **SWAP** and was designed to quantify the relation between fertiliser application rate, soil management and the leaching of nitrogen (N) and phosphorus (P) to groundwater and surface water systems. The **ANIMO** model

is part of the National Dutch modelling system **STONE** for the evaluation of fertiliser policy measures (Wolf et al., 2003). The name of the sequentially coupled models is abbreviated as **SW-ANIM**.

In addition to soil processes also the description of crop development is considered, because the plant related processes such as evaporation, nitrogen and nitrogen supply with crop residues exert a major influence on the water balance and nutrient dynamics in the soil.

Except for **SW-ANIM**, all models simulate the growth of plant biomass. Although **SW-ANIM** has the possibility to calculate the biomass development in a detailed manner, the modellers had chosen to use a simple option of a supposed development of leaf area index, crop height and rooting depth, because the parameters required for detailed simulation of oil pumpkin and catch crops were not available. Except for **EPIC**, the models describe water flow with either the Richards' (1931) equation or the Darcy (1856) - Buckingham (1907) equation, in which the soil water retention and the hydraulic conductivity relations are described according to Mualem (1976) - Van Genuchten (1980). **EPIC** simulates soil water flow as a storage routing process in which percolation occurs when the soil water content of the root zone exceeds the field capacity. In **EPIC** the soil water characteristics are calculated on the basis of texture data and the organic matter content in accordance with Saxton and Rawls (2006).

All models consider ammonium and nitrate as separate mineral nitrogen pools, and simulate organic bounded nitrogen associated with the organic carbon cycle. **SW-ANIM** simulates also the transport and transformation of dissolved organic nitrogen. The method of simulating biological N-fixation is one of the striking differences between the models. The **DAISY** model was applied in a way that biological N-fixation was ignored and the **SW-ANIM** model accounted for this process by the specification of continuous organic material additions representing imposed fixation rates. The other models use relationships based on the crop type, the crop development stage and the soil mineral N status. Ammonia volatilization is not implemented in the **COUP** model code used for this study. Some models consider only the loss of ammonia as a fraction of farmyard manure application (**DAISY**, **SW-ANIM**) while the other models take account for environmental factors as temperature, wind speed and soil moisture. **SIM-STO** uses standardized loss factors that account for the time from the last soil tillage event.

Uptake of ammonium and nitrate depends on the demand for mineral N for crop production and is related to the development stage, by some models expressed by a relationship with the water uptake, and the mineral N content of the soil.

Mineralisation is simulated in close correspondence to the organic matter cycle. All models describe the amount of mineralized nitrogen as the excess nitrogen produced from the organic matter decay and transformations to more stable soil organic matter pools. Nitrification is commonly described as a first order process which rate depends on temperature, soil moisture status and ammonium concentration. Denitrification plays no significant role in the soil of the Wagna lysimeters (Leis, 2009), but can be simulated by the models used. A variety of descriptions are implemented but all assume a relationship with temperature, soil moisture content, nitrate concentration and the potential denitrification rate as a function of organic matter content (Heinen, 2006).

The lysimeter was installed in 2004 and it was ensured that the original soil layers was put back. During the excavation and filling the soil had been in contact with open air. None of the models paid attention to this event in 2004. To establish the starting conditions on 1-1-2005, three of the six models (i.e., **ARMOSA EPIC**, **STO-SIM**, **SW-ANIM**) started in 1987. **COUP** was run for five years prior to the start in 2005 and **DAISY** was run two-years prior to the simulation**.**

2.3 Experimental design of study

The modelling study comprised of: 1) a blind test with non-calibrated models to get an impression of the performance of the models as they are used in situations where extensive data sets are missing, which often occurs in practice, 2) a calibration period, and 3) a validation period. Inter-comparisons were done between measured and simulated leaching of water and nitrate, including nitrate concentration of the percolate. The outcome of the simulations by all models was collected and analysed by a single person.

2.3.1 Step 1: Blind test

The models first performed a simulation based on a minimum set of data: crop rotation, soil cultivation, fertilization rates, meteorological data, soil profile description and soil moisture retention laboratory measurements of some soil samples. The aim is to establish the bandwidth of differences with the observations without an assessment of the individual models. The **SIM-STO** model was excluded from the blind test as the operators of this model were the owners of all data and **SIM-STO** was already partly calibrated for the test site. After all models delivered their outcome, one external operator compared the predictions against the measured data (seasonal cumulated water flux and nitrogen flux at the bottom of the lysimeter, seasonal flow averaged nitrate concentration) for the period 2005 - 2008. It was not the intention of the blind test to qualify or assess the performance of the individual models and, therefore, the outcome of this test will be presented anonymously.

Specifically only data on seasonal percolation, flow-averaged nitrate concentration and seasonal nitrate leaching were considered.

2.3.2 Step 2: Calibration

Each of the six modelling groups calibrated the models for a limited number of parameters. The successive operations, the objective function and the number of parameters were not prescribed, but were chosen freely by the modelling groups, either based on expert judgement or on a sensitivity analysis. Further details of how the calibration has been carried out for the different models can be found in Supplemental Materials.

2.3.3 Step 3: Validation

The validation was performed for the period 2009 - 2011, where only information about crop rotation, application of fertilizers, soil cultivation and meteorology was made available for the modelling groups after step 2 (calibration) was finished. The procedure for the validation is thus similar to that of the blind test, with the difference that the models were calibrated prior to validation and that the **SIM-STO** model was included in the validation.

2.3.4 Step 4: Model comparison

The six models were compared for their performance with respect to 1) the soil moisture retention curves at depths 0.35, 0.90 and 1.8 m; 2) the volumetric water contents at depths 0.35, 0.9 and 1.8 m; 3) the nitrate concentrations at depths 0.35, 0.9 and 1.8 m ; 4) the daily water fluxes at depth 1.8 m; 5) the leached water amounts for the time intervals of collected water samples; 6) the nitrate concentrations of the collected water samples; 7) the nitrate-N fluxes at the bottom of the lysimeter for the time intervals of collected water samples. The comparison of results at the depth of 60 cm was excluded because measurements for this depth were only available up to Sept. 2009. Seasonal leached water amounts, nitrogen yields and nitrate-N fluxes were compared to discuss the predictive power for practice oriented model applications. A nitrogen balance was set up for all models. Water fluxes at 1.8 m depth were evaluated for daily and for seasonal values. Nitrate leaching fluxes and nitrate concentrations in the leachate were evaluated at the time intervals for which the soil water was sampled. The sampling time intervals were irregular in time and the models were not able to present concentrations at these specific time events. Therefore, concentrations values for these time intervals were derived according to a volumetric averaging procedure. The nitrate concentrations at depths 0.35 m and 0.9 m can be used to get an impression whether the transport and transformation processes in soil, which ultimately lead to the leaching at depth 1.8 m, have been described adequately. Due to the nature of the model formulations, **EPIC** was not able to present the concentrations at the depths of measurement. The number of observations at depth 0.35 m in the calibration period was too little and were not considered.

In the models, much knowledge of soil processes is described which all contribute to the nitrate leaching at depth 1.8 m. To understand the similarities and differences between simulation results and measurements, it is important to assess the processes. We have done this through the establishment of nitrogen balances per season.

2.4 Statistical metrics

The behaviour of the main model outputs can be characterized by a number of statistical metrics to indicate the models' ability to capture different aspects. A complete assessment of model performance should include at least one absolute error measure and one goodness-of-fit measure (Legates and McCabe, 1999). There are a wide range of statistical indicators used in studies on soil water and soil nitrogen, but not always a justification is given for the indicators chosen. For *state variables* many authors use mean (absolute) error (*M*(*A*)*E*), root mean square error (*RMSE*), index of agreement (*IoA*; Willmott, 1982), and less often the Nash-Sutcliffe modelling efficiency (*NSE*; Nash-Sutcliffe, 1970) (e.g., Donatelli et al., 2004; Gribb et al., 2009; Herbst et al., 2005; Khodaverdiloo et al., 2011; Patil and Rajput, 2009; Ritter et al., 2003; Vereecken et al., 2010). For *rate variables* authors generally use *MAE*, mean difference (*MD*), absolute maximum error (*AME*), *RMSE*, *IoA*, *NSE*, coefficient of determination (R^2), percentage of error (*PE*), percentage of bias (P_{bias}) (e.g., Akkal-Corfini et al., 2010; Ale et al., 2012; Dawson et al., 2007, 2010; Jabro et al., 2012; Jachner et al., 2007; Kersebaum et al., 2007; Krause et al., 2005; Moriasi et al., 2007; Qi et al., 2012; Reusser et al., 2009; Stumpp et al., 2009; Van der Laan et al., 2011; Wang et al., 2006; Willmott et al., 1985). It appears that a few measures are used both for state as for rate variables, which we have chosen to use here as well: *MAE*, *RMSE*, *IoA*, and *NSE* (only for rates), given by:

1. Mean absolute error: $MAE = \frac{1}{n}\sum_{t=1}^{n}|P_t - O_t|$
2. Root mean squared error: $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(P_t - O_t)^2}$
3. Index of Agreement (Willmott, 1982): $IoA = 1 - \frac{\sum_{t=1}^{n}(P_t - O_t)^2}{\sum_{t=1}^{n}(|P_t - O_t| + |O_t - \bar{O}|)^2}$
4. Nash-Sutcliffe model efficiency (Nash and Sutcliffe, 1970): $NSE = 1 - \frac{\sum_{t=1}^{n}(O_t - P_t)^2}{\sum_{t=1}^{n}(O_t - \bar{O})^2}$

where n is the number of observations, O_t is the observed value, P_t is the model predicted value, and $\overline{O}$ and $\overline{P}$ are the mean values of observations and predictions, respectively. All four measures compare the predictions P_t and observations O_t at the individual level, and try to express the 'spread' in (P_t – O_t) (Janssen and Heuberger, 1995). The *MAE* accounts for the deviations (P_t – O_t) in an absolute value sense. This measure is less sensitive to outliers than *RMSE*, *IoA* and *NSE*. The latter indices measures (P_t – O_t) in a quadratic sense, and, thus, are sensitive to outliers. If model errors are significant, it is more difficult to objectively assess the agreement between model and data on basis of *RMSE*. As an alternative, Willmott (1982) proposed *IoA* to express this agreement more directly. The dimensionless *IoA* has limits 0, indicating no agreement, and 1, indicating perfect agreement. The dimensionless *NSE* ranges between 1 and −∞, where *NSE* = 1 denotes a “perfect” model fit and for *NSE* < 0 the average of the observations would be a better predictor than the model (Krause et al., 2005).

Taylor (2001) presented a graphical method in which several statistical metrics have been combined. Such a Taylor diagram summarizes how closely a set of simulations matches the observations, and it is especially useful in evaluating multiple aspects of complex models. In normalized form, it presents the Pearson’s linear correlation coefficient (*R*) and the root relative square error (*RRSE*) as a function of the ratio of standard deviations of predictions and observations σ_P and σ_O, respectively, where

5. Pearson’s linear correlation coefficient $$R = \frac{\sum_{t=1}^{n}(O_t-\overline{O})(P_t-\overline{P})}{\sqrt{\sum_{t=1}^{n}(O_t-\overline{O})^2}\sqrt{\sum_{t=1}^{n}(P_t-\overline{P})^2}}$$

6. Root relative square error: $$RRSE = \frac{\sqrt{\sigma_P^2+\sigma_O^2+2\sigma_O\sigma_P R}}{\sigma_O}$$

where σ_O and σ_P are the standard deviations of the observations and model predictions, respectively. A value of (1,0) in such a figure indicates a full agreement of model results with observations.

3. Results and discussion

3.1. Blind test

Figure 1 presents the range of predicted seasonal water fluxes, flow-averaged nitrate concentration and nitrate-N fluxes by the five models considered as compared to the observations for the blind test period.

<<Figure 1>>

Maximum deviations between simulated and observed seasonal percolation volumes of almost 400 mm were found. Two of the five models showed a relatively good agreement of the seasonal percolation with the measurements. Three of the five models overestimated the percolation in all seasons. One model underestimated the percolation volume in all seasons and only one model was able to simulate the seasonal percolation accurately. The range of model results was independent of the seasonal percolation.

Seasonal flow averaged nitrate concentrations were underestimated by all models in two of the four seasons. For the first season, all models underestimated the concentration by 10 – 40 mg L^{-1}. The variation of simulated concentrations and N-fluxes was large. Maximum deviations of seasonal nitrate-N leaching of about 25 kg ha^{-1} were found. All models underestimated the leaching rate in 2005 by 8 – 22 kg ha^{-1}. The same holds for the fourth season, but only one model was able to calculate the nitrate-N flux with a reasonable agreement with the measurements. In the second season (maize), four models underestimated and one model overestimated the nitrate concentration and nitrate-N flux. The third season, which was the second season with maize showed a rather different pattern. The measured nitrate concentration and nitrate-N flux under maize in the 3rd season was much lower than for the maize crop in the 2nd season, but the modelled results still showed a large variation with a less skewed distribution of underestimation and overestimation. In the blind test information was lacking about crop-uptake rates and the nitrogen excess per season. The results showed that without this information and without a proper calibration the models were not able to predict nitrate concentrations and leaching rates accurately.

3.2 Calibration and validation

3.2.1 Soil water and soil physical relations

In the blind test the modellers had only laboratory measurements of the water retention curve at their disposal, but in the calibration phase also in situ measured soil moisture contents (θ) and pressure heads (h) were available at four depths. The laboratory measurements were performed for drying samples only, while under field conditions data pairs of $\theta(h)$ were detected during wetting and drying cycles so that these were affected by hysteresis (Basile et al., 2003, 2006). Figure 2 depicts the calibrated $\theta(h)$ curves for three depths. The results at the depth of 0.6 m were comparable to the results of 0.35 m deep and are not shown here. The observed h at depth 0.35 m ranged from -20 cm to -2000 cm. At depth 0.9 m h ranged from -2 cm to -1000 cm and at depth 1.8 m h ranged from -10 to -100 cm. The variation of the $\theta(h)$ observed population is largest at depth 0.35 m.

<<Figure 2>>

Results for the **EPIC** model are represented by three points as EPIC does not use a continuous description of the $\theta(h)$ curve. The greatest value for the saturated water content was obtained by the **EPIC** model with a value greater than 0.3 cm^3 cm^{-3} at depth 1.8 m. This parameter is far outside the range that was established by the other models. A comparison between the calibrated and observed θ(*h*) curves was made by calculating a θ for each value of the measured *h*. The performing indices based on computed θ and measured θ are presented in Table 3.

<<Table 3 >>

In general the resulting *MAE*, *RMSE* and *IoA* showed equal trends. The **ARMOSA** model fitted well at depths 0.35 m and 0.9 m, but performed worse at depth 1.8 m. The performance of the **COUP** model appeared to be weak. At depth 0.9 m the **DAISY** model was better than the **COUP** model, but worse than the other models. The *IoA* for the **SIM-STO** and **SW-ANIM** models was highest at depth 0.9 m and somewhat lower for the other depths. It should be noted that a good match of the calibrated θ(*h*) curves with measured data pairs does not a-priori mean that a good agreement between the time series of measured and calculated θ will be obtained.

The simulated θ was compared with daily averaged values of measured θ (Table 4). For depth 0.35 m an increasing trend was detected from 2008 and onwards which is attributed to the aging of the sensor, and, therefore, the results for this depth were disqualified for the validation period.

<<Table 4>>

Except for **ARMOSA** and **EPIC** in the validation phase, the highest *IoA* values for simulation of the water contents were achieved at depth 0.9 m. For **SIM-STO** and **SW-ANIM**, the *IoA* values were similar to the calibration results of the θ(*h*) curves (Table 3). However, the performance by **COUP** increased and that by **DAISY** decreased compared to Table 3. Except for the **ARMOSA** and the **DAISY** models at depth 0.35 m and the **SW-ANIM** model at depth 1.8 m, in general the resulting performance indices showed a better agreement between simulated and observed values for the period 2005 – 2008 than for the comparison based on soil moisture retention curves. The indices of the validation period 2009 – 2011 were in the same range, or somewhat lower at depth 0.9 m, as for the calibration period (Table 4).

Figure 3 presents the cumulative water fluxes as predicted by the models and as measured as a function of time.

<<Figure 3>>

The pattern of cumulative water fluxes per growing season complies generally with the annual precipitation amounts (Table 2) with the exception of maize in 2006 and its preceding crop in the winter of 2005/2006. During the intermediate period after oil pumpkin in 2005 and before maize in 2006, the precipitation amounted to about

430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative water flux per season corresponded well to the measured water fluxes for most of the models which is also true for the extreme wet year 2009. However, DAISY showed some overestimation in particular seasons due to difficulties in parameterizing maize crop parameters. EPIC performed better in the calibration than in the validation period. **SW-ANIM** underestimated the cumulative water flux in the two first seasons, but overestimated slightly in some other seasons. No model was able to simulate the dry no-flux period during the second half of 2011. Deviations between the simulated and observed soil moisture contents were relatively small and have a limited impact on the cumulative water fluxes. Underestimations and overestimations of the seasonal water fluxes are explained by overestimation and underestimations of the seasonal evapotranspiration. This depends on the difficulty of establishing accurate crop growth parameters. Table 5 presents the statistical performance indices for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

<<Table 5>>

The performance improved for the averaged fluxes per sampling period of the calibration phase relative to the performance of the daily fluxes, but deteriorated for the validation phase. This is counter-intuitive because the peaks of the daily fluxes pattern are flattened by aggregation and one should expect a better performance for the averaged values per sampling interval.

Figure 4 presents the Taylor diagrams for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

<<Figure 4>>

For all models the *R*-values were between 0.5 and 0.9 and the *RRSE*-values were between 0.5 and 1.0. For daily water fluxes the σ_P/σ_O-ratio for the validation period was somewhat higher than for the calibration period, but for the fluxes averaged for the sampling intervals it can be seen that **ARMOSA**, **DAISY**, **COUP** and **EPIC** resulted in lower σ_P/σ_O-ratio's for the validation period than for the calibration period.

The range of seasonal water fluxes for the cultivation periods predicted by the models for all seasons was around the observed values (Figure 5). With respect to the blind test, calibration of the models resulted in a smaller range and in a shift towards the observations.

<<Figure 5>>

The ranges were relatively large for maize and its preceding catch crops in 2006 and 2010. In four of the seven seasons **DAISY** had the lowest value for the leaching and in one season the highest value. Both **COUP** and **EPIC** resulted in three seasons the highest value. **SIM-STO** had the smallest deviation between predicted and measured seasonal water leaching and **DAISY** resulted in the largest deviation

Differences between observed and model predicted water contents, water fluxes and water volumes per sampling interval indicate over- or under-estimation of the water excess in the soil column. Besides uncertainties in soil hydraulic properties and in observations, there was also lack of information about actual plant and root system development as a function of time.

The different modelling groups were not able to find a simultaneous optimal solution which minimizes both water contents deviations and water flux deviations. This may be due to uncertainties in soil hydraulic properties, and the disregarding of hysteresis in the models. The soil at the Wagna experimental station consists of a clayey-sand on top of a gravel layer. Durner et al. (2007) concluded that for layered soils with distinct heterogeneity no unique effective soil hydraulic properties exist. If only fluxes across the boundaries of the system are required, heterogeneous systems can be modelled with quasi-homogeneous ones, even if the internal system state is not matched properly. However, for nutrient dynamics (solute dispersion, biological and chemical reactions) an accurate internal system state description is mandatory (Durner et al., 2007)

3.2.2 Soil temperature

The soil temperature is an important variable determining the rate of biological processes (N dynamics), for the crop development in the period of germination, and for soil moisture flow under winter conditions. A comparison of simulated and measured soil temperatures was carried out as well (data not shown). In general, the models were well able to simulate soil temperatures and resulted in performance indices much higher than for moisture contents. The simulation performance at shallow depth was less than the performance at greater depths: most models showed a delayed warming up in some spring seasons with respect to the measurements, which is attributed to the incomplete description of surface temperatures, for most of the models used the air temperature as the boundary condition.

3.2.3 Nitrate concentrations and nitrate-N fluxes

Figure 3 presents the cumulative nitrate fluxes and the nitrate concentration of the leachate as predicted by the models and as measured as a function of time. Based on a visual inspection the nitrate concentrations are simulated well by **COUP** and **SW-ANIM** for the calibration period. The **SIM-STO** results for this period were poor and the results of the other models were in between. The results for the validation period showed a completely different picture when compared to the corresponding results for the calibration period. The results of **DAISY** and **SIM-STO** were relatively the best, while **EPIC** and **SW-ANIM** results were weak. **ARMOSA**, **COUP** and **SW-ANIM** overestimated the concentration peak in autumn 2009 and **SW-ANIM** simulated a peak for autumn 2010, while there was no peak visible in the measurements.

ARMOSA, **DAISY**, **EPIC** and **SIM-STO** showed more spiky results for the calibration period than the measured values, while **COUP** and **SW-ANIM** showed calmer and more evenly time courses. The results resembled partly the modeller's choice for defining either the nitrate fluxes or the nitrate concentrations in the objective function of the calibration procedure. The **COUP** and **SW-ANIM** modellers used the nitrate concentrations for calibrations, while the **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** modelling groups used the nitrate fluxes. For **DAISY** and **EPIC**, the nitrate concentrations were calculated afterwards by dividing the nitrate flux by the water flux. The nitrate concentrations in the calibration phase simulated by **SIM-STO** showed a bad performance, while the results for the validation phase were much better. The higher peak concentrations during the calibration phase were not approached by **SIM-STO**. On the other hand, **SW-ANIM** showed a good agreement of nitrate concentrations during the calibration phase, while there is a mismatch during the validation phase. The concentration peaks during the validation phase were severely overestimated by **SW-ANIM** due to an overestimation of the biological fixation rates of some non-leguminous catch crops in this period.

The nitrate-N flux at depth 1.8 m represents the nitrogen transport to deeper soil layers and is relevant for predictions of nitrate concentrations in deeper groundwater. **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** underestimated the nitrate N-flux under winter barley preceded by a catch crop in 2007-2008, but **SW-ANIM** overestimated the nitrate N-flux during this period. The **COUP** model was able to calculate the nitrate-N flux in five of the seven seasons that cover the calibration and validation period. **ARMOSA** and **DAISY** calculated the total seasonal nitrate-N flux well in three of the seven seasons, while **EPIC**, **SIM-STO** and **SW-ANIM** calculated this flux well in two of the seven seasons. The last season appeared to be the most difficult one, because of the exceptional dry conditions. The leaching after the 2009 oil pumpkin crop also showed significant deviations between model predictions and measurements. The largest deviations of seasonal nitrate-N fluxes occurred in the results of **COUP** and **SW-ANIM** for the exceptional wet year 2009.

Table 6 presents the statistical indicators for both the nitrate concentrations and the nitrate-N leaching rates, based on the sampling time series. The largest deviations between predicted and simulated nitrate concentrations were found for the **SIM-STO** results in the calibration period for which the *IoA* amounted to 0.43. Remarkably the smallest deviations were found for the same model for the validation period for which *IoA* amounted to 0.78. The underestimation of the nitrate-N flux by **SIM-STO** is most likely due to immobilization processes that are overemphasized for the 2005 and 2008 periods. Thus, less nitrate was released to the soil water phase which led to the underestimation of the nitrate concentration in the leachate. .

<<Table 6>>

The **COUP** model showed the best performance for the nitrate concentrations of the calibration period with *IoA* = 0.97 directly followed by the **SW-ANIM** model. The results from **EPIC** and **SW-ANIM** for concentrations in the validation period were weak with *RMSE* > 20 mg L^{-1}. The statistical indices of the nitrate-N leaching rates showed a similar picture. The **SIM-STO** model performed relatively weak during the calibration phase. For the leaching rates in this period **DAISY** and **SW-ANIM** had the best performance and for the validation period **ARMOSA** and **DAISY** performed relatively the best. The *NSE* values (data not shown) for both the concentration and the leaching rates in the validation period were almost all negative, showing that the calibrated models had great difficulties to predict concentrations and leaching rates for the more extreme conditions of the validation period.

Statistical performance of predicted nitrate concentrations and leaching rates were expressed in Taylor diagrams in Figure 6. Calibrated nitrate concentrations by **COUP** and **SW-ANIM** had *R*-values greater than 0.9 and were closest to the (1,0) point. Except for **SIM-STO**, the models showed σ_P/σ_O ratios for the calibration step that did not deviate much from 1; for **SIM-STO** the σ_P/σ_O ratio was much lower than 1 and $R < 0$.

<<Figure 6>>

The plots clearly show the much weaker performance for the validation period than for the calibration period, expressed by lower *R*-values and higher σ_P/σ_O ratio's. **SIM-STO** showed the best performance for concentrations in the validation period with $R > 0.7$, σ_P/σ_O close to one, and *RRSE* = 0.75, while for the other model *RRSE* > 1. For the nitrate fluxes in the calibration period *RRSE* values were between 0.64 and 0.86, while for the validation period, the values were between 1 and 2 even with a peak of 8.6 for **SW-ANIM** (data point not seen in Figure 6). The *R*-values of the nitrate fluxes in the validation period were in the range 0.18 (**EPIC**) to 0.50 (**COUP**). The σ_P/σ_O ratio ratios were in the range 0.75 to 2.3 with a peak of 8.8 for **SW-ANIM** (data point not seen in Figure 6). The values for σ_P/σ_O ratio greater than 1 for both the concentrations and the nitrate fluxes indicate that the variation of the simulated values is greater than the variation of the observed values.

Table 7 presents the performance indices for the nitrate concentrations at depths 0.35 m and 0.9 m. The *IoA* values indicate that the best agreement between simulated and measured values was achieved for the calibration period, but *MAE*-values and *RMSE*-values were highest for the calibration results at depth 0.9 m and lowest for the validation results at depth 0.9 m. This apparent contradiction is due to the number of measurements on which the indices were calculated. Further analysis was based on *IoA* because the ranking of these values corresponded better to the results of the leaching water at depth 1.80 m.

<<Table 7>>

Calibrated concentrations yielded *IoA*-values ranging from 0.44 (**SIM-STO**) to 0.84 (**SW-ANIM**). The results for the validation period resulted in somewhat lower *IoA* values, except for **SIM-STO** which shows better results for the validation than for the calibration period. The **ARMOSA** results were the most constant for the different depths and periods. Both **COUP** and **SW-ANIM** show significantly poorer *IoA* values for the validation than for the calibration period. **DAISY** and **SIM-STO** showed slightly better results for the concentrations at depth 0.9 m than for the concentrations at depth 0.35 m. The other models performed slightly better for depth 0.35 m. Except for **SW-ANIM**, the *IoA* for the validation period at 0.35 m were in the same range as for the results at depth 0.9 m.

Over- and overestimation of simulated average nitrate concentrations and nitrate-N leaching rates for the calibration period is due to a number of reasons. A formal reason is the formulation of the object function. The calibration method applied for most models attempted to minimize the sum of squared differences $(P_t\text{-}O_t)^2$ for either the nitrate concentrations or the nitrate-N fluxes. A minimal sum does not guarantee a perfect match of the average concentrations. The different modelling groups have chosen different objective functions when calibrating for nitrate observations. Most models based the summation $(P_t\text{-}O_t)^2$ values on the sampling periods but **SIM-STO** used the summed $(P_t\text{-}O_t)^2$ values for the nitrate-N leaching rate per growing season only.

Three out of four models that used nitrate flux in their objective function resulted in *IoA* values in the range 0.76-0.87 for the calibrated nitrate fluxes, while the other model resulted in *IoA* = 0.43 (Table 6). Two out of three models that used nitrate concentration in their objective function resulted in *IoA* values in the range 0.95-0.97, while the third model resulted in *IoA* = 0.87 (Table 6). However, a good calibration on nitrate concentrations did not result in good performance on nitrate fluxes. Both for the calibration and for the validation periods it appeared that all models had difficulties in predicting the nitrate fluxes at the bottom of the lysimeter, even if some of them were calibrated based on the measured nitrate fluxes.

Vereecken et al. (1991) evaluated five complex models from which **SW-ANIM, EPIC** and **DAISY** are also included in our performance assessment**.** A comparison between simulated and observed nitrate leaching rates measured in two sandy soils in Denmark and one sandy soil in the Netherlands revealed that **SW-ANIM, EPIC** and **DAISY** performed similar, although **DAISY** appeared to be a bit superior in behaviour. In general much better statistical metric values were reported than in our study. This may be due to the circumstances of the field trials which were representative for conventional agriculture during the eighties and because the calibration and he comparison was carried out for seasonal values.

Diekkrüger et al, (1995) compared the results produced by 19 simulation models, others than those used in this study, for a loam soil and a sand soil in Southern and Eastern Saxony in Germany. Variation in the leaching rates at 0.9 m depth reflected mainly the differences in soil water fluxes at that depth. Apart from the seasonal differences between the models that were able to simulate a three year period continuously, the cumulative leaching was nearly the same for these models. The results of soil nitrogen simulations were significantly influenced by the results of water flow and plant growth simulations. Diekkrüger et al, (1995) concluded that for long term forecasts the exact determination of the boundary conditions is as important as the model approach itself. Our finding that the unmeasured inputs concerning biological N-fixation are important for the soil nitrogen dynamics is consistent with this conclusion. In our study, differences between model seasonal and long term results are attributed to some extend to different assumptions about fixation rates.

Kersebaum et al., (2007) conducted a comparison of simulation models for 18 different models from which **SW-ANIM** and **SIM-STO** are included in our study. **SW-ANIM** was applied to the Müncheberg data-set (Kroes and Roelsma, 2007) and **SIM-STO** was applied to the data-set of the lysimeter station Berlin-Dahlem for water flow simulation and to the Bad Lauchstädt data-set for simulation of soil nitrogen dynamics (Stenitzer et al., 2007). Results for the mean bias, *RMSE*, *IoA* and *NSE* showed weak performances for the soil mineral nitrogen simulation in the 0-90 cm upper soil layer for nearly all models which were subjected to the Müncheberg data-set. Kersebaum et al. (2007) concluded that comparison of simulated results by models which are intended for field scale and regional scale with measured data often shows unsatisfactory results due to deviating conditions and parameters. It does not automatically mean that the models or the parameters are wrong because the data and parameters are only partly related to the site specific conditions of the measurements. In our study significant amount of data was available, but critical information about rooting depth and pattern, atmospheric deposition rates, mineralization and fixation rates was missing as well as the nitrogen uptake rates and residue amounts of the catch crops. Due to these uncertainties, it is difficult to draw clear conclusions about the predictive power of the models.

3.2.4 Nitrogen balances

Table 8 presents the soil nitrogen balances per season for each of the models.

<<Table 8>>

Exact fertilizer and manure inputs were not represented by **EPIC**, because the model assumes standard compositions which are not equal to the experimental data. This holds also for **SW-ANIM** which assumes fixed nitrogen compositions but this was overcome by introducing new manure types, so that the fertilizer input was close to the observed values.

The estimates for atmospheric deposition ranged from 4.2 kg ha^{-1} a^{-1} (**COUP)** to 23.4 kg ha^{-1} a^{-1} (**DAISY)**, averaged for seven growing seasons. Only literature values were available and most modelling groups have used the model default values or the figure they are familiar with for their own country. **ARMOSA** calculated for the validation phase lower wet deposition rates than for the calibration phase due to lower precipitation amounts. Some models assumed only dry deposition at a constant rate, while other models also imposed nitrogen input by rainfall.

The most stressing differences are for biological N-fixation. Some models do not describe the biological N-fixation process as such but modellers had possibilities to assume fixation rates by introducing a nitrogen rich organic material which was amended continuously during the growing season. The **DAISY** and the **EPIC** modelling groups did not take account for N-fixation, either due to a lack model formulations implemented or to a lack of knowledge about this process. **SIM-STO** assumed only for the first season some biological N-fixation by the crop mixture that included white clover. The **COUP** and the **SW-ANIM** modelling groups took account for N-fixation, including for periods for which one wouldn't expect (English ryegrass). In **SW-ANIM** the biological N-fixation is lumped with the mineralization of some of the crop residues that descended from the most recent and previous catch crops. The model output does not allow to unravel the biological N-fixation as such and mineralisation of earlier catch crop residues.

The **COUP** model did not take account for ammonia volatilization. The other models did, and showed a range of 2% to 35% of the nitrogen in the animal manure amended to the soil. The highest volatilization rates were simulated by **SIM-STO**: 27% and 35% of the animal manure N in 2008 and 2011, respectively. This could possibly explain the underestimation of nitrate leaching in 2008, but not in 2011. For these years, the differences of the model predictions amounted to more than 22 and 37 kg ha^{-1} a^{-1}, respectively, which is higher or in the same range as the measured nitrate-N leaching. Volatilization was calculated by **EPIC** and **ARMOSA** (about 4 kg ha^{-1}) for the first growing season of the validation period, while no farm fertilizer was applied.

The models encountered difficulties with the simulation of nitrogen crop off-take. Deviations of simulated uptake rates from the observed values of more than 50 kg ha^{-1} occurred for three years by **ARMOSA** (2006, 2008, 2009), **EPIC** (2005, 2009, 2010) and **SIM-STO** (2006, 2008, 2010), for two years by **DAISY** (2007, 2010), and for one year (2011) by **COUP** and **SW-ANIM**. The **EPIC** model was not able to simulate nitrogen crop off-take by oil pumpkin, because this crop is unknown in the standard database of crop parameters that comes with the model. The **DAISY** model failed to simulate a reasonable crop off-take by maize in 2007, while the N off-take in the preceding year was overestimated by 60 kg ha^{-1}. The calibrated parameters for crop uptake were not optimal for the maize as is also apparent from the calculated crop off-take in 2010 where the overestimation amounted nearly 100 kg ha^{-1}. Despite the fact that **SW-ANIM** included the N-yield in the object function of the calibration procedure, the modelled crop off-take differed from the measured crop off-take by -14 to +19 kg ha^{-1}. The **SW-ANIM** underestimated crop off-take in the validation period. Crop off-take is governing the soil nitrogen balance to a large extent and an erroneous calculation of the N off-take means that a possible correct nitrate leaching should be considered as little robust.

Denitrification is only of significance for the **DAISY** and **EPIC** results, while other models simulated zero or negligible denitrification rates. For most of the models, these estimates were biased by the opinion of the data holders who made plausible from their analysis of soil nitrogen balances that denitrification is not a significant factor (Leis, 2009). The degree of saturation (*S*) at depth 0.35 m exceeds 80% for most of the time and only **COUP** and **SIM-STO** have default threshold values for *S* higher than 80% while other models use lower default threshold values for *S* (Heinen, 2006). Except for **DAISY** and **EPIC**, also **ARMOSA** and **SW-ANIM** should have calculated some denitrification when using default values. Except for the first year, the denitrification calculated by **EPIC** exceeded the nitrate-N leaching.

The change of the total N amount in soil included both organic and mineral forms and was calculated as the residual from the balance. A positive sign means an increase of the total amount whereas a negative sign indicates a depletion of the stock. The model results showed large differences and the largest difference occurred in 2010 where **DAISY** calculated a depletion of 105 kg ha^{-1} while **SW-ANIM** calculated an increase of 103 kg ha^{-1}. The increase of the amount resulted from the assumed biological fixation and the inputs caused by the cultivation of catch crops. When no additional inputs by fixation or by catch crops was assumed, a depletion will occur (**DAISY** and **EPIC**).

Except for **SIM-STO** in 2005 and 2008, differences between calculated seasonal nitrate-N leaching rates were relatively small for the calibration phase. The deviations were much larger for the validation phase, where **SW-ANIM** overestimated the leaching by 39 and 29 kg ha^{-1} in 2009 and 2010, respectively. The observed small leaching rate in 2010 was not approached by any model. Transport of ammonium, organic dissolved N or by surface runoff was calculated at a maximum of 8 kg ha^{-1} by the **COUP** model for the first year of the validation period.

The long term nitrogen balances were summarized at the bottom of Table 8 to further compare the difference of the modellers perceptions of the plant and soil nitrogen cycle.

The seven year balance depicted the major differences between the models clearly. Despite the crop failure in 2007 simulated by **DAISY**, this model showed the highest summed seven year amount, while the summated crop off-take by **SIM-STO** lagged behind with 200 kg ha^{-1} relative to the recorded amount. For the individual years the **ARMOSA** results differed considerably from the observations, but the summated seven year crop off-take resembled the measured value rather good.

Most models have been designed for the field scale for which an average N-yield is calculated. The spatial scale of the lysimeter (1 m^2) differs from the field scale and the variation of crop off-take rates at this scale is much larger than for the field scale. This is illustrated by the oil pumpkin crop in 2005. Only two seeds were planted in the lysimeter. One of the plants died at the start of the generative phase and no harvest was obtained from this plant. This event influenced the yield at the lysimeter scale pretty much, but the yield at the field scale was barely influenced and it can be expected that field scale models encountered difficulties.

The total nitrogen loss by denitrification ranged from 0 to 249 kg ha^{-1} and was subject to the modellers' perception of the possibility of denitrification in the soil at the Wagna experimental field station.

The low input farming system was capable to produce relatively high yields for maize and grains, and for oil pumpkin a N-yield of 51 to 57 kg ha^{-1} was recorded, but the observed nitrate-N leaching exceeded the N-excess, the latter defined as the total addition of mineral fertilizers and animal manure minus the crop off-take. **ARMOSA**, **DAISY** and **EPIC** predicted higher nitrate N-leaching than the N-excess (Fig. 7), while the other models showed a more or less equal value (**SW-ANIM**) or a lower value (**COUP**, **SIM-STO**). One of the main difficulties was to describe the role of the intermediate catch crops in the crop rotation on the delivery of N. Some of the intermediate crops fixate atmospheric N which leads to an input to the soil and other crops are only able to preserve some of the N excess which remains in soil after the catch crops for the next growing season. No data on the N uptake rates and the quality of the resulting green biomass of these intermediate crops were available. Each of the modellers had to make assumptions for the effect of these crops on the soil N cycle. The estimates of the seven years summed additional input to the soil by biological N-fixation varied from 0-2 kg ha^{-1} (**DAISY, EPIC)** to 371 kg ha^{-1} (**SW-ANIM**) (Table 8).

<<Figure 7>>

None of the models simulated long term soil N-stock at equilibrium. The models that did not take biological N-fixation into consideration showed a decrease of the soil N-stock of -342 kg ha^{-1} (**EPIC**) and -177 kg ha^{-1} (**DAISY**). The other models that take account for this input showed an increase ranging from 165 to 419 kg ha^{-1}.

The comparison of the N mass balance components showed large differences between the models. Despite calibration on nitrate leaching, the nitrate leaching predicted was still different from that measured. Crop off-take, although measured, was only used by two models in the calibration procedure, but even then the predicted off-take differed from the observed one. For the other N processes (deposition, biological fixation, volatilization, other transport processes and denitrification) no measured data were available for comparison and calibration. For these aspects, significant differences between the models were observed, either through differences in process descriptions or in handling input by the modelling groups. The resulting storage change thus was also different for the models. The variation of the mass balance components for each model over the years was large.

A favourable assessment of a good correspondence between a predicted and a measured quantity is difficult, because it may be good for the wrong reasons. For example, **ARMOSA** predicted rather well the overall crop N off-take but was not able to predict the N off-takes of the individual growing seasons.

3.2.5 Performance assessment

In order to compare the performance of models a quantifiable method is needed. The simplest method would be to rank the models based on a performance index. This method is not preferred, as a model may get a high ranking despite a poor performance. Thus, a classification based on some performance index is to be preferred. Any value of *NSE* and *IoA* (except their values 0 and 1) is difficult to interpret (Legates and McCabe, 1999), and thus it is clear that no default classification boundary values exist to evaluate good, moderate and poor model performance for a set of interrelated variables related to water contents, water fluxes, nitrate concentration and nitrate fluxes at te scale of a lysimeter.

Bellocchi et al. (2010) reviewed the methods and different indicators used for the validation of different types of biophysical models. Confalonieri et al. (2010) used *NSE* and *RRMSE*, together with four other indices to assess the quality of simulation of different models in simulating soil water contents. In hydrological studies, it is common practise to assess the model performance on the basis *NSE*, where $NSE > 0.75$ indicates a “good” performance and $NSE < 0.36$ indicates a “weak” similarity of model results with observations (Van Lieuw and Gabrecht, 2003). Moriasi et al. (2007) reviewed the qualification of the model performance of stream discharges and contaminant loads, based on statistical indices for a number of modelling studies. They qualified model simulation on the basis of *NSE* and *PE* but their qualifications are not directly applicable to this study due to differences of spatial scale (catchment versus field) and differences of time scale (month versus day or weekly sample interval). In the literature it is noticeable that classifications and qualifications depend on the considered variables and of the time and space scale. Here we preferred to set up a classification for *IoA*. A number of model studies on the dynamics of soil nitrogen and nitrate leaching have been published that use the *IoA*, alone, or combined with other parameters (Kersebaum et al., 2007; Mantovi et al., 2006; Nolan et al., 2010 ; Sogbedji et al., 2006).

Typical state variables which correspond with instantaneous observations have been distinguished from water fluxes and nitrate concentrations analysed in composed water samples. For the latter we assumed *IoA* values above 0.9 as accurate and *IoA* values below 0.75 as inaccurate. For soil water contents and nitrate concentrations we assume *IoA* values greater than 0.8 as accurate and *IoA* values smaller than 0.6 as inaccurate. Krause et al.

(2005) stated that even for *IoA* > 0.65 models can result in poor performance, they sure will for *IoA* < 0.6, which was here chosen as the lowest boundary. The *IoA* scoring for the calibration and validation periods are listed in Table 9.

<<Table 9>>

The scoring differed for the different models. Two models (**SIM-STO**, **SW-ANIM**) performed well for the calibration of the θ(*h*) curves and the simulated θ at different depths, however, this doesn't guarantee good performance for the other state and rate variables in the calibration and validation periods. For the validation period all models performed weak to moderate on the water volume and weak on the nitrate N-flux per sampling interval, moderate to good on the daily water flux and weak to moderate on the nitrate concentration in the water samples. The models **ARMOSA**, **COUP**, **DAISY** and **EPIC** had more weak qualifications than good qualifications, while **SIM-STO** and **SW-ANIM** had more good qualifications.

We have also assessed the accuracy of the seasonal amounts on the basis of the mean absolute error (*MAE*). The seven seasons included the oil pumpkin crop twice, which was an unknown or a particular crop for most of the modelling groups. The seven year series contained an extremely wet year (2009) and a dry summer (2011). For the performance assessment for average crop and rainfall conditions *MAE* of the five best values (MAE_5) out of seven (MAE_7) are presented in Table 10 to examine if the models perform better for average conditions. In some cases the improvement was more than 50%, and the ranking of the models slightly changed. Despite the fact that *MAE* is less sensitive to outliers than e.g. *IoA*, extreme situations (unknown crop, wet or dry years) can have a large impact on *MAE*.

<<Table 10>>

3.2.6 Methodological aspects for explanation of differences

Data

Experimental data collected from a well-controlled lysimeter were used for the purposes of our study. However, the number of measured state and rate variables were less than those present in the six models. For example, no data were available on field-scale hydraulic conductivity, deposition and biological fixation. This means that the outcome of the models is uncertain as not all components of the internal mass balance could be optimized. We have observed in the blind test that based on a limited availability of data, which resembles situations that would occur in practice, the predictions of the models was poor compared to actual observations. That would imply that usage of such simulation models for predictions on nitrate leaching at unknown, regional scales must be regarded with care. In this study the rainfall excess was positive in most times of the year, such that the imposed bottom boundary condition in the lysimeter resulted in leaching. For other situations where capillary rise may occur, the models have not been inter-compared. Finally, it is noticed that the soil hydraulic properties as determined in the laboratory on small soil samples does not guarantee well-predicted soil water contents and soil water fluxes even for a well-controlled lysimeter situation. Partly, this may be due to the lack of knowledge of hysteresis or its description in the models.

Procedure

Despite the structured set-up of this study (blind test, calibration, validation) there remained flexibility in the approach chosen by the different modelling groups. For example, no formal sensitivity analysis was prescribed, meaning that each group was free to choose a set of parameters to be calibrated. This has introduced a subjective element in this study. Although it was agreed beforehand that the water fluxes and the nitrate concentrations in the lysimeter effluent were the most important parts of the model comparison, the objective function for optimization was chosen freely by the modellers. Some modelling group have chosen to include also the information about soil water contents and crop uptake in the optimization procedure. The comparison is, therefore, not a pure comparison of the model codes, but also a comparison of how modellers used their models.

In this study much effort has been put in calibrating and validating six models for a well-controlled lysimeter situation. Any conclusions of this study are thus at first applicable for these kind of (local) situations. Additional research is required to inter-compare these models for deviant situations, for example, for regional assessments of impact of fertilization strategies.

Decreased performance when averaging

One should expect a better performance for the averaged water fluxes per sampling interval than for the daily water fluxes because peaks of the daily fluxes pattern are flattened by aggregation. This was indeed observed in better performance indices for the calibration period (Table 5). However, the opposite occurred for the validation period (Table 5). This counter-intuitive response of performance indices to the averaging of water fluxes of the validation phase may be due to the following three reasons.

1) The distributions of the time increments of sampling in both phases differed slightly, where in the validation phase samples were taken more frequently with smaller time steps (data not shown). The pattern of sampling

intervals was neither regular nor random. The pattern was more or less dependent on practical circumstances and availability of manpower and on average samples were taken once in seven days. Under extreme rainfall conditions the intervals were shortened and under extreme dry conditions the intervals were longer because no percolation water was present.

2) The probability density distributions of the daily water fluxes and averaged water fluxes for the calibration and validation periods appeared to be unequal (data not shown). This was concluded from a non-parametric analogue of a one-way analysis of variance performed by the one-way analysis of variance by ranks after Kruskal-Wallis (1952). The different statistical behaviour may result in variant effects of volume weighted averaging on the performance indices.

3) Certain days or periods may have had a great effect on the averaging. A leave-one-out calculation procedure was performed to qualitatively explore the effect of certain days and periods on the performance of the models. In the series of data pairs of observed and simulated water fluxes, one data pair is left out and the *IoA* was calculated for the remainder of the population. This procedure is repeated for each of the data pairs and the results are subtracted from the *IoA*-value based on the total series of data pairs belonging to either the daily fluxes of the calibration or the validation phase or to the averaged values of the phases. Only the results greater than 0.001, in absolute sense, haven been plotted in Figure 8.

<<Figure 8>>

The exclusion of a particular data pair can result in both an improvement (negative values) or a deterioration (positive values) of the ΔIoA. Furthermore, it is notable that the ΔIoA of daily fluxes responded differently compared to the ΔIoA for averaged fluxes per sampling interval. For almost all models the exclusion of the value simulated for 19 Sept 2006 would affect the ΔIoA. The effect of excluding the value of this period is much smaller for the ΔIoA based on the averaged values per sampling interval. The maximum effect in the series of daily values occurres for a certain day of the calibration period and the maximum effect in the series of averaged values per sampling interval is calculated for a time interval in Sept. 2010 which belongs to the validation phase The maximal effect of leaving one value out is greater for the validation period than for the calibration period. Based on this analysis, it is plausible that the averaging of water fluxes has a different effect on the performance indices of the calibration phase than on those of the validation phase.

4. Summary and Conclusions

The novel aspect of this study is that six detailed process oriented dynamic models were tested (1) for the Wagna test-site which is known to be highly vulnerable to nitrate leaching, (2) for a crop particular for the Styrian low input agriculture system, (3) for a situation where different catch crops were part of the crop rotation, and (4) for the weather conditions which significantly differed between the calibration and the validation phase..

This study was not performed to determine which model is the best. We like to quote Kersebaum et al. (2007) who stated: "*The comparison of different models applied on the same data set is not suitable to serve as a model contest or to find the best model. Although, the application of different indices for model performance helps to identify strengths and weaknesses of each model, an objective comparison is nearly impossible due to different levels of input requirements, calibration efforts and last but not least the uncertainties and errors within the measured data themselves.*"

We conclude:

a. The blind test showed that simulation results without calibrating the model are generally far from acceptable . Therefore, model calibration is essential.

b. None of the models performed good for the different criteria considered in this study. This may be due to the combined effect of the model structure which is not tuned to the circumstances of the Wagna experimental fields and the lack of knowledge to establish an appropriate set of parameters. Furthermore, not all inputs were measured, so there were too many degrees of freedom.

c. The soil of the Wagna lysimeter is highly vulnerable to nitrate leaching. The seven year summed nitrate leaching rate (123 kg ha^{-1}) exceeds the seven year summed fertilization excess. Models designed for nitrate leaching in high input farming systems have difficulties with an accurate prediction of the nitrate leaching in low input farming systems

d. Judgement of the performance solely on the basis of nitrate concentrations or nitrate fluxes is not sufficient for the assessment of the predictive power of the models. Other results as soil water contents (daily), water and nitrogen fluxes (daily and seasonal), soil temperatures (daily), nitrogen yields (seasonal) should also be taken into account. This should be reflected by the objective function of the model calibration.

e. Traditional Richard's / Darcy Buckingham equation based models that make use of the Mualem-van Genuchten descriptions and disregard phenomena as hysteresis, preferential flow and multiple phase flow encounter difficulties with an accurate and consistent simulation of both water contents and water fluxes for the soil and conditions of the Wagna lysimeter.

f. Some models which performed relatively well in the calibration phase of the study failed to simulate the nitrate concentrations and fluxes in the validation phase (**SW-ANIM**), while other models behaved relatively bad in the calibration phase and showed better results in the validation phase (**SIM-STO**). An accurate calibration does not guarantee a good predictive power of the model.

g. The catch crop mixtures and the non-harvested English ryegrass play an important role in the nutrient dynamics of the soil. This role is addressed weakly by the simulation models: (1) due to a lack of experimental data on nitrogen uptake rates and mineralization of residues of these intermediate crops, and (2) lack of knowledge to describe the relevant processes related to the foreign crops

h. Assessment of future climate and land use changes requires a good predictive power of the models and a certain level of robustness. Although the robustness is not clear for the tested models, the process oriented dynamic models used in this study are useful for hypothesis testing.

5. Acknowledgements

This research was made possible by the GENESIS project of the EU 7th Framework Programme (Project No. 226536; FP7-ENV-2008-1). We are grateful for the experimental data provided by Joanneum Raum (Graz, Austria). The modelling team of Democritus University of Thrace would like to thank Per-Erik Jansson (Royal Institute of Technology, Stockholm, Sweden;) for his valuable help during the application of CoupModel.

6. References

Akkal-Corfini N, Morvan T, Menasseri-Aubry S, Bissuel-Bélaygue C, Poulain D, Orsini F, Leterme P. Nitrogen mineralization, plant uptake and nitrate leaching following the incorporation of (15N)-labeled cauliflower crop residues (Brassica oleracea) into the soil: a 3-year lysimeter study. Plant Soil 2010;. 328:17–26. DOI 10.1007/s11104-009-0104-0

Ale S, Bowling LC, Youssef MA, Brouder SM. Evaluation of simulated strategies for reducing nitrate–nitrogen losses through subsurface drainage systems. J Environ Qual 2012; 41:217-228.

Basile A, Ciollaro G, Coppola A. Hysteresis in soil water characteristics as a key to interpreting comparisons of laboratory and field measured hydraulic properties. Water Resour Res 2003;39:1355-1367.

Basile A., Coppola A, De Mascellis R, Randazzo L. Scaling approach to deduce field unsaturated hydraulic properties and behavior from laboratory measurements on small cores. Vadose Zone J 2006;5:1005-1016.

Bellocchi G, Rivington M, Donatelli M, Matthews K, Validation of biophysical models: issues and methodologies. A review. Agron Sustain Dev 2010;30:109–130.

Bergström L, Johnsson H, Torstensson G. Simulation of soil nitrogen dynamics using the SOILN model. Nutr Cycl Agroecosys 1991;27:181–188.

Bouraoui F, Aloe A. European Agrochemicals Geospatial Loss Estimator: Model development and Applications, EUR – Scientific and Technical Research series, ISSN 1018-5593. Office for Official Publications of the European Communities, Luxembourg. 2007.

Buckingham, E. Studies on the movement of soil moisture. Bull. 38, USDA, Bureau of Soils, Washington, DC, 1907.

Burkart MR, Kolpin DW, James DE. Assessing groundwater vulnerability to agrichemical contamination in the Midwest US. Water Sci Technol 1999;39:103-112.

Confalonieri R, Bregaglio S, Bocchi S, Acutis M. An integrated procedure to evaluate hydrological models. Hydrol Process 2010;24:2762–2770.

Darcy H. Les fontaines publique de la ville de Dijon. Dalmont, Paris, 1856.

Dawson CW, Abrahart RJ, See LM. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Modell Softw 2007;22:1034–1052. http://dx.doi.org/10.1016/j.envsoft.2006.06.008

Dawson CW, Abrahart, RJ, See, LM. HydroTest: Further development of a web resource for the standardised assessment of hydrological models. Environ Modell Softw 2010;25:1481–1482. http://dx.doi.org/10.1016/j.envsoft.2009.01.001

De Willigen P, Neeteson JJ, Comparison of six simulation models for the nitrogen cycle in the soil. Fert Res 1985;8:157-171.

De Willigen P. Nitrogen turnover in the soil-crop system; comparison of fourteen simulation models. Fert Res 1991;27: 141-149.

Diekkrüger B, Söndgerath D, Kersebaum KC, McVoy CW. Validity of agroecosystem models a comparison of results of different models applied to the same data set. Ecol Model 1995:81:3-29. http://www.sciencedirect.com/science/article/pii/030438009400157D

Donatelli M, Wösten JHM, Belocchi G. Evaluation of pedotransfer functions. In: Pachepsky Y. Rawls WJ, editors. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam. 2004. p. 357–362.

Durner W, Jansen U, Iden SC. Effective hydraulic properties of layered soils at the lysimeter scale determined by inverse modelling. Eur. J. Soil Sci. 2007;59:114–124. doi: 10.1111/j.1365-2389.2007.00972.x

EU. 1991. Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off . J. Eur. Commun. L375:1–8. Available at http://ec.europa.eu/environment/water/water-nitrates/directiv.html

Fank J, Fastl G, Kupfersberger H, Rock G. Die Bewirtschaftung des Versuchsfeldes Wagna - Auswirkung auf die Grundwassersituation. Umweltprogramme für die Landwirtschaft 2006. Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft, A-8952 Irdning. Gumpenstein, 2006.

Feichtinger F. STOTRASIM – Ein Modell zur Simulation der Stickstoffdynamik in der ungesättigten Zone eines Ackerstandortes. In: *Modelle für die gesättigte und ungesättigte Bodenzone*. Schriftenreihe des Bundesamtes für Wasserwirtschaft, 7, Wien, 1998, p. 14-41, Ferrara RM, Trevisiol P, Acutis M, Rana G, Richter GM, Baggaley N. Topographic impacts on wheat yields under climate change: two contrasted case studies in Europe. Theor. Appl. Climatol. 2011;99:53–65.

Gribb MM. Hansen FI, Aleshia Chandler, McNamara DG, James P. The effect of various soil hydraulic property estimates on soil moisture simulations. Vadose Zone J 2009; 8:321–331.

Grizzetti B, Bouraoui F, Billen G, Van Grinsven H., Cardoso AC, Thieu V, Garnier J, Curtis C, Howarth R, Johnes P. Nitrogen as a threat to European water quality. In: The European Nitrogen Assessment 17, Cambridge, UK: Cambridge University Press, 2011, p. 379-404.

Groenendijk, P. , L.V. Renaud, and J. Roelsma. 2005. Prediction of Nitrogen and Phosphorus leaching to groundwater and surface waters. Process descriptions of the animo4.0 model. Alterra–Report 983, Alterra, Wageningen. http://content.alterra.wur.nl/Webdocs/PDFFiles/Alterrarapporten/AlterraRapport983.pdf

Hansen S, Jensen HE, Nielsen NE, Svendsen H. **DAISY**: Soil Plant Atmosphere System Model. NPO Report No. A 10. The National Agency for Environmental Protection, Copenhagen, 1990, 272 pp.

Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of nitrogen dynamics and biomass production in winter wheat using the Danish simulation model DAISY, Fert Res 1991a;27: 245-259.

Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of biomass production, nitrogen uptake and nitrogen leaching by using the Daisy model. In: Soil and Groundwater Research Report II: Nitrate in Soils, Final Report on Contracts EV4V-0098-NL and EV4V-00107-C. DG XII. Commission of the European Communities, 1991b; 300–309.

Hansen S. DAISY, a flexible Soil-Plant-Atmosphere system Model. Report. Dept. Agric, Danish Informatics Network in the Agricultural Sciences, 2002. http://www.dina.kvl.dk/~DAISY/ftp/DAISYDescription.pdf.

Heinen, M. Simplified denitrification models: Overview and properties. Geoderma 2006;133:444-463.

Herbst M, Fialkiewicz W, Chen T, Pütz T, Thiéry D, Mouvet C, Vachaaud G, Vereecken H. Intercomparison of flow and transport models applied to vertical drainage in cropped lysimeters. Vadose Zone J 2005;4:240-254.

Jabro JD, Jabro AD, Fales SL. Model performance and robustness for simulating drainage and nitrate-nitrogen fluxes without recalibration. Soil Sci. Soc. Am. J. 2012;76:1957–1964 doi:10.2136/sssaj2012.0172

Jachner S, Van den Boogaart KG, Petzoldt T. Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R package qualV). J Stat Softw 2007;22:(8),1–30. http://www.jstatsoft.org/v22/i08

Janssen PHM, Heuberger PSC. Calibration of process-oriented models. Ecol Model 1995;83: (1-2) 55-66. http://dx.doi.org/10.1016/0304-3800(95)00084-9

Jansson P-E, Karlberg L. Coupled heat and mass transfer model for soil-plant-atmosphere systems Royal Institute of Technology, Dept of Civil and Environmental Engineering, Stockholm, 2004. 435 pp.

Jansson P-E. CoupModel: Model use, calibration and validation, Transactions of the ASABE 2012;55(4):1-11.

Jensen LS, Mueller T, Nielsen NE, Hansen S, Crocker GJ, Grace PR., Klir J, Körschens M, Poulton PR. Simulating trends in soil organic carbon in long-term experiments using the soil-plant-atmosphere model DAISY. Geoderma 1997;81:5-28.

Kersebaum KC, Hecker, J-M, Mirschel W, Wegehenkel M. Modelling water and nutrient dynamics in soil–crop systems: a comparison of simulation models applied on common data sets. In: Kersebaum KC et al., editors, Modelling Water and Nutrient Dynamics in Soil–Crop Systems, Springer, 2007, p. 1–17.

Khodaverdiloo H, Homaee M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions for some calcareous soils, J Hydrol 2011;399:93-99

Klammler G, Fank J.. Determining water and nitrogen balances for beneficial management practices using lysimeters at Wagna test site (Austria). Sci. Tot. Environ 2014; *this issue*.

Krause P, Boyle, DP, Bäse F. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 2005;5:89–97. http://www.adv-geosci.net/5/89/2005/adgeo-5-89-2005.pdf

Kroes J and Roelsma J. Simulation of water and nitrogen flows on field scale: application of the SWAP–ANIMO model for the Müncheberg data set. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient Dynamics in Soil–Crop Systems, 2007, Springer, pp 111–128.

Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.

Legates DR, McCabe GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233-241.

Leis A. Chemical, isotopic, and microbiological evidence for nitrification below the plant root zone from intensive fertilized agricultural area in Austria – Insights from lysimeter studies and soil cores. In: IAEA-TECDOC-1618. Application of Isotopes to the Assessment of Pollutant Behaviour in the Unsaturated Zone for Groundwater Protection. Final report of a coordinated research project 2004-2005. International Atomic Energy Agency,Vienna; 2009. p. 15-30.

Mantovi P, Fumagalli L, Beretta GP, Guermandi M. Nitrate leaching through the unsaturated zone following pig slurry applications, J Hydrol 2006; 316:195-212.

Moreels E., De Neve S, Hoffman G, Van Meirvenne M. Simulating nitrate leaching in bare fallow soils: a model comparison. Nutr Cycl Agroecosys 2003;67:137-144

Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Tansactions of the ASABE 2007;50: 885−900.

Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. 1976;12: 513-522.

Nash JE, Sutcliffe JV. River flow forecasting through conceptual models, 1, A discussion of principles, J. Hydrol., 1970;10:282-290.

Nett L, Feller C, George E, Fink M.Effect of winter catch crops on nitrogen surplus in intensive vegetable crop rotations. Nutr Cycl Agroecosys 2011;91:327-337.

Nolan, BT, Puckett LJ, Ma L, Green TG, Bayless ER, Malone RW. Predicting unsaturated zone nitrogen mass balances in agricultural settings of the United States. J Environ Qual 2010;39:1051-1065.

Oenema O. Governmental policies and measures regulating nitrogen and phosphorus from animal manure in European agriculture. J Anim Sci 2004;82: E196-E206.

Patil N, Rajput G. Evaluation of Water Retention Functions and Computer Program "Rosetta" in Predicting Soil Water Characteristics of Seasonally Impounded Shrink–Swell Soils, J Irrig Drain E-ASCE 2009;135, 286-294. http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29IR.1943-4774.0000007

Perego A, Giussani A, Sanna M, Fumagalli M, Carozzi M, Alfieri L, Brenna S, Acutis M. The ARMOSA simulation crop model: overall features, calibration and validation results. Italian Journal of Agrometeorology 2013;3:23-38.

Qi Z, Ma L, Helmers MJ, Ahuja LR, Malone RW. Simulating nitrate-nitrogen concentration from a subsurface drainage system in response to nitrogen application rates using RZWQM2. J Environ Qual 2012;41:289-295.

Reusser DE, Blume T, Schaefli B, Zehe E. Analysing the temporal dynamics of model performance for hydrological models. Hydrol Earth Syst Sc 2009;13: 999 – 1018. http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf

Richards LA, Capillary conduction of liquids through porous mediums. Physics 1931;1: 318-333.

Richter GM, Acutis M, Trevisiol P, Latiri K, Confalonieri R. Sensitivity analysis for a complex crop model applied to Durum wheat in the Mediterranean. Europ J Agron 2010;32:127-132.

Ritter A, Hupet F, Muñoz-Carpena R, Lambot S, Vanclooster M. Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods. Agr Water Manage 2003;59:77-96.

Saxton KE, Rawls WJ. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci. Soc. Am. J.2006;70:1569-1578. doi:10.2136/sssaj2005.0117

Smith P, Smith JU, Powlson DS, McGill WB, Arah JRM, Chertov OG, Coleman K, Franko U, Frolking S, Jenkinson LS. Jenseng LS, Kellyh RH, Klein-Gunnewiek H,, Komarov AS, Lif C, Molina JAE, Mueller T, Parton WJ, Thornley JHM, Whitmore AP. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma 1997;81: 153-225.

Sogbedji JM, Van Es HM, Melkonian JJ, Schindelbeck RR. Evaluation of the PNM model for simulating drain flow nitrate-N concentration under manure-fertilized maize. Plant Soil 2006;282:343-360.

Sohier C, Degré A, Dautrebande S. From root zone modelling to regional forecasting of nitrate concentration in recharge flows – The case of the Walloon Region (Belgium). J Hydrol 2009;369:350-359.

Stenitzer E. SIMWASER – Ein numerisches Modell zur Simulation des Bodenwasserhaushaltes und des Pflanzenertrages eines Standortes. *Mitt. der Bundesanstalt für Kulturtechnik und Bodenwasserhaushalt*, 31: Petzenkirchen, 1988. p.1-118.

Stenitzer E, Diesel H, Franko U, Schwartengräber R, Zenker T. Performance of the model SIMWASER in two contrasting case studies on soil water movement. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient Dynamics in Soil–Crop Systems, 2007, Springer, pp 27–36.

Stumpp C, Nützmann G, Maciejewski S, Maloszewski P. A comparative modeling study of a dual tracer experiment in a large lysimeter under atmospheric conditions. J Hydrol 2009;375: 566-577.

Svendsen H, Hansen S, Jensen HE. Simulation of crop production, water and nitrogen balances in two German agro-ecosystems using the DAISY model. Ecol Model 1995;81: 197-212.

Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 2001;106: No. D7, P. 7183. doi:10.1029/2000JD900719.

Thorup-Kristensen, K. . Effect of deep and shallow root systems on the dynamics of soil inorganic N during 3-year crop rotations. Plant Soil, 2006;288:233-248.

Van Dam JC,. Field-scale water flow and solute transport: SWAP model concepts, parameter estimation and case studies. PhD thesis Wageningen University,, 2000.

Van Dam JC, Groenendijk P, Hendriks RFA. Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. Vadose Zone J 2008;7:640-653.

Van der Laan M, Miles, N, Annandale JG, Du Preez CC. Identification of opportunities for improved nitrogen management in sugarcane cropping systems using the newly developed Canegro-N model. Nutr Cycl Agroecosys 2011;90,391-404.

Van der Velde M, Bouraoui F, Aloe A. Pan-European regional-scale modelling of water and N efficiencies of rapeseed cultivation for biodiesel production. Glob Change Biol 2009;15:24-37.

Van Genuchten MTh. A closed form for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 1980;44:892-898.

Van Liew MW, Garbrecht J. Hydrologic simulation of the little Washita river experimental watershed using SWAT. J Am Water Resour Ass 2003;39:413-426.

Vereecken, H., E.J. Jansen, M.J.D. Hack-ten Broeke, M. Swerts, M. Engelke, S. Fabrewitz and S. Hansen, 1991. Comparison of simulation results of five nitrogen models using different datasets. In: Commission of

European Communities, editor, Soil and Groundwater Research, Report II Nitrate in Soils, Commision of the European Communities, Luxenbourg, pp 321 – 338.

Vereecken H, Weynants M, Javaux M, Pachepsky Y, Schaap MG, Van Genuchten MTh. Using Pedotransfer Functions to Estimate the van Genuchten–Mualem Soil Hydraulic Properties: A Review Vadose Zone J. 2010;9: 795–820.

Vitousek PM, Naylor, Crews RT, David MB, Drinkwater LE, Holland E, Johnes PJ, Katzenberger J, Martinelli LA, Matson PA, Nziguheba G, Ojima D, Palm CA, Robertson GP, Sanchez PA, Townsend AR, Zhang FS.. Nutrient imbalances in agricultural development. Science 2009; 324, no. 5934: 1519.

Wang X, Mosley CT, Frankenberger JR, Kladivko EJ. Subsurface drain flow and crop yield predictions for different drain spacings using DRAINMOD, Agr Water Manage 2006;79:113-136.

Williams JR, Jones CA, Dyke PTl.. A modeling approach to determining the relationship between erosion and soil productivity. Trans. ASAE 1984;27:129-144.

Williams JR, Jones CA, Kiniry JR, Spanel DA.. The EPIC crop growth model. Trans. ASAE, 1989;32:497–511.

Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteo Soc 1982; 63, 1309-1313.

Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. Statistics for the evaluation and comparison of models. J Geophys Res 1985;90, 8995 – 9005. dx.doi.org/10.1029/JC090iC05p08995

Wolff J, Beusen AHW, Groenendijk P, Kroon T, Rötter R, Van Zeijts H. The integrated modeling system STONE for calculating nutrient emissions from agriculture in the Netherlands. Environ Model Softw 2003;18:597-617. doi:10.1016/S1364-8152(03)00036-7

Table 1. Crop rotation and fertilizer applications on the soil of the KON-lysimeter. CC and MC refer to catch crop and main crop, and FYM and MF refer to farmyard manure and mineral fertilizer, respectively.

Type	Crop	Sowing date	Date of harvesting or amending crop residues to soil	Date of fertilizer application	Type and amount of fertilizer (kg ha^{-1} N)
CC	Mixture: summer common tare, white clover, sunflower	06-Aug-04	06-Apr-05		
MC	Oil pumpkin	30-Apr-05	13-Sep-05	25-Apr-05 03-Jun-05	FYM: 27.4 MF: 35.1
CC	English ryegrass	03-Jun-05	09-Apr-06		
MC	Maize (grain)	24-Apr-06	02-Oct-06	24-Apr-06 08-Jun-06	FYM: 54.5 MF: 75.6
CC	Mixture: forage rye, winter turnip rape	03-Oct-06	09-Apr-07		
MC	Maize (grain)	16-Apr-07	21-Sep-07	16-Apr-07 26-May-07	FYM: 120.7 MF: 59.0
MC	Winter barley	08-Oct-07	30-Jun-08	28-Feb-08 09-Feb-08	FYM: 84.6 MF: 38.0
CC	Mixture: winter turnip rape, mustard, sunflower	04-Aug-08	20-Apr-09		
MC	Oil pumpkin	28-Apr-09	07-Sep-09	22-May-09 01-Jun-09	MF: 36.0 MF: 16.0
CC	English ryegrass	05-Jun-09	31-Dec-09		
MC	Maize (grain)	17-Apr-10	23-Sep-10	16-Apr-10 26-May-10	FYM: 62.6 MF: 81.0
MC	Triticale	09-Oct-10	13-Jul-11	11-Mar-11 11-Apr-11	FYM: 119.1 MF: 62.0
CC	Mixture: mustard, phacelia, sunflower, buckwheat, ryegrass	08-Aug-11	After 31-Dec-11		

Table 2. Annual precipitation rates (mm a^{-1}) and their cumulative probability percentages based on precipitations values of 1961 – 2011.

Phase	Calibration				Validation		
Year	2005	2006	2007	2008	2009	2010	2011
Precipitation (mm a^{-1})	883	839	892	893	1355	1013	739
Cumulative probability	44%	31%	48%	50%	98%	75%	10%

Table 3. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of volumetric water contents derived from calibrated soil moisture retention curves (Figure 2) and observed volumetric water contents at depths 0.35 m (*n* = 922), 0.9 m (*n* = 1413) and 1.8 m (*n* = 1456) depth. **EPIC** is excluded as it does not use soil moisture retention relationships.

Model	*MAE* (cm^3 cm^{-3})			*RMSE* (cm^3 cm^{-3})			*IoA*		
	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m
ARMOSA	0.0064	0.0166	0.0308	0.0112	0.0176	0.0310	0.89	0.79	0.18
COUP	0.0341	0.0753	0.0391	0.0416	0.0775	0.0395	0.59	0.31	0.18
DAISY	0.0295	0.0340	0.0166	0.0326	0.0374	0.0178	0.63	0.62	0.38
SIM-STO	0.0212	0.0119	0.0064	0.0255	0.0130	0.0078	0.75	0.89	0.67
SW-ANIM	0.0072	0.0062	0.0033	0.0117	0.0075	0.0036	0.87	0.96	0.85

Table 4. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of simulated and in situ measured values of volumetric water contents at depths 0.35 m, 0.9 m and 1.8 m for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	*MAE* (cm^3 cm^{-3})			*RMSE* (cm^3 cm^{-3})			*IoA*		
	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m
Calibration 2005 – 2008 (*n* = 1461)									
ARMOSA	0.0119	0.0247	0.0107	0.0168	0.0447	0.0123	0.79	0.75	0.46
COUP	0.0230	0.0104	0.0023	0.0288	0.0363	0.0031	0.74	0.84	0.85
DAISY	0.0956	0.0152	0.0105	0.1083	0.0630	0.0132	0.28	0.65	0.38
EPIC	0.0613	0.1563	0.0909	0.0662	0.0306	0.0925	0.49	0.90	0.07
SIM-STO	0.0180	0.0063	0.0028	0.0249	0.0271	0.0039	0.81	0.92	0.85
SW-ANIM	0.0101	0.0106	0.0072	0.0159	0.0285	0.0082	0.87	0.92	0.59
Validation 2009 – 2011 (*n* = 955)									
ARMOSA	x	0.0260	0.0130	x	0.0291	0.0149	x	0.52	0.47
COUP	x	0.0124	0.0030	x	0.0165	0.0041	x	0.74	0.84
DAISY	x	0.0152	0.0137	x	0.0193	0.0165	x	0.69	0.40
EPIC	x	0.1535	0.0924	x	0.1570	0.0939	x	0.19	0.09
SIM-STO	x	0.0093	0.0039	x	0.0134	0.0054	x	0.87	0.82
SW-ANIM	x	0.0141	0.0075	x	0.0176	0.0088	x	0.74	0.65

x Measurements at depth 0.35 m were disqualified from 2009 onwards due to aging of the sensor, and, therefore, no performance indices were calculated

Table 5. Statistical parameters (*MAE*, *RMSE*, *IoA*, *NSE*) for the comparison of simulated and observed daily fluxes and fluxes averaged per sampling interval at depth 1.8 m for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	Daily water fluxes				Averaged water fluxes per sampling interval			
	MAE (mm d^{-1})	*RMSE* (mm d^{-1})	*IoA*	*NSE*	*MAE* (mm d^{-1})	*RMSE* (mm d^{-1})	*IoA*	*NSE*
	Calibration 2005 – 2008							
	n = 1461				*n* = 199			
ARMOSA	0.45	1.00	0.82	0.41	0.43	0.81	0.84	0.48
COUP	0.45	0.98	0.80	0.44	0.43	0.75	0.85	0.55
DAISY	0.57	1.16	0.68	0.21	0.54	0.90	0.74	0.35
EPIC	0.54	0.99	0.83	0.42	0.46	0.75	0.89	0.55
SIM-STO	0.34	0.87	0.86	0.55	0.30	0.62	0.91	0.69
SW-ANIM	0.38	0.91	0.86	0.51	0.37	0.72	0.88	0.58
	Validation 2009 – 2011							
	n = 1084				*n* = 128			
ARMOSA	0.70	1.75	0.79	0.41	1.66	3.82	0.68	0.39
COUP	0.70	1.57	0.84	0.52	1.41	3.47	0.79	0.50
DAISY	0.73	1.77	0.77	0.39	1.74	4.34	0.56	0.21
EPIC	0.85	1.79	0.77	0.38	1.80	4.00	0.63	0.33
SIM-STO	0.51	1.43	0.90	0.61	1.69	3.94	0.76	0.35
SW-ANIM	0.57	1.59	0.88	0.51	1.77	4.16	0.74	0.27

Table 6. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of observed nitrate concentrations and nitrate N leaching rates with simulated values by calibrated models for the Wagna Lysimeter for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	Nitrate concentrations			Nitrate-N leaching rates		
	MAE (mg L^{-1})	*RMSE*	*IoA*	*MAE* (kg ha^{-1} d^{-1})	*RMSE*	*IoA*
	Calibration 2005 – 2008 (*n* = 199)					
ARMOSA	15.71	20.37	0.78	0.043	0.085	0.77
COUP	6.74	9.60	0.97	0.041	0.085	0.78
DAISY	13.92	16.82	0.87	0.037	0.063	0.87
EPIC	19.55	25.63	0.76	0.049	0.084	0.82
SIM-STO	27.34	34.61	0.43	0.044	0.089	0.60
SW-ANIM	7.88	10.48	0.95	0.035	0.080	0.85
	Validation 2009 – 2011 (*n* = 128)					
ARMOSA	11.17	15.85	0.52	0.058	0.102	0.61
COUP	12.36	18.68	0.52	0.076	0.187	0.53
DAISY	8.54	11.40	0.78	0.045	0.095	0.54
EPIC	18.24	22.07	0.52	0.089	0.155	0.41
SIM-STO	8.88	10.44	0.78	0.058	0.138	0.56
SW-ANIM	19.97	29.37	0.43	0.205	0.800	0.12

Table 7. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of observed nitrate concentrations (mg L^{-1}) in water extracted by suction cups at depths 0.35 m and 0.9 m with simulated concentration.

Model	Calibration (0.9 m; *n* = 47)			Validation (0.35 m; *n* = 91)			Validation (0.9 m; *n* = 108)		
	MAE	*RMSE*	*IoA*	*MAE*	*RMSE*	*IoA*	*MAE*	*RMSE*	*IoA*
ARMOSA	36.8	50.6	0.66	22.7	35.9	0.65	12.7	16.6	0.58
COUP	28.0	35.2	0.80	28.2	44.1	0.38	16.6	24.1	0.37
DAISY	32.2	43.9	0.68	29.1	50.9	0.46	12.9	21.5	0.55
SIM-STO	50.6	66.7	0.44	25.5	36.3	0.68	13.6	15.8	0.71
SW-ANIM	25.5	30.5	0.84	36.4	59.3	0.57	20.8	33.8	0.41

Table 8. Comparison of seasonal soil nitrogen balances observed and calculated by the six benchmark models. For each year the main crop is indicated, but these where preceded by catch crops (including leguminous crops).

Crop and period	Balance term[†] (kg ha^{-1})	Observed	Simulated ARMOSA	COUP	DAISY	EPIC	SIM-STO	SW-ANIM
			Calibration 2005 – 2008					
Oil pumpkin	Fertilization* (+)	35.1+27.4	63.0	62.5	62.9	53.1	62.4	62.5
	Deposition (+)		10.2	3.1	16.9	5.0	6.8	11.5
1.1.2005 –	Biological fixation (+)		41.5	1.7	0.1	1.8	31.3	81.3
	Volatilization (-)		2.7	0.0	1.0	1.5	1.9	2.1
13.9.2005	Crop off-take (-)	50.9	59.7	55.3	83.3	0.0	44.3	70.0
	NO_3-N leaching (-)	22.2	17.2	27.9	25.8	30.3	3.6	15.3
	Other transport[$] (-)		0.0	3.2	0.0	0.9	0.0	0.0
	Denitrification (-)		0.0	0.0	13.0	11.8	0.0	0.1
	Storage change[#]		35.2	-19.1	-43.2	15.4	50.6	67.8
Maize	Fertilization* (+)	75.6+54.5	131.0	130.1	130.7	112.3	130.1	130.1
	Deposition (+)		15.4	4.8	26.5	8.0	10.7	17.8
14.9.2005 –	Biological fixation (+)		28.4	32.7	0.0	0.0	0.0	112.9
2.10.2006	Volatilization (-)		9.6	0.0	9.8	8.8	4.9	2.4
	Crop off-take (-)	137.8	211.6	116.0	197.9	125.5	72.7	134.8
	NO_3-N leaching (-)	25.7	27.9	25.8	22.7	33.6	25.1	29.7
	Other transport[$] (-)		0.0	6.0	0.0	1.2	0.0	0.2
	Denitrification (-)		0.0	0.0	13.6	45.8	0.0	1.3
	Storage change[#]		-74.5	19.9	-86.8	-94.6	38.1	92.4
Maize	Fertilization* (+)	59.0+120.7	185.0	179.7	179.4	136.6	179.7	184.5
	Deposition (+)		14.2	4.3	22.2	6.4	8.7	15.3
3.10.2006 –	Biological fixation (+)		52.9	24.7	0.0	0.0	0.0	32.8
21.9.2007	Volatilization (-)		10.9	0.0	2.7	18.5	5.5	28.5
	Crop off-take (-)	92.7	61.4	107.6	2.1	99.7	75.7	96.7
	NO_3-N leaching (-)	5.9	4.4	7.1	6.3	5.4	8.8	5.8
	Other transport[$] (-)		0.0	3.2	0.0	1.5	0.0	0.0
	Denitrification (-)		0.0	0.0	15.3	33.6	0.0	2.0
	Storage change[#]		175.4	90.8	175.2	-15.7	98.4	99.6
Winter barley	Fertilization* (+)	38.0+84.6	123.0	122.6	123.5	78.2	122.6	123.2
	Deposition (+)		11.3	3.3	15.0	3.9	5.3	10.7
22.9.2007 –	Biological fixation (+)		0.0	0.1	0.0	0.0	0.0	14.0
	Volatilization (-)		0.2	0.0	2.6	5.4	22.7	5.1
30.6.2008	Crop off-take (-)	132.3	66.2	104.7	139.0	114.2	81.8	118.4
	NO_3-N leaching (-)	18.9	13.5	18.5	11.7	12.3	5.7	22.2
	Other transport[$] (-)		0.0	3.4	0.0	0.4	0.0	0.0
	Denitrification (-)		0.0	0.0	11.7	40.6	0.0	1.1
	Storage change[#]		54.4	-0.7	-26.4	-90.8	17.7	1.2
			Validation 2009 – 2011					
Oil pumpkin	Fertilization* (+)	52.0+0.0	52.0	52.0	52.0	51.3	52.0	52.0
	Deposition (+)		12.4	5.9	40.1	13.6	18.4	26.0
1.7.2008 –	Biological fixation (+)		52.1	41.2	0.0	0.0	0.0	22.7
	Volatilization (-)		4.4	0.0	0.0	3.9	0.0	0.0
7.9.2009	Crop off-take (-)	56.9	113.6	59.9	97.2	0.0	72.3	45.7
	NO_3-N leaching (-)	33.1	44.2	61.5	26.4	16.0	32.5	72.1
	Other transport[$] (-)		0.0	8.0	0.1	1.9	0.0	0.2
	Denitrification (-)		0.0	0.2	70.6	31.1	0.0	3.4
	Storage change[#]		-45.8	-30.4	-102.1	11.9	-34.4	-20.7
Maize	Fertilization* (+)	81.0+62.6	144.0	143.6	143.1	112.7	143.6	154.3
	Deposition (+)		7.6	4.7	26.6	8.1	11.0	18.0
8.9.2009 –	Biological fixation (+)		0.0	41.3	0.0	0.0	0.0	88.9
23.9.2010	Volatilization (-)		7.2	0.0	2.2	4.8	4.5	9.2
	Crop off-take (-)	142.4	127.6	96.9	240.3	85.0	78.6	115.5
	NO_3-N leaching (-)	3.6	17.0	14.6	8.7	19.3	13.1	32.9

	Other transport[$] (-)		0.0	5.4	0.0	3.5	0.0	0.2
	Denitrification (-)		0.0	0.0	23.4	47.9	0.0	0.7
	Storage change[#]		-0.2	72.7	-104.9	-39.7	58.3	102.8
Triticale	Fertilization* (+)	62.0+119.1	181.0	180.4	181.8	111.8	181.1	181.7
	Deposition (+)		5.9	3.5	16.7	4.6	6.1	11.7
24.9.2010 – 13.7.2011	Biological fixation (+)		0.0	12.8	0.0	0.0	0.0	18.2
	Volatilization (-)		8.1	0.0	4.6	5.5	41.4	19.8
	Crop off-take (-)	155.8	152.0	44.5	161.5	170.3	143.0	83.6
	NO_3-N leaching (-)	13.9	6.1	3.2	7.6	30.3	13.3	31.0
	Other transport[$] (-)		0.0	2.5	0.0	0.6	0.0	0.2
	Denitrification (-)		0.0	0.0	13.5	38.4	0.0	1.5
	Storage change[#]		20.7	146.5	11.2	-128.8	-10.4	75.5
				Seven year totals 2005– 2011				
All	Fertilization* (+)	871.6	879.0	870.9	873.5	656.1	871.4	888.2
	Deposition (+)		77.0	29.6	164.0	49.6	67.0	111.1
1.1.2005 – 13.7.2011	Biological fixation (+)		174.9	154.6	0.1	1.8	31.3	370.9
	Volatilization (-)		43.2	0.0	22.9	48.5	80.8	67.1
	Crop off-take (-)	768.8	792.1	584.8	921.2	594.8	568.4	664.7
	NO_3-N leaching (-)	123.3	130.3	158.6	109.1	147.3	102.2	209.0
	Other transport[$] (-)		0.0	31.7	0.1	10.0	0.0	0.8
	Denitrification (-)		0.0	0.3	161.3	249.2	0.0	10.0
	Storage change[#]		165.3	279.7	-177.0	-342.2	218.0	418.6

† + indicates input; - indicates output

* Fertilization includes the addition of mineral fertilizer (first number) and the amendment of animal manure (second number)

$ Other transport includes the leaching of NH_4-N and dissolved organic matter and the transport of N-components by surface runoff water flow

A positive value refers to an increase of the nitrogen stock in soil and a negative value indicates its depletion

Table 9. Qualitative assessment of the model performance (*IoA*) for daily or weekly results for the calibration and validation periods.

Phase	Indicator	Item		ARMOSA	COUP	DAISY	EPIC	SIM-STO	SW-ANIM
Calibration	-: $IoA < 0.6$ o: $0.6 \leq IoA < 0.8$ +: $IoA \geq 0.8$	Soil moisture retention relation	0.35 m	+	-	o	n.a.	o	+
			0.9 m	o	-	o	n.a.	+	+
			1.8 m	-	-	-	n.a.	o	+
		Simulated water contents	0.35 m	o	o	-	-	+	+
			0.9 m	o	+	o	+	+	+
			1.8 m	-	+	-	-	+	-
		Nitrate concentration	0.9 m	o	+	o	n.a.	-	+
	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Water flux, daily		o	o	-	o	o	o
		Water volumes per sampling interval		o	o	-	o	+	o
		Nitrate concentration in water samples		o	+	o	o	-	+
		Nitrate-N flux per sampling interval		o	o	o	o	-	o
Validation	-: $IoA < 0.6$ o: $0.6 \leq IoA < 0.8$ +: $IoA \geq 0.8$	Soil water contents	0.9 m	-	o	o	-	+	o
			1.8 m	-	+	-	-	+	o
		Nitrate	0.35 m	o	-	-	n.a.	o	-
			0.9 m	-	-	-	n.a.	o	-
	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Water flux, daily		o	o	o	o	+	o
		Water volume per sampling interval		-	o	-	-	o	-
		Nitrate concentration in water samples		-	-	o	-	o	-
		Nitrate-N flux per sampling interval		-	-	-	-	-	-

n.a.: not applicable

Table 10. Mean absolute errors (*MAE*) of seasonal percolated water, N crop off-take and leached nitrate-N amounts for seven seasons (MAE_7) and for the best five seasons (MAE_5).

Seasonal quantity	Indicators	**ARMOSA**	**COUP**	**DAISY**	**EPIC**	**SIM-STO**	**SW-ANIM**
Percolated water	MAE_7	21.3	24.2	63.9	48.6	14.6	40.3
(mm)	MAE_5	16.0	14.3	30.5	30.5	11.8	32.8
N crop off-take	MAE_7	36.5	32.7	47.7	31.0	33.0	21.5
(kg ha^{-1})	MAE_5	23.1	14.3	29.0	20.6	20.5	10.3
Leached NO_3-N	MAE_7	6.6	8.2	4.6	10.3	6.6	14.2
(kg ha^{-1})	MAE_5	4.4	3.6	3.7	7.8	2.8	6.3

Figures captions

Figure 1 Blind test comparison of seasonal water fluxes, flow averaged nitrate concentration and nitrate-N fluxes simulated by five models (excluding **SIM-STO**) with observations. Results of individual models are indicated by markers.

Figure 2 Measured values and calibrated soil moisture retention curves at depths 0.35 m,0.9 m and 1.8 m.

Figure 3 Comparison of simulated and measured inner season cumulative water fluxes, nitrate concentrations and inner season cumulative nitrate-N fluxes at depth 1.8 m in the low input farming lysimeter at the Wagna experimental field station

Figure 4 Taylor plots of the statistical performance of the simulated water fluxes at depth 1.8 m for daily values (left) and for sampling interval averaged values (right). Circles refer to the calibration results and triangles refer to the validation results. A = **ARMOSA**, C = **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**

Figure 5 Comparison of simulated and measured seasonal water fluxes (mm) at depth 1.8 m in the low input farming lysimeter at the Wagna experimental field station

Figure 6 Taylor plot of the statistical performance parameters for the simulated nitrate concentrations (left) and nitrate-N fluxes (right) at depth 1.8 m. Circles refer to the calibration results and Triangles refer to the validation results. Indicators of **SW-ANIM** nitrate-N fluxes fall outside the range (2.5; 8.5). A = **ARMOSA**, C = **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**

Figure 7 Seven years balances for fertilization minus crop off-take and nitrate-N leaching (all in kg ha^{-1}), summed since the start of the calibration period

Figure 8 Effect of a leave-one-out calculation of a certain data pair of observed and simulated water fluxes on the Index of Agreement, IoA (see text for further explanation).

Performance assessment of nitrate leaching models for highly vulnerable soils used in low input farming based on lysimeter data

Piet Groenendijk[a,*], Marius Heinen[a], Gernot Klammler[b], Johann Fank[b], Hans Kupfersberger[b], Vassilios Pisinaras[c], Alexandra Gemitzi[c], Salvador Peña-Haro[d], Alberto García-Prats[e], Manuel Pulido-Velazquez[f], Alessia Perego[g], Marco Acutis[g], Marco Trevisan[h]

a Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands

b Joanneum Research, Forschungsgesellschaft mbH, Leonhardstraße 59, 8010 Graz, Austria

c Democritus University of Thrace, Department of Environmental Engineering, Vas. Sofias 12, Xanthi, 67100, Greece

d Institute of Environmental Engineering, ETH Zurich, Wolfgang-Pauli-Str. 15, CH-8093 Zurich, Switzerland

Formatted: English (U.K.)

e Universitat Politècnica de València, Department of Hydraulic Engineering and Environment, Camino de Vera, 46022 Valencia, Valencia, Spain

Formatted: English (U.K.)

f Universitat Politècnica de València, Research Institute of Water and Environmental Engineering (IIAMA), Camino de Vera, 46022 Valencia, Valencia, Spain

g University of Milan, Department of Agricultural and Environmental Science, Via G. Celoria 2 20133, Milan, Italy

h Università Cattolica del Sacro Cuore, sede di Piacenza, Via Emilia Parmense, 84 29100, Piacenza, Italy

*Corresponding author: Piet Groenendijk, Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands, Email: piet.groenendijk@wur.nl Tel.: +31 317 486434

Abstract

The agricultural sector faces the challenge of ensuring food security without an excessive burden on the environment. Simulation models provide excellent instruments for researchers to gain more insight into relevant processes and best agricultural practices and provide tools for planners for decision making support. The extent to which models are capable of reliable extrapolation and prediction is important for exploring new farming systems or assessing the impacts of future land and climate changes.

A performance assessment was conducted by testing six detailed state-of-the-art models ~~with capabilities~~ for simulation of nitrate leaching (**ARMOSA, COUPMODEL, DAISY, EPIC, SIMWASER/STOTRASIM, SWAP/ANIMO**) for lysimeter data of the Wagna experimental field station in Eastern Austria, ~~in~~ where the soil is highly vulnerable to nitrate leaching.

Three consecutive phases were distinguished to gain insight in the predictive power of the models: 1) a blind test for 2005 – 2008 in which only soil hydraulic characteristics, meteorological data and information about the agricultural management were accessible; 2) a calibration for the same period in which essential information on field observations was additionally available to the modellers; and 3) a validation for 2009 – 2011 with the corresponding type of data available as for the blind test. A set of statistical metrics (mean absolute error, root mean squared error, index of agreement, model efficiency, root relative squared error, Pearson's linear correlation coefficient) was applied ~~defined~~ for testing the results and comparing the models.

None of the models performed good for all of the statistical metrics. Models designed for nitrate leaching in high input farming systems had difficulties in accurate predicting leaching in low input farming systems that are strongly influenced by the retention of nitrogen in catch crops and nitrogen fixation by legumes. An accurate calibration does not guarantee a good predictive power of the model. Nevertheless all models were able to identify years and crops with high and low leaching rates.

Keywords

Lysimeter, model comparison, nitrate leaching, performance assessment, predictive power, simulation model

1. Introduction

Agriculture is the major land use in Europe (ca. 50% of overall land area) and has strongly increased its use of external inputs (fertiliser, pesticides and water) over the last 50 years. The environmental effects of intensive agriculture include a decline in biodiversity, eutrophication of ecosystems and surface waters, acidification, global warming, air pollution and diffuse nitrate pollution of groundwater. A global challenge is to produce enough food for the ever-growing population and at the same time minimizing the loss of reactive nitrogen (N) to the environment. Since the 1980s, agriculture in Western Europe has managed to reduce its N surpluses, owing to stringent national and European community policies (Vitousek et al., 2009; Grizzetti et al., 2011).

The main aim of the Nitrates Directive (EU, 1991: Directive 91/676/EEC) is to reduce water pollution caused or induced by nitrates and phosphorus from agricultural sources. The Nitrates Directive legally restricts farm application of manure to 170 kg ha^{-1} of nitrogen, or in case of derogation to inputs up to 250 kg ha^{-1} (Oenema, 2004). An implementation measure of the Nitrates Directive is the establishment of codes of Good Agricultural Practice. Recommended measures include, among others, the application of crop rotations, the cultivation of a soil winter cover and catch crops to prevent nitrate leaching and run-off during wet seasons. Catch crops create a new challenge in the assessment of environmental effects of crop rotations. In theory, catch crops take up N that would otherwise be lost, and, after incorporation of the crop residues into the soil, make this N available to the succeeding crop via mineralization. However, the influence of a catch crop on the nitrogen supply to the succeeding crop can vary greatly and range from a positive to a negative effect (Nett et al., 2011). The effect is determined by the N uptake capacity, the rooting depth of a catch crop, the weather and soil conditions as well as the rooting depth of the succeeding crop (Thorup-Kristensen, 2006).

Models are an important tool for assessment of environmental impacts of a certain agricultural practice and are also an instrument for increasing the understanding of the biological, pedological and hydrological factors that affect productivity and the risk of nitrate leaching. For this reason, for more than 30 years simulation models have been developed and applied in the research on nitrate leaching. The different model descriptions are a reflection of the intended purpose, the physical conditions and the available data for model application and the knowledge and skill of the model developer. Technical implementations have evolved from stand-alone model codes to modelling platforms comprising modular models able to include and compare different process descriptions.

Calibration and validation of models contributes to their reliability. In addition also an analysis of the implemented process descriptions and the mutual comparison of models provides information on the predictive power. Several model comparison studies have been conducted in which nitrate leaching models were compared (De Willigen and Neeteson, 1985; Vereecken et al., 1991; De Willigen, 1991; Diekkrüger et al., 1995; Moreels et al., 2003; Kersebaum et al., 2007; Jabro et al., 2012). Most of them were related to ordinary agricultural conditions with a single crop on a typical agricultural soil. Thus, there is no information (comparison) available for situations in soils that are highly vulnerable to nitrate leaching in combination with low-input conditions and the use of catch crops.

It is widely recognised that despite the deterministic nature of process oriented models they often have a limited validity range for certain climatic, pedological, hydrological and agronomic circumstances characterised by high inputs. It is not clear whether the models are able to produce relatively reliable predictions for low input conditions. A better insight into the model performance for such uncommon circumstances underpins conclusions about the predictive power.

In this study a number of models were inter-compared for low input conditions of one of the lysimeters of the Wagna experimental research station, Austria (~~Fank and~~ Klammler and Fank, 2014; this issue) for three typical conditions for which they were not designed: 1) the crop rotation which included an uncommon crop (oil pumpkin), 2) catch crops for which the N-uptake was not measured, and 3) the soil consisted of a shallow soil vulnerable to nitrate leaching on top of a high conductive gravel layer. The objectives of this study were: 1) to assess the performance of state-of-the-art nitrate leaching models as they are used in the scientific research community, for the above mentioned conditions, 2) to inter-compare the models for analysing their predictive power, and 3) to identify strengths and weaknesses of bio-physically based models.

2. Materials and Methods

2.1 Description of the lysimeter

Observations were used of a lysimeter located in the agricultural experimental field station in Wagna in Eastern Austria (46° 46.113'N, 15° 33.140'E; altitude 265 m; Klammler and Fank, 2014 (this issue)). Since 1987 different cultivation strategies are investigated concerning nitrogen-fertilizer input, nitrate leaching and crop yields. In 2004, the cultivation changed into comparing low-input farming and organic farming, each covering 50% of the test site. Since then, two of the test plots have been equipped with two weighable, monolithic, high-precision lysimeters (2 m depth, 1 m² surface). The lysimeter in the conventional tillage test plot (KON-system) is subject for this study. Cultivation practices including crop species, sowing and harvest dates, and fertilizer applications in the test plot are presented in Table 1.

<<Table 1 >>

The lysimeters are equipped with soil water samplers, soil moisture probes, matrix sensors/tensiometer and soil temperature probes at four measuring depths (0.35, 0.6, 0.9, 1.8 m). An accompanied measuring profile for soil moisture, matrix potential and soil temperature is also installed outside the lysimeters (same depths as inside the lysimeter) to determine if the conditions inside the lysimeter are representative for the rest of the field. At the bottom of the lysimeter (depth 1.8 m) a suction cups rake was installed which kept the pressure head at this depth equal to that outside the lysimeter. The water sucked off was collected, weighted and sampled for the determination of the nitrate concentration. While quantity of seepage water was recorded automatically in 0.1 mm resolution by a tipping bucket, nitrogen concentration in the accumulated leachate was analysed in an approximately weekly interval. Furthermore, a weather station is installed at agricultural test site in Wagna for the recording of air temperature, relative humidity, shortwave solar radiation, wind speed, wind direction, precipitation, sunshine duration and atmospheric pressure at high temporal resolution (~~Fank and~~ Klammler and Fank, 2014; this issue). Annual precipitation rates and cumulative probabilities of the rates relative to the values of the period 1961 – 2011 are presented in Table 2.

<<Table 2>>

Annual rainfall amounts during the calibration years can be considered as moderate, the first year of the validation period is characterised by an extreme high rainfall and during the last year of the validation a low precipitation amount was recorded.

2.2 Description of models

This performance assessment study was conducted as part of the EU-FP7 GENESIS project (2009 – 2014) by six partners. Six well-known detailed models for European research on field-scale crop and soil water and soil nitrogen dynamics were chosen: **ARMOSA**, **CoupModel (COUP)**, **DAISY**, **EPIC**, **SIMWASER-STOTRASIM** and **SWAP-ANIMO**. It goes beyond the scope of this paper to give full details on the process descriptions of the six models used. Brief descriptions will be given in text and inter-comparison of processes and various other characteristics can be found in Supplemental Materials. All models are one-dimensional.

- **ARMOSA** has recently been developed specifically for the Lombardy region in Italy to assess the regional soil vulnerability to nitrate leaching (Perego et al., 2013). The model allows the simulation at field and multi-field level. The model is based on the **SWAP** (version 2.07) approach for simulating the water flow (Van Dam, 2000), on **STAMINA** for simulating the crop development and growth (Ferrara et al., 2011; Richter et al., 2010) and on **SOILN** for simulation of the soil organic matter and nitrogen cycle and nitrate leaching (Bergström et al., 1991).

- **CoupModel (COUP),** a coupled heat and mass transfer model for soil plant-atmosphere systems, was initially developed to simulate conditions in forest soils, but it has been further developed to simulate conditions in any type of soil, independent of plant cover (Jansson and Karlberg, 2004). COUP applicability is very wide as it includes water, heat, tracer, chloride, nitrogen and carbon modules that can be incorporated in the modelling process. COUP development, calibration procedures and applications are presented by Jansson (2012).
- **DAISY** is a soil-plant-atmosphere system model designed to simulate crop production, soil water dynamics, and nitrogen dynamics in crop production at various agricultural management practices and strategies (Hansen et al., 1990). The agricultural management model allows for building complex management scenarios (Hansen, 2002). The model has been validated in a number of major comparative tests (Diekkrüger et al., 1995; Hansen et al., 1991a,b; Jensen et al., 1997; Smith et al., 1997; Svendsen et al., 1995; Vereecken et al., 1991; De Willigen, 1991).
- **EPIC** (Williams et al., 1984; 1989) is a cropping systems simulation model, which was developed to estimate soil productivity as affected by erosion throughout the United States during the 1980's. **EPIC** is a field scale model, but linked to a GIS it has been applied in several regional model applications (Burkart et al., 1999; Sohier et al., 2009). Furthermore the **EPIC** model has been applied to study the effect of agricultural practices and biofuels cultivation on N leaching at the European scale (Bouraoui and Aloe, 2007; Van der Velde et al., 2009).
- **SIMWASER** (Stenitzer, 1988) simulates the water flow in soil. A unique feature of the model is the description of actual rooting depths based on both root biomass simulated for a crop and on the penetration resistance of the soil. **STOTRASIM** (Feichtinger, 1998) is fully coupled to **SIMWASER** and simulates nitrogen and basic carbon dynamics of agriculturally used soils. The model has already been applied to the region of southeast Styria (Fank et al., 2006). The name of these coupled models is abbreviated as **SIM-STO**.
- The **SWAP** model, version 3.2 (Van Dam et al., 2008) simulates water flow in the soil – plant – atmosphere domain in an integrated manner. The **ANIMO** model (Groenendijk et al., 2005) is sequentially coupled to **SWAP** and was designed to quantify the relation between fertiliser application rate, soil management and the leaching of nitrogen (N) and phosphorus (P) to groundwater and surface water systems. The **ANIMO** model

is part of the National Dutch modelling system **STONE** for the evaluation of fertiliser policy measures (Wolf et al., 2003). The name of the sequentially coupled models is abbreviated as **SW-ANIM**.

In addition to soil processes also the description of crop development is considered, because the plant related processes such as evaporation, nitrogen and nitrogen supply with crop residues exert a major influence on the water balance and nutrient dynamics in the soil.

Except for **SW-ANIM**, all models simulate the growth of plant biomass. Although **SW-ANIM** has the possibility to calculate the biomass development in a detailed manner, the modellers had chosen to use a simple option of a supposed development of leaf area index, crop height and rooting depth, because the parameters required for detailed simulation of oil pumpkin and catch crops were not available. Except for **EPIC**, the models describe water flow with either the Richards' (1931) equation or the Darcy (1856) - Buckingham (1907) equation, in which the soil water retention and the hydraulic conductivity relations are described according to Mualem (1976) - Van Genuchten (1980). **EPIC** simulates soil water flow as a storage routing process in which percolation occurs when the soil water content of the root zone exceeds the field capacity. In **EPIC** the soil water characteristics are calculated on the basis of texture data and the organic matter content in accordance with Saxton and Rawls (2006).

All models consider ammonium and nitrate as separate mineral nitrogen pools, and simulate organic bounded nitrogen associated with the organic carbon cycle. **SW-ANIM** simulates also the transport and transformation of dissolved organic nitrogen. The method of simulating biological N-fixation is one of the striking differences between the models. The **DAISY** model was applied in a way that biological N-fixation was ignored and the **SW-ANIM** model accounted for this process by the specification of continuous organic material additions representing imposed fixation rates. The other models use relationships based on the crop type, the crop development stage and the soil mineral N status. Ammonia volatilization is not implemented in the **COUP** model code used for this study. Some models consider only the loss of ammonia as a fraction of farmyard manure application (**DAISY**, **SW-ANIM**) while the other models take account for environmental factors as temperature, wind speed and soil moisture. **SIM-STO** uses standardized loss factors that account for the time from the last soil tillage event.

Uptake of ammonium and nitrate depends on the demand for mineral N for crop production and is related to the development stage, by some models expressed by a relationship with the water uptake, and the mineral N content of the soil.

Mineralisation is simulated in close correspondence to the organic matter cycle. All models describe the amount of mineralized nitrogen as the excess nitrogen produced from the organic matter decay and transformations to more stable soil organic matter pools. Nitrification is commonly described as a first order process which rate depends on temperature, soil moisture status and ammonium concentration. Denitrification plays no significant role in the soil of the Wagna lysimeters (Leis, 2009), but can be simulated by the models used. A variety of descriptions are implemented but all assume a relationship with temperature, soil moisture content, nitrate concentration and the potential denitrification rate as a function of organic matter content (Heinen, 2006).

The lysimeter was installed in 2004 and it was ensured that the original soil layers was put back. During the excavation and filling the soil had been in contact with open air. None of the models paid attention to this event in 2004. To establish the starting conditions on 1-1-2005, three of the six models (i.e., **ARMOSA EPIC**, **STO-SIM**, **SW-ANIM**) started in 1987. **COUP** was run for five years prior to the start in 2005 and **DAISY** was run two-years prior to the simulation**.**

2.3 Experimental design of study

The modelling study comprised of: 1) a blind test with non-calibrated models to get an impression of the performance of the models as they are used in situations where extensive data sets are missing, which often occurs in practice, 2) a calibration period, and 3) a validation period. Inter-comparisons were done between measured and simulated leaching of water and nitrate, including nitrate concentration of the percolate. The outcome of the simulations by all models was collected and analysed by a single person.

2.3.1 Step 1: Blind test

The models first performed a simulation based on a minimum set of data: crop rotation, soil cultivation, fertilization rates, meteorological data, soil profile description and soil moisture retention laboratory measurements of some soil samples. The aim is to establish the bandwidth of differences with the observations without an assessment of the individual models. The **SIM-STO** model was excluded from the blind test as the operators of this model were the owners of all data and **SIM-STO** was already partly calibrated for the test site. After all models delivered their outcome, one external operator compared the predictions against the measured data (seasonal cumulated water flux and nitrogen flux at the bottom of the lysimeter, seasonal flow averaged nitrate concentration) for the period 2005 - 2008. It was not the intention of the blind test to qualify or assess the performance of the individual models and, therefore, the outcome of this test will be presented anonymously.

Specifically only data on seasonal percolation, flow-averaged nitrate concentration and seasonal nitrate leaching were considered.

2.3.2 Step 2: Calibration

Each of the six modelling groups calibrated the models for a limited number of parameters. The successive operations, the objective function and the number of parameters were not prescribed, but were chosen freely by the modelling groups, either based on expert judgement or on a sensitivity analysis. Further details of how the calibration has been carried out for the different models can be found in Supplemental~~ry~~ Materials~~ 2~~.

2.3.3 Step 3: Validation

The validation was performed for the period 2009 - 2011, where only information about crop rotation, application of fertilizers, soil cultivation and meteorology was made available for the modelling groups after step 2 (calibration) was finished. The procedure for the validation is thus similar to that of the blind test, with the difference that the models were calibrated prior to validation and that the **SIM-STO** model was included in the validation.

2.3.4 Step 4: Model comparison

The six models were compared for their performance with respect to 1) the soil moisture retention curves at depths 0.35, 0.90 and 1.8 m; 2) the volumetric water contents at depths 0.35, 0.9 and 1.8 m; 3) the nitrate concentrations at depths 0.35, 0.9 and 1.8 m ; 4) the daily water fluxes at depth 1.8 m; 5) the leached water amounts for the time intervals of collected water samples; 6) the nitrate concentrations of the collected water samples; 7) the nitrate-N fluxes at the bottom of the lysimeter for the time intervals of collected water samples. The comparison of results at the depth of 60 cm was excluded because measurements for this depth were only available up to Sept. 2009. Seasonal leached water amounts, nitrogen yields and nitrate-N fluxes were compared to discuss the predictive power for practice oriented model applications. A nitrogen balance was set up for all models. Water fluxes at 1.8 m depth were evaluated for daily and for seasonal values. Nitrate leaching fluxes and nitrate concentrations in the leachate were evaluated at the time intervals for which the soil water was sampled. The sampling time intervals were irregular in time and the models were not able to present concentrations at these specific time events. Therefore, concentrations values for these time intervals were derived according to a volumetric averaging procedure. The nitrate concentrations at depths 0.35 m and 0.9 m can be used to get an impression whether the transport and transformation processes in soil, which ultimately lead to the leaching at depth 1.8 m, have been described adequately. Due to the nature of the model formulations, **EPIC** was not able to present the concentrations at the depths of measurement. The number of observations at depth 0.35 m in the calibration period was too little and were not considered.

In the models, much knowledge of soil processes is described which all contribute to the nitrate leaching at depth 1.8 m. To understand the similarities and differences between simulation results and measurements, it is important to assess the processes. We have done this through the establishment of nitrogen balances per season.

2.4 Statistical metrics

The behaviour of the main model outputs can be characterized by a number of statistical metrics to indicate the models' ability to capture different aspects. A complete assessment of model performance should include at least one absolute error measure and one goodness-of-fit measure (Legates and McCabe, 1999). There are a wide range of statistical indicators used in studies on soil water and soil nitrogen, but not always a justification is given for the indicators chosen. For *state variables* many authors use mean (absolute) error (*M*(*A*)*E*), root mean square error (*RMSE*), index of agreement (*IoA*; Willmott, 1982), and less often the Nash-Sutcliffe modelling efficiency (*NSE*; Nash-Sutcliffe, 1970) (e.g., Donatelli et al., 2004; Gribb et al., 2009; Herbst et al., 2005; Khodaverdiloo et al., 2011; Patil and Rajput, 2009; Ritter et al., 2003; Vereecken et al., 2010). For *rate variables* authors generally use *MAE*, mean difference (*MD*), absolute maximum error (*AME*), *RMSE*, *IoA*, *NSE*, coefficient of determination (R^2), percentage of error (*PE*), percentage of bias (P_{bias}) (e.g., Akkal-Corfini et al., 2010; Ale et al., 2012; Dawson et al., 2007, 2010; Jabro et al., 2012; Jachner et al., 2007; Kersebaum et al., 2007; Krause et al., 2005; Moriasi et al., 2007; Qi et al., 2012; Reusser et al., 2009; Stumpp et al., 2009; Van der Laan et al., 2011; Wang et al., 2006; Willmott et al., 1985). It appears that a few measures are used both for state as for rate variables, which we have chosen to use here as well: *MAE*, *RMSE*, *IoA*, and *NSE* (only for rates), given by:

1. Mean absolute error: $MAE = \frac{1}{n}\sum_{t=1}^{n}|P_t - O_t|$
2. Root mean squared error: $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(P_t - O_t)^2}$
3. Index of Agreement (Willmott, 1982): $IoA = 1 - \frac{\sum_{t=1}^{n}(P_t-O_t)^2}{\sum_{t=1}^{n}(|P_t-O_t|+|O_t-\bar{O}|)^2}$
4. Nash-Sutcliffe model efficiency (Nash and Sutcliffe, 1970): $NSE = 1 - \frac{\sum_{t=1}^{n}(O_t-P_t)^2}{\sum_{t=1}^{n}(O_t-\bar{O})^2}$

where *n* is the number of observations, O_t is the observed value, P_t is the model predicted value, and $\bar{O}$ and $\bar{P}$ are the mean values of observations and predictions, respectively. All four measures compare the predictions P_t and observations O_t at the individual level, and try to express the 'spread' in ($P_t – O_t$) (Janssen and Heuberger, 1995). The *MAE* accounts for the deviations ($P_t – O_t$) in an absolute value sense. This measure is less sensitive to outliers than *RMSE*, *IoA* and *NSE*. The latter indices measures ($P_t – O_t$) in a quadratic sense, and, thus, are sensitive to outliers. If model errors are significant, it is more difficult to objectively assess the agreement between model and data on basis of *RMSE*. As an alternative, Willmott (1982) proposed *IoA* to express this agreement more directly. The dimensionless *IoA* has limits 0, indicating no agreement, and 1, indicating perfect agreement. The dimensionless *NSE* ranges between 1 and −∞, where *NSE* = 1 denotes a "perfect" model fit and for *NSE* < 0 the average of the observations would be a better predictor than the model (Krause et al., 2005).

Taylor (2001) presented a graphical method in which several statistical metrics have been combined. Such a Taylor diagram summarizes how closely a set of simulations matches the observations, and it is especially useful in evaluating multiple aspects of complex models. In normalized form, it presents the Pearson's linear correlation coefficient (*R*) and the root relative square error (*RRSE*) as a function of the ratio of standard deviations of predictions and observations σ_P and σ_O, respectively, where

5. Pearson's linear correlation coefficient $$R = \frac{\sum_{t=1}^{n}(O_t-\overline{O})(P_t-\overline{P})}{\sqrt{\sum_{t=1}^{n}(O_t-\overline{O})^2}\sqrt{\sum_{t=1}^{n}(P_t-\overline{P})^2}}$$

6. Root relative square error: $$RRSE = \frac{\sqrt{\sigma_P^2+\sigma_O^2+2\sigma_O\sigma_P R}}{\sigma_O}$$

where σ_O and σ_P are the standard deviations of the observations and model predictions, respectively. A value of (1,0) in such a figure indicates a full agreement of model results with observations.

3. Results and discussion

3.1. Blind test

Figure 1 presents the range of predicted seasonal water fluxes, flow-averaged nitrate concentration and nitrate-N fluxes by the five models considered as compared to the observations for the blind test period.

<<Figure 1>>

Maximum deviations between simulated and observed seasonal percolation volumes of almost 400 mm were found. Two of the five models showed a relatively good agreement of the seasonal percolation with the measurements. Three of the five models overestimated the percolation in all seasons. One model underestimated the percolation volume in all seasons and only one model was able to simulate the seasonal percolation accurately. The range of model results was independent of the seasonal percolation.

Seasonal flow averaged nitrate concentrations were underestimated by all models in two of the four seasons. For the first season, all models underestimated the concentration by 10 – 40 mg L^{-1}. The variation of simulated concentrations and N-fluxes was large. Maximum deviations of seasonal nitrate-N leaching of about 25 kg ha^{-1} were found. All models underestimated the leaching rate in 2005 by 8 – 22 kg ha^{-1}. The same holds for the fourth season, but only one model was able to calculate the nitrate-N flux with a reasonable agreement with the measurements. In the second season (maize), four models underestimated and one model overestimated the nitrate concentration and nitrate-N flux. The third season, which was the second season with maize showed a rather different pattern. The measured nitrate concentration and nitrate-N flux under maize in the 3rd season was much lower than for the maize crop in the 2nd season, but the modelled results still showed a large variation with a less skewed distribution of underestimation and overestimation. In the blind test information was lacking about crop-uptake rates and the nitrogen excess per season. The results showed that without this information and without a proper calibration the models were not able to predict nitrate concentrations and leaching rates accurately.

3.2 Calibration and validation

3.2.1 Soil water ~~moisture~~ and soil physical relations

In the blind test the modellers had only laboratory measurements of the water retention curve at their disposal, but in the calibration phase also in situ measured soil moisture contents (θ) and pressure heads (*h*) were available at four depths. The laboratory measurements were performed for drying samples only, while under field conditions data pairs of θ(*h*) were detected during wetting and drying cycles so that these were affected by hysteresis (Basile et al., 2003, 2006). Figure 2 depicts the calibrated θ(*h*) curves for three depths. The results at the depth of 0.6 m were comparable to the results of 0.35 m deep and are not shown here. The observed *h* at depth 0.35 m ranged from -20 cm to -2000 cm. At depth 0.9 m *h* ranged from -2 cm to -1000 cm and at depth 1.8 m *h* ranged from -10 to -100 cm. The variation of the θ(*h*) observed population is largest at depth 0.35 m.

<<Figure 2>>

Results for the **EPIC** model are represented by three points as EPIC does not use a continuous description of the θ(*h*) curve. The greatest value for the saturated water content was obtained by the **EPIC** model with a value greater than 0.3 cm^3 cm^{-3} at depth 1.8 m. This parameter is far outside the range that was established by the other models. A comparison between the calibrated and observed θ(*h*) curves was made by calculating a θ for each value of the measured *h*. The performing indices based on computed θ and measured θ are presented in Table 3.

<<Table 3 >>

In general the resulting *MAE*, *RMSE* and *IoA* showed equal trends. The **ARMOSA** model fitted well at depths 0.35 m and 0.9 m, but performed worse at depth 1.8 m. The performance of the **COUP** model appeared to be weak. At depth 0.9 m the **DAISY** model was better than the **COUP** model, but worse than the other models. The *IoA* for the **SIM-STO** and **SW-ANIM** models was highest at depth 0.9 m and somewhat lower for the other depths. It should be noted that a good match of the calibrated θ(*h*) curves with measured data pairs does not a-priori mean that a good agreement between the time series of measured and calculated θ will be obtained.

The simulated θ was compared with daily averaged values of measured θ (Table 4). For depth 0.35 m an increasing trend was detected from 2008 and onwards which is attributed to the aging of the sensor, and, therefore, the results for this depth were disqualified for the validation period.

<<Table 4>>

Except for **ARMOSA** and **EPIC** in the validation phase, the highest *IoA* values for simulation of the water contents were achieved at depth 0.9 m. For **SIM-STO** and **SW-ANIM**, the *IoA* values were similar to the calibration results of the θ(*h*) curves (Table 3). However, the performance by **COUP** increased and that by **DAISY** decreased compared to Table 3. Except for the **ARMOSA** and the **DAISY** models at depth 0.35 m and the **SW-ANIM** model at depth 1.8 m, in general the resulting performance indices showed a better agreement between simulated and observed values for the period 2005 – 2008 than for the comparison based on soil moisture retention curves. The indices of the validation period 2009 – 2011 were in the same range, or somewhat lower at depth 0.9 m, as for the calibration period (Table 4).

Figure 3 presents the cumulative water fluxes as predicted by the models and as measured as a function of time.

<<Figure 3>>

The pattern of cumulative water fluxes per growing season complies generally with the annual precipitation amounts (Table 2) with the exception of maize in 2006 and its preceding crop in the winter of 2005/2006. During the intermediate period after oil pumpkin in 2005 and before maize in 2006, the precipitation amounted to about

430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative water flux per season corresponded well to the measured water fluxes for most of the models which is also true for the extreme wet year 2009. However, DAISY showed some overestimation in particular seasons due to difficulties in parameterizing maize crop parameters. EPIC performed better in the calibration than in the validation period~~The simulated cumulative water flux per season corresponded well to the measured water fluxes for most of the models: **DAISY** underestimated the water flux under maize in 2006 and 2010, while **EPIC** overestimated the water flux for most of the growing seasons.~~. **SW-ANIM** underestimated the cumulative water flux in the two first seasons, but overestimated slightly in some other seasons. ~~.~~

~~Except for the **EPIC** model, the cumulative water fluxes in the extreme wet year 2009 were simulated well by the models.~~ No model was able to simulate the dry no-flux period during the second half of 2011. Deviations between the simulated and observed soil moisture contents were relatively small and have a limited impact on the cumulative water fluxes. Underestimations and overestimations of the seasonal water fluxes are explained by overestimation and underestimations of the seasonal evapotranspiration. This depends on the difficulty of establishing accurate crop growth parameters. Table 5 presents the statistical performance indices for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

<<Table 5>>

The performance improved for the averaged fluxes per sampling period of the calibration phase relative to the performance of the daily fluxes, but deteriorated for the validation phase. This is counter-intuitive because the peaks of the daily fluxes pattern are flattened by aggregation and one should expect a better performance for the averaged values per sampling interval.

Figure 4 presents the Taylor diagrams for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

<<Figure 4>>

For all models the *R*-values were between 0.5 and 0.9 and the *RRSE*-values were between 0.5 and 1.0. For daily water fluxes the σ_P/σ_O-ratio for the validation period was somewhat higher than for the calibration period, but for the fluxes averaged for the sampling intervals it can be seen that **ARMOSA**, **DAISY**, **COUP** and **EPIC** resulted in lower σ_P/σ_O-ratio's for the validation period than for the calibration period.

The range of seasonal water fluxes for the cultivation periods predicted by the models for all seasons was around the observed values (Figure 5). With respect to the blind test, calibration of the models resulted in a smaller range and in a shift towards the observations.

<<Figure 5>>

The ranges were relatively large for maize and its preceding catch crops in 2006 and 2010. In four of the seven seasons **DAISY** had the lowest value for the leaching and in one season the highest value. Both **COUP** and **EPIC** resulted in three seasons the highest value. **SIM-STO** had the smallest deviation between predicted and measured seasonal water leaching and **DAISY** resulted in the largest deviation

Differences between observed and model predicted water contents, water fluxes and water volumes per sampling interval indicate over- or under-estimation of the water excess in the soil column. Besides uncertainties in soil hydraulic properties and in observations, there was also lack of information about actual plant and root system development as a function of time.

The different modelling groups were not able to find a simultaneous optimal solution which minimizes both water contents deviations and water flux deviations. This may be due to uncertainties in soil hydraulic properties, and the disregarding of hysteresis in the models. The soil at the Wagna experimental station consists of a clayey-sand on top of a gravel layer. Durner et al. (2007) concluded that for layered soils with distinct heterogeneity no unique effective soil hydraulic properties exist. If only fluxes across the boundaries of the system are required, heterogeneous systems can be modelled with quasi-homogeneous ones, even if the internal system state is not matched properly. However, for nutrient dynamics (solute dispersion, biological and chemical reactions) an accurate internal system state description is mandatory (Durner et al., 2007)

3.2.2 Soil temperature

The soil temperature is an important variable determining the rate of biological processes (N dynamics), for the crop development in the period of germination, and for soil moisture flow under winter conditions. A comparison of simulated and measured soil temperatures was carried out as well (data not shown). In general, the models were well able to simulate soil temperatures and resulted in performance indices much higher than for moisture contents. The simulation performance at shallow depth was less than the performance at greater depths: most models showed a delayed warming up in some spring seasons with respect to the measurements, which is attributed to the incomplete description of surface temperatures, for most of the models used the air temperature as the boundary condition.

3.2.3 Water fluxes

Figure 3 presents the cumulative water fluxes as predicted by the models and as measured as a function of time.

<<Figure 3>>

The pattern of cumulative water fluxes per growing season complies generally with the annual precipitation amounts (Table 2) with the exception of maize in 2006 and its preceding crop in the winter of 2005/2006. During the intermediate period after oil pumpkin in 2005 and before maize in 2006, the precipitation amounted to about 430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative water flux per season corresponded well to the measured water fluxes for most of the models: **DAISY** underestimated the water flux under maize in 2006 and 2010, while **EPIC** overestimated the water flux for most of the growing seasons. **SW-ANIM** underestimated the cumulative water flux in the two first seasons, but overestimated slightly in some other seasons.

Except for the **EPIC** model, the cumulative water fluxes in the extreme wet year 2009 were simulated well by the models. No model was able to simulate the dry no-flux period during the second half of 2011.

Table 5 presents the statistical performance indices for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

<<Table 5>>

The performance improved for the averaged fluxes per sampling period of the calibration phase relative to the performance of the daily fluxes, but deteriorated for the validation phase. This is counter-intuitive because the peaks of the daily fluxes pattern are flattened by aggregation and one should expect a better performance for the averaged values per sampling interval.

Figure 4 presents the Taylor diagrams for the daily water fluxes and for averaged water fluxes per sampling interval for both the calibration and the validation periods.

~~<<Figure 4>>~~

~~For all models the *R*-values were between 0.5 and 0.9 and the *RRSE*-values were between 0.5 and 1.0. For daily water fluxes the σ_P/σ_O-ratio for the validation period was somewhat higher than for the calibration period, but for the fluxes averaged for the sampling intervals it can be seen that **ARMOSA**, **DAISY**, **COUP** and **EPIC** resulted in lower σ_P/σ_O-ratio's for the validation period than for the calibration period.~~

~~The range of seasonal water fluxes for the cultivation periods predicted by the models for all seasons was around the observed values (Figure 5). With respect to the blind test, calibration of the models resulted in a smaller range and in a shift towards the observations.~~

~~<<Figure 5>>~~

~~The ranges were relatively large for maize and its preceding catch crops in 2006 and 2010. In four of the seven seasons **DAISY** had the lowest value for the leaching and in one season the highest value. Both **COUP** and **EPIC** resulted in three seasons the highest value. **SIM-STO** had the smallest deviation between predicted and measured seasonal water leaching and **DAISY** resulted in the larges~~

3.2.34 Nitrate concentrations and nitrate-N fluxes

Figure 3 presents the cumulative nitrate fluxes and the nitrate concentration of the leachate as predicted by the models and as measured as a function of time. Based on a visual inspection the nitrate concentrations are simulated well by **COUP** and **SW-ANIM** for the calibration period. The **SIM-STO** results for this period were poor and the results of the other models were in between. The results for the validation period showed a completely different picture when compared to the corresponding results for the calibration period. The results of **DAISY** and **SIM-STO** were relatively the best, while **EPIC** and **SW-ANIM** results were weak. **ARMOSA**, **COUP** and **SW-ANIM** overestimated the concentration peak in autumn 2009 and **SW-ANIM** simulated a peak for autumn 2010, while there was no peak visible in the measurements.

ARMOSA, **DAISY**, **EPIC** and **SIM-STO** showed more spiky results for the calibration period than the measured values, while **COUP** and **SW-ANIM** showed calmer and more evenly time courses. The results resembled partly the modeller's choice for defining either the nitrate fluxes or the nitrate concentrations in the objective function of the calibration procedure. The **COUP** and **SW-ANIM** modellers used the nitrate concentrations for calibrations, while the **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** modelling groups used the nitrate fluxes. For **DAISY** and **EPIC**, the nitrate concentrations were calculated afterwards by dividing the nitrate flux by the water flux. The nitrate concentrations in the calibration phase simulated by **SIM-STO** showed a bad performance, while the results for the validation phase were much better. The higher peak concentrations during the calibration phase were not approached by **SIM-STO**. On the other hand, **SW-ANIM** showed a good agreement of nitrate concentrations during the calibration phase, while there is a mismatch during the validation phase. The concentration peaks during the validation phase were severely overestimated by **SW-ANIM** due to an overestimation of the biological fixation rates of some non-leguminous catch crops in this period.

The nitrate-N flux at depth 1.8 m represents the nitrogen transport to deeper soil layers and is relevant for predictions of nitrate concentrations in deeper groundwater. **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** underestimated the nitrate N-flux under winter barley preceded by a catch crop in 2007-2008, but **SW-ANIM** overestimated the nitrate N-flux during this period. The **COUP** model was able to calculate the nitrate-N flux in five of the seven seasons that cover the calibration and validation period. **ARMOSA** and **DAISY** calculated the total seasonal nitrate-N flux well in three of the seven seasons, while **EPIC**, **SIM-STO** and **SW-ANIM** calculated this flux well in two of the seven seasons. The last season appeared to be the most difficult one, because of the exceptional dry conditions. The leaching after the 2009 oil pumpkin crop also showed significant deviations between model predictions and measurements. The largest deviations of seasonal nitrate-N fluxes occurred in the results of **COUP** and **SW-ANIM** for the exceptional wet year 2009.

Table 6 presents the statistical indicators for both the nitrate concentrations and the nitrate-N leaching rates, based on the sampling time series. The largest deviations between predicted and simulated nitrate concentrations were found for the **SIM-STO** results in the calibration period for which the *IoA* ~~end the *NSE*~~ amounted to 0.43. ~~and -0.76, respectively.~~ Remarkably the smallest deviations were found for the same model for the validation period for which *IoA* ~~and *NSE*~~ amounted to 0.78. The underestimation of the nitrate-N flux by **SIM-STO** is most likely due to immobilization processes that are overemphasized for the 2005 and 2008 periods. Thus, less nitrate was released to the soil water phase which led to the underestimation of the nitrate concentration in the leachate. ~~and 0.08, respectivel.y~~

<<Table 6>>

The **COUP** model showed the best performance for the nitrate concentrations of the calibration period with *IoA* = 0.97 ~~and *NSE* = 0.86~~ directly followed by the **SW-ANIM** model. The results from **EPIC** and **SW-ANIM** for concentrations in the validation period were weak with *RMSE* > 20 mg L^{-1} ~~and *NSE* < -2~~. The statistical indices of the nitrate-N leaching rates showed a similar picture. The **SIM-STO** model performed relatively weak during the calibration phase. For the leaching rates in this period **DAISY** and **SW-ANIM** had the best performance and for the validation period **ARMOSA** and **DAISY** performed relatively the best. The *NSE* values (data not shown) for both the concentration and the leaching rates in the validation period were almost all negative, showing that the calibrated models had great difficulties to predict concentrations and leaching rates for the more extreme conditions of the validation period.

Statistical performance of predicted nitrate concentrations and leaching rates were expressed in Taylor diagrams in Figure 6. Calibrated nitrate concentrations by **COUP** and **SW-ANIM** had *R*-values greater than 0.9 and were closest to the (1,0) point. Except for **SIM-STO**, the models showed σ_P/σ_O ratios for the calibration step that did not deviate much from 1; for **SIM-STO** the σ_P/σ_O ratio was much lower than 1 and $R < 0$.

<<Figure 6>>

The plots clearly show the much weaker performance for the validation period than for the calibration period, expressed by lower *R*-values and higher σ_P/σ_O ratio's. **SIM-STO** showed the best performance for concentrations in the validation period with $R > 0.7$, σ_P/σ_O close to one, and $RRSE = 0.75$, while for the other model $RRSE > 1$. For the nitrate fluxes in the calibration period *RRSE* values were between 0.64 and 0.86, while for the validation period, the values were between 1 and 2 even with a peak of 8.6 for **SW-ANIM** (data point not seen in Figure 6). The *R*-values of the nitrate fluxes in the validation period were in the range 0.18 (**EPIC**) to 0.50 (**COUP**). The σ_P/σ_O ratio ratios were in the range 0.75 to 2.3 with a peak of 8.8 for **SW-ANIM** (data point not seen in Figure 6). The values for σ_P/σ_O ratio greater than 1 for both the concentrations and the nitrate fluxes indicate that the variation of the simulated values is greater than the variation of the observed values.

Table 7 presents the performance indices for the nitrate concentrations at depths 0.35 m and 0.9 m. The *IoA* values indicate that the best agreement between simulated and measured values was achieved for the calibration period, but *MAE*-values and *RMSE*-values were highest for the calibration results at depth 0.9 m and lowest for the validation results at depth 0.9 m. This apparent contradiction is due to the number of measurements on which the indices were calculated. Further analysis was based on *IoA* because the ranking of these values corresponded better to the results of the leaching water at depth 1.80 m.

<<Table 7>>

Calibrated concentrations yielded *IoA*-values ranging from 0.44 (**SIM-STO**) to 0.84 (**SW-ANIM**). The results for the validation period resulted in somewhat lower *IoA* values, except for **SIM-STO** which shows better results for the validation than for the calibration period. The **ARMOSA** results were the most constant for the different depths and periods. Both **COUP** and **SW-ANIM** show significantly poorer *IoA* values for the validation than for the calibration period. **DAISY** and **SIM-STO** showed slightly better results for the concentrations at depth 0.9 m than for the concentrations at depth 0.35 m. The other models performed slightly better for depth 0.35 m. Except for **SW-ANIM**, the *IoA* for the validation period at 0.35 m were in the same range as for the results at depth 0.9 m.

Over- and overestimation of simulated average nitrate concentrations and nitrate-N leaching rates for the calibration period is due to a number of reasons. A formal reason is the formulation of the object function. The calibration method applied for most models attempted to minimize the sum of squared differences $(P_t\text{-}O_t)^2$ for either the nitrate concentrations or the nitrate-N fluxes. A minimal sum does not guarantee a perfect match of the average concentrations. The different modelling groups have chosen different objective functions when calibrating for nitrate observations. Most models based the summation $(P_t\text{-}O_t)^2$ values on the sampling periods but **SIM-STO** used the summed $(P_t\text{-}O_t)^2$ values for the nitrate-N leaching rate per growing season only.

Three~~wo~~ out of four models that used nitrate flux in their objective function resulted in ~~moderate~~ *IoA* values in the range 0.76-0.87 for the calibrated nitrate fluxes, while the other~~s~~ model resulted in ~~poor~~ *IoA* = 0.43~~values~~ (Table ~~9~~6). Two out of three models that used nitrate concentration in their objective function resulted in ~~good~~ *IoA* values in the range 0.95-0.97, while the third model resulted in ~~a moderate~~ *IoA* = 0.87 ~~value~~ (Table ~~9~~6). However, a good calibration on nitrate concentrations did not result in good performance on nitrate fluxes. Both for the calibration and for the validation periods it appeared that all models had difficulties in predicting the nitrate fluxes at the bottom of the lysimeter, even if some of them were calibrated based on the measured nitrate fluxes.

Vereecken et al. (1991) evaluated five complex models from which **SW-ANIM, EPIC** and **DAISY** are also included in our performance assessment**.** A comparison between simulated and observed nitrate leaching rates measured in two sandy soils in Denmark and one sandy soil in the Netherlands revealed that **SW-ANIM, EPIC** and **DAISY** performed similar, although **DAISY** appeared to be a bit superior in behaviour. In general much better statistical metric values were reported than in our study. This may be due to the circumstances of the field trials which were representative for conventional agriculture during the eighties and because the calibration and he comparison was carried out for seasonal values.

Diekkrüger et al, (1995) compared the results produced by 19 simulation models, others than those used in this study, for a loam soil and a sand soil in Southern and Eastern Saxony in Germany. Variation in the leaching rates at 0.9 m depth reflected mainly the differences in soil water fluxes at that depth. Apart from the seasonal differences between the models that were able to simulate a three year period continuously, the cumulative leaching was nearly the same for these models. The results of soil nitrogen simulations were significantly influenced by the results of water flow and plant growth simulations. Diekkrüger et al, (1995) concluded that for long term forecasts the exact determination of the boundary conditions is as important as the model approach itself. Our finding that the unmeasured inputs concerning biological N-fixation are important for the soil nitrogen dynamics is consistent with this conclusion. In our study, differences between model seasonal and long term results are attributed to some extend to different assumptions about fixation rates.

Kersebaum et al., (2007) conducted a comparison of simulation models for 18 different models from which **SW-ANIM** and **SIM-STO** are included in our study. **SW-ANIM** was applied to the Müncheberg data-set (Kroes and Roelsma, 2007) and **SIM-STO** was applied to the data-set of the lysimeter station Berlin-Dahlem for water flow simulation and to the Bad Lauchstädt data-set for simulation of soil nitrogen dynamics (Stenitzer et al., 2007). Results for the mean bias, *RMSE*, *IoA* and *NSE* showed weak performances for the soil mineral nitrogen simulation in the 0-90 cm upper soil layer for nearly all models which were subjected to the Müncheberg data-set. Kersebaum et al. (2007) concluded that comparison of simulated results by models which are intended for field scale and regional scale with measured data often shows unsatisfactory results due to deviating conditions and parameters. It does not automatically mean that the models or the parameters are wrong because the data and parameters are only partly related to the site specific conditions of the measurements. In our study significant amount of data was available, but critical information about rooting depth and pattern, atmospheric deposition rates, mineralization and fixation rates was missing as well as the nitrogen uptake rates and residue amounts of the catch crops. Due to these uncertainties, it is difficult to draw clear conclusions about the predictive power of the models.

3.2.45 Nitrogen balances

Table 8 presents the soil nitrogen balances per season for each of the models.

<<Table 8>>

Exact fertilizer and manure inputs were not represented by **EPIC**, because the model assumes standard compositions which are not equal to the experimental data. This holds also for **SW-ANIM** which assumes fixed nitrogen compositions but this was overcome by introducing new manure types, so that the fertilizer input was close to the observed values.

The estimates for atmospheric deposition ranged from 4.2 kg ha^{-1} a^{-1} (**COUP)** to 23.4 kg ha^{-1} a^{-1} (**DAISY)**, averaged for seven growing seasons. Only literature values were available and most modelling groups have used the model default values or the figure they are familiar with for their own country. **ARMOSA** calculated for the validation phase lower wet deposition rates than for the calibration phase due to lower precipitation amounts. Some models assumed only dry deposition at a constant rate, while other models also imposed nitrogen input by rainfall.

The most stressing differences are for biological N-fixation. Some models do not describe the biological N-fixation process as such but modellers had possibilities to assume fixation rates by introducing a nitrogen rich organic material which was amended continuously during the growing season. The **DAISY** and the **EPIC** modelling groups did not take account for N-fixation, either due to a lack model formulations implemented or to a lack of knowledge about this process. **SIM-STO** assumed only for the first season some biological N-fixation by the crop mixture that included white clover. The **COUP** and the **SW-ANIM** modelling groups took account for N-fixation, including for periods for which one wouldn't expect (English ryegrass). In **SW-ANIM** the biological N-fixation is lumped with the mineralization of some of the crop residues that descended from the most recent and previous catch crops. The model output does not allow to unravel the biological N-fixation as such and mineralisation of earlier catch crop residues.

The **COUP** model did not take account for ammonia volatilization. The other models did, and showed a range of 2% to 35% of the nitrogen in the animal manure amended to the soil. The highest volatilization rates were simulated by **SIM-STO**: 27% and 35% of the animal manure N in 2008 and 2011, respectively. This could possibly explain the underestimation of nitrate leaching in 2008, but not in 2011. For these years, the differences of the model predictions amounted to more than 22 and 37 kg ha^{-1} a^{-1}, respectively, which is higher or in the same range as the measured nitrate-N leaching. Volatilization was calculated by **EPIC** and **ARMOSA** (about 4 kg ha^{-1}) for the first growing season of the validation period, while no farm fertilizer was applied.

The models encountered difficulties with the simulation of nitrogen crop off-take. Deviations of simulated uptake rates from the observed values of more than 50 kg ha^{-1} occurred for three years by **ARMOSA** (2006, 2008, 2009), **EPIC** (2005, 2009, 2010) and **SIM-STO** (2006, 2008, 2010), for two years by **DAISY** (2007, 2010), and for one year (2011) by **COUP** and **SW-ANIM**. The **EPIC** model was not able to simulate nitrogen crop off-take by oil pumpkin, because this crop is unknown in the standard database of crop parameters that comes with the model. The **DAISY** model failed to simulate a reasonable crop off-take by maize in 2007, while the N off-take in the preceding year was overestimated by 60 kg ha^{-1}. The calibrated parameters for crop uptake were not optimal for the maize as is also apparent from the calculated crop off-take in 2010 where the overestimation amounted nearly 100 kg ha^{-1}. Despite the fact that **SW-ANIM** included the N-yield in the object function of the calibration procedure, the modelled crop off-take differed from the measured crop off-take by -14 to +19 kg ha^{-1}. The **SW-ANIM** underestimated crop off-take in the validation period. Crop off-take is governing the soil nitrogen balance to a large extent and an erroneous calculation of the N off-take means that a possible correct nitrate leaching should be considered as little robust.

Denitrification is only of significance for the **DAISY** and **EPIC** results, while other models simulated zero or negligible denitrification rates. For most of the models, these estimates were biased by the opinion of the data holders who made plausible from their analysis of soil nitrogen balances that denitrification is not a significant factor (Leis, 2009). The degree of saturation (S) at depth 0.35 m exceeds 80% for most of the time and only **COUP** and **SIM-STO** have default threshold values for S higher than 80% while other models use lower default threshold values for S (Heinen, 2006). Except for **DAISY** and **EPIC**, also **ARMOSA** and **SW-ANIM** should have calculated some denitrification when using default values. Except for the first year, the denitrification calculated by **EPIC** exceeded the nitrate-N leaching.

The change of the total N amount in soil included both organic and mineral forms and was calculated as the residual from the balance. A positive sign means an increase of the total amount whereas a negative sign indicates a depletion of the stock. The model results showed large differences and the largest difference occurred in 2010 where **DAISY** calculated a depletion of 105 kg ha^{-1} while **SW-ANIM** calculated an increase of 103 kg ha^{-1}. The increase of the amount resulted from the assumed biological fixation and the inputs caused by the cultivation of catch crops. When no additional inputs by fixation or by catch crops was assumed, a depletion will occur (**DAISY** and **EPIC**).

Except for **SIM-STO** in 2005 and 2008, differences between calculated seasonal nitrate-N leaching rates were relatively small for the calibration phase. The deviations were much larger for the validation phase, where **SW-ANIM** overestimated the leaching by 39 and 29 kg ha^{-1} in 2009 and 2010, respectively. The observed small leaching rate in 2010 was not approached by any model. Transport of ammonium, organic dissolved N or by surface runoff was calculated at a maximum of 8 kg ha^{-1} by the **COUP** model for the first year of the validation period.

The long term nitrogen balances were summarized at the bottom of Table 8 to further compare the difference of the modellers perceptions of the plant and soil nitrogen cycle.

The seven year balance depicted the major differences between the models clearly. Despite the crop failure in 2007 simulated by **DAISY**, this model showed the highest summed seven year amount, while the summated crop off-take by **SIM-STO** lagged behind with 200 kg ha^{-1} relative to the recorded amount. For the individual years the **ARMOSA** results differed considerably from the observations, but the summated seven year crop off-take resembled the measured value rather good.

Most models have been designed for the field scale for which an average N-yield is calculated. The spatial scale of the lysimeter (1 m^2) differs from the field scale and the variation of crop off-take rates at this scale is much larger than for the field scale. This is illustrated by the oil pumpkin crop in 2005. Only two seeds were planted in the lysimeter. One of the plants died at the start of the generative phase and no harvest was obtained from this plant. This event influenced the yield at the lysimeter scale pretty much, but the yield at the field scale was barely influenced and it can be expected that field scale models encountered difficulties.

The total nitrogen loss by denitrification ranged from 0 to 249 kg ha^{-1} and was subject to the modellers' perception of the possibility of denitrification in the soil at the Wagna experimental field station.

The low input farming system was capable to produce relatively high yields for maize and grains, and for oil pumpkin a N-yield of 51 to 57 kg ha^{-1} was recorded, but the observed nitrate-N leaching exceeded the N-excess, the latter defined as the total addition of mineral fertilizers and animal manure minus the crop off-take. **ARMOSA**, **DAISY** and **EPIC** predicted higher nitrate N-leaching than the N-excess (Fig. 7), while the other models showed a more or less equal value (**SW-ANIM**) or a lower value (**COUP**, **SIM-STO**). One of the main difficulties was to describe the role of the intermediate catch crops in the crop rotation on the delivery of N. Some of the intermediate crops fixate atmospheric N which leads to an input to the soil and other crops are only able to preserve some of the N excess which remains in soil after the catch crops for the next growing season. No data on the N uptake rates and the quality of the resulting green biomass of these intermediate crops were available. Each of the modellers had to make assumptions for the effect of these crops on the soil N cycle. The estimates of the seven years summed additional input to the soil by biological N-fixation varied from 0-2 kg ha^{-1} (**DAISY, EPIC)** to 371 kg ha^{-1} (**SW-ANIM**) (Table 8).

<<Figure 7>>

None of the models simulated long term soil N-stock at equilibrium. The models that did not take biological N-fixation into consideration showed a decrease of the soil N-stock of -342 kg ha^{-1} (**EPIC**) and -177 kg ha^{-1} (**DAISY**). The other models that take account for this input showed an increase ranging from 165 to 419 kg ha^{-1}.

The comparison of the N mass balance components showed large differences between the models. Despite calibration on nitrate leaching, the nitrate leaching predicted was still different from that measured. Crop off-take, although measured, was only used by two models in the calibration procedure, but even then the predicted off-take differed from the observed one. For the other N processes (deposition, biological fixation, volatilization, other transport processes and denitrification) no measured data were available for comparison and calibration. For these aspects, significant differences between the models were observed, either through differences in process descriptions or in handling input by the modelling groups. The resulting storage change thus was also different for the models. The variation of the mass balance components for each model over the years was large. A favourable assessment of a good correspondence between a predicted and a measured quantity is difficult, because it may be good for the wrong reasons. For example, **ARMOSA** predicted rather well the overall crop N off-take but was not able to predict the N off-takes of the individual growing seasons.

3.2.5 Performance assessment

In order to compare the performance of models a quantifiable method is needed. The simplest method would be to rank the models based on a performance index. This method is not preferred, as a model may get a high ranking despite a poor performance. Thus, a classification based on some performance index is to be preferred. Any value of *NSE* and *IoA* (except their values 0 and 1) is difficult to interpret (Legates and McCabe, 1999), and thus it is clear that no default classification boundary values exist to evaluate good, moderate and poor model performance for a set of interrelated variables related to water contents, water fluxes, nitrate concentration and nitrate fluxes at te scale of a lysimeter. ~~One of the difficulties of statistical metrics for model assessment is the judgement of values, whether they indicate a “good”, “moderate” or “weak” performance.~~

Bellocchi et al. (2010) reviewed the methods and different indicators used for the validation of different types of biophysical models. Confalonieri et al. (2010) used *NSE* and *RRMSE*, together with four other indices to assess the quality of simulation of different models in simulating soil water contents. In hydrological studies, it is common practise to assess the model performance on the basis *NSE*, where *NSE* > 0.75 indicates a "good" performance and *NSE* < 0.36 indicates a "weak" similarity of model results with observations (Van Lieuw and Gabrecht, 2003). Moriasi et al. (2007) reviewed the qualification of the model performance of stream discharges and contaminant loads, based on statistical indices for a number of modelling studies. They qualified model simulation on the basis of *NSE* and *PE* but their qualifications are not directly applicable to this study due to differences of spatial scale (catchment versus field) and differences of time scale (month versus day or weekly sample interval). In the literature it is noticeable that classifications and qualifications depend on the considered variables and of the time and space scale. Here we preferred to set up a classification for *IoA*. A number of model studies on the dynamics of soil nitrogen and nitrate leaching have been published that use the *IoA*, alone, or combined with other parameters (Kersebaum et al., 2007; Mantovi et al., 2006; Nolan et al., 2010 ; Sogbedji et al., 2006).

~~Following these authors, we have chosen the *IoA* for a qualitative assessment of the different model outputs.~~ Typical state variables which correspond with instantaneous observations have been distinguished from water fluxes and nitrate concentrations analysed in composed water samples. For the latter we assumed *IoA* values above 0.9 as accurate and *IoA* values below 0.75 as inaccurate. For soil water contents and nitrate concentrations we assume *IoA* values greater than 0.8 as accurate and *IoA* values smaller than 0.6 as inaccurate. Krause et al. (2005) stated that even for *IoA* > 0.65 models can result in poor performance, they sure will for *IoA* < 0.6, which was here chosen as the lowest boundary. The *IoA* scoring for the calibration and validation periods are listed in Table 9.

<<Table 9>>

The scoring differed for the different models. Two models (**SIM-STO**, **SW-ANIM**) performed well for the calibration of the θ(*h*) curves and the simulated θ at different depths, however, this doesn't guarantee good performance for the other state and rate variables in the calibration and validation periods. For the validation period all models performed weak to moderate on the water volume and weak on the nitrate N-flux per sampling interval, ~~and~~ moderate to good on the daily water flux and weak to moderate on the nitrate concentration in the water samples. The models **ARMOSA**, **COUP**, **DAISY** and **EPIC** had more weak qualifications than good qualifications, while **SIM-STO** and **SW-ANIM** had more good qualifications.

We have also assessed the accuracy of the seasonal amounts on the basis of the mean absolute error (*MAE*). The seven seasons included the oil pumpkin crop twice, which was an unknown or a particular crop for most of the modelling groups. The seven year series contained an extremely wet year (2009) and a dry summer (2011). For the performance assessment for average crop and rainfall conditions *MAE* of the five best values (MAE_5) out of seven (MAE_7) are presented in Table 10 to examine if the models perform better for average conditions. In some cases the improvement was more than 50%, and the ranking of the models slightly changed. Despite the fact that *MAE* is less sensitive to outliers than e.g. *IoA*, extreme situations (unknown crop, wet or dry years) can have a large impact on *MAE*.

<<Table 10>>

4. General discussion

5.1 Water contents and water fluxes

Differences between observed and model predicted water contents, water fluxes and water volumes per sampling interval indicate over- or under-estimation of the water excess in the soil column. Besides uncertainties in soil hydraulic properties and in observations, there was also lack of information about actual plant and root system development as a function of time.

The different modelling groups were not able to find a simultaneous optimal solution which minimizes both water contents deviations and water flux deviations. This may be due to uncertainties in soil hydraulic properties, and the disregarding of hysteresis in the models. The soil at the Wagna experimental station consists of a clayey-sand on top of a gravel layer. Durner et al. (2007) concluded that for layered soils with distinct heterogeneity no unique effective soil hydraulic properties exist. If only fluxes across the boundaries of the system are required, heterogeneous systems can be modelled with quasi-homogeneous ones, even if the internal system state is not matched properly. However, for nutrient dynamics (solute dispersion, biological and chemical reactions) an accurate internal system state description is mandatory (Durner et al., 2007).

5.2 Nitrate concentrations and fluxes

The different modelling groups have chosen different objective functions when calibrating for nitrate observations. Two out of four models that used nitrate flux in their objective function resulted in moderate *IoA* values for the nitrate fluxes, while the others resulted in poor *IoA* values (Table 9). Two out of three models that used nitrate concentration in their objective function resulted in good *IoA* values, while the third model resulted in a moderate *IoA* value (Table 9). However, a good calibration on nitrate concentrations did not result in good performance on nitrate fluxes. Both for the calibration and for the validation periods it appeared that all models had difficulties in predicting the nitrate fluxes at the bottom of the lysimeter, even if some of them were calibrated based on the measured nitrate fluxes.

Vereecken et al. (1991) evaluated five complex models from which SW-ANIM, EPIC and DAISY are also included in our performance assessment. A comparison between simulated and observed nitrate leaching rates measured in two sandy soils in Denmark and one sandy soil in the Netherlands revealed that SW-ANIM, EPIC and DAISY performed similar, although DAISY appeared to be a bit superior in behaviour. In general much better statistical metric values were reported than in our study. This may be due to the circumstances of the field trials which were representative for conventional agriculture during the eighties and because the calibration and he comparison was carried out for seasonal values.

Diekkrüger et al, (1995) compared the results produced by 19 simulation models, others than those used in this study, for a loam soil and a sand soil in Southern and Eastern Saxony in Germany. Variation in the leaching rates at 0.9 m depth reflected mainly the differences in soil water fluxes at that depth. Apart from the seasonal differences between the models that were able to simulate a three year period continuously, the cumulative leaching was nearly the same for these models. The results of soil nitrogen simulations were significantly influenced by the results of water flow and plant growth simulations. Diekkrüger et al, (1995) concluded that for long term forecasts the exact determination of the boundary conditions is as important as the model approach itself. Our finding that the unmeasured inputs concerning biological N-fixation are important for the soil nitrogen dynamics is consistent with this conclusion. In our study, differences between model seasonal and long term results are attributed to some extend to different assumptions about fixation rates.

Kersebaum et al., (2007) conducted a comparison of simulation models for 18 different models from which SW-ANIM and SIM-STO are included in our study. SW-ANIM was applied to the Müncheberg data-set (Kroes and Roelsma, 2007) and SIM-STO was applied to the data-set of the the lysimeter station Berlin-Dahlem for water flow simulation and to the Bad Lauchstädt data-set for simulation of soil nitrogen dynamics (Stenitzer et al., 2007). Results for the mean bias, *RMSE*, *IoA* and *NSE* showed weak performances for the soil mineral nitrogen simulation in the 0-90 cm upper soil layer for nearly all models which were subjected to the Müncheberg data-set. Kersebaum et al. (2007) concluded that comparison of simulated results by models which are intended for field scale and regional scale with measured data often shows unsatisfactory results due to deviating conditions and parameters. It does not automatically mean that the models or the parameters are wrong because the data and parameters are only partly related to the site specific conditions of the measurements. In our study significant amount of data was available, but critical information about rooting depth and pattern, atmospheric deposition rates, mineralization and fixation rates was missing as well as the nitrogen uptake rates and residue amounts of the catch crops. Due to these uncertainties, it is difficult to draw clear conclusions about the predictive power of the models.

5.3 Seasonal nitrogen balances

The comparison of the N mass balance components showed large differences between the models. Despite calibration on nitrate leaching, the nitrate leaching predicted was still different from that measured. Crop off-take, although measured, was only used by two models in the calibration procedure, but even then the predicted off-take differed from the observed one. For the other N processes (deposition, biological fixation, volatilization, other transport processes and denitrification) no measured data were available for comparison and calibration. For these aspects, significant differences between the models were observed, either through differences in process descriptions or in handling input by the modelling groups. The resulting storage change thus was also different for the models. The variation of the mass balance components for each model over the years was large. A favourable assessment of a good correspondence between a predicted and a measured quantity is difficult, because it may be good for the wrong reasons. For example, ARMOSA predicted rather well the overall crop N off-take but was not able to predict the N off-takes of the individual growing seasons.

3.2.6 5.4 Methodological aspects for explanation of differences

5.4.1 Data

Experimental data collected from a well-controlled lysimeter were used for the purposes of our study. However, the number of measured state and rate variables were less than those present in the six models. For example, no data were available on field-scale hydraulic conductivity, deposition and biological fixation. This means that the outcome of the models is uncertain as not all components of the internal mass balance could be optimized. We have observed in the blind test that based on a limited availability of data, which resembles situations that would occur in practice, the predictions of the models was poor compared to actual observations. That would imply that usage of such simulation models for predictions on nitrate leaching at unknown, regional scales must be regarded with care. In this study the rainfall excess was positive in most times of the year, such that the imposed bottom boundary condition in the lysimeter resulted in leaching. For other situations where capillary rise may occur, the models have not been inter-compared. Finally, it is noticed that the soil hydraulic properties as determined in the laboratory on small soil samples does not guarantee well-predicted soil water contents and soil water fluxes even for a well-controlled lysimeter situation. Partly, this may be due to the lack of knowledge of hysteresis or its description in the models.

5.4.2 Procedure

Despite the structured set-up of this study (blind test, calibration, validation) there remained flexibility in the approach chosen by the different modelling groups. For example, no formal sensitivity analysis was prescribed, meaning that each group was free to choose a set of parameters to be calibrated. This has introduced a subjective element in this study. Although it was agreed beforehand that the water fluxes and the nitrate concentrations in the lysimeter effluent were the most important parts of the model comparison, the objective function for optimization was chosen freely by the modellers. Some modelling group have chosen to include also the information about soil water contents and crop uptake in the optimization procedure. The comparison is, therefore, not a pure comparison of the model codes, but also a comparison of how modellers used their models.

In this study much effort has been put in calibrating and validating six models for a well-controlled lysimeter situation. Any conclusions of this study are thus at first applicable for these kind of (local) situations. Additional research is required to inter-compare these models for deviant situations, for example, for regional assessments of impact of fertilization strategies.

5.4.3 Decreased performance when averaging

One should expect a better performance for the averaged water fluxes per sampling interval than for the daily water fluxes because peaks of the daily fluxes pattern are flattened by aggregation. This was indeed observed in better performance indices for the calibration period (Table 5). However, the opposite occurred for the validation period (Table 5). This counter-intuitive response of performance indices to the averaging of water fluxes of the validation phase may be due to the following three reasons.

1) The distributions of the time increments of sampling in both phases differed slightly, where in the validation phase samples were taken more frequently with smaller time steps (data not shown). The pattern of sampling intervals was neither regular nor random. The pattern was more or less dependent on practical circumstances and availability of manpower and on average samples were taken once in seven days. Under extreme rainfall

conditions the intervals were shortened and under extreme dry conditions the intervals were longer because no percolation water was present.

2) The probability density distributions of the daily water fluxes and averaged water fluxes for the calibration and validation periods appeared to be unequal (data not shown). This was concluded from a non-parametric analogue of a one-way analysis of variance performed by the one-way analysis of variance by ranks after Kruskal-Wallis (1952). The different statistical behaviour may result in variant effects of volume weighted averaging on the performance indices.

3) Certain days or periods may have had a great effect on the averaging. A leave-one-out calculation procedure was performed to qualitatively explore the effect of certain days and periods on the performance of the models. In the series of data pairs of observed and simulated water fluxes, one data pair is left out and the *IoA* was calculated for the remainder of the population. This procedure is repeated for each of the data pairs and the results are subtracted from the *IoA*-value based on the total series of data pairs belonging to either the daily fluxes of the calibration or the validation phase or to the averaged values of the phases. Only the results greater than 0.001, in absolute sense, haven been plotted in Figure 8.

<<Figure 8>>

The exclusion of a particular data pair can result in both an improvement (negative values) or a deterioration (positive values) of the Δ*IoA*. Furthermore, it is notable that the Δ*IoA* of daily fluxes responded differently compared to the Δ*IoA* for averaged fluxes per sampling interval. For almost all models the exclusion of the value simulated for 19 Sept 2006 would affect the Δ*IoA*. The effect of excluding the value of this period is much smaller for the Δ*IoA* based on the averaged values per sampling interval. The maximum effect in the series of daily values occurres for a certain day of the calibration period and the maximum effect in the series of averaged values per sampling interval is calculated for a time interval in Sept. 2010 which belongs to the validation phase The maximal effect of leaving one value out is greater for the validation period than for the calibration period. Based on this analysis, it is plausible that the averaging of water fluxes has a different effect on the performance indices of the calibration phase than on those of the validation phase.

~~5.~~4. Summary and Conclusions

The novel aspect of this study is that six detailed process oriented dynamic models were tested (1) for the Wagna test-site which is known to be highly vulnerable to nitrate leaching, (2) for a crop particular for the Styrian low input agriculture system, (3) for a situation where different catch crops were part of the crop rotation, and (4) for the weather conditions which significantly differed between the calibration and the validation phase..

This study was not performed to determine which model is the best. We like to quote Kersebaum et al. (2007) who stated: "*The comparison of different models applied on the same data set is not suitable to serve as a model contest or to find the best model. Although, the application of different indices for model performance helps to identify strengths and weaknesses of each model, an objective comparison is nearly impossible due to different levels of input requirements, calibration efforts and last but not least the uncertainties and errors within the measured data themselves.*"

We conclude:

a. The blind test showed that simulation results without calibrating the model are generally far from acceptable . Therefore, model calibration is essential.
b. None of the models performed good for the different criteria considered in this study. This may be due to the combined effect of the model structure which is not tuned to the circumstances of the Wagna experimental fields and the lack of knowledge to establish an appropriate set of parameters. Furthermore, not all inputs were measured, so there were too many degrees of freedom.
c. The soil of the Wagna lysimeter is highly vulnerable to nitrate leaching. The seven year summed nitrate leaching rate (123 kg ha^{-1}) exceeds the seven year summed fertilization excess. Models designed for nitrate leaching in high input farming systems have difficulties with an accurate prediction of the nitrate leaching in low input farming systems
d. Judgement of the performance solely on the basis of nitrate concentrations or nitrate fluxes is not sufficient for the assessment of the predictive power of the models. Other results as soil water contents (daily), water and nitrogen fluxes (daily and seasonal), soil temperatures (daily), nitrogen yields (seasonal) should also be taken into account. This should be reflected by the objective function of the model calibration.

e. Traditional Richard's / Darcy Buckingham equation based models that make use of the Mualem-van Genuchten descriptions and disregard phenomena as hysteresis, preferential flow and multiple phase flow encounter difficulties with an accurate and consistent simulation of both water contents and water fluxes for the soil and conditions of the Wagna lysimeter.

f. Some models which performed relatively well in the calibration phase of the study failed to simulate the nitrate concentrations and fluxes in the validation phase (**SW-ANIM**), while other models behaved relatively bad in the calibration phase and showed better results in the validation phase (**SIM-STO**). An accurate calibration does not guarantee a good predictive power of the model.

g. The catch crop mixtures and the non-harvested English ryegrass play an important role in the nutrient dynamics of the soil. This role is addressed weakly by the simulation models: (1) due to a lack of experimental data on nitrogen uptake rates and mineralization of residues of these intermediate crops, and (2) lack of knowledge to describe the relevant processes related to the foreign crops

h. Assessment of future climate and land use changes requires a good predictive power of the models and a certain level of robustness. Although the robustness is not clear for the tested models, the process oriented dynamic models used in this study are useful for hypothesis testing.

~~6.~~5. Acknowledgements

This research was made possible by the GENESIS project of the EU 7th Framework Programme (Project No. 226536; FP7-ENV-2008-1). We are grateful for the experimental data provided by Joanneum Raum (Graz, Austria). The modelling team of Democritus University of Thrace would like to thank Per-Erik Jansson (Royal Institute of Technology, Stockholm, Sweden;) for his valuable help during the application of CoupModel.

~~7.~~6. References

~~Acutis M, Confalonieri R. Optimization algorithms for calibrating cropping systems simulation models. A case study with simplex-derived methods integrated in the WARM simulation environment. Ital J Agrometeorol 2006;11:26-34.~~

Akkal-Corfini N, Morvan T, Menasseri-Aubry S, Bissuel-Bélaygue C, Poulain D, Orsini F, Leterme P. Nitrogen mineralization, plant uptake and nitrate leaching following the incorporation of (15N)-labeled cauliflower crop residues (Brassica oleracea) into the soil: a 3-year lysimeter study. Plant Soil 2010;. 328:17–26. DOI 10.1007/s11104-009-0104-0

Ale S, Bowling LC, Youssef MA, Brouder SM. Evaluation of simulated strategies for reducing nitrate–nitrogen losses through subsurface drainage systems. J Environ Qual 2012; 41:217-228.

Basile A, Ciollaro G, Coppola A. Hysteresis in soil water characteristics as a key to interpreting comparisons of laboratory and field measured hydraulic properties. Water Resour Res 2003;39:1355-1367.

Basile A., Coppola A, De Mascellis R, Randazzo L. Scaling approach to deduce field unsaturated hydraulic properties and behavior from laboratory measurements on small cores. Vadose Zone J 2006;5:1005-1016.

Bellocchi G, Rivington M, Donatelli M, Matthews K, Validation of biophysical models: issues and methodologies. A review. Agron Sustain Dev 2010;30:109–130.

Bergström L, Johnsson H, Torstensson G. Simulation of soil nitrogen dynamics using the SOILN model. Nutr Cycl Agroecosys 1991;27:181–188.

Bouraoui F, Aloe A. European Agrochemicals Geospatial Loss Estimator: Model development and Applications, EUR – Scientific and Technical Research series, ISSN 1018-5593. Office for Official Publications of the European Communities, Luxembourg. 2007.

Buckingham, E. Studies on the movement of soil moisture. Bull. 38, USDA, Bureau of Soils, Washington, DC, 1907.

Burkart MR, Kolpin DW, James DE. Assessing groundwater vulnerability to agrichemical contamination in the Midwest US. Water Sci Technol 1999;39:103-112.

Confalonieri R, Bregaglio S, Bocchi S, Acutis M. An integrated procedure to evaluate hydrological models. Hydrol Process 2010;24:2762–2770.

Darcy H. Les fontaines publique de la ville de Dijon. Dalmont, Paris, 1856.

Dawson CW, Abrahart RJ, See LM. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Modell Softw 2007;22:1034–1052. http://dx.doi.org/10.1016/j.envsoft.2006.06.008

Dawson CW, Abrahart, RJ, See, LM. HydroTest: Further development of a web resource for the standardised assessment of hydrological models. Environ Modell Softw 2010;25:1481–1482. http://dx.doi.org/10.1016/j.envsoft.2009.01.001

De Willigen P, Neeteson JJ, Comparison of six simulation models for the nitrogen cycle in the soil. Fert Res 1985;8:157-171.

De Willigen P. Nitrogen turnover in the soil-crop system; comparison of fourteen simulation models. Fert Res 1991;27: 141-149.

Diekkrüger B, Söndgerath D, Kersebaum KC, McVoy CW. Validity of agroecosystem models a comparison of results of different models applied to the same data set. Ecol Model 1995:81:3-29. http://www.sciencedirect.com/science/article/pii/030438009400157D

~~Doherty J. PEST Model-independent parameter estimation user manual: 5th edition. Watermark Numerical Computing, 2005.~~

Donatelli M, Wösten JHM, Belocchi G. Evaluation of pedotransfer functions. In: Pachepsky Y. Rawls WJ, editors. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam. 2004. p. 357–362.

Formatted: English (U.K.)

Durner W, Jansen U, Iden SC. Effective hydraulic properties of layered soils at the lysimeter scale determined by inverse modelling. Eur. J. Soil Sci. 2007;59:114–124. doi: 10.1111/j.1365-2389.2007.00972.x

EU. 1991. Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off . J. Eur. Commun. L375:1–8. Available at http://ec.europa.eu/environment/water/water-nitrates/directiv.html

Fank J, Fastl G, Kupfersberger H, Rock G. Die Bewirtschaftung des Versuchsfeldes Wagna - Auswirkung auf die Grundwassersituation. Umweltprogramme für die Landwirtschaft 2006. Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft, A-8952 Irdning. Gumpenstein, 2006.

Feichtinger F. STOTRASIM – Ein Modell zur Simulation der Stickstoffdynamik in der ungesättigten Zone eines Ackerstandortes. In: *Modelle für die gesättigte und ungesättigte Bodenzone*. Schriftenreihe des Bundesamtes für Wasserwirtschaft, 7, Wien, 1998, p. 14-41,

Ferrara RM, Trevisiol P, Acutis M, Rana G, Richter GM, Baggaley N. Topographic impacts on wheat yields under climate change: two contrasted case studies in Europe. Theor. Appl. Climatol. 2011;99:53–65.

Gribb MM. Hansen FI, Aleshia Chandler, McNamara DG, James P. The effect of various soil hydraulic property estimates on soil moisture simulations. Vadose Zone J 2009; 8:321–331.

Grizzetti B, Bouraoui F, Billen G, Van Grinsven H., Cardoso AC, Thieu V, Garnier J, Curtis C, Howarth R, Johnes P. Nitrogen as a threat to European water quality. In: The European Nitrogen Assessment 17, Cambridge, UK: Cambridge University Press, 2011, p. 379-404.

Groenendijk, P. , L.V. Renaud, and J. Roelsma. 2005. Prediction of Nitrogen and Phosphorus leaching to groundwater and surface waters. Process descriptions of the animo4.0 model. Alterra–Report 983, Alterra, Wageningen. http://content.alterra.wur.nl/Webdocs/PDFFiles/Alterrarapporten/AlterraRapport983.pdf

Hansen S, Jensen HE, Nielsen NE, Svendsen H. **DAISY**: Soil Plant Atmosphere System Model. NPO Report No. A 10. The National Agency for Environmental Protection, Copenhagen, 1990, 272 pp.

Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of nitrogen dynamics and biomass production in winter wheat using the Danish simulation model DAISY, Fert Res 1991a;27: 245-259.

Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of biomass production, nitrogen uptake and nitrogen leaching by using the Daisy model. In: Soil and Groundwater Research Report II: Nitrate in Soils, Final Report on Contracts EV4V-0098-NL and EV4V-00107-C. DG XII. Commission of the European Communities, 1991b; 300–309.

Hansen S. DAISY, a flexible Soil-Plant-Atmosphere system Model. Report. Dept. Agric, Danish Informatics Network in the Agricultural Sciences, 2002. http://www.dina.kvl.dk/~DAISY/ftp/DAISYDescription.pdf.

Heinen, M. Simplified denitrification models: Overview and properties. Geoderma 2006;133:444-463.

Herbst M, Fialkiewicz W, Chen T, Pütz T, Thiéry D, Mouvet C, Vachaaud G, Vereecken H. Intercomparison of flow and transport models applied to vertical drainage in cropped lysimeters. Vadose Zone J 2005;4:240-254.

Jabro JD, Jabro AD, Fales SL. Model performance and robustness for simulating drainage and nitrate-nitrogen fluxes without recalibration. Soil Sci. Soc. Am. J. 2012;76:1957–1964 doi:10.2136/sssaj2012.0172

Jachner S, Van den Boogaart KG, Petzoldt T. Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R package qualV). J Stat Softw 2007;22:(8),1–30. http://www.jstatsoft.org/v22/i08

Janssen PHM, Heuberger PSC. Calibration of process-oriented models. Ecol Model 1995;83: (1-2) 55-66. http://dx.doi.org/10.1016/0304-3800(95)00084-9

Jansson P-E, Karlberg L. Coupled heat and mass transfer model for soil-plant-atmosphere systems Royal Institute of Technology, Dept of Civil and Environmental Engineering, Stockholm, 2004. 435 pp.

Jansson P-E. CoupModel: Model use, calibration and validation, Transactions of the ASABE 2012;55(4):1-11.

Jensen LS, Mueller T, Nielsen NE, Hansen S, Crocker GJ, Grace PR., Klir J, Körschens M, Poulton PR. Simulating trends in soil organic carbon in long-term experiments using the soil-plant-atmosphere model DAISY. Geoderma 1997;81:5-28.

~~Justes E, Mary B, Meynard JM, Machet JM, Thelier-Huches L. Determination of a critical nitrogen dilution curve for winter wheat crops. Ann Bot-London 1994; 74:397-407.~~

Kersebaum KC, Hecker, J-M, Mirschel W, Wegehenkel M. Modelling water and nutrient dynamics in soil–crop systems: a comparison of simulation models applied on common data sets. In: Kersebaum KC et al., editors, Modelling Water and Nutrient Dynamics in Soil–Crop Systems, Springer, 2007, p. 1–17.

Khodaverdiloo H, Homaee M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions for some calcareous soils, J Hydrol 2011;399:93-99

Klammler G, Fank J.. Determining Measuring water and nitrogen balances for beneficial management practices using lysimeters at Wagna test site (Austria). Sci. Tot. Environ 2014; *submittedthis issue*.

Krause P, Boyle, DP, Bäse F. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 2005;5:89–97. http://www.adv-geosci.net/5/89/2005/adgeo-5-89-2005.pdf

Kroes J and Roelsma J. Simulation of water and nitrogen flows on field scale: application of the SWAP–ANIMO model for the Müncheberg data set. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient Dynamics in Soil–Crop Systems, 2007, Springer, pp 111–128.

Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.

Khodaverdiloo H, Homaee M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions for some calcareous soils, J Hydrol 2011;399:93-99

Legates DR, McCabe GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233-241.

Leis A. Chemical, isotopic, and microbiological evidence for nitrification below the plant root zone from intensive fertilized agricultural area in Austria – Insights from lysimeter studies and soil cores. In: IAEA-TECDOC-1618. Application of Isotopes to the Assessment of Pollutant Behaviour in the Unsaturated Zone for Groundwater Protection. Final report of a coordinated research project 2004-2005. International Atomic Energy Agency,Vienna; 2009. p. 15-30.

Mantovi P, Fumagalli L, Beretta GP, Guermandi M. Nitrate leaching through the unsaturated zone following pig slurry applications, J Hydrol 2006; 316:195-212.

Moreels E., De Neve S, Hoffman G, Van Meirvenne M. Simulating nitrate leaching in bare fallow soils: a model comparison. Nutr Cycl Agroecosys 2003;67:137-144

Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Tansactions of the ASABE 2007;50: 885−900.

Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. 1976;12: 513-522.

Nash JE, Sutcliffe JV. River flow forecasting through conceptual models, 1, A discussion of principles, J. Hydrol., 1970;10:282-290.

Nett L, Feller C, George E, Fink M.Effect of winter catch crops on nitrogen surplus in intensive vegetable crop rotations. Nutr Cycl Agroecosys 2011;91:327-337.

Nolan, BT, Puckett LJ, Ma L, Green TG, Bayless ER, Malone RW. Predicting unsaturated zone nitrogen mass balances in agricultural settings of the United States. J Environ Qual 2010;39:1051-1065.

Oenema O. Governmental policies and measures regulating nitrogen and phosphorus from animal manure in European agriculture. J Anim Sci 2004;82: E196-E206.

Patil N, Rajput G. Evaluation of Water Retention Functions and Computer Program "Rosetta" in Predicting Soil Water Characteristics of Seasonally Impounded Shrink–Swell Soils, J Irrig Drain E-ASCE 2009;135, 286-294. http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29IR.1943-4774.0000007

Perego A, Giussani A, Sanna M, Fumagalli M, Carozzi M, Alfieri L, Brenna S, Acutis M. The ARMOSA simulation crop model: overall features, calibration and validation results. Italian Journal of Agrometeorology ~~(in press).~~ 2013;3:23-38..

Qi Z, Ma L, Helmers MJ, Ahuja LR, Malone RW. Simulating nitrate-nitrogen concentration from a subsurface drainage system in response to nitrogen application rates using RZWQM2. J Environ Qual 2012;41:289-295.

Reusser DE, Blume T, Schaefli B, Zehe E. Analysing the temporal dynamics of model performance for hydrological models. Hydrol Earth Syst Sc 2009;13: 999 – 1018. http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf

Richards LA, Capillary conduction of liquids through porous mediums. Physics 1931;1: 318-333.

Richter GM, Acutis M, Trevisiol P, Latiri K, Confalonieri R. Sensitivity analysis for a complex crop model applied to Durum wheat in the Mediterranean. Europ J Agron 2010;32:127-132.

Ritter A, Hupet F, Muñoz-Carpena R, Lambot S, Vanclooster M. Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods. Agr Water Manage 2003;59:77-96.

Saxton KE, Rawls WJ. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci. Soc. Am. J.2006;70:1569-1578. doi:10.2136/sssaj2005.0117

~~SimLab, SimLab ver. 3.2.6. Development Framework for Uncertainty and Sensitivity Analysis. Joint Research Centre of the European Commission, Econometrics and Applied Statistics, Ispra, Italy, 2009. http://simlab.jrc.ec.europa.eu~~

Smith P, Smith JU, Powlson DS, McGill WB, Arah JRM, Chertov OG, Coleman K, Franko U, Frolking S, Jenkinson LS. Jenseng LS, Kellyh RH, Klein-Gunnewiek H,, Komarov AS, Lif C, Molina JAE, Mueller T, Parton WJ, Thornley JHM, Whitmore AP. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma 1997;81: 153-225.

Sogbedji JM, Van Es HM, Melkonian JJ, Schindelbeck RR. Evaluation of the PNM model for simulating drain flow nitrate-N concentration under manure-fertilized maize. Plant Soil 2006;282:343-360.

Sohier C, Degré A, Dautrebande S. From root zone modelling to regional forecasting of nitrate concentration in recharge flows – The case of the Walloon Region (Belgium). J Hydrol 2009;369:350-359.

Stenitzer E. SIMWASER – Ein numerisches Modell zur Simulation des Bodenwasserhaushaltes und des Pflanzenertrages eines Standortes. *Mitt. der Bundesanstalt für Kulturtechnik und Bodenwasserhaushalt*, 31: Petzenkirchen, 1988. p.1-118.

Stenitzer E, Diesel H, Franko U, Schwartengräber R, Zenker T. Performance of the model SIMWASER in two contrasting case studies on soil water movement. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient Dynamics in Soil–Crop Systems, 2007, Springer, pp 27–36.

Stumpp C, Nützmann G, Maciejewski S, Maloszewski P. A comparative modeling study of a dual tracer experiment in a large lysimeter under atmospheric conditions. J Hydrol 2009;375: 566-577.

Svendsen H, Hansen S, Jensen HE. Simulation of crop production, water and nitrogen balances in two German agro-ecosystems using the DAISY model. Ecol Model 1995;81: 197-212.

Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 2001;106: No. D7, P. 7183. doi:10.1029/2000JD900719.

Thorup-Kristensen, K. . Effect of deep and shallow root systems on the dynamics of soil inorganic N during 3-year crop rotations. Plant Soil, 2006;288:233-248.

Van Dam JC,. Field-scale water flow and solute transport: SWAP model concepts, parameter estimation and case studies. PhD thesis Wageningen University,, 2000.

Van Dam JC, Groenendijk P, Hendriks RFA. Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. Vadose Zone J 2008;7:640-653.

Van der Laan M, Miles, N, Annandale JG, Du Preez CC. Identification of opportunities for improved nitrogen management in sugarcane cropping systems using the newly developed Canegro-N model. Nutr Cycl Agroecosys 2011;90,391-404.

Van der Velde M, Bouraoui F, Aloe A. Pan-European regional-scale modelling of water and N efficiencies of rapeseed cultivation for biodiesel production. Glob Change Biol 2009;15:24-37.

Van Genuchten MTh. A closed form for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 1980;44:892-898.

Van Liew MW, Garbrecht J. Hydrologic simulation of the little Washita river experimental watershed using SWAT. J Am Water Resour Ass 2003;39:413-426.

Vereecken, H., E.J. Jansen, M.J.D. Hack-ten Broeke, M. Swerts, M. Engelke, S. Fabrewitz and S. Hansen, 1991. Comparison of simulation results of five nitrogen models using different datasets. In: Commission of European Communities, editor, Soil and Groundwater Research, Report II Nitrate in Soils, Commision of the European Communities, Luxenbourg, pp 321 – 338.

Vereecken H, Weynants M, Javaux M, Pachepsky Y, Schaap MG, Van Genuchten MTh. Using Pedotransfer Functions to Estimate the van Genuchten–Mualem Soil Hydraulic Properties: A Review Vadose Zone J. 2010;9: 795–820.

Vitousek PM, Naylor, Crews RT, David MB, Drinkwater LE, Holland E, Johnes PJ, Katzenberger J, Martinelli LA, Matson PA, Nziguheba G, Ojima D, Palm CA, Robertson GP, Sanchez PA, Townsend AR, Zhang FS.. Nutrient imbalances in agricultural development. Science 2009; 324, no. 5934: 1519.

Wang X, Mosley CT, Frankenberger JR, Kladivko EJ. Subsurface drain flow and crop yield predictions for different drain spacings using DRAINMOD, Agr Water Manage 2006;79:113-136.

Williams JR, Jones CA, Dyke PTl.. A modeling approach to determining the relationship between erosion and soil productivity. Trans. ASAE 1984;27:129-144.

Williams JR, Jones CA, Kiniry JR, Spanel DA.. The EPIC crop growth model. Trans. ASAE, 1989;32:497–511.

Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteo Soc 1982; 63, 1309-1313.

Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. Statistics for the evaluation and comparison of models. J Geophys Res 1985;90, 8995 – 9005. dx.doi.org/10.1029/JC090iC05p08995

Wolff J, Beusen AHW, Groenendijk P, Kroon T, Rötter R, Van Zeijts H. The integrated modeling system STONE for calculating nutrient emissions from agriculture in the Netherlands. Environ Model Softw 2003;18:597-617. doi:10.1016/S1364-8152(03)00036-7

Table 1. Crop rotation and fertilizer applications on the soil of the KON-lysimeter. CC and MC refer to catch crop and main crop, and FYM and MF refer to farmyard manure and mineral fertilizer, respectively.

Type	Crop	Sowing date	Date of harvesting or amending crop residues to soil	Date of fertilizer application	Type and amount of fertilizer (kg ha^{-1} N)	
CC	Mixture: summer common tare, white clover, sunflower	06-Aug-04	06-Apr-05			
MC	Oil pumpkin	30-Apr-05	13-Sep-05	25-Apr-05 03-Jun-05	FYM: MF:	27.4 35.1
CC	English ryegrass	03-Jun-05	09-Apr-06			
MC	Maize (grain)	24-Apr-06	02-Oct-06	24-Apr-06 08-Jun-06	FYM: MF:	54.5 75.6
CC	Mixture: forage rye, winter turnip rape	03-Oct-06	09-Apr-07			
MC	Maize (grain)	16-Apr-07	21-Sep-07	16-Apr-07 26-May-07	FYM: MF:	120.7 59.0
MC	Winter barley	08-Oct-07	30-Jun-08	28-Feb-08 09-Feb-08	FYM: MF:	84.6 38.0
CC	Mixture: winter turnip rape, mustard, sunflower	04-Aug-08	20-Apr-09			
MC	Oil pumpkin	28-Apr-09	07-Sep-09	22-May-09 01-Jun-09	MF: MF:	36.0 16.0
CC	English ryegrass	05-Jun-09	31-Dec-09			
MC	Maize (grain)	17-Apr-10	23-Sep-10	16-Apr-10 26-May-10	FYM: MF:	62.6 81.0
MC	Triticale	09-Oct-10	13-Jul-11	11-Mar-11 11-Apr-11	FYM: MF:	119.1 62.0
CC	Mixture: mustard, phacelia, sunflower, buckwheat, ryegrass	08-Aug-11	After 31-Dec-11			

Table 2. Annual precipitation rates (mm a^{-1}) and their cumulative probability percentages based on precipitations values of 1961 – 2011.

Phase	Calibration				Validation		
Year	2005	2006	2007	2008	2009	2010	2011
Precipitation (mm a^{-1})	883	839	892	893	1355	1013	739
Cumulative probability	44%	31%	48%	50%	98%	75%	10%

Table 3. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of volumetric water contents derived from calibrated soil moisture retention curves (Figure 2) and observed volumetric water contents at depths 0.35 m (*n* = 922), 0.9 m (*n* = 1413) and 1.8 m (*n* = 1456) depth. **EPIC** is excluded as it does not use soil moisture retention relationships.

Model	*MAE* ($cm^3 cm^{-3}$)			*RMSE* ($cm^3 cm^{-3}$)			*IoA*		
	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m
ARMOSA	0.0064	0.0166	0.0308	0.0112	0.0176	0.0310	0.89	0.79	0.18
COUP	0.0341	0.0753	0.0391	0.0416	0.0775	0.0395	0.59	0.31	0.18
DAISY	0.0295	0.0340	0.0166	0.0326	0.0374	0.0178	0.63	0.62	0.38
SIM-STO	0.0212	0.0119	0.0064	0.0255	0.0130	0.0078	0.75	0.89	0.67
SW-ANIM	0.0072	0.0062	0.0033	0.0117	0.0075	0.0036	0.87	0.96	0.85

Table 4. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of simulated and in situ measured values of volumetric water contents at depths 0.35 m, 0.9 m and 1.8 m for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	*MAE* ($cm^3 cm^{-3}$)			*RMSE* ($cm^3 cm^{-3}$)			*IoA*		
	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m	0.35 m	0.9 m	1.8 m
				Calibration 2005 – 2008 (*n* = 1461)					
ARMOSA	0.0119	0.0247	0.0107	0.0168	0.0447	0.0123	0.79	0.75	0.46
COUP	0.0230	0.0104	0.0023	0.0288	0.0363	0.0031	0.74	0.84	0.85
DAISY	0.0956	0.0152	0.0105	0.1083	0.0630	0.0132	0.28	0.65	0.38
EPIC	0.0613	0.1563	0.0909	0.0662	0.0306	0.0925	0.49	0.90	0.07
SIM-STO	0.0180	0.0063	0.0028	0.0249	0.0271	0.0039	0.81	0.92	0.85
SW-ANIM	0.0101	0.0106	0.0072	0.0159	0.0285	0.0082	0.87	0.92	0.59
				Validation 2009 – 2011 (*n* = 955)					
ARMOSA	x	0.0260	0.0130	x	0.0291	0.0149	x	0.52	0.47
COUP	x	0.0124	0.0030	x	0.0165	0.0041	x	0.74	0.84
DAISY	x	0.0152	0.0137	x	0.0193	0.0165	x	0.69	0.40
EPIC	x	0.1535	0.0924	x	0.1570	0.0939	x	0.19	0.09
SIM-STO	x	0.0093	0.0039	x	0.0134	0.0054	x	0.87	0.82
SW-ANIM	x	0.0141	0.0075	x	0.0176	0.0088	x	0.74	0.65

x Measurements at depth 0.35 m were disqualified from 2009 onwards due to aging of the sensor, and, therefore, no performance indices were calculated

Formatted: Font: Italic

Formatted: Font: Italic

Table 5. Statistical parameters (*MAE*, *RMSE*, *IoA*, *NSE*) for the comparison of simulated and observed daily fluxes and fluxes averaged per sampling interval at depth 1.8 m for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	Daily water fluxes				Averaged water fluxes per sampling interval			
	MAE (mm d^{-1})	*RMSE* (mm d^{-1})	*IoA*	*NSE*	*MAE* (mm d^{-1})	*RMSE* (mm d^{-1})	*IoA*	*NSE*
	Calibration 2005 – 2008							
	n = 1461				*n* = 199			
ARMOSA	0.45	1.00	0.82	0.41	0.43	0.81	0.84	0.48
COUP	0.45	0.98	0.80	0.44	0.43	0.75	0.85	0.55
DAISY	0.57	1.16	0.68	0.21	0.54	0.90	0.74	0.35
EPIC	0.54	0.99	0.83	0.42	0.46	0.75	0.89	0.55
SIM-STO	0.34	0.87	0.86	0.55	0.30	0.62	0.91	0.69
SW-ANIM	0.38	0.91	0.86	0.51	0.37	0.72	0.88	0.58
	Validation 2009 – 2011							
	n = 1084				*n* = 128			
ARMOSA	0.70	1.75	0.79	0.41	1.66	3.82	0.68	0.39
COUP	0.70	1.57	0.84	0.52	1.41	3.47	0.79	0.50
DAISY	0.73	1.77	0.77	0.39	1.74	4.34	0.56	0.21
EPIC	0.85	1.79	0.77	0.38	1.80	4.00	0.63	0.33
SIM-STO	0.51	1.43	0.90	0.61	1.69	3.94	0.76	0.35
SW-ANIM	0.57	1.59	0.88	0.51	1.77	4.16	0.74	0.27

Table 6. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of observed nitrate concentrations and nitrate N leaching rates with simulated values by calibrated models for the Wagna Lysimeter for periods 2005 – 2008 (calibration) and 2009 – 2011 (validation).

Model	Nitrate concentrations			Nitrate-N leaching rates		
	MAE (mg L^{-1})	*RMSE*	*IoA*	*MAE* (kg ha^{-1} d^{-1})	*RMSE*	*IoA*
	Calibration 2005 – 2008 (*n* = 199)					
ARMOSA	15.71	20.37	0.78	0.043	0.085	0.77
COUP	6.74	9.60	0.97	0.041	0.085	0.78
DAISY	13.92	16.82	0.87	0.037	0.063	0.87
EPIC	19.55	25.63	0.76	0.049	0.084	0.82
SIM-STO	27.34	34.61	0.43	0.044	0.089	0.60
SW-ANIM	7.88	10.48	0.95	0.035	0.080	0.85
	Validation 2009 – 2011 (*n* = 128)					
ARMOSA	11.17	15.85	0.52	0.058	0.102	0.61
COUP	12.36	18.68	0.52	0.076	0.187	0.53
DAISY	8.54	11.40	0.78	0.045	0.095	0.54
EPIC	18.24	22.07	0.52	0.089	0.155	0.41
SIM-STO	8.88	10.44	0.78	0.058	0.138	0.56
SW-ANIM	19.97	29.37	0.43	0.205	0.800	0.12

Table 7. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of observed nitrate concentrations (mg L^{-1}) in water extracted by suction cups at depths 0.35 m and 0.9 m with simulated concentration.

Model	Calibration (0.9 m; *n* = 47)			Validation (0.35 m; *n* = 91)			Validation (0.9 m; *n* = 108)		
	MAE	*RMSE*	*IoA*	*MAE*	*RMSE*	*IoA*	*MAE*	*RMSE*	*IoA*
ARMOSA	36.8	50.6	0.66	22.7	35.9	0.65	12.7	16.6	0.58
COUP	28.0	35.2	0.80	28.2	44.1	0.38	16.6	24.1	0.37
DAISY	32.2	43.9	0.68	29.1	50.9	0.46	12.9	21.5	0.55
SIM-STO	50.6	66.7	0.44	25.5	36.3	0.68	13.6	15.8	0.71
SW-ANIM	25.5	30.5	0.84	36.4	59.3	0.57	20.8	33.8	0.41

Table 8. Comparison of seasonal soil nitrogen balances observed and calculated by the six benchmark models.

For each year the main crop is indicated, but these where preceded by catch crops (including leguminous crops).

Crop and period	Balance term[†] (kg ha^{-1})	Observed	Simulated ARMOSA	COUP	DAISY	EPIC	SIM-STO	SW-ANIM
			Calibration 2005 – 2008					
Oil	Fertilization* (+)	35.1+27.4	63.0	62.5	62.9	53.1	62.4	62.5
pumpkin	Deposition (+)		10.2	3.1	16.9	5.0	6.8	11.5
1.1.2005	Biological fixation (+)		41.5	1.7	0.1	1.8	31.3	81.3
–	Volatilization (-)		2.7	0.0	1.0	1.5	1.9	2.1
13.9.2005	Crop off-take (-)	50.9	59.7	55.3	83.3	0.0	44.3	70.0
	NO_3-N leaching (-)	22.2	17.2	27.9	25.8	30.3	3.6	15.3
	Other transport[$] (-)		0.0	3.2	0.0	0.9	0.0	0.0
	Denitrification (-)		0.0	0.0	13.0	11.8	0.0	0.1
	Storage change[#]		35.2	-19.1	-43.2	15.4	50.6	67.8
Maize	Fertilization* (+)	75.6+54.5	131.0	130.1	130.7	112.3	130.1	130.1
	Deposition (+)		15.4	4.8	26.5	8.0	10.7	17.8
14.9.2005	Biological fixation (+)		28.4	32.7	0.0	0.0	0.0	112.9
–	Volatilization (-)		9.6	0.0	9.8	8.8	4.9	2.4
2.10.2006	Crop off-take (-)	137.8	211.6	116.0	197.9	125.5	72.7	134.8
	NO_3-N leaching (-)	25.7	27.9	25.8	22.7	33.6	25.1	29.7
	Other transport[$] (-)		0.0	6.0	0.0	1.2	0.0	0.2
	Denitrification (-)		0.0	0.0	13.6	45.8	0.0	1.3
	Storage change[#]		-74.5	19.9	-86.8	-94.6	38.1	92.4
Maize	Fertilization* (+)	59.0+120.7	185.0	179.7	179.4	136.6	179.7	184.5
	Deposition (+)		14.2	4.3	22.2	6.4	8.7	15.3
3.10.2006	Biological fixation (+)		52.9	24.7	0.0	0.0	0.0	32.8
–	Volatilization (-)		10.9	0.0	2.7	18.5	5.5	28.5
21.9.2007	Crop off-take (-)	92.7	61.4	107.6	2.1	99.7	75.7	96.7
	NO_3-N leaching (-)	5.9	4.4	7.1	6.3	5.4	8.8	5.8
	Other transport[$] (-)		0.0	3.2	0.0	1.5	0.0	0.0
	Denitrification (-)		0.0	0.0	15.3	33.6	0.0	2.0
	Storage change[#]		175.4	90.8	175.2	-15.7	98.4	99.6
Winter	Fertilization* (+)	38.0+84.6	123.0	122.6	123.5	78.2	122.6	123.2
barley	Deposition (+)		11.3	3.3	15.0	3.9	5.3	10.7
22.9.2007	Biological fixation (+)		0.0	0.1	0.0	0.0	0.0	14.0
–	Volatilization (-)		0.2	0.0	2.6	5.4	22.7	5.1
30.6.2008	Crop off-take (-)	132.3	66.2	104.7	139.0	114.2	81.8	118.4
	NO_3-N leaching (-)	18.9	13.5	18.5	11.7	12.3	5.7	22.2
	Other transport[$] (-)		0.0	3.4	0.0	0.4	0.0	0.0
	Denitrification (-)		0.0	0.0	11.7	40.6	0.0	1.1
	Storage change[#]		54.4	-0.7	-26.4	-90.8	17.7	1.2
			Validation 2009 – 2011					
Oil	Fertilization* (+)	52.0+0.0	52.0	52.0	52.0	51.3	52.0	52.0
pumpkin	Deposition (+)		12.4	5.9	40.1	13.6	18.4	26.0
1.7.2008	Biological fixation (+)		52.1	41.2	0.0	0.0	0.0	22.7
–	Volatilization (-)		4.4	0.0	0.0	3.9	0.0	0.0
7.9.2009	Crop off-take (-)	56.9	113.6	59.9	97.2	0.0	72.3	45.7
	NO_3-N leaching (-)	33.1	44.2	61.5	26.4	16.0	32.5	72.1
	Other transport[$] (-)		0.0	8.0	0.1	1.9	0.0	0.2
	Denitrification (-)		0.0	0.2	70.6	31.1	0.0	3.4
	Storage change[#]		-45.8	-30.4	-102.1	11.9	-34.4	-20.7
Maize	Fertilization* (+)	81.0+62.6	144.0	143.6	143.1	112.7	143.6	154.3
	Deposition (+)		7.6	4.7	26.6	8.1	11.0	18.0
8.9.2009	Biological fixation (+)		0.0	41.3	0.0	0.0	0.0	88.9
–	Volatilization (-)		7.2	0.0	2.2	4.8	4.5	9.2
23.9.2010	Crop off-take (-)	142.4	127.6	96.9	240.3	85.0	78.6	115.5
	NO_3-N leaching (-)	3.6	17.0	14.6	8.7	19.3	13.1	32.9

	Other transport[$] (-)		0.0	5.4	0.0	3.5	0.0	0.2
	Denitrification (-)		0.0	0.0	23.4	47.9	0.0	0.7
	Storage change[#]		-0.2	72.7	-104.9	-39.7	58.3	102.8
Triticale	Fertilization* (+)	62.0+119.1	181.0	180.4	181.8	111.8	181.1	181.7
	Deposition (+)		5.9	3.5	16.7	4.6	6.1	11.7
24.9.2010 – 13.7.2011	Biological fixation (+)		0.0	12.8	0.0	0.0	0.0	18.2
	Volatilization (-)		8.1	0.0	4.6	5.5	41.4	19.8
	Crop off-take (-)	155.8	152.0	44.5	161.5	170.3	143.0	83.6
	NO_3-N leaching (-)	13.9	6.1	3.2	7.6	30.3	13.3	31.0
	Other transport[$] (-)		0.0	2.5	0.0	0.6	0.0	0.2
	Denitrification (-)		0.0	0.0	13.5	38.4	0.0	1.5
	Storage change[#]		20.7	146.5	11.2	-128.8	-10.4	75.5
			Seven year totals 2005– 2011					
All	Fertilization* (+)	871.6	879.0	870.9	873.5	656.1	871.4	888.2
	Deposition (+)		77.0	29.6	164.0	49.6	67.0	111.1
1.1.2005 – 13.7.2011	Biological fixation (+)		174.9	154.6	0.1	1.8	31.3	370.9
	Volatilization (-)		43.2	0.0	22.9	48.5	80.8	67.1
	Crop off-take (-)	768.8	792.1	584.8	921.2	594.8	568.4	664.7
	NO_3-N leaching (-)	123.3	130.3	158.6	109.1	147.3	102.2	209.0
	Other transport[$] (-)		0.0	31.7	0.1	10.0	0.0	0.8
	Denitrification (-)		0.0	0.3	161.3	249.2	0.0	10.0
	Storage change[#]		165.3	279.7	-177.0	-342.2	218.0	418.6

† + indicates input; - indicates output

* Fertilization includes the addition of mineral fertilizer (first number) and the amendment of animal manure (second number)

$ Other transport includes the leaching of NH_4-N and dissolved organic matter and the transport of N-components by surface runoff water flow

A positive value refers to an increase of the nitrogen stock in soil and a negative value indicates its depletion

Table 9. Qualitative assessment of the model performance (*IoA*) for daily or weekly results for the calibration and validation periods.

Phase	Indicator	Item		**ARMOSA**	**COUP**	**DAISY**	**EPIC**	**SIM-STO**	**SW-ANIM**
Calibration	-: $IoA < 0.6$ o: $0.6 \leq IoA < 0.8$ +: $IoA \geq 0.8$	Soil moisture retention relation	0.35 m	+	-	o	n.a.	o	+
			0.9 m	o	-	o	n.a.	+	+
			1.8 m	-	-	-	n.a.	o	+
		Simulated water contents	0.35 m	o	o	-	-	+	+
			0.9 m	o	+	o	+	+	+
			1.8 m	-	+	-	-	+	-
		Nitrate concentration	0.9 m	o	+	o	n.a.	-	+
	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Water flux, daily		o	o	-	o	o	o
		Water volumes per sampling interval		o	o	-	o	+	o
		Nitrate concentration in water samples		o	+	o	o	-	+
		Nitrate-N flux per sampling interval		o	o	o	o	-	o
Validation	-: $IoA < 0.6$ o: $0.6 \leq IoA < 0.8$ +: $IoA \geq 0.8$	Soil water contents	0.9 m	-	o	o	-	+	o
			1.8 m	-	+	-	-	+	o
		Nitrate	0.35 m	o	-	-	n.a.	o	-
			0.9 m	-	-	-	n.a.	o	-
	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Water flux, daily		o	o	o	o	+	o
		Water volume per sampling interval		-	o	-	-	o	-
		Nitrate concentration in water samples		-	-	o	-	o	-
		Nitrate-N flux per sampling interval		-	-	-	-	-	-

n.a.: not applicable

Table 10. Mean absolute errors (*MAE*) of seasonal percolated water, N crop off-take and leached nitrate-N amounts for seven seasons (MAE_7) and for the best five seasons (MAE_5).

Seasonal quantity	Indicators	**ARMOSA**	**COUP**	**DAISY**	**EPIC**	**SIM-STO**	**SW-ANIM**
Percolated water (mm)	MAE_7	21.3	24.2	63.9	48.6	14.6	40.3
	MAE_5	16.0	14.3	30.5	30.5	11.8	32.8
N crop off-take (kg ha^{-1})	MAE_7	36.5	32.7	47.7	31.0	33.0	21.5
	MAE_5	23.1	14.3	29.0	20.6	20.5	10.3
Leached NO_3-N (kg ha^{-1})	MAE_7	6.6	8.2	4.6	10.3	6.6	14.2
	MAE_5	4.4	3.6	3.7	7.8	2.8	6.3

Figures captions

Figure 1 Blind test comparison of seasonal water fluxes, flow averaged nitrate concentration and nitrate-N fluxes simulated by five models (excluding **SIM-STO**) with observations. Results of individual models are indicated by markers.

Figure 2 Measured values and calibrated soil moisture retention curves at depths 0.35 m,0.9 m and 1.8 m.

Figure 3 Comparison of simulated and measured inner season cumulative water fluxes, nitrate concentrations and inner season cumulative nitrate-N fluxes at depth 1.8 m in the low input farming lysimeter at the Wagna experimental field station

Figure 4 Taylor plots of the statistical performance of the simulated water fluxes at depth 1.8 m for daily values (left) and for sampling interval averaged values (right). Circles refer to the calibration results and triangles refer to the validation results. A = **ARMOSA**, C = **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**

Figure 5 Comparison of simulated and measured seasonal water fluxes (mm) at depth 1.8 m in the low input farming lysimeter at the Wagna experimental field station

Figure 6 Taylor plot of the statistical performance parameters for the simulated nitrate concentrations (left) and nitrate-N fluxes (right) at depth 1.8 m. Circles refer to the calibration results and Triangles refer to the validation results. Indicators of **SW-ANIM** nitrate-N fluxes fall outside the range (2.5; 8.5). A = **ARMOSA**, C = **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**

Figure 7 Seven years balances for fertilization minus crop off-take and nitrate-N leaching (all in kg ha^{-1}), summed since the start of the calibration period

Figure 8 Effect of a leave-one-out calculation of a certain data pair of observed and simulated water fluxes on the Index of Agreement, IoA (see text for further explanation).

Figure 1
Click here to download Figure: Fig_01_STOTEN-D-14-00814.pdf

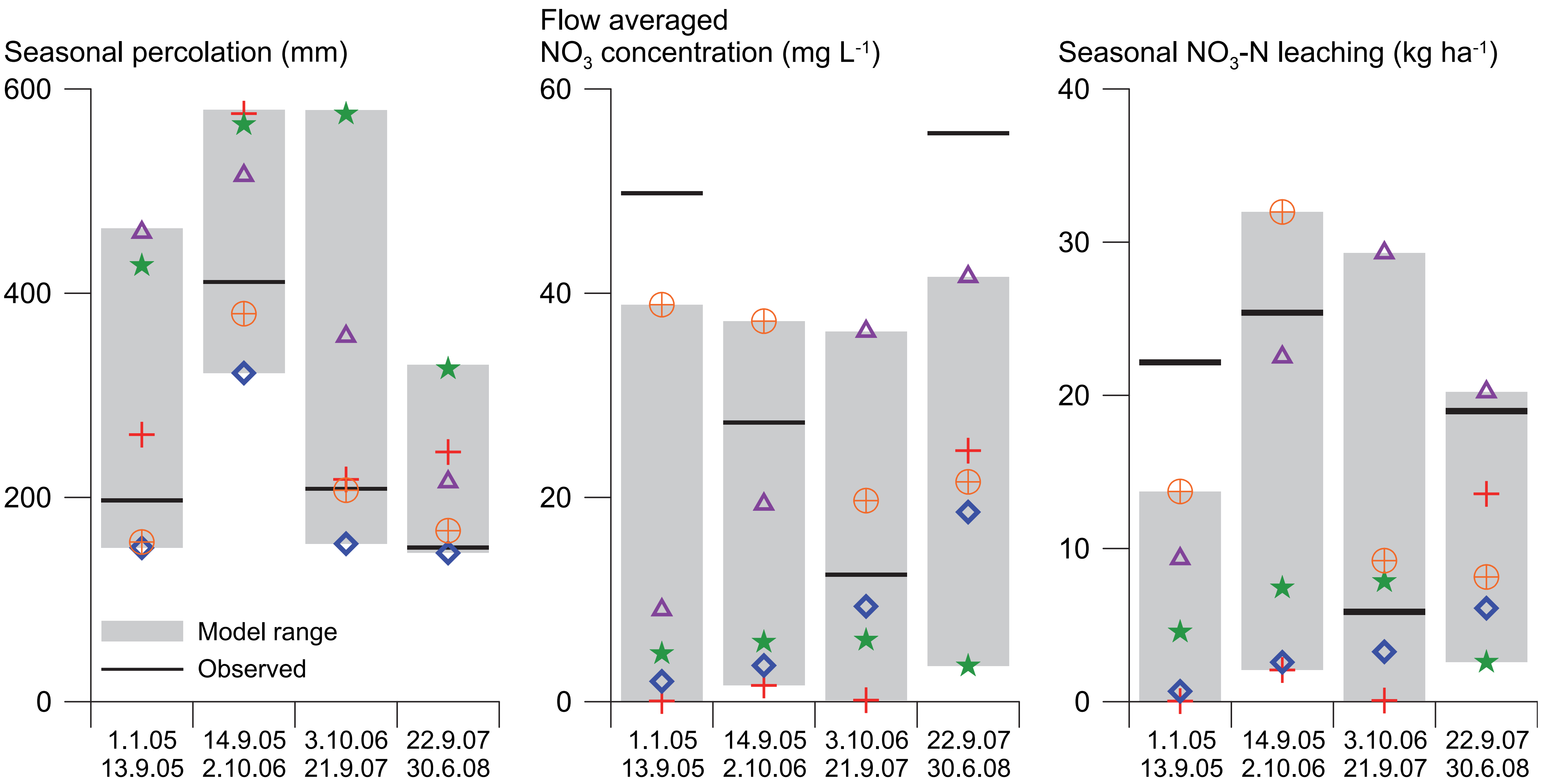

Figure 2
Click here to download Figure: Fig_02_STOTEN-D-14-00814.pdf

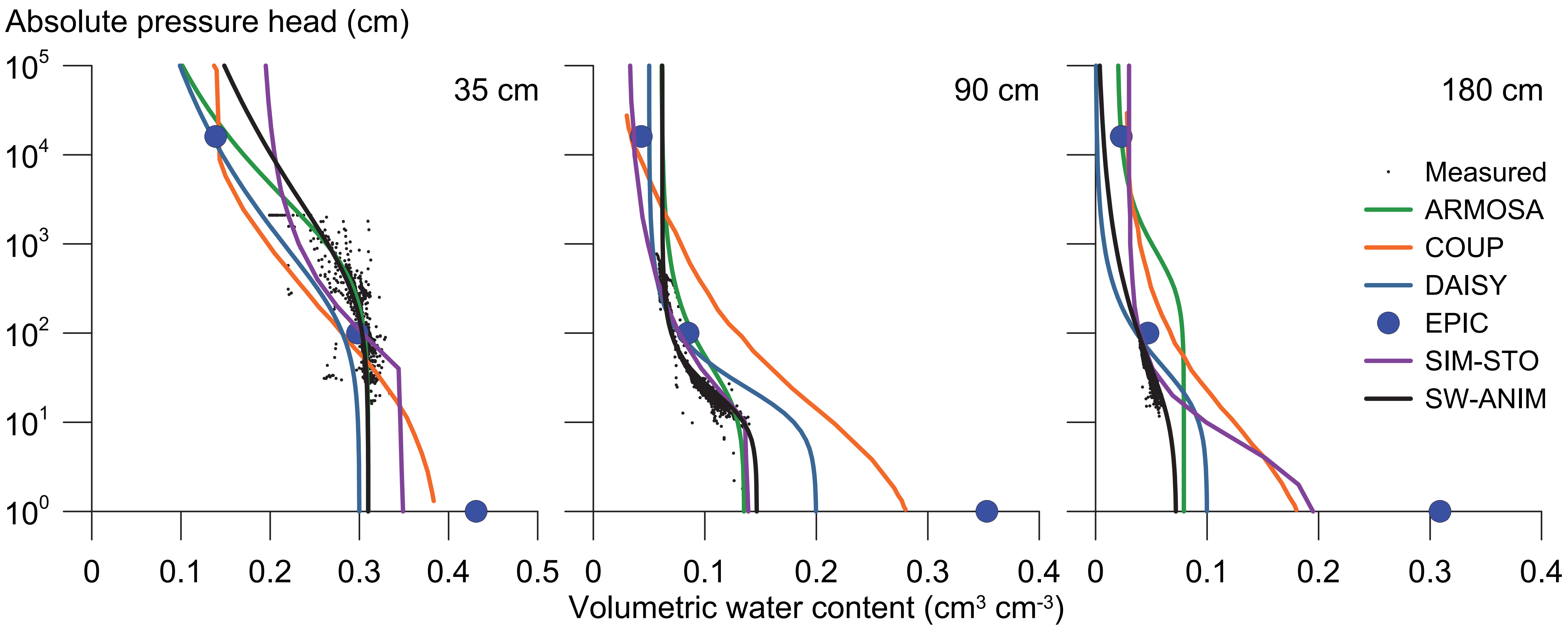

Figure 3

Click here to download Figure: Fig_03_STOTEN-D-14-00814.pdf

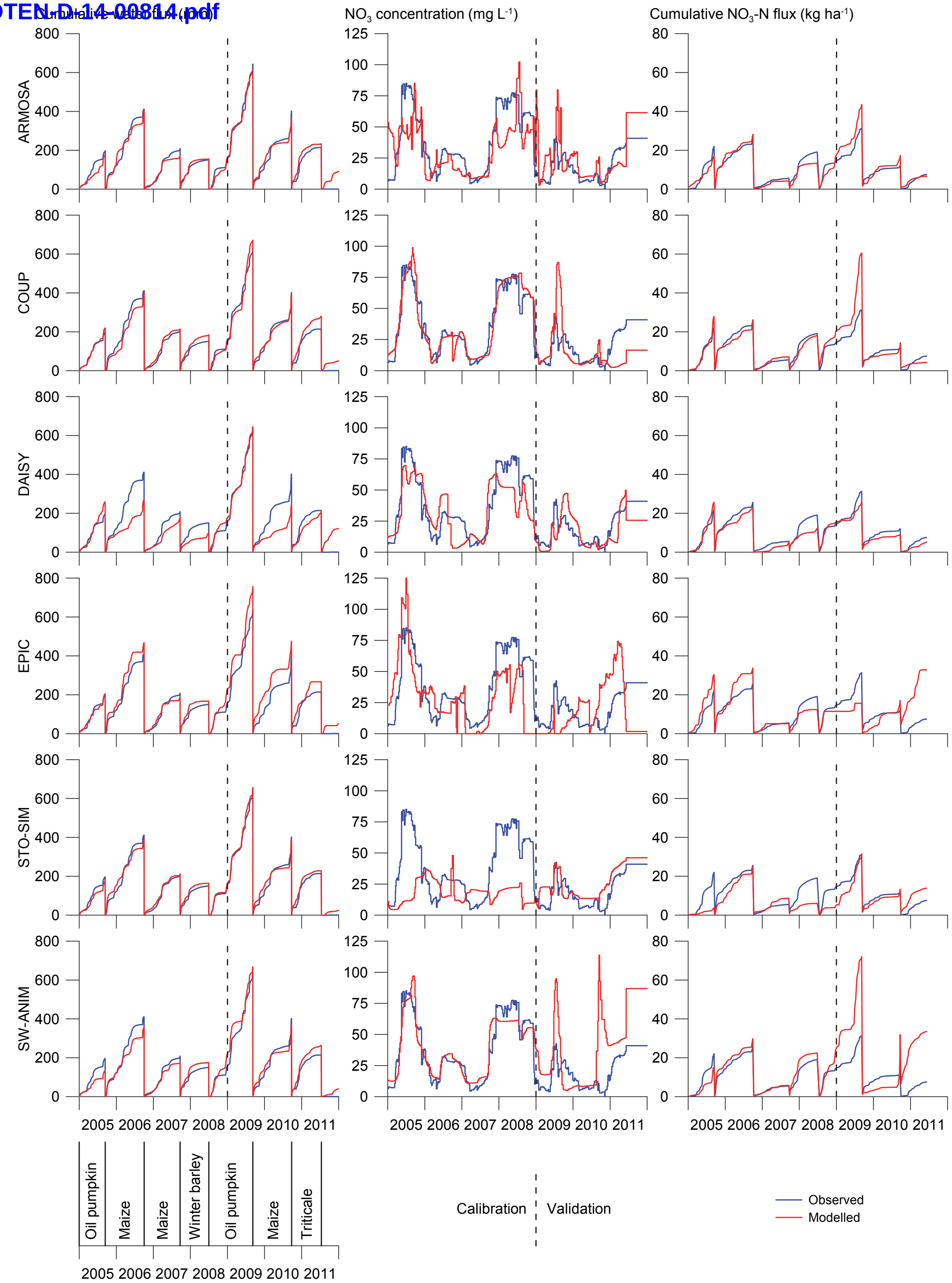

Figure 4

Click here to download Figure: Fig_04_STOTEN-D-14-00814.pdf

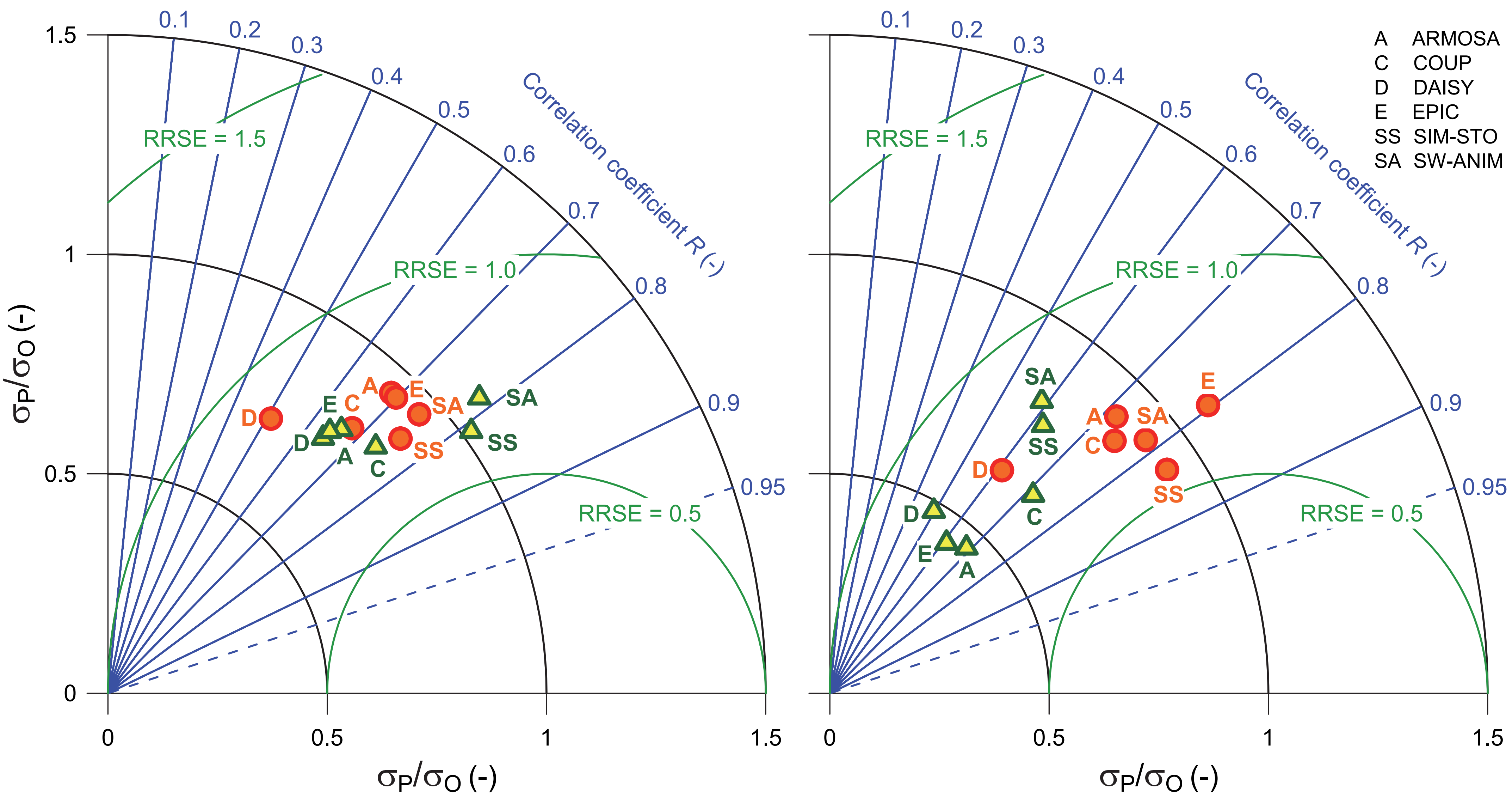

Figure 5

Click here to download Figure: Fig_05_STOTEN-D-14-00814.pdf

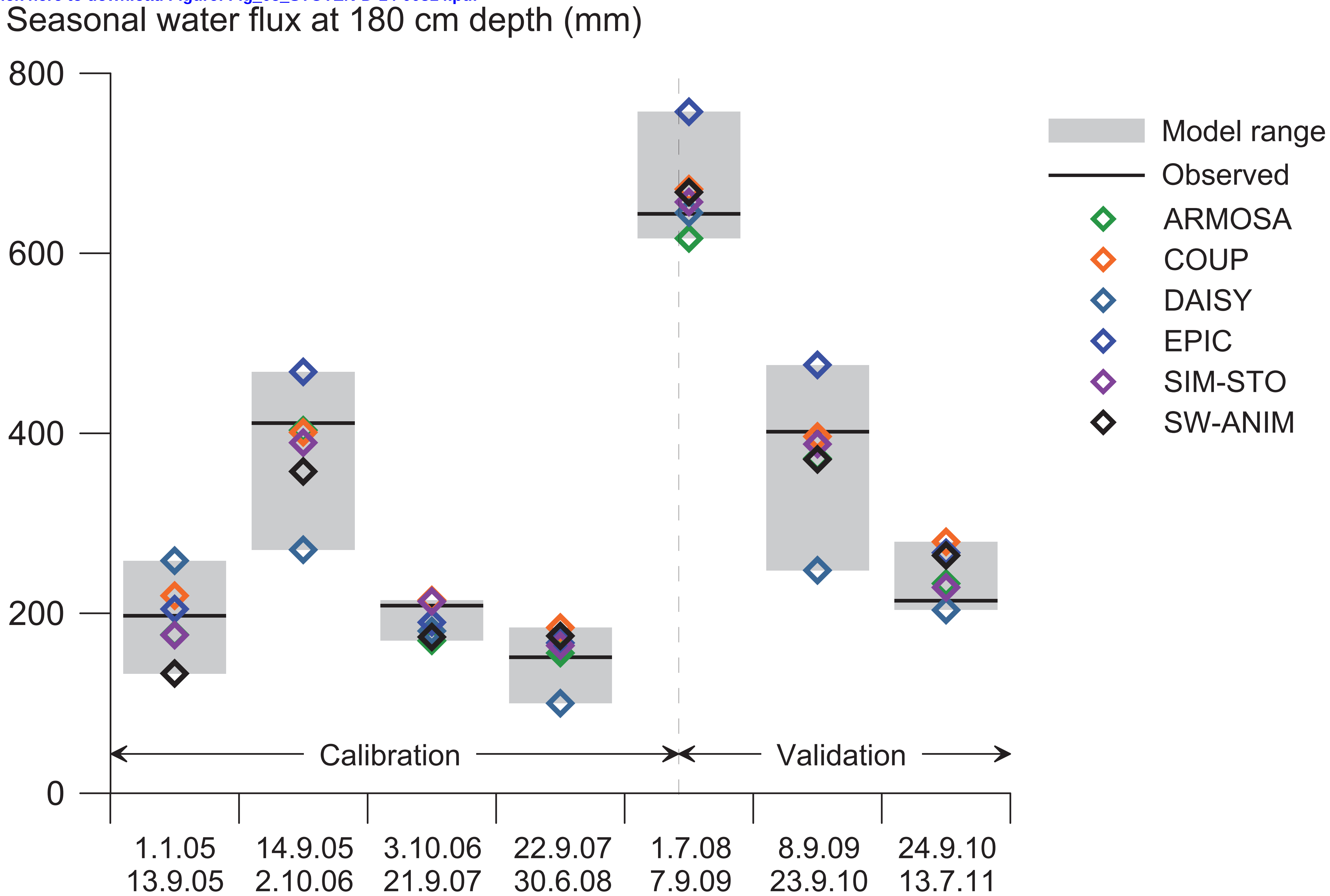

Figure 6

Click here to download Figure: Fig_06_STOTEN-D-14-00814.pdf

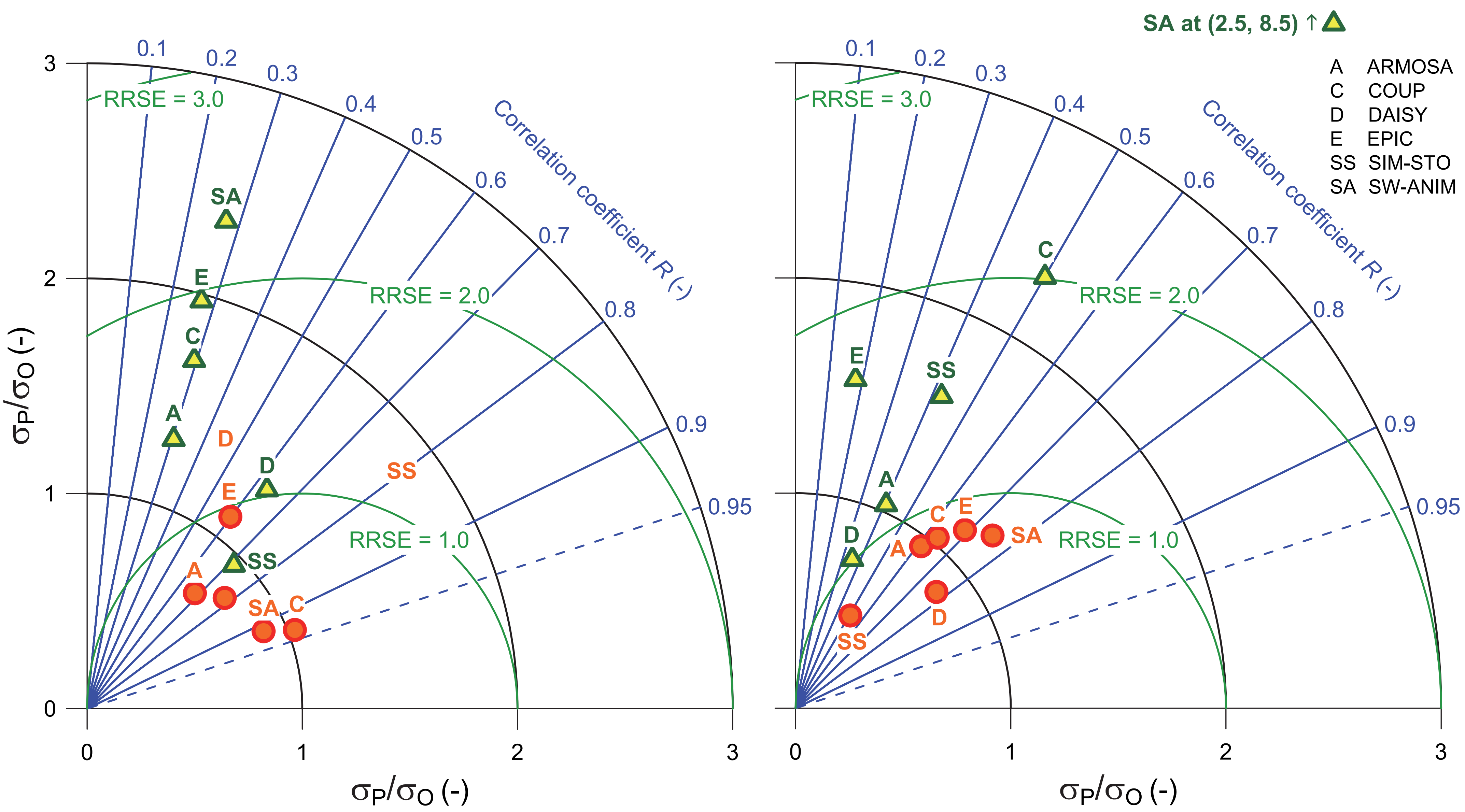

Figure 7
Click here to download Figure: Fig_07_STOTEN-D-14-00814.pdf

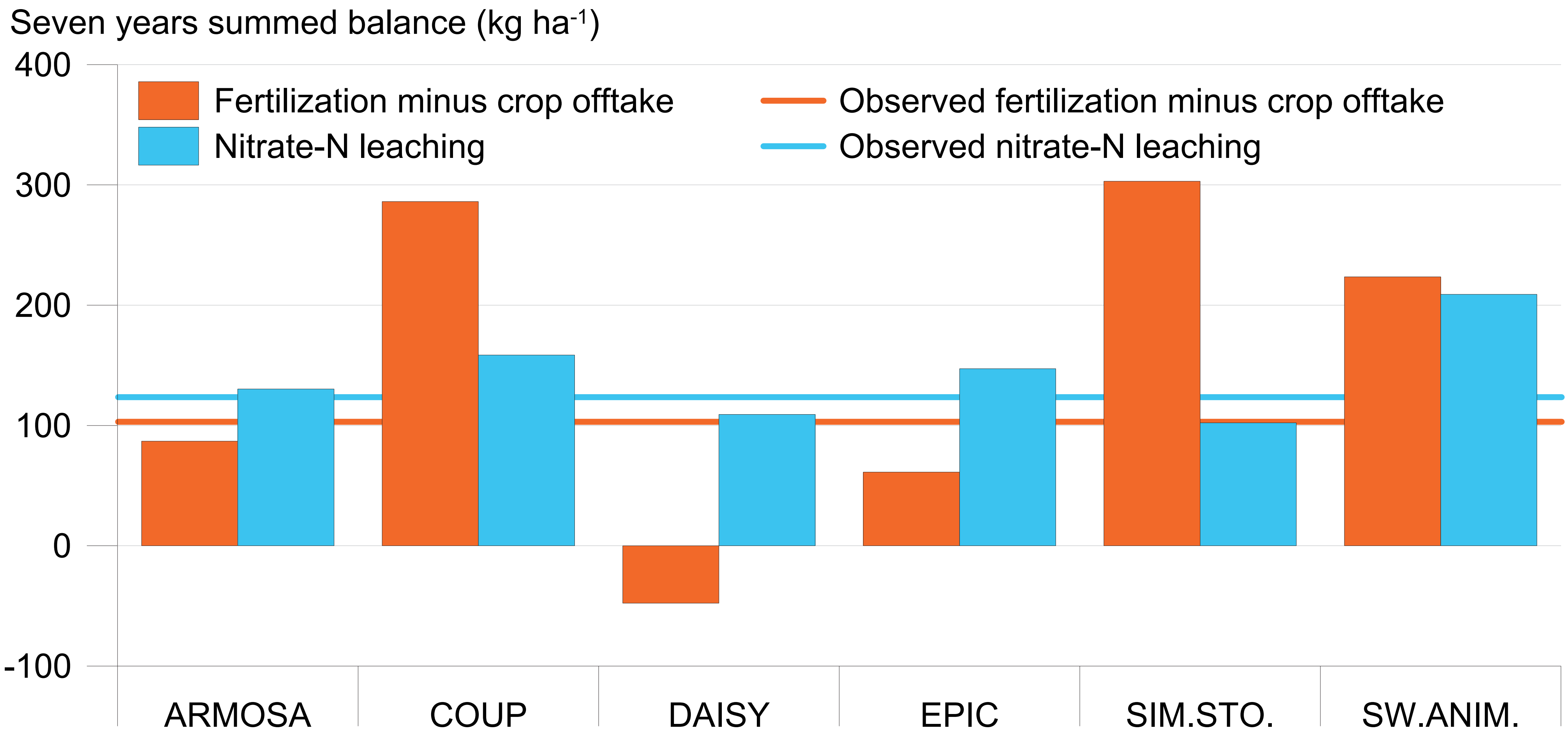

Figure 8

Click here to download Figure: Fig_08_STOTEN-D-14-00814.pdf