

Document downloaded from:

<http://hdl.handle.net/10251/61621>

This paper must be cited as:

Groenendijk, P.; Heinen, M.; Klammler, G.; Fank, J.; Kupfersberger, H.; Pisinaras, V.; Gemitzi, A.... (2014). Performance assessment of nitrate leaching models for highly vulnerable soils used in low-input farming based on lysimeter data. *Science of the Total Environment*. 499:463-480. doi:10.1016/j.scitotenv.2014.07.002.



The final publication is available at

<http://dx.doi.org/10.1016/j.scitotenv.2014.07.002>

Copyright Elsevier

Additional Information

1  
2  
3  
4 1 **Performance assessment of nitrate leaching models for highly**  
5  
6  
7 2 **vulnerable soils used in low input farming based on lysimeter data**  
8  
9

10 3 Piet Groenendijk<sup>a,\*</sup>, Marius Heinen<sup>a</sup>, Gernot Klammler<sup>b</sup>, Johann Fank<sup>b</sup>, Hans Kupfersberger<sup>b</sup>, Vassilios  
11 4 Pisinaras<sup>c</sup>, Alexandra Gemitzi<sup>c</sup>, Salvador Peña-Haro<sup>d</sup>, Alberto García-Prats<sup>e</sup>, Manuel Pulido-Velazquez<sup>f</sup>, Alessia  
12 5 Perego<sup>g</sup>, Marco Acutis<sup>g</sup>, Marco Trevisan<sup>h</sup>  
13  
14  
15  
16  
17  
18  
19

20 7 a Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands  
21

22 8 b Joanneum Research, Forschungsgesellschaft mbH, Leonhardstraße 59, 8010 Graz, Austria  
23  
24

25 9 c Democritus University of Thrace, Department of Environmental Engineering, Vas. Sofias 12, Xanthi, 67100,  
26 10 Greece  
27  
28  
29

30 11 d Institute of Environmental Engineering, ETH Zurich, Wolfgang-Pauli-Str. 15, CH-8093 Zurich, Switzerland  
31

32 12 e Universitat Politècnica de València, Department of Hydraulic Engineering and Environment, Camino de  
33 13 Vera, 46022 Valencia, Valencia, Spain  
34  
35  
36

37 14 f Universitat Politècnica de València, Research Institute of Water and Environmental Engineering (IIAMA),  
38 15 Camino de Vera, 46022 Valencia, Valencia, Spain  
39  
40  
41

42 16 g University of Milan, Department of Agricultural and Environmental Science, Via G. Celoria 2 20133, Milan,  
43 17 Italy  
44  
45  
46

47 18 h Università Cattolica del Sacro Cuore, sede di Piacenza, Via Emilia Parmense, 84 29100, Piacenza, Italy  
48  
49  
50  
51

52 20 \*Corresponding author: Piet Groenendijk, Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands, Email:  
53 21 [piet.groenendijk@wur.nl](mailto:piet.groenendijk@wur.nl) Tel.: +31 317 486434  
54  
55  
56

57 22  
58  
59 23 **Abstract**  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 24 The agricultural sector faces the challenge of ensuring food security without an excessive burden on the  
5  
6 25 environment. Simulation models provide excellent instruments for researchers to gain more insight into relevant  
7  
8 26 processes and best agricultural practices and provide tools for planners for decision making support. The extent  
9  
10 27 to which models are capable of reliable extrapolation and prediction is important for exploring new farming  
11  
12 28 systems or assessing the impacts of future land and climate changes.

13  
14 29 A performance assessment was conducted by testing six detailed state-of-the-art models for simulation of nitrate  
15  
16 30 leaching (**ARMOSA, COUPMODEL, DAISY, EPIC, SIMWASER/STOTRASIM, SWAP/ANIMO**) for  
17  
18 31 lysimeter data of the Wagna experimental field station in Eastern Austria, where the soil is highly vulnerable to  
19  
20 32 nitrate leaching.

21  
22 33 Three consecutive phases were distinguished to gain insight in the predictive power of the models: 1) a blind test  
23  
24 34 for 2005 – 2008 in which only soil hydraulic characteristics, meteorological data and information about the  
25  
26 35 agricultural management were accessible; 2) a calibration for the same period in which essential information on  
27  
28 36 field observations was additionally available to the modellers; and 3) a validation for 2009 – 2011 with the  
29  
30 37 corresponding type of data available as for the blind test. A set of statistical metrics (mean absolute error, root  
31  
32 38 mean squared error, index of agreement, model efficiency, root relative squared error, Pearson's linear  
33  
34 39 correlation coefficient) was applied for testing the results and comparing the models.

35  
36  
37 40 None of the models performed good for all of the statistical metrics. Models designed for nitrate leaching in high  
38  
39 41 input farming systems had difficulties in accurate predicting leaching in low input farming systems that are  
40  
41 42 strongly influenced by the retention of nitrogen in catch crops and nitrogen fixation by legumes. An accurate  
42  
43 43 calibration does not guarantee a good predictive power of the model. Nevertheless all models were able to  
44  
45 44 identify years and crops with high and low leaching rates.

#### 46 47 **Keywords**

48  
49 46 Lysimeter, model comparison, nitrate leaching, performance assessment, predictive power, simulation model  
50  
51  
52

## 53 54 **1. Introduction**

55  
56 48 Agriculture is the major land use in Europe (ca. 50% of overall land area) and has strongly increased its use of  
57  
58 49 external inputs (fertiliser, pesticides and water) over the last 50 years. The environmental effects of intensive  
59  
60 50 agriculture include a decline in biodiversity, eutrophication of ecosystems and surface waters, acidification,  
61  
62  
63  
64  
65

1  
2  
3  
4 51 global warming, air pollution and diffuse nitrate pollution of groundwater. A global challenge is to produce  
5  
6 52 enough food for the ever-growing population and at the same time minimizing the loss of reactive nitrogen (N)  
7  
8 53 to the environment. Since the 1980s, agriculture in Western Europe has managed to reduce its N surpluses,  
9  
10 54 owing to stringent national and European community policies (Vitousek et al., 2009; Grizzetti et al., 2011).

11  
12 55 The main aim of the Nitrates Directive (EU, 1991: Directive 91/676/EEC) is to reduce water pollution caused or  
13  
14 56 induced by nitrates and phosphorus from agricultural sources. The Nitrates Directive legally restricts farm  
15  
16 57 application of manure to 170 kg ha<sup>-1</sup> of nitrogen, or in case of derogation to inputs up to 250 kg ha<sup>-1</sup> (Oenema,  
17  
18 58 2004). An implementation measure of the Nitrates Directive is the establishment of codes of Good Agricultural  
19  
20 59 Practice. Recommended measures include, among others, the application of crop rotations, the cultivation of a  
21  
22 60 soil winter cover and catch crops to prevent nitrate leaching and run-off during wet seasons. Catch crops create a  
23  
24 61 new challenge in the assessment of environmental effects of crop rotations. In theory, catch crops take up N that  
25  
26 62 would otherwise be lost, and, after incorporation of the crop residues into the soil, make this N available to the  
27  
28 63 succeeding crop via mineralization. However, the influence of a catch crop on the nitrogen supply to the  
29  
30 64 succeeding crop can vary greatly and range from a positive to a negative effect (Nett et al., 2011). The effect is  
31  
32 65 determined by the N uptake capacity, the rooting depth of a catch crop, the weather and soil conditions as well as  
33  
34 66 the rooting depth of the succeeding crop (Thorup-Kristensen, 2006).

35  
36 67 Models are an important tool for assessment of environmental impacts of a certain agricultural practice and are  
37  
38 68 also an instrument for increasing the understanding of the biological, pedological and hydrological factors that  
39  
40 69 affect productivity and the risk of nitrate leaching. For this reason, for more than 30 years simulation models  
41  
42 70 have been developed and applied in the research on nitrate leaching. The different model descriptions are a  
43  
44 71 reflection of the intended purpose, the physical conditions and the available data for model application and the  
45  
46 72 knowledge and skill of the model developer. Technical implementations have evolved from stand-alone model  
47  
48 73 codes to modelling platforms comprising modular models able to include and compare different process  
49  
50 74 descriptions.

51  
52 75 Calibration and validation of models contributes to their reliability. In addition also an analysis of the  
53  
54 76 implemented process descriptions and the mutual comparison of models provides information on the predictive  
55  
56 77 power. Several model comparison studies have been conducted in which nitrate leaching models were compared  
57  
58 78 (De Willigen and Neeteson, 1985; Vereecken et al., 1991; De Willigen, 1991; Diekkrüger et al., 1995; Moreels  
59  
60 79 et al., 2003; Kersebaum et al., 2007; Jabro et al., 2012). Most of them were related to ordinary agricultural

1  
2  
3  
4 80 conditions with a single crop on a typical agricultural soil. Thus, there is no information (comparison) available  
5  
6 81 for situations in soils that are highly vulnerable to nitrate leaching in combination with low-input conditions and  
7  
8 82 the use of catch crops.  
9

10 83 It is widely recognised that despite the deterministic nature of process oriented models they often have a limited  
11  
12 84 validity range for certain climatic, pedological, hydrological and agronomic circumstances characterised by high  
13  
14 85 inputs. It is not clear whether the models are able to produce relatively reliable predictions for low input  
15  
16 86 conditions. A better insight into the model performance for such uncommon circumstances underpins  
17  
18 87 conclusions about the predictive power.

19  
20  
21 88 In this study a number of models were inter-compared for low input conditions of one of the lysimeters of the  
22  
23 89 Wagna experimental research station, Austria (Klammler and Fank, 2014; this issue) for three typical conditions  
24  
25 90 for which they were not designed: 1) the crop rotation which included an uncommon crop (oil pumpkin), 2)  
26  
27 91 catch crops for which the N-uptake was not measured, and 3) the soil consisted of a shallow soil vulnerable to  
28  
29 92 nitrate leaching on top of a high conductive gravel layer. The objectives of this study were: 1) to assess the  
30  
31 93 performance of state-of-the-art nitrate leaching models as they are used in the scientific research community, for  
32  
33 94 the above mentioned conditions, 2) to inter-compare the models for analysing their predictive power, and 3) to  
34  
35 95 identify strengths and weaknesses of bio-physically based models.  
36  
37

## 38 96 **2. Materials and Methods**

### 39 40 41 97 **2.1 Description of the lysimeter**

42  
43 98 Observations were used of a lysimeter located in the agricultural experimental field station in Wagna in Eastern  
44  
45 99 Austria (46° 46.113'N, 15° 33.140'E; altitude 265 m; Klammler and Fank, 2014 (this issue)). Since 1987  
46  
47 100 different cultivation strategies are investigated concerning nitrogen-fertilizer input, nitrate leaching and crop  
48  
49 101 yields. In 2004, the cultivation changed into comparing low-input farming and organic farming, each covering  
50  
51 102 50% of the test site. Since then, two of the test plots have been equipped with two weighable, monolithic, high-  
52  
53 103 precision lysimeters (2 m depth, 1 m<sup>2</sup> surface). The lysimeter in the conventional tillage test plot (KON-system)  
54  
55 104 is subject for this study. Cultivation practices including crop species, sowing and harvest dates, and fertilizer  
56  
57 105 applications in the test plot are presented in Table 1.  
58

59  
60 106 <<Table 1 >>  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

107 The lysimeters are equipped with soil water samplers, soil moisture probes, matrix sensors/tensiometer and soil  
108 temperature probes at four measuring depths (0.35, 0.6, 0.9, 1.8 m). An accompanied measuring profile for soil  
109 moisture, matrix potential and soil temperature is also installed outside the lysimeters (same depths as inside the  
110 lysimeter) to determine if the conditions inside the lysimeter are representative for the rest of the field. At the  
111 bottom of the lysimeter (depth 1.8 m) a suction cups rake was installed which kept the pressure head at this depth  
112 equal to that outside the lysimeter. The water sucked off was collected, weighted and sampled for the  
113 determination of the nitrate concentration. While quantity of seepage water was recorded automatically in  
114 0.1 mm resolution by a tipping bucket, nitrogen concentration in the accumulated leachate was analysed in an  
115 approximately weekly interval. Furthermore, a weather station is installed at agricultural test site in Wagna for  
116 the recording of air temperature, relative humidity, shortwave solar radiation, wind speed, wind direction,  
117 precipitation, sunshine duration and atmospheric pressure at high temporal resolution (Klammler and Fank,  
118 2014; this issue). Annual precipitation rates and cumulative probabilities of the rates relative to the values of the  
119 period 1961 – 2011 are presented in Table 2.

<<Table 2>>

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

124 **2.2 Description of models**

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

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

135 al., 2010) and on **SOILN** for simulation of the soil organic matter and nitrogen cycle and nitrate leaching  
136 (Bergström et al., 1991).

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

1  
2  
3  
4 164 is part of the National Dutch modelling system **STONE** for the evaluation of fertiliser policy measures (Wolf  
5  
6 165 et al., 2003). The name of the sequentially coupled models is abbreviated as **SW-ANIM**.

7  
8  
9 166 In addition to soil processes also the description of crop development is considered, because the plant related  
10  
11 167 processes such as evaporation, nitrogen and nitrogen supply with crop residues exert a major influence on the  
12  
13 168 water balance and nutrient dynamics in the soil.

14  
15 169 Except for **SW-ANIM**, all models simulate the growth of plant biomass. Although **SW-ANIM** has the  
16  
17 170 possibility to calculate the biomass development in a detailed manner, the modellers had chosen to use a simple  
18  
19 171 option of a supposed development of leaf area index, crop height and rooting depth, because the parameters  
20  
21 172 required for detailed simulation of oil pumpkin and catch crops were not available. Except for **EPIC**, the models  
22  
23 173 describe water flow with either the Richards' (1931) equation or the Darcy (1856) - Buckingham (1907)  
24  
25 174 equation, in which the soil water retention and the hydraulic conductivity relations are described according to  
26  
27 175 Mualem (1976) - Van Genuchten (1980). **EPIC** simulates soil water flow as a storage routing process in which  
28  
29 176 percolation occurs when the soil water content of the root zone exceeds the field capacity. In **EPIC** the soil water  
30  
31 177 characteristics are calculated on the basis of texture data and the organic matter content in accordance with  
32  
33 178 Saxton and Rawls (2006).

34  
35 179 All models consider ammonium and nitrate as separate mineral nitrogen pools, and simulate organic bounded  
36  
37 180 nitrogen associated with the organic carbon cycle. **SW-ANIM** simulates also the transport and transformation of  
38  
39 181 dissolved organic nitrogen. The method of simulating biological N-fixation is one of the striking differences  
40  
41 182 between the models. The **DAISY** model was applied in a way that biological N-fixation was ignored and the  
42  
43 183 **SW-ANIM** model accounted for this process by the specification of continuous organic material additions  
44  
45 184 representing imposed fixation rates. The other models use relationships based on the crop type, the crop  
46  
47 185 development stage and the soil mineral N status. Ammonia volatilization is not implemented in the **COUP**  
48  
49 186 model code used for this study. Some models consider only the loss of ammonia as a fraction of farmyard  
50  
51 187 manure application (**DAISY**, **SW-ANIM**) while the other models take account for environmental factors as  
52  
53 188 temperature, wind speed and soil moisture. **SIM-STO** uses standardized loss factors that account for the time  
54  
55 189 from the last soil tillage event.

56  
57 190 Uptake of ammonium and nitrate depends on the demand for mineral N for crop production and is related to the  
58  
59 191 development stage, by some models expressed by a relationship with the water uptake, and the mineral N content  
60  
61 192 of the soil.



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

205 **2.3 Experimental design of study**

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

211 **2.3.1 Step 1: Blind test**

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

223 **2.3.2 Step 2: Calibration**

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

228 **2.3.3 Step 3: Validation**

229 The validation was performed for the period 2009 - 2011, where only information about crop rotation,  
230 application of fertilizers, soil cultivation and meteorology was made available for the modelling groups after step  
231 2 (calibration) was finished. The procedure for the validation is thus similar to that of the blind test, with the  
232 difference that the models were calibrated prior to validation and that the **SIM-STO** model was included in the  
233 validation.

234 **2.3.4 Step 4: Model comparison**

235 The six models were compared for their performance with respect to 1) the soil moisture retention curves at  
236 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  
237 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  
238 amounts for the time intervals of collected water samples; 6) the nitrate concentrations of the collected water  
239 samples; 7) the nitrate-N fluxes at the bottom of the lysimeter for the time intervals of collected water samples.  
240 The comparison of results at the depth of 60 cm was excluded because measurements for this depth were only  
241 available up to Sept. 2009. Seasonal leached water amounts, nitrogen yields and nitrate-N fluxes were compared  
242 to discuss the predictive power for practice oriented model applications. A nitrogen balance was set up for all  
243 models. Water fluxes at 1.8 m depth were evaluated for daily and for seasonal values. Nitrate leaching fluxes and  
244 nitrate concentrations in the leachate were evaluated at the time intervals for which the soil water was sampled.  
245 The sampling time intervals were irregular in time and the models were not able to present concentrations at  
246 these specific time events. Therefore, concentrations values for these time intervals were derived according to a  
247 volumetric averaging procedure. The nitrate concentrations at depths 0.35 m and 0.9 m can be used to get an  
248 impression whether the transport and transformation processes in soil, which ultimately lead to the leaching at  
249 depth 1.8 m, have been described adequately. Due to the nature of the model formulations, **EPIC** was not able to

1  
2  
3  
4 250 present the concentrations at the depths of measurement. The number of observations at depth 0.35 m in the  
5  
6 251 calibration period was too little and were not considered.

7  
8 252 In the models, much knowledge of soil processes is described which all contribute to the nitrate leaching at depth  
9  
10 253 1.8 m. To understand the similarities and differences between simulation results and measurements, it is  
11  
12 254 important to assess the processes. We have done this through the establishment of nitrogen balances per season.

## 15 255 **2.4 Statistical metrics**

16  
17 256 The behaviour of the main model outputs can be characterized by a number of statistical metrics to indicate the  
18  
19 257 models' ability to capture different aspects. A complete assessment of model performance should include at least  
20  
21 258 one absolute error measure and one goodness-of-fit measure (Legates and McCabe, 1999). There are a wide  
22  
23 259 range of statistical indicators used in studies on soil water and soil nitrogen, but not always a justification is  
24  
25 260 given for the indicators chosen. For *state variables* many authors use mean (absolute) error (*MAE*), root mean  
26  
27 261 square error (*RMSE*), index of agreement (*IoA*; Willmott, 1982), and less often the Nash-Sutcliffe modelling  
28  
29 262 efficiency (*NSE*; Nash-Sutcliffe, 1970) (e.g., Donatelli et al., 2004; Gribb et al., 2009; Herbst et al., 2005;  
30  
31 263 Khodaverdiloo et al., 2011; Patil and Rajput, 2009; Ritter et al., 2003; Vereecken et al., 2010). For *rate variables*  
32  
33 264 authors generally use *MAE*, mean difference (*MD*), absolute maximum error (*AME*), *RMSE*, *IoA*, *NSE*,  
34  
35 265 coefficient of determination ( $R^2$ ), percentage of error (*PE*), percentage of bias ( $P_{\text{bias}}$ ) (e.g., Akkal-Corfini et al.,  
36  
37 266 2010; Ale et al., 2012; Dawson et al., 2007, 2010; Jabro et al., 2012; Jachner et al., 2007; Kersebaum et al.,  
38  
39 267 2007; Krause et al., 2005; Moriasi et al., 2007; Qi et al., 2012; Reusser et al., 2009; Stumpp et al., 2009; Van der  
40  
41 268 Laan et al., 2011; Wang et al., 2006; Willmott et al., 1985). It appears that a few measures are used both for state  
42  
43 269 as for rate variables, which we have chosen to use here as well: *MAE*, *RMSE*, *IoA*, and *NSE* (only for rates),  
44  
45 270 given by:

46  
47 271 1. Mean absolute error: 
$$MAE = \frac{1}{n} \sum_{t=1}^n |P_t - O_t|$$

48  
49 272 2. Root mean squared error: 
$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (P_t - O_t)^2}$$

50  
51 273 3. Index of Agreement (Willmott, 1982): 
$$IoA = 1 - \frac{\sum_{t=1}^n (P_t - O_t)^2}{\sum_{t=1}^n (|P_t - O_t| + |O_t - \bar{O}|)^2}$$

52  
53 274 4. Nash-Sutcliffe model efficiency (Nash and Sutcliffe, 1970): 
$$NSE = 1 - \frac{\sum_{t=1}^n (O_t - P_t)^2}{\sum_{t=1}^n (O_t - \bar{O})^2}$$

54  
55 275 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  
56  
57 276 the mean values of observations and predictions, respectively. All four measures compare the predictions  $P_t$  and

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

277 observations  $O_t$  at the individual level, and try to express the 'spread' in  $(P_t - O_t)$  (Janssen and Heuberger, 1995).  
278 The *MAE* accounts for the deviations  $(P_t - O_t)$  in an absolute value sense. This measure is less sensitive to  
279 outliers than *RMSE*, *IoA* and *NSE*. The latter indices measures  $(P_t - O_t)$  in a quadratic sense, and, thus, are  
280 sensitive to outliers. If model errors are significant, it is more difficult to objectively assess the agreement  
281 between model and data on basis of *RMSE*. As an alternative, Willmott (1982) proposed *IoA* to express this  
282 agreement more directly. The dimensionless *IoA* has limits 0, indicating no agreement, and 1, indicating perfect  
283 agreement. The dimensionless *NSE* ranges between 1 and  $-\infty$ , where  $NSE = 1$  denotes a “perfect” model fit and  
284 for  $NSE < 0$  the average of the observations would be a better predictor than the model (Krause et al., 2005).  
285 Taylor (2001) presented a graphical method in which several statistical metrics have been combined. Such a  
286 Taylor diagram summarizes how closely a set of simulations matches the observations, and it is especially useful  
287 in evaluating multiple aspects of complex models. In normalized form, it presents the Pearson’s linear  
288 correlation coefficient (*R*) and the root relative square error (*RRSE*) as a function of the ratio of standard  
289 deviations of predictions and observations  $\sigma_P$  and  $\sigma_O$ , respectively, where

290 5. Pearson’s linear correlation coefficient

$$R = \frac{\sum_{t=1}^n (O_t - \bar{O})(P_t - \bar{P})}{\sqrt{\sum_{t=1}^n (O_t - \bar{O})^2} \sqrt{\sum_{t=1}^n (P_t - \bar{P})^2}}$$

291 6. Root relative square error:

$$RRSE = \frac{\sqrt{\sigma_P^2 + \sigma_O^2 + 2\sigma_O\sigma_P R}}{\sigma_O}$$

292 where  $\sigma_O$  and  $\sigma_P$  are the standard deviations of the observations and model predictions, respectively. A value of  
293 (1,0) in such a figure indicates a full agreement of model results with observations.

### 294 3. Results and discussion

#### 295 3.1. Blind test

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

298 <<Figure 1>>

299 Maximum deviations between simulated and observed seasonal percolation volumes of almost 400 mm were  
300 found. Two of the five models showed a relatively good agreement of the seasonal percolation with the  
301 measurements. Three of the five models overestimated the percolation in all seasons. One model underestimated

1  
2  
3  
4 302 the percolation volume in all seasons and only one model was able to simulate the seasonal percolation  
5  
6 303 accurately. The range of model results was independent of the seasonal percolation.  
7  
8 304 Seasonal flow averaged nitrate concentrations were underestimated by all models in two of the four seasons. For  
9  
10 305 the first season, all models underestimated the concentration by 10 – 40 mg L<sup>-1</sup>. The variation of simulated  
11  
12 306 concentrations and N-fluxes was large. Maximum deviations of seasonal nitrate-N leaching of about 25 kg ha<sup>-1</sup>  
13  
14 307 were found. All models underestimated the leaching rate in 2005 by 8 – 22 kg ha<sup>-1</sup>. The same holds for the fourth  
15  
16 308 season, but only one model was able to calculate the nitrate-N flux with a reasonable agreement with the  
17  
18 309 measurements. In the second season (maize), four models underestimated and one model overestimated the  
19  
20 310 nitrate concentration and nitrate-N flux. The third season, which was the second season with maize showed a  
21  
22 311 rather different pattern. The measured nitrate concentration and nitrate-N flux under maize in the 3<sup>rd</sup> season was  
23  
24 312 much lower than for the maize crop in the 2<sup>nd</sup> season, but the modelled results still showed a large variation with  
25  
26 313 a less skewed distribution of underestimation and overestimation. In the blind test information was lacking about  
27  
28 314 crop-uptake rates and the nitrogen excess per season. The results showed that without this information and  
29  
30 315 without a proper calibration the models were not able to predict nitrate concentrations and leaching rates  
31  
32 316 accurately.

## 317 **3.2 Calibration and validation**

### 318 **3.2.1 Soil water and soil physical relations**

319 In the blind test the modellers had only laboratory measurements of the water retention curve at their disposal,  
40  
41 320 but in the calibration phase also in situ measured soil moisture contents ( $\theta$ ) and pressure heads ( $h$ ) were available  
42  
43 321 at four depths. The laboratory measurements were performed for drying samples only, while under field  
44  
45 322 conditions data pairs of  $\theta(h)$  were detected during wetting and drying cycles so that these were affected by  
46  
47 323 hysteresis (Basile et al., 2003, 2006). Figure 2 depicts the calibrated  $\theta(h)$  curves for three depths. The results at  
48  
49 324 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  
50  
51 325 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  
52  
53 326 m  $h$  ranged from -10 to -100 cm. The variation of the  $\theta(h)$  observed population is largest at depth 0.35 m.

54  
55  
56 327 <<Figure 2>>

57  
58 328 Results for the **EPIC** model are represented by three points as EPIC does not use a continuous description of the  
59  
60 329  $\theta(h)$  curve. The greatest value for the saturated water content was obtained by the **EPIC** model with a value

1  
2  
3  
4 330 greater than  $0.3 \text{ cm}^3 \text{ cm}^{-3}$  at depth 1.8 m. This parameter is far outside the range that was established by the other  
5  
6 331 models. A comparison between the calibrated and observed  $\theta(h)$  curves was made by calculating a  $\theta$  for each  
7  
8 332 value of the measured  $h$ . The performing indices based on computed  $\theta$  and measured  $\theta$  are presented in Table 3.  
9

10 333 <<Table 3 >>  
11  
12

13 334 In general the resulting *MAE*, *RMSE* and *IoA* showed equal trends. The **ARMOSA** model fitted well at depths  
14  
15 335 0.35 m and 0.9 m, but performed worse at depth 1.8 m. The performance of the **COUP** model appeared to be  
16  
17 336 weak. At depth 0.9 m the **DAISY** model was better than the **COUP** model, but worse than the other models. The  
18  
19 337 *IoA* for the **SIM-STO** and **SW-ANIM** models was highest at depth 0.9 m and somewhat lower for the other  
20  
21 338 depths. It should be noted that a good match of the calibrated  $\theta(h)$  curves with measured data pairs does not a-  
22  
23 339 priori mean that a good agreement between the time series of measured and calculated  $\theta$  will be obtained.  
24  
25  
26 340 The simulated  $\theta$  was compared with daily averaged values of measured  $\theta$  (Table 4). For depth 0.35 m an  
27  
28 341 increasing trend was detected from 2008 and onwards which is attributed to the aging of the sensor, and,  
29  
30 342 therefore, the results for this depth were disqualified for the validation period.  
31

32 343 <<Table 4>>  
33  
34

35 344 Except for **ARMOSA** and **EPIC** in the validation phase, the highest *IoA* values for simulation of the water  
36  
37 345 contents were achieved at depth 0.9 m. For **SIM-STO** and **SW-ANIM**, the *IoA* values were similar to the  
38  
39 346 calibration results of the  $\theta(h)$  curves (Table 3). However, the performance by **COUP** increased and that by  
40  
41 347 **DAISY** decreased compared to Table 3. Except for the **ARMOSA** and the **DAISY** models at depth 0.35 m and  
42  
43 348 the **SW-ANIM** model at depth 1.8 m, in general the resulting performance indices showed a better agreement  
44  
45 349 between simulated and observed values for the period 2005 – 2008 than for the comparison based on soil  
46  
47 350 moisture retention curves. The indices of the validation period 2009 – 2011 were in the same range, or somewhat  
48  
49 351 lower at depth 0.9 m, as for the calibration period (Table 4).  
50

51 352 Figure 3 presents the cumulative water fluxes as predicted by the models and as measured as a function of time.  
52

53 353 <<Figure 3>>  
54  
55  
56

57 354 The pattern of cumulative water fluxes per growing season complies generally with the annual precipitation  
58  
59 355 amounts (Table 2) with the exception of maize in 2006 and its preceding crop in the winter of 2005/2006. During  
60  
61 356 the intermediate period after oil pumpkin in 2005 and before maize in 2006, the precipitation amounted to about  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

357 430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low  
358 which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative  
359 water flux per season corresponded well to the measured water fluxes for most of the models which is also true  
360 for the extreme wet year 2009. However, DAISY showed some overestimation in particular seasons due to  
361 difficulties in parameterizing maize crop parameters. EPIC performed better in the calibration than in the  
362 validation period. **SW-ANIM** underestimated the cumulative water flux in the two first seasons, but  
363 overestimated slightly in some other seasons. No model was able to simulate the dry no-flux period during the  
364 second half of 2011. Deviations between the simulated and observed soil moisture contents were relatively  
365 small and have a limited impact on the cumulative water fluxes. Underestimations and overestimations of the  
366 seasonal water fluxes are explained by overestimation and underestimations of the seasonal evapotranspiration.  
367 This depends on the difficulty of establishing accurate crop growth parameters. Table 5 presents the statistical  
368 performance indices for the daily water fluxes and for averaged water fluxes per sampling interval for both the  
369 calibration and the validation periods.

<<Table 5>>

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

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

<<Figure 4>>

378 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  
379 water fluxes the  $\sigma_p/\sigma_o$ -ratio for the validation period was somewhat higher than for the calibration period, but for  
380 the fluxes averaged for the sampling intervals it can be seen that **ARMOSA**, **DAISY**, **COUP** and **EPIC** resulted  
381 in lower  $\sigma_p/\sigma_o$ -ratio's for the validation period than for the calibration period.

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

385 <<Figure 5>>

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

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

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

402 **3.2.2 Soil temperature**

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

411



1  
2  
3  
4 412 **3.2.3 Nitrate concentrations and nitrate-N fluxes**

5 413 Figure 3 presents the cumulative nitrate fluxes and the nitrate concentration of the leachate as predicted by the  
6  
7 414 models and as measured as a function of time. Based on a visual inspection the nitrate concentrations are  
8  
9 415 simulated well by **COUP** and **SW-ANIM** for the calibration period. The **SIM-STO** results for this period were  
10  
11 416 poor and the results of the other models were in between. The results for the validation period showed a  
12  
13 417 completely different picture when compared to the corresponding results for the calibration period. The results  
14  
15 418 of **DAISY** and **SIM-STO** were relatively the best, while **EPIC** and **SW-ANIM** results were weak. **ARMOSA**,  
16  
17 419 **COUP** and **SW-ANIM** overestimated the concentration peak in autumn 2009 and **SW-ANIM** simulated a peak  
18  
19 420 for autumn 2010, while there was no peak visible in the measurements.

20  
21 421 **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** showed more spiky results for the calibration period than the  
22  
23 422 measured values, while **COUP** and **SW-ANIM** showed calmer and more evenly time courses. The results  
24  
25 423 resembled partly the modeller's choice for defining either the nitrate fluxes or the nitrate concentrations in the  
26  
27 424 objective function of the calibration procedure. The **COUP** and **SW-ANIM** modellers used the nitrate  
28  
29 425 concentrations for calibrations, while the **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO** modelling groups used the  
30  
31 426 nitrate fluxes. For **DAISY** and **EPIC**, the nitrate concentrations were calculated afterwards by dividing the  
32  
33 427 nitrate flux by the water flux. The nitrate concentrations in the calibration phase simulated by **SIM-STO** showed  
34  
35 428 a bad performance, while the results for the validation phase were much better. The higher peak concentrations  
36  
37 429 during the calibration phase were not approached by **SIM-STO**. On the other hand, **SW-ANIM** showed a good  
38  
39 430 agreement of nitrate concentrations during the calibration phase, while there is a mismatch during the validation  
40  
41 431 phase. The concentration peaks during the validation phase were severely overestimated by **SW-ANIM** due to  
42  
43 432 an overestimation of the biological fixation rates of some non-leguminous catch crops in this period.

44  
45 433 The nitrate-N flux at depth 1.8 m represents the nitrogen transport to deeper soil layers and is relevant for  
46  
47 434 predictions of nitrate concentrations in deeper groundwater. **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO**  
48  
49 435 underestimated the nitrate N-flux under winter barley preceded by a catch crop in 2007-2008, but **SW-ANIM**  
50  
51 436 overestimated the nitrate N-flux during this period. The **COUP** model was able to calculate the nitrate-N flux in  
52  
53 437 five of the seven seasons that cover the calibration and validation period. **ARMOSA** and **DAISY** calculated the  
54  
55 438 total seasonal nitrate-N flux well in three of the seven seasons, while **EPIC**, **SIM-STO** and **SW-ANIM**  
56  
57 439 calculated this flux well in two of the seven seasons. The last season appeared to be the most difficult one,  
58  
59 440 because of the exceptional dry conditions. The leaching after the 2009 oil pumpkin crop also showed significant

1  
2  
3  
4 441 deviations between model predictions and measurements. The largest deviations of seasonal nitrate-N fluxes  
5  
6 442 occurred in the results of **COUP** and **SW-ANIM** for the exceptional wet year 2009.

7  
8 443 Table 6 presents the statistical indicators for both the nitrate concentrations and the nitrate-N leaching rates,  
9  
10 444 based on the sampling time series. The largest deviations between predicted and simulated nitrate concentrations  
11  
12 445 were found for the **SIM-STO** results in the calibration period for which the *I<sub>oA</sub>* amounted to 0.43. Remarkably  
13  
14 446 the smallest deviations were found for the same model for the validation period for which *I<sub>oA</sub>* amounted to 0.78.  
15  
16 447 The underestimation of the nitrate-N flux by **SIM-STO** is most likely due to immobilization processes that are  
17  
18 448 overemphasized for the 2005 and 2008 periods. Thus, less nitrate was released to the soil water phase which led  
19  
20 449 to the underestimation of the nitrate concentration in the leachate. .

21  
22 450 <<Table 6>>

23  
24  
25 451 The **COUP** model showed the best performance for the nitrate concentrations of the calibration period with *I<sub>oA</sub>*  
26  
27 452 = 0.97 directly followed by the **SW-ANIM** model. The results from **EPIC** and **SW-ANIM** for concentrations in  
28  
29 453 the validation period were weak with *RMSE* > 20 mg L<sup>-1</sup>. The statistical indices of the nitrate-N leaching rates  
30  
31 454 showed a similar picture. The **SIM-STO** model performed relatively weak during the calibration phase. For the  
32  
33 455 leaching rates in this period **DAISY** and **SW-ANIM** had the best performance and for the validation period  
34  
35 456 **ARMOSA** and **DAISY** performed relatively the best. The *NSE* values (data not shown) for both the  
36  
37 457 concentration and the leaching rates in the validation period were almost all negative, showing that the calibrated  
38  
39 458 models had great difficulties to predict concentrations and leaching rates for the more extreme conditions of the  
40  
41 459 validation period.

42  
43 460 Statistical performance of predicted nitrate concentrations and leaching rates were expressed in Taylor diagrams  
44  
45 461 in Figure 6. Calibrated nitrate concentrations by **COUP** and **SW-ANIM** had *R*-values greater than 0.9 and were  
46  
47 462 closest to the (1,0) point. Except for **SIM-STO**, the models showed  $\sigma_p/\sigma_o$  ratios for the calibration step that did  
48  
49 463 not deviate much from 1; for **SIM-STO** the  $\sigma_p/\sigma_o$  ratio was much lower than 1 and *R* < 0.

50  
51  
52 464 <<Figure 6>>

53  
54 465 The plots clearly show the much weaker performance for the validation period than for the calibration period,  
55  
56 466 expressed by lower *R*-values and higher  $\sigma_p/\sigma_o$  ratio's. **SIM-STO** showed the best performance for  
57  
58 467 concentrations in the validation period with *R* > 0.7,  $\sigma_p/\sigma_o$  close to one, and *RRSE* = 0.75, while for the other  
59  
60 468 model *RRSE* > 1. For the nitrate fluxes in the calibration period *RRSE* values were between 0.64 and 0.86, while

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

469 for the validation period, the values were between 1 and 2 even with a peak of 8.6 for **SW-ANIM** (data point not  
470 seen in Figure 6). The *R*-values of the nitrate fluxes in the validation period were in the range 0.18 (**EPIC**) to  
471 0.50 (**COUP**). The  $\sigma_P/\sigma_O$  ratio ratios were in the range 0.75 to 2.3 with a peak of 8.8 for **SW-ANIM** (data point  
472 not seen in Figure 6). The values for  $\sigma_P/\sigma_O$  ratio greater than 1 for both the concentrations and the nitrate fluxes  
473 indicate that the variation of the simulated values is greater than the variation of the observed values.

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

480 <<Table 7>>

481 Calibrated concentrations yielded *IoA*-values ranging from 0.44 (**SIM-STO**) to 0.84 (**SW-ANIM**). The results  
482 for the validation period resulted in somewhat lower *IoA* values, except for **SIM-STO** which shows better results  
483 for the validation than for the calibration period. The **ARMOSA** results were the most constant for the different  
484 depths and periods. Both **COUP** and **SW-ANIM** show significantly poorer *IoA* values for the validation than for  
485 the calibration period. **DAISY** and **SIM-STO** showed slightly better results for the concentrations at depth 0.9 m  
486 than for the concentrations at depth 0.35 m. The other models performed slightly better for depth 0.35 m. Except  
487 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  
488 m.

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

1  
2  
3  
4 496 Three out of four models that used nitrate flux in their objective function resulted in *IoA* values in the range 0.76-  
5  
6 497 0.87 for the calibrated nitrate fluxes, while the other model resulted in *IoA* = 0.43 (Table 6). Two out of three  
7  
8 498 models that used nitrate concentration in their objective function resulted in *IoA* values in the range 0.95-0.97,  
9  
10 499 while the third model resulted in *IoA* = 0.87 (Table 6). However, a good calibration on nitrate concentrations did  
11  
12 500 not result in good performance on nitrate fluxes. Both for the calibration and for the validation periods it  
13  
14 501 appeared that all models had difficulties in predicting the nitrate fluxes at the bottom of the lysimeter, even if  
15  
16 502 some of them were calibrated based on the measured nitrate fluxes.

17  
18 503 Vereecken et al. (1991) evaluated five complex models from which **SW-ANIM**, **EPIC** and **DAISY** are also  
19  
20 504 included in our performance assessment. A comparison between simulated and observed nitrate leaching rates  
21  
22 505 measured in two sandy soils in Denmark and one sandy soil in the Netherlands revealed that **SW-ANIM**, **EPIC**  
23  
24 506 and **DAISY** performed similar, although **DAISY** appeared to be a bit superior in behaviour. In general much  
25  
26 507 better statistical metric values were reported than in our study. This may be due to the circumstances of the field  
27  
28 508 trials which were representative for conventional agriculture during the eighties and because the calibration and  
29  
30 509 the comparison was carried out for seasonal values.

31  
32 510 Diekkrüger et al, (1995) compared the results produced by 19 simulation models, others than those used in this  
33  
34 511 study, for a loam soil and a sand soil in Southern and Eastern Saxony in Germany. Variation in the leaching  
35  
36 512 rates at 0.9 m depth reflected mainly the differences in soil water fluxes at that depth. Apart from the seasonal  
37  
38 513 differences between the models that were able to simulate a three year period continuously, the cumulative  
39  
40 514 leaching was nearly the same for these models. The results of soil nitrogen simulations were significantly  
41  
42 515 influenced by the results of water flow and plant growth simulations. Diekkrüger et al, (1995) concluded that for  
43  
44 516 long term forecasts the exact determination of the boundary conditions is as important as the model approach  
45  
46 517 itself. Our finding that the unmeasured inputs concerning biological N-fixation are important for the soil nitrogen  
47  
48 518 dynamics is consistent with this conclusion. In our study, differences between model seasonal and long term  
49  
50 519 results are attributed to some extent to different assumptions about fixation rates.

51  
52 520 Kersebaum et al., (2007) conducted a comparison of simulation models for 18 different models from which **SW-**  
53  
54 521 **ANIM** and **SIM-STO** are included in our study. **SW-ANIM** was applied to the Müncheberg data-set (Kroes and  
55  
56 522 Roelsma, 2007) and **SIM-STO** was applied to the data-set of the lysimeter station Berlin-Dahlem for water flow  
57  
58 523 simulation and to the Bad Lauchstädt data-set for simulation of soil nitrogen dynamics (Stenitzer et al., 2007).  
59  
60 524 Results for the mean bias, *RMSE*, *IoA* and *NSE* showed weak performances for the soil mineral nitrogen

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

525 simulation in the 0-90 cm upper soil layer for nearly all models which were subjected to the Müncheberg data-  
526 set. Kersebaum et al. (2007) concluded that comparison of simulated results by models which are intended for  
527 field scale and regional scale with measured data often shows unsatisfactory results due to deviating conditions  
528 and parameters. It does not automatically mean that the models or the parameters are wrong because the data and  
529 parameters are only partly related to the site specific conditions of the measurements. In our study significant  
530 amount of data was available, but critical information about rooting depth and pattern, atmospheric deposition  
531 rates, mineralization and fixation rates was missing as well as the nitrogen uptake rates and residue amounts of  
532 the catch crops. Due to these uncertainties, it is difficult to draw clear conclusions about the predictive power of  
533 the models.

534 **3.2.4 Nitrogen balances**

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

536 <<Table 8>>

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

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

547 The most stressing differences are for biological N-fixation. Some models do not describe the biological N-  
548 fixation process as such but modellers had possibilities to assume fixation rates by introducing a nitrogen rich  
549 organic material which was amended continuously during the growing season. The **DAISY** and the **EPIC**  
550 modelling groups did not take account for N-fixation, either due to a lack model formulations implemented or to  
551 a lack of knowledge about this process. **SIM-STO** assumed only for the first season some biological N-fixation  
552 by the crop mixture that included white clover. The **COUP** and the **SW-ANIM** modelling groups took account

1  
2  
3  
4 553 for N-fixation, including for periods for which one wouldn't expect (English ryegrass). In **SW-ANIM** the  
5  
6 554 biological N-fixation is lumped with the mineralization of some of the crop residues that descended from the  
7  
8 555 most recent and previous catch crops. The model output does not allow to unravel the biological N-fixation as  
9  
10 556 such and mineralisation of earlier catch crop residues.

11  
12 557 The **COUP** model did not take account for ammonia volatilization. The other models did, and showed a range of  
13  
14 558 2% to 35% of the nitrogen in the animal manure amended to the soil. The highest volatilization rates were  
15  
16 559 simulated by **SIM-STO**: 27% and 35% of the animal manure N in 2008 and 2011, respectively. This could  
17  
18 560 possibly explain the underestimation of nitrate leaching in 2008, but not in 2011. For these years, the differences  
19  
20 561 of the model predictions amounted to more than 22 and 37 kg ha<sup>-1</sup> a<sup>-1</sup>, respectively, which is higher or in the  
21  
22 562 same range as the measured nitrate-N leaching. Volatilization was calculated by **EPIC** and **ARMOSA** (about 4  
23  
24 563 kg ha<sup>-1</sup>) for the first growing season of the validation period, while no farm fertilizer was applied.

25  
26 564 The models encountered difficulties with the simulation of nitrogen crop off-take. Deviations of simulated  
27  
28 565 uptake rates from the observed values of more than 50 kg ha<sup>-1</sup> occurred for three years by **ARMOSA** (2006,  
29  
30 566 2008, 2009), **EPIC** (2005, 2009, 2010) and **SIM-STO** (2006, 2008, 2010), for two years by **DAISY** (2007,  
31  
32 567 2010), and for one year (2011) by **COUP** and **SW-ANIM**. The **EPIC** model was not able to simulate nitrogen  
33  
34 568 crop off-take by oil pumpkin, because this crop is unknown in the standard database of crop parameters that  
35  
36 569 comes with the model. The **DAISY** model failed to simulate a reasonable crop off-take by maize in 2007, while  
37  
38 570 the N off-take in the preceding year was overestimated by 60 kg ha<sup>-1</sup>. The calibrated parameters for crop uptake  
39  
40 571 were not optimal for the maize as is also apparent from the calculated crop off-take in 2010 where the  
41  
42 572 overestimation amounted nearly 100 kg ha<sup>-1</sup>. Despite the fact that **SW-ANIM** included the N-yield in the object  
43  
44 573 function of the calibration procedure, the modelled crop off-take differed from the measured crop off-take by -14  
45  
46 574 to +19 kg ha<sup>-1</sup>. The **SW-ANIM** underestimated crop off-take in the validation period. Crop off-take is governing  
47  
48 575 the soil nitrogen balance to a large extent and an erroneous calculation of the N off-take means that a possible  
49  
50 576 correct nitrate leaching should be considered as little robust.

51  
52 577 Denitrification is only of significance for the **DAISY** and **EPIC** results, while other models simulated zero or  
53  
54 578 negligible denitrification rates. For most of the models, these estimates were biased by the opinion of the data  
55  
56 579 holders who made plausible from their analysis of soil nitrogen balances that denitrification is not a significant  
57  
58 580 factor (Leis, 2009). The degree of saturation (*S*) at depth 0.35 m exceeds 80% for most of the time and only  
59  
60 581 **COUP** and **SIM-STO** have default threshold values for *S* higher than 80% while other models use lower default

1  
2  
3  
4 582 threshold values for *S* (Heinen, 2006). Except for **DAISY** and **EPIC**, also **ARMOSA** and **SW-ANIM** should  
5  
6 583 have calculated some denitrification when using default values. Except for the first year, the denitrification  
7  
8 584 calculated by **EPIC** exceeded the nitrate-N leaching.

9  
10 585 The change of the total N amount in soil included both organic and mineral forms and was calculated as the  
11  
12 586 residual from the balance. A positive sign means an increase of the total amount whereas a negative sign  
13  
14 587 indicates a depletion of the stock. The model results showed large differences and the largest difference occurred  
15  
16 588 in 2010 where **DAISY** calculated a depletion of 105 kg ha<sup>-1</sup> while **SW-ANIM** calculated an increase of 103 kg  
17  
18 589 ha<sup>-1</sup>. The increase of the amount resulted from the assumed biological fixation and the inputs caused by the  
19  
20 590 cultivation of catch crops. When no additional inputs by fixation or by catch crops was assumed, a depletion will  
21  
22 591 occur (**DAISY** and **EPIC**).

23  
24 592 Except for **SIM-STO** in 2005 and 2008, differences between calculated seasonal nitrate-N leaching rates were  
25  
26 593 relatively small for the calibration phase. The deviations were much larger for the validation phase, where **SW-**  
27  
28 594 **ANIM** overestimated the leaching by 39 and 29 kg ha<sup>-1</sup> in 2009 and 2010, respectively. The observed small  
29  
30 595 leaching rate in 2010 was not approached by any model. Transport of ammonium, organic dissolved N or by  
31  
32 596 surface runoff was calculated at a maximum of 8 kg ha<sup>-1</sup> by the **COUP** model for the first year of the validation  
33  
34 597 period.

35  
36  
37 598 The long term nitrogen balances were summarized at the bottom of Table 8 to further compare the difference of  
38  
39 599 the modellers perceptions of the plant and soil nitrogen cycle.

40  
41 600 The seven year balance depicted the major differences between the models clearly. Despite the crop failure in  
42  
43 601 2007 simulated by **DAISY**, this model showed the highest summed seven year amount, while the summated crop  
44  
45 602 off-take by **SIM-STO** lagged behind with 200 kg ha<sup>-1</sup> relative to the recorded amount. For the individual years  
46  
47 603 the **ARMOSA** results differed considerably from the observations, but the summated seven year crop off-take  
48  
49 604 resembled the measured value rather good.

50  
51 605 Most models have been designed for the field scale for which an average N-yield is calculated. The spatial scale  
52  
53 606 of the lysimeter (1 m<sup>2</sup>) differs from the field scale and the variation of crop off-take rates at this scale is much  
54  
55 607 larger than for the field scale. This is illustrated by the oil pumpkin crop in 2005. Only two seeds were planted in  
56  
57 608 the lysimeter. One of the plants died at the start of the generative phase and no harvest was obtained from this  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 609 plant. This event influenced the yield at the lysimeter scale pretty much, but the yield at the field scale was  
5  
6 610 barely influenced and it can be expected that field scale models encountered difficulties.  
7  
8 611 The total nitrogen loss by denitrification ranged from 0 to 249 kg ha<sup>-1</sup> and was subject to the modellers'  
9  
10 612 perception of the possibility of denitrification in the soil at the Wagna experimental field station.  
11  
12  
13 613 The low input farming system was capable to produce relatively high yields for maize and grains, and for oil  
14  
15 614 pumpkin a N-yield of 51 to 57 kg ha<sup>-1</sup> was recorded, but the observed nitrate-N leaching exceeded the N-excess,  
16  
17 615 the latter defined as the total addition of mineral fertilizers and animal manure minus the crop off-take.  
18  
19 616 **ARMOSA, DAISY** and **EPIC** predicted higher nitrate N-leaching than the N-excess (Fig. 7), while the other  
20  
21 617 models showed a more or less equal value (**SW-ANIM**) or a lower value (**COUP, SIM-STO**). One of the main  
22  
23 618 difficulties was to describe the role of the intermediate catch crops in the crop rotation on the delivery of N.  
24  
25 619 Some of the intermediate crops fixate atmospheric N which leads to an input to the soil and other crops are only  
26  
27 620 able to preserve some of the N excess which remains in soil after the catch crops for the next growing season. No  
28  
29 621 data on the N uptake rates and the quality of the resulting green biomass of these intermediate crops were  
30  
31 622 available. Each of the modellers had to make assumptions for the effect of these crops on the soil N cycle. The  
32  
33 623 estimates of the seven years summed additional input to the soil by biological N-fixation varied from 0-2 kg ha<sup>-1</sup>  
34  
35 624 (**DAISY, EPIC**) to 371 kg ha<sup>-1</sup> (**SW-ANIM**) (Table 8).

36  
37 625 <<Figure 7>>

38  
39 626 None of the models simulated long term soil N-stock at equilibrium. The models that did not take biological N-  
40  
41 627 fixation into consideration showed a decrease of the soil N-stock of -342 kg ha<sup>-1</sup> (**EPIC**) and -177 kg ha<sup>-1</sup>  
42  
43 628 (**DAISY**). The other models that take account for this input showed an increase ranging from 165 to 419 kg ha<sup>-1</sup>.  
44  
45  
46 629 The comparison of the N mass balance components showed large differences between the models. Despite  
47  
48 630 calibration on nitrate leaching, the nitrate leaching predicted was still different from that measured. Crop off-  
49  
50 631 take, although measured, was only used by two models in the calibration procedure, but even then the predicted  
51  
52 632 off-take differed from the observed one. For the other N processes (deposition, biological fixation, volatilization,  
53  
54 633 other transport processes and denitrification) no measured data were available for comparison and calibration.  
55  
56 634 For these aspects, significant differences between the models were observed, either through differences in  
57  
58 635 process descriptions or in handling input by the modelling groups. The resulting storage change thus was also  
59  
60 636 different for the models. The variation of the mass balance components for each model over the years was large.



1  
2  
3  
4 637 A favourable assessment of a good correspondence between a predicted and a measured quantity is difficult,  
5  
6 638 because it may be good for the wrong reasons. For example, **ARMOSA** predicted rather well the overall crop N  
7  
8 639 off-take but was not able to predict the N off-takes of the individual growing seasons.

### 640 **3.2.5 Performance assessment**

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

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

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

1  
2  
3  
4 666 (2005) stated that even for  $IoA > 0.65$  models can result in poor performance, they sure will for  $IoA < 0.6$ , which  
5  
6 667 was here chosen as the lowest boundary. The  $IoA$  scoring for the calibration and validation periods are listed in  
7  
8 668 Table 9.

9  
10 669 <<Table 9>>

11  
12  
13 670 The scoring differed for the different models. Two models (**SIM-STO**, **SW-ANIM**) performed well for the  
14  
15 671 calibration of the  $\theta(h)$  curves and the simulated  $\theta$  at different depths, however, this doesn't guarantee good  
16  
17 672 performance for the other state and rate variables in the calibration and validation periods. For the validation  
18  
19 673 period all models performed weak to moderate on the water volume and weak on the nitrate N-flux per sampling  
20  
21 674 interval, moderate to good on the daily water flux and weak to moderate on the nitrate concentration in the water  
22  
23 675 samples. The models **ARMOSA**, **COUP**, **DAISY** and **EPIC** had more weak qualifications than good  
24  
25 676 qualifications, while **SIM-STO** and **SW-ANIM** had more good qualifications.

26  
27 677 We have also assessed the accuracy of the seasonal amounts on the basis of the mean absolute error ( $MAE$ ). The  
28  
29 678 seven seasons included the oil pumpkin crop twice, which was an unknown or a particular crop for most of the  
30  
31 679 modelling groups. The seven year series contained an extremely wet year (2009) and a dry summer (2011). For  
32  
33 680 the performance assessment for average crop and rainfall conditions  $MAE$  of the five best values ( $MAE_5$ ) out of  
34  
35 681 seven ( $MAE_7$ ) are presented in Table 10 to examine if the models perform better for average conditions. In some  
36  
37 682 cases the improvement was more than 50%, and the ranking of the models slightly changed. Despite the fact that  
38  
39 683  $MAE$  is less sensitive to outliers than e.g.  $IoA$ , extreme situations (unknown crop, wet or dry years) can have a  
40  
41 684 large impact on  $MAE$ .

42  
43 685 <<Table 10>>

### 44 45 46 686 **3.2.6 Methodological aspects for explanation of differences**

#### 47 687 **Data**

48  
49  
50  
51 688 Experimental data collected from a well-controlled lysimeter were used for the purposes of our study. However,  
52  
53 689 the number of measured state and rate variables were less than those present in the six models. For example, no  
54  
55 690 data were available on field-scale hydraulic conductivity, deposition and biological fixation. This means that the  
56  
57 691 outcome of the models is uncertain as not all components of the internal mass balance could be optimized. We  
58  
59 692 have observed in the blind test that based on a limited availability of data, which resembles situations that would  
60  
61 693 occur in practice, the predictions of the models was poor compared to actual observations. That would imply that

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

694 usage of such simulation models for predictions on nitrate leaching at unknown, regional scales must be regarded  
695 with care. In this study the rainfall excess was positive in most times of the year, such that the imposed bottom  
696 boundary condition in the lysimeter resulted in leaching. For other situations where capillary rise may occur, the  
697 models have not been inter-compared. Finally, it is noticed that the soil hydraulic properties as determined in the  
698 laboratory on small soil samples does not guarantee well-predicted soil water contents and soil water fluxes even  
699 for a well-controlled lysimeter situation. Partly, this may be due to the lack of knowledge of hysteresis or its  
700 description in the models.

701 **Procedure**

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

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

714 **Decreased performance when averaging**

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

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

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

738 <<Figure 8>>

739 The exclusion of a particular data pair can result in both an improvement (negative values) or a deterioration  
740 (positive values) of the  $\Delta IoA$ . Furthermore, it is notable that the  $\Delta IoA$  of daily fluxes responded differently  
741 compared to the  $\Delta IoA$  for averaged fluxes per sampling interval. For almost all models the exclusion of the value  
742 simulated for 19 Sept 2006 would affect the  $\Delta IoA$ . The effect of excluding the value of this period is much  
743 smaller for the  $\Delta IoA$  based on the averaged values per sampling interval. The maximum effect in the series of  
744 daily values occurs for a certain day of the calibration period and the maximum effect in the series of averaged  
745 values per sampling interval is calculated for a time interval in Sept. 2010 which belongs to the validation phase  
746 The maximal effect of leaving one value out is greater for the validation period than for the calibration period.  
747 Based on this analysis, it is plausible that the averaging of water fluxes has a different effect on the performance  
748 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 ( $123 \text{ kg ha}^{-1}$ ) 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.

- 1  
2  
3  
4 777 e. Traditional Richard's / Darcy Buckingham equation based models that make use of the Mualem-van  
5  
6 778 Genuchten descriptions and disregard phenomena as hysteresis, preferential flow and multiple  
7  
8 779 phase flow encounter difficulties with an accurate and consistent simulation of both water contents  
9  
10 780 and water fluxes for the soil and conditions of the Wagna lysimeter.
- 11 781 f. Some models which performed relatively well in the calibration phase of the study failed to  
12  
13 782 simulate the nitrate concentrations and fluxes in the validation phase (**SW-ANIM**), while other  
14  
15 783 models behaved relatively bad in the calibration phase and showed better results in the validation  
16  
17 784 phase (**SIM-STO**). An accurate calibration does not guarantee a good predictive power of the  
18  
19 785 model.
- 20  
21 786 g. The catch crop mixtures and the non-harvested English ryegrass play an important role in the  
22  
23 787 nutrient dynamics of the soil. This role is addressed weakly by the simulation models: (1) due to a  
24  
25 788 lack of experimental data on nitrogen uptake rates and mineralization of residues of these  
26  
27 789 intermediate crops, and (2) lack of knowledge to describe the relevant processes related to the  
28  
29 790 foreign crops
- 30  
31 791 h. Assessment of future climate and land use changes requires a good predictive power of the models  
32  
33 792 and a certain level of robustness. Although the robustness is not clear for the tested models, the  
34  
35 793 process oriented dynamic models used in this study are useful for hypothesis testing.

## 38 794 **5. Acknowledgements**

39  
40  
41 795 This research was made possible by the GENESIS project of the EU 7<sup>th</sup> Framework Programme (Project No.  
42  
43 796 226536; FP7-ENV-2008-1). We are grateful for the experimental data provided by Joanneum Raum (Graz,  
44  
45 797 Austria). The modelling team of Democritus University of Thrace would like to thank Per-Erik Jansson (Royal  
46  
47 798 Institute of Technology, Stockholm, Sweden;) for his valuable help during the application of CoupModel.

## 50 51 799 **6. References**

- 52  
53 800 Akkal-Corfini N, Morvan T, Menasseri-Aubry S, Bissuel-Bélaygue C, Poulain D, Orsini F, Leterme P. Nitrogen  
54 801 mineralization, plant uptake and nitrate leaching following the incorporation of (15N)-labeled cauliflower  
55 802 crop residues (*Brassica oleracea*) into the soil: a 3-year lysimeter study. *Plant Soil* 2010;. 328:17–26. DOI  
56 803 10.1007/s11104-009-0104-0
- 57  
58 804 Ale S, Bowling LC, Youssef MA, Brouder SM. Evaluation of simulated strategies for reducing nitrate–nitrogen  
59 805 losses through subsurface drainage systems. *J Environ Qual* 2012; 41:217-228.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

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

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

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

817 Buckingham, E. Studies on the movement of soil moisture. Bull. 38, USDA, Bureau of Soils, Washington, DC,  
818 1907.

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

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

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

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

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

830 De Willigen P, Neeteson JJ, Comparison of six simulation models for the nitrogen cycle in the soil. *Fert Res*  
831 1985;8:157-171.

832 De Willigen P. Nitrogen turnover in the soil-crop system; comparison of fourteen simulation models. *Fert Res*  
833 1991;27: 141-149.

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

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

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

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

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

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

1  
2  
3  
4 850 Ferrara RM, Trevisiol P, Acutis M, Rana G, Richter GM, Baggaley N. Topographic impacts on wheat yields  
5 851 under climate change: two contrasted case studies in Europe. *Theor. Appl. Climatol.* 2011;99:53–65.  
6  
7 852 Gribb MM, Hansen FI, Aleshia Chandler, McNamara DG, James P. The effect of various soil hydraulic property  
8 853 estimates on soil moisture simulations. *Vadose Zone J* 2009; 8:321–331.  
9  
10 854 Grizzetti B, Bouraoui F, Billen G, Van Grinsven H., Cardoso AC, Thieu V, Garnier J, Curtis C, Howarth R,  
11 855 Johnes P. Nitrogen as a threat to European water quality. In: *The European Nitrogen Assessment 17*,  
12 856 Cambridge, UK: Cambridge University Press, 2011, p. 379-404.  
13  
14 857 Groenendijk, P., L.V. Renaud, and J. Roelsma. 2005. Prediction of Nitrogen and Phosphorus leaching to  
15 858 groundwater and surface waters. Process descriptions of the animo4.0 model. Alterra–Report 983, Alterra,  
16 859 Wageningen. <http://content.alterra.wur.nl/Webdocs/PDFFiles/Alterrapporten/AlterraRapport983.pdf>  
17  
18 860 Hansen S, Jensen HE, Nielsen NE, Svendsen H. **DAISY**: Soil Plant Atmosphere System Model. NPO Report  
19 861 No. A 10. The National Agency for Environmental Protection, Copenhagen, 1990, 272 pp.  
20  
21 862 Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of nitrogen dynamics and biomass production in  
22 863 winter wheat using the Danish simulation model DAISY, *Fert Res* 1991a;27: 245-259.  
23  
24 864 Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of biomass production, nitrogen uptake and nitrogen  
25 865 leaching by using the Daisy model. In: *Soil and Groundwater Research Report II: Nitrate in Soils, Final*  
26 866 *Report on Contracts EV4V-0098-NL and EV4V-00107-C. DG XII. Commission of the European*  
27 867 *Communities, 1991b; 300–309.*  
28  
29 868 Hansen S. DAISY, a flexible Soil-Plant-Atmosphere system Model. Report. Dept. Agric, Danish Informatics  
30 869 Network in the Agricultural Sciences, 2002. <http://www.dina.kvl.dk/~DAISY/ftp/DAISYDescription.pdf>.  
31  
32 870 Heinen, M. Simplified denitrification models: Overview and properties. *Geoderma* 2006;133:444-463.  
33  
34 871 Herbst M, Fialkiewicz W, Chen T, Pütz T, Thiéry D, Mouvet C, Vachaud G, Vereecken H. Intercomparison of  
35 872 flow and transport models applied to vertical drainage in cropped lysimeters. *Vadose Zone J* 2005;4:240-254.  
36  
37 873 Jabro JD, Jabro AD, Fales SL. Model performance and robustness for simulating drainage and nitrate-nitrogen  
38 874 fluxes without recalibration. *Soil Sci. Soc. Am. J.* 2012;76:1957–1964 [doi:10.2136/sssaj2012.0172](https://doi.org/10.2136/sssaj2012.0172)  
39  
40 875 Jachner S, Van den Boogaart KG, Petzoldt T. Statistical Methods for the Qualitative Assessment of Dynamic  
41 876 Models with Time Delay (R package qualV). *J Stat Softw* 2007;22:(8),1–30. <http://www.jstatsoft.org/v22/i08>  
42  
43 877 Janssen PHM, Heuberger PSC. Calibration of process-oriented models. *Ecol Model* 1995;83: (1-2) 55-66.  
44 878 [http://dx.doi.org/10.1016/0304-3800\(95\)00084-9](http://dx.doi.org/10.1016/0304-3800(95)00084-9)  
45  
46 879 Jansson P-E, Karlberg L. Coupled heat and mass transfer model for soil-plant-atmosphere systems Royal  
47 880 Institute of Technology, Dept of Civil and Environmental Engineering, Stockholm, 2004. 435 pp.  
48  
49 881 Jansson P-E. CoupModel: Model use, calibration and validation, *Transactions of the ASABE* 2012;55(4):1-11.  
50  
51 882 Jensen LS, Mueller T, Nielsen NE, Hansen S, Crocker GJ, Grace PR., Klir J, Körschens M, Poulton PR.  
52 883 Simulating trends in soil organic carbon in long-term experiments using the soil-plant-atmosphere model  
53 884 DAISY. *Geoderma* 1997;81:5-28.  
54  
55 885 Kersebaum KC, Hecker, J-M, Mirschel W, Wegehenkel M. Modelling water and nutrient dynamics in soil–crop  
56 886 systems: a comparison of simulation models applied on common data sets. In: Kersebaum KC et al., editors,  
57 887 *Modelling Water and Nutrient Dynamics in Soil–Crop Systems*, Springer, 2007, p. 1–17.  
58  
59 888 Khodaverdilo H, Homaee M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions  
60 889 for some calcareous soils, *J Hydrol* 2011;399:93-99  
61  
62 890 Klammler G, Fank J.. Determining water and nitrogen balances for beneficial management practices using  
63 891 lysimeters at Wagna test site (Austria). *Sci. Tot. Environ* 2014; *this issue*.



- 1  
2  
3  
4 892 Krause P, Boyle, DP, Båse F. Comparison of different efficiency criteria for hydrological model assessment.  
5 893 Advances in Geosciences, 2005;5:89–97. <http://www.adv-geosci.net/5/89/2005/adgeo-5-89-2005.pdf>  
6  
7 894 Kroes J and Roelsma J. Simulation of water and nitrogen flows on field scale: application of the SWAP–ANIMO  
8 895 model for the Müncheberg data set. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient  
9 896 Dynamics in Soil–Crop Systems, 2007, Springer, pp 111–128.  
10  
11 897 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.  
12  
13 898 Legates DR, McCabe GJ. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic  
14 899 model validation. Water Resour Res 1999;35:233-241.  
15  
16 900 Leis A. Chemical, isotopic, and microbiological evidence for nitrification below the plant root zone from  
17 901 intensive fertilized agricultural area in Austria – Insights from lysimeter studies and soil cores. In: IAEA-  
18 902 TECDOC-1618. Application of Isotopes to the Assessment of Pollutant Behaviour in the Unsaturated Zone  
19 903 for Groundwater Protection. Final report of a coordinated research project 2004-2005. International Atomic  
20 904 Energy Agency, Vienna; 2009. p. 15-30.  
21  
22 905 Mantovi P, Fumagalli L, Beretta GP, Guermandi M. Nitrate leaching through the unsaturated zone following pig  
23 906 slurry applications, J Hydrol 2006; 316:195-212.  
24  
25 907 Moreels E., De Neve S, Hoffman G, Van Meirvenne M. Simulating nitrate leaching in bare fallow soils: a model  
26 908 comparison. Nutr Cycl Agroecosys 2003;67:137-144  
27  
28 909 Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL. Model evaluation guidelines for  
29 910 systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 2007;50:  
30 911 885–900.  
31  
32 912 Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour.  
33 913 Res. 1976;12: 513-522.  
34  
35 914 Nash JE, Sutcliffe JV. River flow forecasting through conceptual models, 1, A discussion of principles, J.  
36 915 Hydrol., 1970;10:282-290.  
37  
38 916 Nett L, Feller C, George E, Fink M. Effect of winter catch crops on nitrogen surplus in intensive vegetable crop  
39 917 rotations. Nutr Cycl Agroecosys 2011;91:327-337.  
40  
41 918 Nolan, BT, Puckett LJ, Ma L, Green TG, Bayless ER, Malone RW. Predicting unsaturated zone nitrogen mass  
42 919 balances in agricultural settings of the United States. J Environ Qual 2010;39:1051-1065.  
43  
44 920 Oenema O. Governmental policies and measures regulating nitrogen and phosphorus from animal manure in  
45 921 European agriculture. J Anim Sci 2004;82: E196-E206.  
46  
47 922 Patil N, Rajput G. Evaluation of Water Retention Functions and Computer Program “Rosetta” in Predicting Soil  
48 923 Water Characteristics of Seasonally Impounded Shrink–Swell Soils, J Irrig Drain E-ASCE 2009;135, 286-  
49 924 294. <http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29IR.1943-4774.0000007>  
50  
51 925 Perego A, Giussani A, Sanna M, Fumagalli M, Carozzi M, Alfieri L, Brenna S, Acutis M. The ARMOSA  
52 926 simulation crop model: overall features, calibration and validation results. Italian Journal of  
53 927 Agrometeorology 2013;3:23-38.  
54  
55 928 Qi Z, Ma L, Helmers MJ, Ahuja LR, Malone RW. Simulating nitrate-nitrogen concentration from a subsurface  
56 929 drainage system in response to nitrogen application rates using RZWQM2. J Environ Qual 2012;41:289-295.  
57  
58 930 Reusser DE, Blume T, Schaefli B, Zehe E. Analysing the temporal dynamics of model performance for  
59 931 hydrological models. Hydrol Earth Syst Sc 2009;13: 999 – 1018. [http://www.hydrol-earth-syst-](http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf)  
60 932 [sci.net/13/999/2009/hess-13-999-2009.pdf](http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf)  
61  
62 933 Richards LA, Capillary conduction of liquids through porous mediums. Physics 1931;1: 318-333.  
63  
64  
65

- 1  
2  
3  
4 934 Richter GM, Acutis M, Trevisiol P, Latiri K, Confalonieri R. Sensitivity analysis for a complex crop model  
5 935 applied to Durum wheat in the Mediterranean. *Europ J Agron* 2010;32:127-132.
- 6  
7 936 Ritter A, Hupet F, Muñoz-Carpena R, Lambot S, Vanclooster M. Using inverse methods for estimating soil  
8 937 hydraulic properties from field data as an alternative to direct methods. *Agr Water Manage* 2003;59:77-96.
- 9  
10 938 Saxton KE, Rawls WJ. Soil water characteristic estimates by texture and organic matter for hydrologic solutions.  
11 939 *Soil Sci. Soc. Am. J.* 2006;70:1569-1578. doi:10.2136/sssaj2005.0117
- 12 940 Smith P, Smith JU, Powlson DS, McGill WB, Arah JRM, Chertov OG, Coleman K, Franko U, Frohling S,  
13 941 Jenkinson LS, Jenseng LS, Kellyh RH, Klein-Gunnewiek H., Komarov AS, Lif C, Molina JAE, Mueller T,  
14 942 Parton WJ, Thornley JHM, Whitmore AP. A comparison of the performance of nine soil organic matter  
15 943 models using datasets from seven long-term experiments. *Geoderma* 1997;81: 153-225.
- 16  
17 944 Sogbedji JM, Van Es HM, Melkonian JJ, Schindelbeck RR. Evaluation of the PNM model for simulating drain  
18 945 flow nitrate-N concentration under manure-fertilized maize. *Plant Soil* 2006;282:343-360.
- 19  
20 946 Sohier C, Degré A, Dautrebande S. From root zone modelling to regional forecasting of nitrate concentration in  
21 947 recharge flows – The case of the Walloon Region (Belgium). *J Hydrol* 2009;369:350-359.
- 22  
23 948 Stenitzer E. SIMWASER – Ein numerisches Modell zur Simulation des Bodenwasserhaushaltes und des  
24 949 Pflanzenertrages eines Standortes. *Mitt. der Bundesanstalt für Kulturtechnik und Bodenwasserhaushalt*, 31:  
25 950 Petzenkirchen, 1988. p.1-118.
- 26  
27 951 Stenitzer E, Diesel H, Franko U, Schwartengraber R, Zenker T. Performance of the model SIMWASER in two  
28 952 contrasting case studies on soil water movement. In: K. Ch. Kersebaum et al. (eds.), *Modelling Water and*  
29 953 *Nutrient Dynamics in Soil–Crop Systems*, 2007, Springer, pp 27–36.
- 30  
31 954 Stumpp C, Nützmann G, Maciejewski S, Maloszewski P. A comparative modeling study of a dual tracer  
32 955 experiment in a large lysimeter under atmospheric conditions. *J Hydrol* 2009;375: 566-577.
- 33  
34 956 Svendsen H, Hansen S, Jensen HE. Simulation of crop production, water and nitrogen balances in two German  
35 957 agro-ecosystems using the DAISY model. *Ecol Model* 1995;81: 197-212.
- 36  
37 958 Taylor KE. Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res* 2001;106:  
38 959 No. D7, P. 7183. doi:10.1029/2000JD900719.
- 39  
40 960 Thorup-Kristensen, K. . Effect of deep and shallow root systems on the dynamics of soil inorganic N during 3-  
41 961 year crop rotations. *Plant Soil*, 2006;288:233-248.
- 42  
43 962 Van Dam JC,. Field-scale water flow and solute transport: SWAP model concepts, parameter estimation and  
44 963 case studies. PhD thesis Wageningen University,, 2000.
- 45  
46 964 Van Dam JC, Groenendijk P, Hendriks RFA. Advances of Modeling Water Flow in Variably Saturated Soils  
47 965 with SWAP. *Vadose Zone J* 2008;7:640-653.
- 48  
49 966 Van der Laan M, Miles, N, Annandale JG, Du Preez CC. Identification of opportunities for improved nitrogen  
50 967 management in sugarcane cropping systems using the newly developed Canegro-N model. *Nutr Cycl*  
51 968 *Agroecosys* 2011;90,391-404.
- 52  
53 969 Van der Velde M, Bouraoui F, Aloe A. Pan-European regional-scale modelling of water and N efficiencies of  
54 970 rapeseed cultivation for biodiesel production. *Glob Change Biol* 2009;15:24-37.
- 55  
56 971 Van Genuchten MTh. A closed form for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc.*  
57 972 *Am. J.* 1980;44:892-898.
- 58  
59 973 Van Liew MW, Garbrecht J. Hydrologic simulation of the little Washita river experimental watershed using  
60 974 SWAT. *J Am Water Resour Ass* 2003;39:413-426.
- 61  
62 975 Vereecken, H., E.J. Jansen, M.J.D. Hack-ten Broeke, M. Swerts, M. Engelke, S. Fabrewitz and S. Hansen, 1991.  
63 976 Comparison of simulation results of five nitrogen models using different datasets. In: Commission of

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

977 European Communities, editor, Soil and Groundwater Research, Report II Nitrate in Soils, Commission of the  
978 European Communities, Luxembourg, pp 321 – 338.

979 Vereecken H, Weynants M, Javaux M, Pachepsky Y, Schaap MG, Van Genuchten MTh. Using Pedotransfer  
980 Functions to Estimate the van Genuchten–Mualem Soil Hydraulic Properties: A Review Vadose Zone J.  
981 2010;9: 795–820.

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

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

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

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

990 Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteor Soc 1982; 63, 1309-  
991 1313.

992 Willmott CJ, Ackleson SG, Davis RE, Feddesma JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. Statistics  
993 for the evaluation and comparison of models. J Geophys Res 1985;90, 8995 – 9005.  
994 [dx.doi.org/10.1029/JC090iC05p08995](https://doi.org/10.1029/JC090iC05p08995)

995 Wolff J, Beusen AHW, Groenendijk P, Kroon T, Rötter R, Van Zeijts H. The integrated modeling system  
996 STONE for calculating nutrient emissions from agriculture in the Netherlands. Environ Model Softw  
997 2003;18:597-617. [doi:10.1016/S1364-8152\(03\)00036-7](https://doi.org/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 <sup>-1</sup> 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		

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 2. Annual precipitation rates (mm a<sup>-1</sup>) 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 <sup>-1</sup> )	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	<i>MAE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>RMSE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>IoA</i>		
	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
<b>ARMOSA</b>	0.0064	0.0166	0.0308	0.0112	0.0176	0.0310	0.89	0.79	0.18
<b>COUP</b>	0.0341	0.0753	0.0391	0.0416	0.0775	0.0395	0.59	0.31	0.18
<b>DAISY</b>	0.0295	0.0340	0.0166	0.0326	0.0374	0.0178	0.63	0.62	0.38
<b>SIM-STO</b>	0.0212	0.0119	0.0064	0.0255	0.0130	0.0078	0.75	0.89	0.67
<b>SW-ANIM</b>	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	<i>MAE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>RMSE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>IoA</i>		
	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 ( $n = 1461$ )									
<b>ARMOSA</b>	0.0119	0.0247	0.0107	0.0168	0.0447	0.0123	0.79	0.75	0.46
<b>COUP</b>	0.0230	0.0104	0.0023	0.0288	0.0363	0.0031	0.74	0.84	0.85
<b>DAISY</b>	0.0956	0.0152	0.0105	0.1083	0.0630	0.0132	0.28	0.65	0.38
<b>EPIC</b>	0.0613	0.1563	0.0909	0.0662	0.0306	0.0925	0.49	0.90	0.07
<b>SIM-STO</b>	0.0180	0.0063	0.0028	0.0249	0.0271	0.0039	0.81	0.92	0.85
<b>SW-ANIM</b>	0.0101	0.0106	0.0072	0.0159	0.0285	0.0082	0.87	0.92	0.59
Validation 2009 – 2011 ( $n = 955$ )									
<b>ARMOSA</b>	x	0.0260	0.0130	x	0.0291	0.0149	x	0.52	0.47
<b>COUP</b>	x	0.0124	0.0030	x	0.0165	0.0041	x	0.74	0.84
<b>DAISY</b>	x	0.0152	0.0137	x	0.0193	0.0165	x	0.69	0.40
<b>EPIC</b>	x	0.1535	0.0924	x	0.1570	0.0939	x	0.19	0.09
<b>SIM-STO</b>	x	0.0093	0.0039	x	0.0134	0.0054	x	0.87	0.82
<b>SW-ANIM</b>	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			
	<i>MAE</i> (mm d <sup>-1</sup> )	<i>RMSE</i> (mm d <sup>-1</sup> )	<i>IoA</i>	<i>NSE</i>	<i>MAE</i> (mm d <sup>-1</sup> )	<i>RMSE</i> (mm d <sup>-1</sup> )	<i>IoA</i>	<i>NSE</i>
Calibration 2005 – 2008								
	<i>n</i> = 1461				<i>n</i> = 199			
<b>ARMOSA</b>	0.45	1.00	0.82	0.41	0.43	0.81	0.84	0.48
<b>COUP</b>	0.45	0.98	0.80	0.44	0.43	0.75	0.85	0.55
<b>DAISY</b>	0.57	1.16	0.68	0.21	0.54	0.90	0.74	0.35
<b>EPIC</b>	0.54	0.99	0.83	0.42	0.46	0.75	0.89	0.55
<b>SIM-STO</b>	0.34	0.87	0.86	0.55	0.30	0.62	0.91	0.69
<b>SW-ANIM</b>	0.38	0.91	0.86	0.51	0.37	0.72	0.88	0.58
Validation 2009 – 2011								
	<i>n</i> = 1084				<i>n</i> = 128			
<b>ARMOSA</b>	0.70	1.75	0.79	0.41	1.66	3.82	0.68	0.39
<b>COUP</b>	0.70	1.57	0.84	0.52	1.41	3.47	0.79	0.50
<b>DAISY</b>	0.73	1.77	0.77	0.39	1.74	4.34	0.56	0.21
<b>EPIC</b>	0.85	1.79	0.77	0.38	1.80	4.00	0.63	0.33
<b>SIM-STO</b>	0.51	1.43	0.90	0.61	1.69	3.94	0.76	0.35
<b>SW-ANIM</b>	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		
	<i>MAE</i> (mg L <sup>-1</sup> )	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i> (kg ha <sup>-1</sup> d <sup>-1</sup> )	<i>RMSE</i>	<i>IoA</i>
Calibration 2005 – 2008 ( <i>n</i> = 199)						
<b>ARMOSA</b>	15.71	20.37	0.78	0.043	0.085	0.77
<b>COUP</b>	6.74	9.60	0.97	0.041	0.085	0.78
<b>DAISY</b>	13.92	16.82	0.87	0.037	0.063	0.87
<b>EPIC</b>	19.55	25.63	0.76	0.049	0.084	0.82
<b>SIM-STO</b>	27.34	34.61	0.43	0.044	0.089	0.60
<b>SW-ANIM</b>	7.88	10.48	0.95	0.035	0.080	0.85
Validation 2009 – 2011 ( <i>n</i> = 128)						
<b>ARMOSA</b>	11.17	15.85	0.52	0.058	0.102	0.61
<b>COUP</b>	12.36	18.68	0.52	0.076	0.187	0.53
<b>DAISY</b>	8.54	11.40	0.78	0.045	0.095	0.54
<b>EPIC</b>	18.24	22.07	0.52	0.089	0.155	0.41
<b>SIM-STO</b>	8.88	10.44	0.78	0.058	0.138	0.56
<b>SW-ANIM</b>	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 ( $\text{mg L}^{-1}$ ) 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; $n = 108$ )		
	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>
<b>ARMOSA</b>	36.8	50.6	0.66	22.7	35.9	0.65	12.7	16.6	0.58
<b>COUP</b>	28.0	35.2	0.80	28.2	44.1	0.38	16.6	24.1	0.37
<b>DAISY</b>	32.2	43.9	0.68	29.1	50.9	0.46	12.9	21.5	0.55
<b>SIM-STO</b>	50.6	66.7	0.44	25.5	36.3	0.68	13.6	15.8	0.71
<b>SW-ANIM</b>	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 were preceded by catch crops (including leguminous crops).

Crop and period	Balance term <sup>†</sup> (kg ha <sup>-1</sup> )	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
	Biological fixation (+)		41.5	1.7	0.1	1.8	31.3	81.3
1.1.2005	Volatilization (-)		2.7	0.0	1.0	1.5	1.9	2.1
–	Crop off-take (-)		59.7	55.3	83.3	0.0	44.3	70.0
13.9.2005	NO <sub>3</sub> -N leaching (-)	50.9	17.2	27.9	25.8	30.3	3.6	15.3
	Other transport <sup>§</sup> (-)	22.2	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 <sup>#</sup>		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 <sub>3</sub> -N leaching (-)	25.7	27.9	25.8	22.7	33.6	25.1	29.7
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-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 <sub>3</sub> -N leaching (-)	5.9	4.4	7.1	6.3	5.4	8.8	5.8
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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
	Biological fixation (+)		0.0	0.1	0.0	0.0	0.0	14.0
22.9.2007	Volatilization (-)		0.2	0.0	2.6	5.4	22.7	5.1
–	Crop off-take (-)	132.3	66.2	104.7	139.0	114.2	81.8	118.4
30.6.2008	NO <sub>3</sub> -N leaching (-)	18.9	13.5	18.5	11.7	12.3	5.7	22.2
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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
	Biological fixation (+)		52.1	41.2	0.0	0.0	0.0	22.7
1.7.2008	Volatilization (-)		4.4	0.0	0.0	3.9	0.0	0.0
–	Crop off-take (-)	56.9	113.6	59.9	97.2	0.0	72.3	45.7
7.9.2009	NO <sub>3</sub> -N leaching (-)	33.1	44.2	61.5	26.4	16.0	32.5	72.1
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-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 <sub>3</sub> -N leaching (-)	3.6	17.0	14.6	8.7	19.3	13.1	32.9

	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-0.2	72.7	-104.9	-39.7	58.3	102.8	
	<hr/>								
	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	Biological fixation (+)		0.0	12.8	0.0	0.0	0.0	18.2
	-								
	13.7.2011	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 <sub>3</sub> -N leaching (-)	13.9	6.1	3.2	7.6	30.3	13.3	31.0
		Other transport <sup>§</sup> (-)		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 <sup>#</sup>		20.7	146.5	11.2	-128.8	-10.4	75.5
	<hr/>								
	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	Biological fixation (+)		174.9	154.6	0.1	1.8	31.3	370.9
	-								
	13.7.2011	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 <sub>3</sub> -N leaching (-)	123.3	130.3	158.6	109.1	147.3	102.2	209.0
		Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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<sub>4</sub>-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	+	
		0.35 m	o	o	-	-	+	+	
		0.9 m	o	+	o	+	+	+	
		1.8 m	-	+	-	-	+	-	
		0.9 m	o	+	o	n.a.	-	+	
		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	
		0.35 m	o	-	-	n.a.	o	-	
		0.9 m	-	-	-	n.a.	o	-	
		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*<sub>7</sub>) and for the best five seasons (*MAE*<sub>5</sub>).

Seasonal quantity	Indicators	ARMOSA	COUP	DAISY	EPIC	SIM-STO	SW-ANIM
Percolated water (mm)	<i>MAE</i> <sub>7</sub>	21.3	24.2	63.9	48.6	14.6	40.3
	<i>MAE</i> <sub>5</sub>	16.0	14.3	30.5	30.5	11.8	32.8
N crop off-take (kg ha <sup>-1</sup> )	<i>MAE</i> <sub>7</sub>	36.5	32.7	47.7	31.0	33.0	21.5
	<i>MAE</i> <sub>5</sub>	23.1	14.3	29.0	20.6	20.5	10.3
Leached NO <sub>3</sub> -N (kg ha <sup>-1</sup> )	<i>MAE</i> <sub>7</sub>	6.6	8.2	4.6	10.3	6.6	14.2
	<i>MAE</i> <sub>5</sub>	4.4	3.6	3.7	7.8	2.8	6.3

1  
2  
3  
4 **Figures captions**  
5  
6

7  
8 Figure 1 Blind test comparison of seasonal water fluxes, flow averaged nitrate concentration and nitrate-N fluxes simulated by five models  
9 (excluding **SIM-STO**) with observations. Results of individual models are indicated by markers.  
10

11 Figure 2 Measured values and calibrated soil moisture retention curves at depths 0.35 m, 0.9 m and 1.8 m.  
12

13 Figure 3 Comparison of simulated and measured inner season cumulative water fluxes, nitrate concentrations and inner season cumulative  
14 nitrate-N fluxes at depth 1.8 m in the low input farming lysimeter at the Wagna experimental field station  
15  
16

17 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  
18 interval averaged values (right). Circles refer to the calibration results and triangles refer to the validation results. A = **ARMOSA**, C =  
19 **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**  
20  
21

22 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  
23 experimental field station  
24  
25

26 Figure 6 Taylor plot of the statistical performance parameters for the simulated nitrate concentrations (left) and nitrate-N fluxes (right) at  
27 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  
28 outside the range (2.5; 8.5). A = **ARMOSA**, C = **COUP**, D = **DAISY**, E = **EPIC**, SS = **SIM-STO**, SA = **SW-ANIM**  
29  
30

31 Figure 7 Seven years balances for fertilization minus crop off-take and nitrate-N leaching (all in kg ha<sup>-1</sup>), summed since the start of the  
32 calibration period  
33  
34

35 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  
36 (see text for further explanation).  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

3 Piet Groenendijk<sup>a,\*</sup>, Marius Heinen<sup>a</sup>, Gernot Klammler<sup>b</sup>, Johann Fank<sup>b</sup>, Hans Kupfersberger<sup>b</sup>, Vassilios  
4 Pisinaras<sup>c</sup>, Alexandra Gemitzi<sup>c</sup>, Salvador Peña-Haro<sup>d</sup>, Alberto García-Prats<sup>c</sup>, Manuel Pulido-Velazquez<sup>f</sup>, Alessia  
5 Perego<sup>g</sup>, Marco Acutis<sup>g</sup>, Marco Trevisan<sup>h</sup>

- 6  
7 a Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands  
8 b Joanneum Research, Forschungsgesellschaft mbH, Leonhardstraße 59, 8010 Graz, Austria  
9 c Democritus University of Thrace, Department of Environmental Engineering, Vas. Sofias 12, Xanthi, 67100,  
10 Greece  
11 d Institute of Environmental Engineering, ETH Zurich, Wolfgang-Pauli-Str. 15, CH-8093 Zurich, Switzerland  
12 e Universitat Politècnica de València, Department of Hydraulic Engineering and Environment, Camino de  
13 Vera, 46022 Valencia, Valencia, Spain  
14 f Universitat Politècnica de València, Research Institute of Water and Environmental Engineering (IIAMA),  
15 Camino de Vera, 46022 Valencia, Valencia, Spain  
16 g University of Milan, Department of Agricultural and Environmental Science, Via G. Celoria 2 20133, Milan,  
17 Italy  
18 h Università Cattolica del Sacro Cuore, sede di Piacenza, Via Emilia Parmense, 84 29100, Piacenza, Italy

Formatted: English (U.K.)

Formatted: English (U.K.)

19  
20 \*Corresponding author: Piet Groenendijk, Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands, Email:  
21 [piet.groenendijk@wur.nl](mailto:piet.groenendijk@wur.nl) Tel.: +31 317 486434

22  
23 **Abstract**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

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

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

#### 45 **Keywords**

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

## 51 **1. Introduction**

52  
53 Agriculture is the major land use in Europe (ca. 50% of overall land area) and has strongly increased its use of  
54 external inputs (fertiliser, pesticides and water) over the last 50 years. The environmental effects of intensive  
55 agriculture include a decline in biodiversity, eutrophication of ecosystems and surface waters, acidification,  
56  
57

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

51 global warming, air pollution and diffuse nitrate pollution of groundwater. A global challenge is to produce  
52 enough food for the ever-growing population and at the same time minimizing the loss of reactive nitrogen (N)  
53 to the environment. Since the 1980s, agriculture in Western Europe has managed to reduce its N surpluses,  
54 owing to stringent national and European community policies (Vitousek et al., 2009; Grizzetti et al., 2011).

55 The main aim of the Nitrates Directive (EU, 1991: Directive 91/676/EEC) is to reduce water pollution caused or  
56 induced by nitrates and phosphorus from agricultural sources. The Nitrates Directive legally restricts farm  
57 application of manure to 170 kg ha<sup>-1</sup> of nitrogen, or in case of derogation to inputs up to 250 kg ha<sup>-1</sup> (Oenema,  
58 2004). An implementation measure of the Nitrates Directive is the establishment of codes of Good Agricultural  
59 Practice. Recommended measures include, among others, the application of crop rotations, the cultivation of a  
60 soil winter cover and catch crops to prevent nitrate leaching and run-off during wet seasons. Catch crops create a  
61 new challenge in the assessment of environmental effects of crop rotations. In theory, catch crops take up N that  
62 would otherwise be lost, and, after incorporation of the crop residues into the soil, make this N available to the  
63 succeeding crop via mineralization. However, the influence of a catch crop on the nitrogen supply to the  
64 succeeding crop can vary greatly and range from a positive to a negative effect (Nett et al., 2011). The effect is  
65 determined by the N uptake capacity, the rooting depth of a catch crop, the weather and soil conditions as well as  
66 the rooting depth of the succeeding crop (Thorup-Kristensen, 2006).

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

75 Calibration and validation of models contributes to their reliability. In addition also an analysis of the  
76 implemented process descriptions and the mutual comparison of models provides information on the predictive  
77 power. Several model comparison studies have been conducted in which nitrate leaching models were compared  
78 (De Willigen and Neeteson, 1985; Vereecken et al., 1991; De Willigen, 1991; Diekkrüger et al., 1995; Moreels  
79 et al., 2003; Kersebaum et al., 2007; Jabro et al., 2012). [Most of them were related to ordinary agricultural](#)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

80 | [conditions with a single crop on a typical agricultural soil. Thus, there is no information \(comparison\) available](#)  
81 | [for situations in soils that are highly vulnerable to nitrate leaching in combination with low-input conditions and](#)  
82 | [the use of catch crops.](#)

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

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

## 96 | **2. Materials and Methods**

### 97 | **2.1 Description of the lysimeter**

98 | Observations were used of a lysimeter located in the agricultural experimental field station in Wagna in Eastern  
99 | Austria (46° 46.113'N, 15° 33.140'E; altitude 265 m; Klammler and Fank, 2014 ([this issue](#))). Since 1987  
100 | different cultivation strategies are investigated concerning nitrogen-fertilizer input, nitrate leaching and crop  
101 | yields. In 2004, the cultivation changed into comparing low-input farming and organic farming, each covering  
102 | 50% of the test site. Since then, two of the test plots have been equipped with two weighable, monolithic, high-  
103 | precision lysimeters (2 m depth, 1 m<sup>2</sup> surface). The lysimeter in the conventional tillage test plot (KON-system)  
104 | is subject for this study. Cultivation practices including crop species, sowing and harvest dates, and fertilizer  
105 | applications in the test plot are presented in Table 1.

106 | <<Table 1 >>



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

107 The lysimeters are equipped with soil water samplers, soil moisture probes, matrix sensors/tensiometer and soil  
108 temperature probes at four measuring depths (0.35, 0.6, 0.9, 1.8 m). An accompanied measuring profile for soil  
109 moisture, matrix potential and soil temperature is also installed outside the lysimeters (same depths as inside the  
110 lysimeter) to determine if the conditions inside the lysimeter are representative for the rest of the field. At the  
111 bottom of the lysimeter (depth 1.8 m) a suction cups rake was installed which kept the pressure head at this depth  
112 equal to that outside the lysimeter. The water sucked off was collected, weighted and sampled for the  
113 determination of the nitrate concentration. While quantity of seepage water was recorded automatically in  
114 0.1 mm resolution by a tipping bucket, nitrogen concentration in the accumulated leachate was analysed in an  
115 approximately weekly interval. Furthermore, a weather station is installed at agricultural test site in Wagna for  
116 the recording of air temperature, relative humidity, shortwave solar radiation, wind speed, wind direction,  
117 precipitation, sunshine duration and atmospheric pressure at high temporal resolution (~~Fank and Klammler~~ and  
118 ~~Fank~~, 2014; [this issue](#)). Annual precipitation rates and cumulative probabilities of the rates relative to the values  
119 of the period 1961 – 2011 are presented in Table 2.

<<Table 2>>

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

## 2.2 Description of models

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

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

1  
2  
3  
4  
5  
6  
7 135 al., 2010) and on **SOILN** for simulation of the soil organic matter and nitrogen cycle and nitrate leaching  
8  
9 136 (Bergström et al., 1991).

10 137 • **CoupModel (COUP)**, a coupled heat and mass transfer model for soil plant-atmosphere systems, was  
11  
12 138 initially developed to simulate conditions in forest soils, but it has been further developed to simulate  
13  
14 139 conditions in any type of soil, independent of plant cover (Jansson and Karlberg, 2004). COUP applicability  
15  
16 140 is very wide as it includes water, heat, tracer, chloride, nitrogen and carbon modules that can be incorporated  
17  
18 141 in the modelling process. COUP development, calibration procedures and applications are presented by  
19 142 Jansson (2012).

20  
21 143 • **DAISY** is a soil-plant-atmosphere system model designed to simulate crop production, soil water dynamics,  
22  
23 144 and nitrogen dynamics in crop production at various agricultural management practices and strategies  
24 145 (Hansen et al., 1990). The agricultural management model allows for building complex management  
25  
26 146 scenarios (Hansen, 2002). The model has been validated in a number of major comparative tests (Diekkrüger  
27  
28 147 et al., 1995; Hansen et al., 1991a,b; Jensen et al., 1997; Smith et al., 1997; Svendsen et al., 1995; Vereecken  
29 148 et al., 1991; De Willigen, 1991).

30  
31 149 • **EPIC** (Williams et al., 1984; 1989) is a cropping systems simulation model, which was developed to  
32  
33 150 estimate soil productivity as affected by erosion throughout the United States during the 1980's. **EPIC** is a  
34  
35 151 field scale model, but linked to a GIS it has been applied in several regional model applications (Burkart et  
36 152 al., 1999; Sohier et al., 2009). Furthermore the **EPIC** model has been applied to study the effect of  
37  
38 153 agricultural practices and biofuels cultivation on N leaching at the European scale (Bouraoui and Aloe, 2007;  
39 154 Van der Velde et al., 2009).

40  
41 155 • **SIMWASER** (Stenitzer, 1988) simulates the water flow in soil. A unique feature of the model is the  
42  
43 156 description of actual rooting depths based on both root biomass simulated for a crop and on the penetration  
44  
45 157 resistance of the soil. **STOTRASIM** (Feichtinger, 1998) is fully coupled to **SIMWASER** and simulates  
46 158 nitrogen and basic carbon dynamics of agriculturally used soils. The model has already been applied to the  
47  
48 159 region of southeast Styria (Fank et al., 2006). The name of these coupled models is abbreviated as **SIM-STO**.  
49

50 160 • The **SWAP** model, version 3.2 (Van Dam et al., 2008) simulates water flow in the soil – plant – atmosphere  
51  
52 161 domain in an integrated manner. The **ANIMO** model (Groenendijk et al., 2005) is sequentially coupled to  
53 162 **SWAP** and was designed to quantify the relation between fertiliser application rate, soil management and the  
54  
55 163 leaching of nitrogen (N) and phosphorus (P) to groundwater and surface water systems. The **ANIMO** model  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

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

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

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

1  
2  
3  
4  
5  
6  
7 193 Mineralisation is simulated in close correspondence to the organic matter cycle. All models describe the amount  
8  
9 194 of mineralized nitrogen as the excess nitrogen produced from the organic matter decay and transformations to  
10  
11 195 more stable soil organic matter pools. Nitrification is commonly described as a first order process which rate  
12  
13 196 depends on temperature, soil moisture status and ammonium concentration. Denitrification plays no significant  
14  
15 197 role in the soil of the Wagna lysimeters (Leis, 2009), but can be simulated by the models used. A variety of  
16  
17 198 descriptions are implemented but all assume a relationship with temperature, soil moisture content, nitrate  
18  
19 199 concentration and the potential denitrification rate as a function of organic matter content (Heinen, 2006).  
20  
21 200 The lysimeter was installed in 2004 and it was ensured that the original soil layers was put back. During the  
22  
23 201 excavation and filling the soil had been in contact with open air. None of the models paid attention to this event  
24  
25 202 in 2004. To establish the starting conditions on 1-1-2005, three of the six models (i.e., **ARMOSA EPIC**, **STO-**  
26  
27 203 **SIM**, **SW-ANIM**) started in 1987. **COUP** was run for five years prior to the start in 2005 and **DAISY** was run  
28  
29 204 two-years prior to the simulation.

## 30 205 **2.3 Experimental design of study**

31 206 The modelling study comprised of: 1) a blind test with non-calibrated models to get an impression of the  
32  
33 207 performance of the models as they are used in situations where extensive data sets are missing, which often  
34  
35 208 occurs in practice, 2) a calibration period, and 3) a validation period. Inter-comparisons were done between  
36  
37 209 measured and simulated leaching of water and nitrate, including nitrate concentration of the percolate. The  
38  
39 210 outcome of the simulations by all models was collected and analysed by a single person.

### 40 211 **2.3.1 Step 1: Blind test**

41 212 The models first performed a simulation based on a minimum set of data: crop rotation, soil cultivation,  
42  
43 213 fertilization rates, meteorological data, soil profile description and soil moisture retention laboratory  
44  
45 214 measurements of some soil samples. The aim is to establish the bandwidth of differences with the observations  
46  
47 215 without an assessment of the individual models. The **SIM-STO** model was excluded from the blind test as the  
48  
49 216 operators of this model were the owners of all data and **SIM-STO** was already partly calibrated for the test site.  
50  
51 217 After all models delivered their outcome, one external operator compared the predictions against the measured  
52  
53 218 data (seasonal cumulated water flux and nitrogen flux at the bottom of the lysimeter, seasonal flow averaged  
54  
55 219 nitrate concentration) for the period 2005 - 2008. It was not the intention of the blind test to qualify or assess the  
56  
57 220 performance of the individual models and, therefore, the outcome of this test will be presented anonymously.

1  
2  
3  
4  
5  
6  
7 221 Specifically only data on seasonal percolation, flow-averaged nitrate concentration and seasonal nitrate leaching  
8  
9 222 were considered.

10  
11 223 **2.3.2 Step 2: Calibration**  
12  
13 224 Each of the six modelling groups calibrated the models for a limited number of parameters. The successive  
14  
15 225 operations, the objective function and the number of parameters were not prescribed, but were chosen freely by  
16  
17 226 the modelling groups, either based on expert judgement or on a sensitivity analysis. Further details of how the  
18 227 calibration has been carried out for the different models can be found in Supplemental~~ly~~ Material~~s~~<sup>s</sup>.  
19

20 228 **2.3.3 Step 3: Validation**  
21  
22 229 The validation was performed for the period 2009 - 2011, where only information about crop rotation,  
23  
24 230 application of fertilizers, soil cultivation and meteorology was made available for the modelling groups after step  
25  
26 231 2 (calibration) was finished. The procedure for the validation is thus similar to that of the blind test, with the  
27 232 difference that the models were calibrated prior to validation and that the **SIM-STO** model was included in the  
28  
29 233 validation.  
30

31 234 **2.3.4 Step 4: Model comparison**  
32  
33 235 The six models were compared for their performance with respect to 1) the soil moisture retention curves at  
34  
35 236 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  
36  
37 237 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  
38 238 amounts for the time intervals of collected water samples; 6) the nitrate concentrations of the collected water  
39  
40 239 samples; 7) the nitrate-N fluxes at the bottom of the lysimeter for the time intervals of collected water samples.  
41  
42 240 The comparison of results at the depth of 60 cm was excluded because measurements for this depth were only  
43 241 available up to Sept. 2009. Seasonal leached water amounts, nitrogen yields and nitrate-N fluxes were compared  
44  
45 242 to discuss the predictive power for practice oriented model applications. A nitrogen balance was set up for all  
46  
47 243 models. Water fluxes at 1.8 m depth were evaluated for daily and for seasonal values. Nitrate leaching fluxes and  
48 244 nitrate concentrations in the leachate were evaluated at the time intervals for which the soil water was sampled.  
49  
50 245 The sampling time intervals were irregular in time and the models were not able to present concentrations at  
51  
52 246 these specific time events. Therefore, concentrations values for these time intervals were derived according to a  
53  
54 247 volumetric averaging procedure. The nitrate concentrations at depths 0.35 m and 0.9 m can be used to get an  
55 248 impression whether the transport and transformation processes in soil, which ultimately lead to the leaching at  
56  
57 249 depth 1.8 m, have been described adequately. Due to the nature of the model formulations, **EPIC** was not able to  
58

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

250 present the concentrations at the depths of measurement. The number of observations at depth 0.35 m in the  
251 calibration period was too little and were not considered.

252 In the models, much knowledge of soil processes is described which all contribute to the nitrate leaching at depth  
253 1.8 m. To understand the similarities and differences between simulation results and measurements, it is  
254 important to assess the processes. We have done this through the establishment of nitrogen balances per season.

## 2.4 Statistical metrics

255 The behaviour of the main model outputs can be characterized by a number of statistical metrics to indicate the  
256 models' ability to capture different aspects. A complete assessment of model performance should include at least  
257 one absolute error measure and one goodness-of-fit measure (Legates and McCabe, 1999). There are a wide  
258 range of statistical indicators used in studies on soil water and soil nitrogen, but not always a justification is  
259 given for the indicators chosen. For *state variables* many authors use mean (absolute) error ( $MAE$ ), root mean  
260 square error ( $RMSE$ ), index of agreement ( $IoA$ ; Willmott, 1982), and less often the Nash-Sutcliffe modelling  
261 efficiency ( $NSE$ ; Nash-Sutcliffe, 1970) (e.g., Donatelli et al., 2004; Gribb et al., 2009; Herbst et al., 2005;  
262 Khodaverdiloo et al., 2011; Patil and Rajput, 2009; Ritter et al., 2003; Vereecken et al., 2010). For *rate variables*  
263 authors generally use  $MAE$ , mean difference ( $MD$ ), absolute maximum error ( $AME$ ),  $RMSE$ ,  $IoA$ ,  $NSE$ ,  
264 coefficient of determination ( $R^2$ ), percentage of error ( $PE$ ), percentage of bias ( $P_{bias}$ ) (e.g., Akkal-Corfini et al.,  
265 2010; Ale et al., 2012; Dawson et al., 2007, 2010; Jabro et al., 2012; Jachner et al., 2007; Kersebaum et al.,  
266 2007; Krause et al., 2005; Moriasi et al., 2007; Qi et al., 2012; Reusser et al., 2009; Stumpp et al., 2009; Van der  
267 Laan et al., 2011; Wang et al., 2006; Willmott et al., 1985). It appears that a few measures are used both for state  
268 as for rate variables, which we have chosen to use here as well:  $MAE$ ,  $RMSE$ ,  $IoA$ , and  $NSE$  (only for rates),  
269 given by:  
270

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

275 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  
276 the mean values of observations and predictions, respectively. All four measures compare the predictions  $P_t$  and

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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  $-\infty$ , 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  $\sigma_P$  and  $\sigma_O$ , respectively, where

5. Pearson's linear correlation coefficient

$$R = \frac{\sum_{t=1}^n (O_t - \bar{O})(P_t - \bar{P})}{\sqrt{\sum_{t=1}^n (O_t - \bar{O})^2} \sqrt{\sum_{t=1}^n (P_t - \bar{P})^2}}$$

6. Root relative square error:

$$RRSE = \frac{\sqrt{\sigma_P^2 + \sigma_O^2 + 2\sigma_O\sigma_P R}}{\sigma_O}$$

where  $\sigma_O$  and  $\sigma_P$  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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

302 the percolation volume in all seasons and only one model was able to simulate the seasonal percolation  
303 accurately. The range of model results was independent of the seasonal percolation.

304 Seasonal flow averaged nitrate concentrations were underestimated by all models in two of the four seasons. For  
305 the first season, all models underestimated the concentration by 10 – 40 mg L<sup>-1</sup>. The variation of simulated  
306 concentrations and N-fluxes was large. Maximum deviations of seasonal nitrate-N leaching of about 25 kg ha<sup>-1</sup>  
307 were found. All models underestimated the leaching rate in 2005 by 8 – 22 kg ha<sup>-1</sup>. The same holds for the fourth  
308 season, but only one model was able to calculate the nitrate-N flux with a reasonable agreement with the  
309 measurements. In the second season (maize), four models underestimated and one model overestimated the  
310 nitrate concentration and nitrate-N flux. The third season, which was the second season with maize showed a  
311 rather different pattern. The measured nitrate concentration and nitrate-N flux under maize in the 3<sup>rd</sup> season was  
312 much lower than for the maize crop in the 2<sup>nd</sup> season, but the modelled results still showed a large variation with  
313 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

319 In the blind test the modellers had only laboratory measurements of the water retention curve at their disposal,  
320 but in the calibration phase also in situ measured soil moisture contents ( $\theta$ ) and pressure heads ( $h$ ) were available  
321 at four depths. The laboratory measurements were performed for drying samples only, while under field  
322 conditions data pairs of  $\theta(h)$  were detected during wetting and drying cycles so that these were affected by  
323 hysteresis (Basile et al., 2003, 2006). Figure 2 depicts the calibrated  $\theta(h)$  curves for three depths. The results at  
324 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  
325 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  
326 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>>

328 Results for the **EPIC** model are represented by three points as EPIC does not use a continuous description of the  
329  $\theta(h)$  curve. The greatest value for the saturated water content was obtained by the **EPIC** model with a value



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

330 greater than  $0.3 \text{ cm}^3 \text{ cm}^{-3}$  at depth 1.8 m. This parameter is far outside the range that was established by the other  
331 models. A comparison between the calibrated and observed  $\theta(h)$  curves was made by calculating a  $\theta$  for each  
332 value of the measured  $h$ . The performing indices based on computed  $\theta$  and measured  $\theta$  are presented in Table 3.

<<Table 3 >>

334 In general the resulting *MAE*, *RMSE* and *IoA* showed equal trends. The **ARMOSA** model fitted well at depths  
335 0.35 m and 0.9 m, but performed worse at depth 1.8 m. The performance of the **COUP** model appeared to be  
336 weak. At depth 0.9 m the **DAISY** model was better than the **COUP** model, but worse than the other models. The  
337 *IoA* for the **SIM-STO** and **SW-ANIM** models was highest at depth 0.9 m and somewhat lower for the other  
338 depths. It should be noted that a good match of the calibrated  $\theta(h)$  curves with measured data pairs does not a-  
339 priori mean that a good agreement between the time series of measured and calculated  $\theta$  will be obtained.

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

<<Table 4>>

344 Except for **ARMOSA** and **EPIC** in the validation phase, the highest *IoA* values for simulation of the water  
345 contents were achieved at depth 0.9 m. For **SIM-STO** and **SW-ANIM**, the *IoA* values were similar to the  
346 calibration results of the  $\theta(h)$  curves (Table 3). However, the performance by **COUP** increased and that by  
347 **DAISY** decreased compared to Table 3. Except for the **ARMOSA** and the **DAISY** models at depth 0.35 m and  
348 the **SW-ANIM** model at depth 1.8 m, in general the resulting performance indices showed a better agreement  
349 between simulated and observed values for the period 2005 – 2008 than for the comparison based on soil  
350 moisture retention curves. The indices of the validation period 2009 – 2011 were in the same range, or somewhat  
351 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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

357 430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low  
358 which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative  
359 water flux per season corresponded well to the measured water fluxes for most of the models which is also true  
360 for the extreme wet year 2009. However, DAISY showed some overestimation in particular seasons due to  
361 difficulties in parameterizing maize crop parameters. EPIC performed better in the calibration than in the  
362 validation period. The simulated cumulative water flux per season corresponded well to the measured water fluxes  
363 for most of the models: DAISY underestimated the water flux under maize in 2006 and 2010, while EPIC  
364 overestimated the water flux for most of the growing seasons. SW-ANIM underestimated the cumulative water  
365 flux in the two first seasons, but overestimated slightly in some other seasons.

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

374 <<Table 5>>

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

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

381 <<Figure 4>>

382 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  
383 water fluxes the  $\sigma_p/\sigma_o$ -ratio for the validation period was somewhat higher than for the calibration period, but for

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

384 the fluxes averaged for the sampling intervals it can be seen that **ARMOSA**, **DAISY**, **COUP** and **EPIC** resulted  
385 in lower  $\sigma_p/\sigma_o$ -ratio's for the validation period than for the calibration period.

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

389 <<Figure 5>>

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

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

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

### 3.2.2 Soil temperature

407 The soil temperature is an important variable determining the rate of biological processes (N dynamics), for the  
408 crop development in the period of germination, and for soil moisture flow under winter conditions. A  
409 comparison of simulated and measured soil temperatures was carried out as well (data not shown). In general,  
410 the models were well able to simulate soil temperatures and resulted in performance indices much higher than  
411 for moisture contents. The simulation performance at shallow depth was less than the performance at greater

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

412 depths: most models showed a delayed warming up in some spring seasons with respect to the measurements,  
413 which is attributed to the incomplete description of surface temperatures, for most of the models used the air  
414 temperature as the boundary condition.

### 3.2.3 Water fluxes

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

416 ~~————— <<Figure 3>>~~

417 ~~The pattern of cumulative water fluxes per growing season complies generally with the annual precipitation~~  
418 ~~amounts (Table 2) with the exception of maize in 2006 and its preceding crop in the winter of 2005/2006. During~~  
419 ~~the intermediate period after oil pumpkin in 2005 and before maize in 2006, the precipitation amounted to about~~  
420 ~~430 mm. It appears that the evapotranspiration of the intermediate crop (English ryegrass) was relatively low~~  
421 ~~which resulted in a relatively high leaching volume at the start of the maize crop. The simulated cumulative~~  
422 ~~water flux per season corresponded well to the measured water fluxes for most of the models: **DAISY**~~  
423 ~~underestimated the water flux under maize in 2006 and 2010, while **EPIC** overestimated the water flux for most~~  
424 ~~of the growing seasons. **SW-ANIM** underestimated the cumulative water flux in the two first seasons, but~~  
425 ~~overestimated slightly in some other seasons.~~

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

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

430 ~~————— <<Table 5>>~~

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

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465

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  $\sigma_p/\sigma_o$ -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  $\sigma_p/\sigma_o$ -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

### 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

1  
2  
3  
4  
5  
6  
7 466 nitrate flux by the water flux. The nitrate concentrations in the calibration phase simulated by **SIM-STO** showed  
8  
9 467 a bad performance, while the results for the validation phase were much better. The higher peak concentrations  
10  
11 468 during the calibration phase were not approached by **SIM-STO**. On the other hand, **SW-ANIM** showed a good  
12  
13 469 agreement of nitrate concentrations during the calibration phase, while there is a mismatch during the validation  
14  
15 470 phase. The concentration peaks during the validation phase were severely overestimated by **SW-ANIM** due to  
16  
17 471 an overestimation of the biological fixation rates of some non-leguminous catch crops in this period.

18 472 The nitrate-N flux at depth 1.8 m represents the nitrogen transport to deeper soil layers and is relevant for  
19  
20 473 predictions of nitrate concentrations in deeper groundwater. **ARMOSA**, **DAISY**, **EPIC** and **SIM-STO**  
21  
22 474 underestimated the nitrate N-flux under winter barley preceded by a catch crop in 2007-2008, but **SW-ANIM**  
23  
24 475 overestimated the nitrate N-flux during this period. The **COUP** model was able to calculate the nitrate-N flux in  
25  
26 476 five of the seven seasons that cover the calibration and validation period. **ARMOSA** and **DAISY** calculated the  
27  
28 477 total seasonal nitrate-N flux well in three of the seven seasons, while **EPIC**, **SIM-STO** and **SW-ANIM**  
29  
30 478 calculated this flux well in two of the seven seasons. The last season appeared to be the most difficult one,  
31  
32 479 because of the exceptional dry conditions. The leaching after the 2009 oil pumpkin crop also showed significant  
33  
34 480 deviations between model predictions and measurements. The largest deviations of seasonal nitrate-N fluxes  
35  
36 481 occurred in the results of **COUP** and **SW-ANIM** for the exceptional wet year 2009.

37 482 Table 6 presents the statistical indicators for both the nitrate concentrations and the nitrate-N leaching rates,  
38  
39 483 based on the sampling time series. The largest deviations between predicted and simulated nitrate concentrations  
40  
41 484 were found for the **SIM-STO** results in the calibration period for which the *IoA* ~~and the *NSE*~~ amounted to 0.43,  
42  
43 485 ~~and -0.76, respectively.~~ Remarkably the smallest deviations were found for the same model for the validation  
44  
45 486 period for which *IoA* ~~and *NSE*~~ amounted to 0.78, The underestimation of the nitrate-N flux by **SIM-STO** is  
46  
47 487 most likely due to immobilization processes that are overemphasized for the 2005 and 2008 periods. Thus, less  
48  
49 488 nitrate was released to the soil water phase which led to the underestimation of the nitrate concentration in the  
50  
51 489 leachate, and 0.08, respectively.

50 490 <<Table 6>>

52 491 The **COUP** model showed the best performance for the nitrate concentrations of the calibration period with *IoA*  
53  
54 492 = 0.97 ~~and *NSE* = 0.86~~ directly followed by the **SW-ANIM** model. The results from **EPIC** and **SW-ANIM** for  
55  
56 493 concentrations in the validation period were weak with *RMSE* > 20 mg L<sup>-1</sup> ~~and *NSE* < -2~~. The statistical indices  
57  
58 494 of the nitrate-N leaching rates showed a similar picture. The **SIM-STO** model performed relatively weak during

1  
2  
3  
4  
5  
6  
7 495 the calibration phase. For the leaching rates in this period **DAISY** and **SW-ANIM** had the best performance and  
8  
9 496 for the validation period **ARMOSA** and **DAISY** performed relatively the best. The *NSE* values (data not shown)  
10 497 for both the concentration and the leaching rates in the validation period were almost all negative, showing that  
11  
12 498 the calibrated models had great difficulties to predict concentrations and leaching rates for the more extreme  
13  
14 499 conditions of the validation period.

15  
16 500 Statistical performance of predicted nitrate concentrations and leaching rates were expressed in Taylor diagrams  
17  
18 501 in Figure 6. Calibrated nitrate concentrations by **COUP** and **SW-ANIM** had *R*-values greater than 0.9 and were  
19  
20 502 closest to the (1,0) point. Except for **SIM-STO**, the models showed  $\sigma_p/\sigma_o$  ratios for the calibration step that did  
21  
22 503 not deviate much from 1; for **SIM-STO** the  $\sigma_p/\sigma_o$  ratio was much lower than 1 and *R* < 0.

23  
24 504 <<Figure 6>>

25  
26 505 The plots clearly show the much weaker performance for the validation period than for the calibration period,  
27  
28 506 expressed by lower *R*-values and higher  $\sigma_p/\sigma_o$  ratio's. **SIM-STO** showed the best performance for  
29  
30 507 concentrations in the validation period with *R* > 0.7,  $\sigma_p/\sigma_o$  close to one, and *RRSE* = 0.75, while for the other  
31  
32 508 model *RRSE* > 1. For the nitrate fluxes in the calibration period *RRSE* values were between 0.64 and 0.86, while  
33  
34 509 for the validation period, the values were between 1 and 2 even with a peak of 8.6 for **SW-ANIM** (data point not  
35  
36 510 seen in Figure 6). The *R*-values of the nitrate fluxes in the validation period were in the range 0.18 (**EPIC**) to  
37  
38 511 0.50 (**COUP**). The  $\sigma_p/\sigma_o$  ratio ratios were in the range 0.75 to 2.3 with a peak of 8.8 for **SW-ANIM** (data point  
39  
40 512 not seen in Figure 6). The values for  $\sigma_p/\sigma_o$  ratio greater than 1 for both the concentrations and the nitrate fluxes  
41  
42 513 indicate that the variation of the simulated values is greater than the variation of the observed values.

43  
44 514 Table 7 presents the performance indices for the nitrate concentrations at depths 0.35 m and 0.9 m. The *IoA*  
45  
46 515 values indicate that the best agreement between simulated and measured values was achieved for the calibration  
47  
48 516 period, but *MAE*-values and *RMSE*-values were highest for the calibration results at depth 0.9 m and lowest for  
49  
50 517 the validation results at depth 0.9 m. This apparent contradiction is due to the number of measurements on which  
51  
52 518 the indices were calculated. Further analysis was based on *IoA* because the ranking of these values corresponded  
53  
54 519 better to the results of the leaching water at depth 1.80 m.

55  
56 520 <<Table 7>>

57  
58 521 Calibrated concentrations yielded *IoA*-values ranging from 0.44 (**SIM-STO**) to 0.84 (**SW-ANIM**). The results  
59  
60 522 for the validation period resulted in somewhat lower *IoA* values, except for **SIM-STO** which shows better results

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

523 for the validation than for the calibration period. The **ARMOSA** results were the most constant for the different  
524 depths and periods. Both **COUP** and **SW-ANIM** show significantly poorer *IoA* values for the validation than for  
525 the calibration period. **DAISY** and **SIM-STO** showed slightly better results for the concentrations at depth 0.9 m  
526 than for the concentrations at depth 0.35 m. The other models performed slightly better for depth 0.35 m. Except  
527 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  
528 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  $(P_i - O_i)^2$  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  $(P_i - O_i)^2$  values on the sampling periods  
but **SIM-STO** used the summed  $(P_i - O_i)^2$  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 moderate-*IoA* values in  
the range 0.76-0.87 for the calibrated nitrate fluxes, while the others model resulted in poor-*IoA* = 0.43 values  
(Table 96). 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 96).  
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  
the comparison was carried out for seasonal values.



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

1  
2  
3  
4  
5  
6  
7 580 nitrogen compositions but this was overcome by introducing new manure types, so that the fertilizer input was  
8  
9 581 close to the observed values.

10  
11 582 The estimates for atmospheric deposition ranged from 4.2 kg ha<sup>-1</sup> a<sup>-1</sup> (**COUP**) to 23.4 kg ha<sup>-1</sup> a<sup>-1</sup> (**DAISY**),  
12  
13 583 averaged for seven growing seasons. Only literature values were available and most modelling groups have used  
14  
15 584 the model default values or the figure they are familiar with for their own country. **ARMOSA** calculated for the  
16  
17 585 validation phase lower wet deposition rates than for the calibration phase due to lower precipitation amounts.  
18 586 Some models assumed only dry deposition at a constant rate, while other models also imposed nitrogen input by  
19  
20 587 rainfall.

21  
22 588 The most stressing differences are for biological N-fixation. Some models do not describe the biological N-  
23  
24 589 fixation process as such but modellers had possibilities to assume fixation rates by introducing a nitrogen rich  
25  
26 590 organic material which was amended continuously during the growing season. The **DAISY** and the **EPIC**  
27 591 modelling groups did not take account for N-fixation, either due to a lack model formulations implemented or to  
28  
29 592 a lack of knowledge about this process. **SIM-STO** assumed only for the first season some biological N-fixation  
30  
31 593 by the crop mixture that included white clover. The **COUP** and the **SW-ANIM** modelling groups took account  
32  
33 594 for N-fixation, including for periods for which one wouldn't expect (English ryegrass). In **SW-ANIM** the  
34 595 biological N-fixation is lumped with the mineralization of some of the crop residues that descended from the  
35  
36 596 most recent and previous catch crops. The model output does not allow to unravel the biological N-fixation as  
37  
38 597 such and mineralisation of earlier catch crop residues.

39  
40 598 The **COUP** model did not take account for ammonia volatilization. The other models did, and showed a range of  
41  
42 599 2% to 35% of the nitrogen in the animal manure amended to the soil. The highest volatilization rates were  
43 600 simulated by **SIM-STO**: 27% and 35% of the animal manure N in 2008 and 2011, respectively. This could  
44  
45 601 possibly explain the underestimation of nitrate leaching in 2008, but not in 2011. For these years, the differences  
46  
47 602 of the model predictions amounted to more than 22 and 37 kg ha<sup>-1</sup> a<sup>-1</sup>, respectively, which is higher or in the  
48  
49 603 same range as the measured nitrate-N leaching. Volatilization was calculated by **EPIC** and **ARMOSA** (about 4  
50 604 kg ha<sup>-1</sup>) for the first growing season of the validation period, while no farm fertilizer was applied.

51  
52 605 The models encountered difficulties with the simulation of nitrogen crop off-take. Deviations of simulated  
53  
54 606 uptake rates from the observed values of more than 50 kg ha<sup>-1</sup> occurred for three years by **ARMOSA** (2006,  
55  
56 607 2008, 2009), **EPIC** (2005, 2009, 2010) and **SIM-STO** (2006, 2008, 2010), for two years by **DAISY** (2007,  
57  
58 608 2010), and for one year (2011) by **COUP** and **SW-ANIM**. The **EPIC** model was not able to simulate nitrogen

1  
2  
3  
4  
5  
6  
7 609 crop off-take by oil pumpkin, because this crop is unknown in the standard database of crop parameters that  
8  
9 610 comes with the model. The **DAISY** model failed to simulate a reasonable crop off-take by maize in 2007, while  
10 611 the N off-take in the preceding year was overestimated by 60 kg ha<sup>-1</sup>. The calibrated parameters for crop uptake  
11  
12 612 were not optimal for the maize as is also apparent from the calculated crop off-take in 2010 where the  
13  
14 613 overestimation amounted nearly 100 kg ha<sup>-1</sup>. Despite the fact that **SW-ANIM** included the N-yield in the object  
15  
16 614 function of the calibration procedure, the modelled crop off-take differed from the measured crop off-take by -14  
17 615 to +19 kg ha<sup>-1</sup>. The **SW-ANIM** underestimated crop off-take in the validation period. Crop off-take is governing  
18  
19 616 the soil nitrogen balance to a large extent and an erroneous calculation of the N off-take means that a possible  
20  
21 617 correct nitrate leaching should be considered as little robust.

22  
23 618 Denitrification is only of significance for the **DAISY** and **EPIC** results, while other models simulated zero or  
24  
25 619 negligible denitrification rates. For most of the models, these estimates were biased by the opinion of the data  
26  
27 620 holders who made plausible from their analysis of soil nitrogen balances that denitrification is not a significant  
28  
29 621 factor (Leis, 2009). The degree of saturation (*S*) at depth 0.35 m exceeds 80% for most of the time and only  
30 622 **COUP** and **SIM-STO** have default threshold values for *S* higher than 80% while other models use lower default  
31  
32 623 threshold values for *S* (Heinen, 2006). Except for **DAISY** and **EPIC**, also **ARMOSA** and **SW-ANIM** should  
33  
34 624 have calculated some denitrification when using default values. Except for the first year, the denitrification  
35 625 calculated by **EPIC** exceeded the nitrate-N leaching.

36  
37 626 The change of the total N amount in soil included both organic and mineral forms and was calculated as the  
38  
39 627 residual from the balance. A positive sign means an increase of the total amount whereas a negative sign  
40  
41 628 indicates a depletion of the stock. The model results showed large differences and the largest difference occurred  
42  
43 629 in 2010 where **DAISY** calculated a depletion of 105 kg ha<sup>-1</sup> while **SW-ANIM** calculated an increase of 103 kg  
44 630 ha<sup>-1</sup>. The increase of the amount resulted from the assumed biological fixation and the inputs caused by the  
45  
46 631 cultivation of catch crops. When no additional inputs by fixation or by catch crops was assumed, a depletion will  
47  
48 632 occur (**DAISY** and **EPIC**).

49  
50 633 Except for **SIM-STO** in 2005 and 2008, differences between calculated seasonal nitrate-N leaching rates were  
51  
52 634 relatively small for the calibration phase. The deviations were much larger for the validation phase, where **SW-**  
53  
54 635 **ANIM** overestimated the leaching by 39 and 29 kg ha<sup>-1</sup> in 2009 and 2010, respectively. The observed small  
55 636 leaching rate in 2010 was not approached by any model. Transport of ammonium, organic dissolved N or by

1  
2  
3  
4  
5  
6  
7 637 surface runoff was calculated at a maximum of 8 kg ha<sup>-1</sup> by the **COUP** model for the first year of the validation  
8  
9 638 period.

10  
11 639 The long term nitrogen balances were summarized at the bottom of Table 8 to further compare the difference of  
12  
13 640 the modellers perceptions of the plant and soil nitrogen cycle.

14  
15 641 The seven year balance depicted the major differences between the models clearly. Despite the crop failure in  
16  
17 642 2007 simulated by **DAISY**, this model showed the highest summed seven year amount, while the summated crop  
18  
19 643 off-take by **SIM-STO** lagged behind with 200 kg ha<sup>-1</sup> relative to the recorded amount. For the individual years  
20  
21 644 the **ARMOSA** results differed considerably from the observations, but the summated seven year crop off-take  
22  
23 645 resembled the measured value rather good.

24  
25 646 Most models have been designed for the field scale for which an average N-yield is calculated. The spatial scale  
26  
27 647 of the lysimeter (1 m<sup>2</sup>) differs from the field scale and the variation of crop off-take rates at this scale is much  
28  
29 648 larger than for the field scale. This is illustrated by the oil pumpkin crop in 2005. Only two seeds were planted in  
30  
31 649 the lysimeter. One of the plants died at the start of the generative phase and no harvest was obtained from this  
32  
33 650 plant. This event influenced the yield at the lysimeter scale pretty much, but the yield at the field scale was  
34  
35 651 barely influenced and it can be expected that field scale models encountered difficulties.

36  
37 652 The total nitrogen loss by denitrification ranged from 0 to 249 kg ha<sup>-1</sup> and was subject to the modellers'  
38  
39 653 perception of the possibility of denitrification in the soil at the Wagna experimental field station.

40  
41 654 The low input farming system was capable to produce relatively high yields for maize and grains, and for oil  
42  
43 655 pumpkin a N-yield of 51 to 57 kg ha<sup>-1</sup> was recorded, but the observed nitrate-N leaching exceeded the N-excess,  
44  
45 656 the latter defined as the total addition of mineral fertilizers and animal manure minus the crop off-take.

46  
47 657 **ARMOSA**, **DAISY** and **EPIC** predicted higher nitrate N-leaching than the N-excess (Fig. 7), while the other  
48  
49 658 models showed a more or less equal value (**SW-ANIM**) or a lower value (**COUP**, **SIM-STO**). One of the main  
50  
51 659 difficulties was to describe the role of the intermediate catch crops in the crop rotation on the delivery of N.

52  
53 660 Some of the intermediate crops fixate atmospheric N which leads to an input to the soil and other crops are only  
54  
55 661 able to preserve some of the N excess which remains in soil after the catch crops for the next growing season. No  
56  
57 662 data on the N uptake rates and the quality of the resulting green biomass of these intermediate crops were  
58  
59 663 available. Each of the modellers had to make assumptions for the effect of these crops on the soil N cycle. The  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

664 estimates of the seven years summed additional input to the soil by biological N-fixation varied from 0-2 kg ha<sup>-1</sup>  
665 (**DAISY**, **EPIC**) to 371 kg ha<sup>-1</sup> (**SW-ANIM**) (Table 8).

<<Figure 7>>

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

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 the 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.

690 Bellocchi et al. (2010) reviewed the methods and different indicators used for the validation of different types of  
691 biophysical models. Confalonieri et al. (2010) used *NSE* and *RRMSE*, together with four other indices to assess

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

~~Following these authors, we have chosen the *IoA* for a qualitative assessment of the different model outputs.~~  
705 Typical state variables which correspond with instantaneous observations have been distinguished from water  
706 fluxes and nitrate concentrations analysed in composed water samples. For the latter we assumed *IoA* values  
707 above 0.9 as accurate and *IoA* values below 0.75 as inaccurate. For soil water contents and nitrate concentrations  
708 we assume *IoA* values greater than 0.8 as accurate and *IoA* values smaller than 0.6 as inaccurate. Krause et al.  
709 (2005) stated that even for  $IoA > 0.65$  models can result in poor performance, they sure will for  $IoA < 0.6$ , which  
710 was here chosen as the lowest boundary. The *IoA* scoring for the calibration and validation periods are listed in  
711 Table 9.

<<Table 9>>

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

720 We have also assessed the accuracy of the seasonal amounts on the basis of the mean absolute error (*MAE*). The  
721 seven seasons included the oil pumpkin crop twice, which was an unknown or a particular crop for most of the  
722 modelling groups. The seven year series contained an extremely wet year (2009) and a dry summer (2011). For  
723 the performance assessment for average crop and rainfall conditions *MAE* of the five best values (*MAE<sub>5</sub>*) out of  
724 seven (*MAE<sub>7</sub>*) are presented in Table 10 to examine if the models perform better for average conditions. In some  
725 cases the improvement was more than 50%, and the ranking of the models slightly changed. Despite the fact that  
726 *MAE* is less sensitive to outliers than e.g. *IoA*, extreme situations (unknown crop, wet or dry years) can have a  
727 large impact on *MAE*.

728 <<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~~

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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 the comparison was carried out for seasonal values.

Dieckrüger 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. Dieckrüger 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 extent 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 Roelmsma, 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



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

~~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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

805 boundary condition in the lysimeter resulted in leaching. For other situations where capillary rise may occur, the  
806 models have not been inter-compared. Finally, it is noticed that the soil hydraulic properties as determined in the  
807 laboratory on small soil samples does not guarantee well-predicted soil water contents and soil water fluxes even  
808 for a well-controlled lysimeter situation. Partly, this may be due to the lack of knowledge of hysteresis or its  
809 description in the models.

#### 5.4.2 Procedure

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

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

#### 5.4.3 Decreased performance when averaging

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

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

1  
2  
3  
4  
5  
6  
7 833 conditions the intervals were shortened and under extreme dry conditions the intervals were longer because no  
8  
9 834 percolation water was present.

10  
11 835 2) The probability density distributions of the daily water fluxes and averaged water fluxes for the calibration  
12  
13 836 and validation periods appeared to be unequal (data not shown). This was concluded from a non-parametric  
14  
15 837 analogue of a one-way analysis of variance performed by the one-way analysis of variance by ranks after  
16  
17 838 Kruskal-Wallis (1952). The different statistical behaviour may result in variant effects of volume weighted  
18  
19 839 averaging on the performance indices.

20  
21 840 3) Certain days or periods may have had a great effect on the averaging. A leave-one-out calculation procedure  
22  
23 841 was performed to qualitatively explore the effect of certain days and periods on the performance of the models.  
24  
25 842 In the series of data pairs of observed and simulated water fluxes, one data pair is left out and the *IoA* was  
26  
27 843 calculated for the remainder of the population. This procedure is repeated for each of the data pairs and the  
28  
29 844 results are subtracted from the *IoA*-value based on the total series of data pairs belonging to either the daily  
30  
31 845 fluxes of the calibration or the validation phase or to the averaged values of the phases. Only the results greater  
32  
33 846 than 0.001, in absolute sense, haven been plotted in Figure 8.

33 847 <<Figure 8>>

34  
35 848 The exclusion of a particular data pair can result in both an improvement (negative values) or a deterioration  
36  
37 849 (positive values) of the  $\Delta IoA$ . Furthermore, it is notable that the  $\Delta IoA$  of daily fluxes responded differently  
38  
39 850 compared to the  $\Delta IoA$  for averaged fluxes per sampling interval. For almost all models the exclusion of the value  
40  
41 851 simulated for 19 Sept 2006 would affect the  $\Delta IoA$ . The effect of excluding the value of this period is much  
42  
43 852 smaller for the  $\Delta IoA$  based on the averaged values per sampling interval. The maximum effect in the series of  
44  
45 853 daily values occurs for a certain day of the calibration period and the maximum effect in the series of averaged  
46  
47 854 values per sampling interval is calculated for a time interval in Sept. 2010 which belongs to the validation phase  
48  
49 855 The maximal effect of leaving one value out is greater for the validation period than for the calibration period.  
50  
51 856 Based on this analysis, it is plausible that the averaging of water fluxes has a different effect on the performance  
52  
53 857 indices of the calibration phase than on those of the validation phase.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

#### 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 ( $123 \text{ kg ha}^{-1}$ ) 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.

- 1  
2  
3  
4  
5  
6  
7 886 e. Traditional Richard's / Darcy Buckingham equation based models that make use of the Mualem-van  
8  
9 887 Genuchten descriptions and disregard phenomena as hysteresis, preferential flow and multiple  
10 888 phase flow encounter difficulties with an accurate and consistent simulation of both water contents  
11  
12 889 and water fluxes for the soil and conditions of the Wagna lysimeter.  
13  
14 890 f. Some models which performed relatively well in the calibration phase of the study failed to  
15  
16 891 simulate the nitrate concentrations and fluxes in the validation phase (**SW-ANIM**), while other  
17 892 models behaved relatively bad in the calibration phase and showed better results in the validation  
18  
19 893 phase (**SIM-STO**). An accurate calibration does not guarantee a good predictive power of the  
20  
21 894 model.  
22 895 g. The catch crop mixtures and the non-harvested English ryegrass play an important role in the  
23  
24 896 nutrient dynamics of the soil. This role is addressed weakly by the simulation models: (1) due to a  
25  
26 897 lack of experimental data on nitrogen uptake rates and mineralization of residues of these  
27 898 intermediate crops, and (2) lack of knowledge to describe the relevant processes related to the  
28  
29 899 foreign crops  
30  
31 900 h. Assessment of future climate and land use changes requires a good predictive power of the models  
32  
33 901 and a certain level of robustness. Although the robustness is not clear for the tested models, the  
34 902 process oriented dynamic models used in this study are useful for hypothesis testing.  
35  
36  
37

## 38 903 **6.5. Acknowledgements**

39  
40 904 This research was made possible by the GENESIS project of the EU 7<sup>th</sup> Framework Programme (Project No.  
41  
42 905 226536; FP7-ENV-2008-1). We are grateful for the experimental data provided by Joanneum Raum (Graz,  
43 906 Austria). [The modelling team of Democritus University of Thrace would like to thank Per-Erik Jansson \(Royal](#)  
44  
45 907 [Institute of Technology, Stockholm, Sweden;\) for his valuable help during the application of CoupModel.](#)  
46  
47  
48

## 49 908 **7.6. References**

- 50  
51 909 [Acutis M, Confalonieri R. Optimization algorithms for calibrating cropping systems simulation models. A case](#)  
52 910 [study with simplex derived methods integrated in the WARM simulation environment. Ital J Agrometeorol](#)  
53 911 [2006;11:26-34.](#)  
54  
55 912 Akkal-Corfini N, Morvan T, Menasseri-Aubry S, Bissuel-Bélaygue C, Poulain D, Orsini F, Leterme P. Nitrogen  
56 913 mineralization, plant uptake and nitrate leaching following the incorporation of (15N)-labeled cauliflower  
57 914 crop residues (Brassica oleracea) into the soil: a 3-year lysimeter study. Plant Soil 2010;. 328:17–26. DOI  
58 915 10.1007/s11104-009-0104-0  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7 916 Ale S, Bowling LC, Youssef MA, Brouder SM. Evaluation of simulated strategies for reducing nitrate–nitrogen  
8 917 losses through subsurface drainage systems. *J Environ Qual* 2012; 41:217-228.  
9  
10 918 Basile A, Ciollaro G, Coppola A. Hysteresis in soil water characteristics as a key to interpreting comparisons of  
11 919 laboratory and field measured hydraulic properties. *Water Resour Res* 2003;39:1355-1367.  
12 920 Basile A., Coppola A, De Mascellis R, Randazzo L. Scaling approach to deduce field unsaturated hydraulic  
13 921 properties and behavior from laboratory measurements on small cores. *Vadose Zone J* 2006;5:1005-1016.  
14 922 Bellocchi G, Rivington M, Donatelli M, Matthews K, Validation of biophysical models: issues and  
15 923 methodologies. A review. *Agron Sustain Dev* 2010;30:109–130.  
16  
17 924 Bergström L, Johnsson H, Torstensson G. Simulation of soil nitrogen dynamics using the SOILN model. *Nutr  
18 925 Cycl Agroecosys* 1991;27:181–188.  
19 926 Bouraoui F, Aloe A. European Agrochemicals Geospatial Loss Estimator: Model development and Applications,  
20 927 EUR – Scientific and Technical Research series, ISSN 1018-5593. Office for Official Publications of the  
21 928 European Communities, Luxembourg. 2007.  
22  
23 929 Buckingham, E. Studies on the movement of soil moisture. *Bull.* 38, USDA, Bureau of Soils, Washington, DC,  
24 930 1907.  
25 931 Burkart MR, Kolpin DW, James DE. Assessing groundwater vulnerability to agrichemical contamination in the  
26 932 Midwest US. *Water Sci Technol* 1999;39:103-112.  
27  
28 933 Confalonieri R, Bregaglio S, Bocchi S, Acutis M. An integrated procedure to evaluate hydrological models.  
29 934 *Hydrol Process* 2010;24:2762–2770.  
30 935 Darcy H. Les fontaines publique de la ville de Dijon. Dalmont, Paris, 1856.  
31  
32 936 Dawson CW, Abrahart RJ, See LM. HydroTest: A web-based toolbox of evaluation metrics for the standardised  
33 937 assessment of hydrological forecasts. *Environ Modell Softw* 2007;22:1034–1052.  
34 938 <http://dx.doi.org/10.1016/j.envsoft.2006.06.008>  
35 939 Dawson CW, Abrahart, RJ, See, LM. HydroTest: Further development of a web resource for the standardised  
36 940 assessment of hydrological models. *Environ Modell Softw* 2010;25:1481–1482.  
37 941 <http://dx.doi.org/10.1016/j.envsoft.2009.01.001>  
38 942 De Willigen P, Neeteson JJ, Comparison of six simulation models for the nitrogen cycle in the soil. *Fert Res*  
39 943 1985;8:157-171.  
40  
41 944 De Willigen P. Nitrogen turnover in the soil-crop system; comparison of fourteen simulation models. *Fert Res*  
42 945 1991;27: 141-149.  
43 946 Diekkrüger B, Söndgerath D, Kersebaum KC, McVoy CW. Validity of agroecosystem models a comparison of  
44 947 results of different models applied to the same data set. *Ecol Model* 1995;81:3-29.  
45 948 <http://www.sciencedirect.com/science/article/pii/030438009400157D>  
46  
47 949 ~~Doherty J. PEST Model independent parameter estimation user manual: 5th edition. Watermark Numerical  
48 950 Computing, 2005.~~  
49 951 Donatelli M, Wösten JHM, Bellocchi G. Evaluation of pedotransfer functions. In: Pachepsky Y, Rawls WJ,  
50 952 editors. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam. 2004. p. 357–362.  
51 953 Durner W, Jansen U, Iden SC. Effective hydraulic properties of layered soils at the lysimeter scale determined  
52 954 by inverse modelling. *Eur. J. Soil Sci.* 2007;59:114–124. doi: 10.1111/j.1365-2389.2007.00972.x  
53  
54 955 EU. 1991. Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against  
55 956 pollution caused by nitrates from agricultural sources. *Off. J. Eur. Commun.* L375:1–8. Available at  
56 957 <http://ec.europa.eu/environment/water/water-nitrates/directiv.html>

Formatted: English (U.K.)

1  
2  
3  
4  
5  
6  
7 958 Fank J, Fastl G, Kupfersberger H, Rock G. Die Bewirtschaftung des Versuchsfeldes Wagna - Auswirkung auf  
8 959 die Grundwassersituation. Umweltprogramme für die Landwirtschaft 2006. Höhere Bundeslehr- und  
9 960 Forschungsanstalt für Landwirtschaft, A-8952 Irdning. Gumpenstein, 2006.

10 961 Feichtinger F. STOTRASIM – Ein Modell zur Simulation der Stickstoffdynamik in der ungesättigten Zone eines  
11 962 Ackerstandortes. In: *Modelle für die gesättigte und ungesättigte Bodenzone*. Schriftenreihe des Bundesamtes  
12 963 für Wasserwirtschaft, 7, Wien, 1998, p. 14-41,

13 964 Ferrara RM, Trevisiol P, Acutis M, Rana G, Richter GM, Baggaley N. Topographic impacts on wheat yields  
14 965 under climate change: two contrasted case studies in Europe. *Theor. Appl. Climatol.* 2011;99:53–65.

15 966 Gribb MM, Hansen FI, Aleshia Chandler, McNamara DG, James P. The effect of various soil hydraulic property  
16 967 estimates on soil moisture simulations. *Vadose Zone J* 2009; 8:321–331.

17 968 Grizzetti B, Bouraoui F, Billen G, Van Grinsven H., Cardoso AC, Thieu V, Garnier J, Curtis C, Howarth R,  
18 969 Johnes P. Nitrogen as a threat to European water quality. In: *The European Nitrogen Assessment 17*,  
19 970 Cambridge, UK: Cambridge University Press, 2011, p. 379-404.

20 971 Groenendijk, P., L.V. Renaud, and J. Roelmsa. 2005. Prediction of Nitrogen and Phosphorus leaching to  
21 972 groundwater and surface waters. Process descriptions of the animo4.0 model. Alterra–Report 983, Alterra,  
22 973 Wageningen. <http://content.alterra.wur.nl/Webdocs/PDFFiles/Alterrarapporten/AlterraRapport983.pdf>

23 974 Hansen S, Jensen HE, Nielsen NE, Svendsen H. **DAISY**: Soil Plant Atmosphere System Model. NPO Report  
24 975 No. A 10. The National Agency for Environmental Protection, Copenhagen, 1990, 272 pp.

25 976 Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of nitrogen dynamics and biomass production in  
26 977 winter wheat using the Danish simulation model DAISY. *Fert Res* 1991a;27: 245-259.

27 978 Hansen S, Jensen HE, Nielsen NE, Svendsen H. Simulation of biomass production, nitrogen uptake and nitrogen  
28 979 leaching by using the Daisy model. In: *Soil and Groundwater Research Report II: Nitrate in Soils, Final*  
29 980 *Report on Contracts EV4V-0098-NL and EV4V-00107-C. DG XII. Commission of the European*  
30 981 *Communities, 1991b; 300–309.*

31 982 Hansen S. DAISY, a flexible Soil-Plant-Atmosphere system Model. Report. Dept. Agric, Danish Informatics  
32 983 Network in the Agricultural Sciences, 2002. <http://www.dina.kvl.dk/~DAISY/ftp/DAISYDescription.pdf>.

33 984 Heinen, M. Simplified denitrification models: Overview and properties. *Geoderma* 2006;133:444-463.

34 985 Herbst M, Fialkiewicz W, Chen T, Pütz T, Thiéry D, Mouvet C, Vachaud G, Vereecken H. Intercomparison of  
35 986 flow and transport models applied to vertical drainage in cropped lysimeters. *Vadose Zone J* 2005;4:240-254.

36 987 Jabro JD, Jabro AD, Fales SL. Model performance and robustness for simulating drainage and nitrate-nitrogen  
37 988 fluxes without recalibration. *Soil Sci. Soc. Am. J.* 2012;76:1957–1964 [doi:10.2136/sssaj2012.0172](https://doi.org/10.2136/sssaj2012.0172)

38 989 Jachner S, Van den Boogaart KG, Petzoldt T. Statistical Methods for the Qualitative Assessment of Dynamic  
39 990 Models with Time Delay (R package qualV). *J Stat Softw* 2007;22:(8),1–30. <http://www.jstatsoft.org/v22/i08>

40 991 Janssen PHM, Heuberger PSC. Calibration of process-oriented models. *Ecol Model* 1995;83: (1-2) 55-66.  
41 992 [http://dx.doi.org/10.1016/0304-3800\(95\)00084-9](http://dx.doi.org/10.1016/0304-3800(95)00084-9)

42 993 Jansson P-E, Karlberg L. Coupled heat and mass transfer model for soil-plant-atmosphere systems Royal  
43 994 Institute of Technology, Dept of Civil and Environmental Engineering, Stockholm, 2004. 435 pp.

44 995 Jansson P-E. CoupModel: Model use, calibration and validation, *Transactions of the ASABE* 2012;55(4):1-11.

45 996 Jensen LS, Mueller T, Nielsen NE, Hansen S, Crocker GJ, Grace PR., Klir J, Körschens M, Poulton PR.  
46 997 Simulating trends in soil organic carbon in long-term experiments using the soil-plant-atmosphere model  
47 998 DAISY. *Geoderma* 1997;81:5-28.

48 999 **Justes E, Mary B, Meynard JM, Machet JM, Thelier Huches L. Determination of a critical nitrogen dilution**  
50 1000 **curve for winter wheat crops. *Ann Bot London* 1994; 74:397-407.**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1001 Kersebaum KC, Hecker, J-M, Mirschel W, Wegehenkel M. Modelling water and nutrient dynamics in soil–crop  
1002 systems: a comparison of simulation models applied on common data sets. In: Kersebaum KC et al., editors,  
1003 Modelling Water and Nutrient Dynamics in Soil–Crop Systems, Springer, 2007, p. 1–17.

1004 [Khodaverdilo H, Homae M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions](#)  
1005 [for some calcareous soils, J Hydrol 2011;399:93-99](#)

1006 Klammler G, Fank J. [Determining Measuring](#) water and nitrogen balances for beneficial management practices  
1007 using lysimeters at Wagna test site (Austria). Sci. Tot. Environ 2014; *submitted this issue*.

1008 Krause P, Boyle, DP, Båse F. Comparison of different efficiency criteria for hydrological model assessment.  
1009 Advances in Geosciences, 2005;5:89–97. <http://www.adv-geosci.net/5/89/2005/adgeo-5-89-2005.pdf>

1010 Kroes J and Roelmsa J. Simulation of water and nitrogen flows on field scale: application of the SWAP–ANIMO  
1011 model for the Müncheberg data set. In: K. Ch. Kersebaum et al. (eds.), Modelling Water and Nutrient  
1012 Dynamics in Soil–Crop Systems, 2007, Springer, pp 111–128.

1013 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.

1014 [Khodaverdilo H, Homae M, Van Genuchten MT, Dashtaki SG. Deriving and validating pedotransfer functions](#)  
1015 [for some calcareous soils, J Hydrol 2011;399:93-99](#)

1016 Legates DR, McCabe GJ. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic  
1017 model validation. Water Resour Res 1999;35:233-241.

1018 Leis A. Chemical, isotopic, and microbiological evidence for nitrification below the plant root zone from  
1019 intensive fertilized agricultural area in Austria – Insights from lysimeter studies and soil cores. In: IAEA-  
1020 TECDOC-1618. Application of Isotopes to the Assessment of Pollutant Behaviour in the Unsaturated Zone  
1021 for Groundwater Protection. Final report of a coordinated research project 2004-2005. International Atomic  
1022 Energy Agency, Vienna; 2009. p. 15-30.

1023 Mantovi P, Fumagalli L, Beretta GP, Guermandi M. Nitrate leaching through the unsaturated zone following pig  
1024 slurry applications, J Hydrol 2006; 316:195-212.

1025 Moreels E., De Neve S, Hoffman G, Van Meirvenne M. Simulating nitrate leaching in bare fallow soils: a model  
1026 comparison. Nutr Cycl Agroecosys 2003;67:137-144

1027 Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL. Model evaluation guidelines for  
1028 systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 2007;50:  
1029 885–900.

1030 Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour.  
1031 Res. 1976;12: 513-522.

1032 Nash JE, Sutcliffe JV. River flow forecasting through conceptual models, 1, A discussion of principles, J.  
1033 Hydrol., 1970;10:282-290.

1034 Nett L, Feller C, George E, Fink M. Effect of winter catch crops on nitrogen surplus in intensive vegetable crop  
1035 rotations. Nutr Cycl Agroecosys 2011;91:327-337.

1036 Nolan, BT, Puckett LJ, Ma L, Green TG, Bayless ER, Malone RW. Predicting unsaturated zone nitrogen mass  
1037 balances in agricultural settings of the United States. J Environ Qual 2010;39:1051-1065.

1038 Oenema O. Governmental policies and measures regulating nitrogen and phosphorus from animal manure in  
1039 European agriculture. J Anim Sci 2004;82: E196-E206.

1040 Patil N, Rajput G. Evaluation of Water Retention Functions and Computer Program “Rosetta” in Predicting Soil  
1041 Water Characteristics of Seasonally Impounded Shrink–Swell Soils, J Irrig Drain E-ASCE 2009;135, 286-  
1042 294. <http://ascelibrary.org/doi/abs/10.1061/%28ASCE%29IR.1943-4774.0000007>



1  
2  
3  
4  
5  
6  
7 1043 Perego A, Giussani A, Sanna M, Fumagalli M, Carozzi M, Alfieri L, Brenna S, Acutis M. The ARMOSA  
8 1044 simulation crop model: overall features, calibration and validation results. Italian Journal of  
9 1045 Agrometeorology (*in press*): 2013;3:23-38.

10  
11 1046 Qi Z, Ma L, Helmers MJ, Ahuja LR, Malone RW. Simulating nitrate-nitrogen concentration from a subsurface  
12 1047 drainage system in response to nitrogen application rates using RZWQM2. J Environ Qual 2012;41:289-295.

13 1048 Reusser DE, Blume T, Schaefli B, Zehe E. Analysing the temporal dynamics of model performance for  
14 1049 hydrological models. Hydrol Earth Syst Sc 2009;13: 999 – 1018. [http://www.hydrol-earth-syst-](http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf)  
15 1050 [sci.net/13/999/2009/hess-13-999-2009.pdf](http://www.hydrol-earth-syst-sci.net/13/999/2009/hess-13-999-2009.pdf)

16  
17 1051 Richards LA, Capillary conduction of liquids through porous mediums. Physics 1931;1: 318-333.

18 1052 Richter GM, Acutis M, Trevisiol P, Latiri K, Confalonieri R. Sensitivity analysis for a complex crop model  
19 1053 applied to Durum wheat in the Mediterranean. Europ J Agron 2010;32:127-132.

20  
21 1054 Ritter A, Hupet F, Muñoz-Carpena R, Lambot S, Vanclooster M. Using inverse methods for estimating soil  
22 1055 hydraulic properties from field data as an alternative to direct methods. Agr Water Manage 2003;59:77-96.

23 1056 Saxton KE, Rawls WJ. Soil water characteristic estimates by texture and organic matter for hydrologic solutions.  
24 1057 Soil Sci. Soc. Am. J.2006;70:1569-1578. doi:10.2136/sssaj2005.0117

25 1058 **SimLab, SimLab ver. 3.2.6. Development Framework for Uncertainty and Sensitivity Analysis. Joint Research**  
26 1059 **Centre of the European Commission, Econometrics and Applied Statistics, Ispra, Italy, 2009.**  
27 1060 <http://simlab.jrc.ec.europa.eu>

28  
29 1061 Smith P, Smith JU, Powelson DS, McGill WB, Arah JRM, Chertov OG, Coleman K, Franko U, Frolking S,  
30 1062 Jenkinson LS, Jenseng LS, Kellyh RH, Klein-Gunnewiek H., Komarov AS, Lif C, Molina JAE, Mueller T,  
31 1063 Parton WJ, Thornley JHM, Whitmore AP. A comparison of the performance of nine soil organic matter  
32 1064 models using datasets from seven long-term experiments. Geoderma 1997;81: 153-225.

33 1065 Sogbedji JM, Van Es HM, Melkonian JJ, Schindelbeck RR. Evaluation of the PNM model for simulating drain  
34 1066 flow nitrate-N concentration under manure-fertilized maize. Plant Soil 2006;282:343-360.

35  
36 1067 Sohier C, Degré A, Dautrebande S. From root zone modelling to regional forecasting of nitrate concentration in  
37 1068 recharge flows – The case of the Walloon Region (Belgium). J Hydrol 2009;369:350-359.

38 1069 Stenitzer E. SIMWASER – Ein numerisches Modell zur Simulation des Bodenwasserhaushaltes und des  
39 1070 Pflanzenertrages eines Standortes. *Mitt. der Bundesanstalt für Kulturtechnik und Bodenwasserhaushalt*, 31:  
40 1071 Petzenkirchen, 1988. p.1-118.

41  
42 1072 Stenitzer E, Diesel H, Franko U, Schwartengraber R, Zenker T. Performance of the model SIMWASER in two  
43 1073 contrasting case studies on soil water movement. In: K. Ch. Kersebaum et al. (eds.), *Modelling Water and*  
44 1074 *Nutrient Dynamics in Soil–Crop Systems*, 2007, Springer, pp 27–36.

45 1075 Stump C, Nützmann G, Maciejewski S, Maloszewski P. A comparative modeling study of a dual tracer  
46 1076 experiment in a large lysimeter under atmospheric conditions. J Hydrol 2009;375: 566-577.

47 1077 Svendsen H, Hansen S, Jensen HE. Simulation of crop production, water and nitrogen balances in two German  
48 1078 agro-ecosystems using the DAISY model. *Ecol Model* 1995;81: 197-212.

49  
50 1079 Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 2001;106:  
51 1080 No. D7, P. 7183. doi:10.1029/2000JD900719.

52 1081 Thorup-Kristensen, K. . Effect of deep and shallow root systems on the dynamics of soil inorganic N during 3-  
53 1082 year crop rotations. *Plant Soil*, 2006;288:233-248.

54  
55 1083 Van Dam JC,. Field-scale water flow and solute transport: SWAP model concepts, parameter estimation and  
56 1084 case studies. PhD thesis Wageningen University,, 2000.

1  
2  
3  
4  
5  
6  
7 1085 Van Dam JC, Groenendijk P, Hendriks RFA. Advances of Modeling Water Flow in Variably Saturated Soils  
8 1086 with SWAP. *Vadose Zone J* 2008;7:640-653.  
9  
10 1087 Van der Laan M, Miles, N, Annandale JG, Du Preez CC. Identification of opportunities for improved nitrogen  
11 1088 management in sugarcane cropping systems using the newly developed Canegro-N model. *Nutr Cycl*  
12 1089 *Agroecosys* 2011;90:391-404.  
13 1090 Van der Velde M, Bouraoui F, Aloe A. Pan-European regional-scale modelling of water and N efficiencies of  
14 1091 rapeseed cultivation for biodiesel production. *Glob Change Biol* 2009;15:24-37.  
15  
16 1092 Van Genuchten MTh. A closed form for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc.*  
17 1093 *Am. J.* 1980;44:892-898.  
18 1094 Van Liew MW, Garbrecht J. Hydrologic simulation of the little Washita river experimental watershed using  
19 1095 SWAT. *J Am Water Resour Ass* 2003;39:413-426.  
20  
21 1096 Vereecken, H., E.J. Jansen, M.J.D. Hack-ten Broeke, M. Swerts, M. Engelke, S. Fabrewitz and S. Hansen, 1991.  
22 1097 Comparison of simulation results of five nitrogen models using different datasets. In: Commission of  
23 1098 European Communities, editor, *Soil and Groundwater Research, Report II Nitrate in Soils*, Commission of the  
24 1099 European Communities, Luxembourg, pp 321 – 338.  
25 1100 Vereecken H, Weynants M, Javaux M, Pachepsky Y, Schaap MG, Van Genuchten MTh. Using Pedotransfer  
26 1101 Functions to Estimate the van Genuchten–Mualem Soil Hydraulic Properties: A Review *Vadose Zone J.*  
27 1102 2010;9: 795–820.  
28 1103 Vitousek PM, Naylor, Crews RT, David MB, Drinkwater LE, Holland E, Johnes PJ, Katzenberger J, Martinelli  
29 1104 LA, Matson PA, Nziguheba G, Ojima D, Palm CA, Robertson GP, Sanchez PA, Townsend AR, Zhang FS..  
30 1105 Nutrient imbalances in agricultural development. *Science* 2009; 324, no. 5934: 1519.  
31  
32 1106 Wang X, Mosley CT, Frankenberger JR, Klavdivko EJ. Subsurface drain flow and crop yield predictions for  
33 1107 different drain spacings using DRAINMOD. *Agr Water Manage* 2006;79:113-136.  
34 1108 Williams JR, Jones CA, Dyke PTL. A modeling approach to determining the relationship between erosion and  
35 1109 soil productivity. *Trans. ASAE* 1984;27:129-144.  
36  
37 1110 Williams JR, Jones CA, Kiniry JR, Spaul DA.. The EPIC crop growth model. *Trans. ASAE*, 1989;32:497–511.  
38 1111 Willmott CJ. Some comments on the evaluation of model performance. *Bull Am Meteo Soc* 1982; 63, 1309-  
39 1112 1313.  
40  
41 1113 Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. Statistics  
42 1114 for the evaluation and comparison of models. *J Geophys Res* 1985;90, 8995 – 9005.  
43 1115 [dx.doi.org/10.1029/JC090iC05p08995](https://doi.org/10.1029/JC090iC05p08995)  
44 1116 Wolff J, Beusen AHW, Groenendijk P, Kroon T, Rötter R, Van Zeijts H. The integrated modeling system  
45 1117 STONE for calculating nutrient emissions from agriculture in the Netherlands. *Environ Model Softw*  
46 1118 2003;18:597-617. [doi:10.1016/S1364-8152\(03\)00036-7](https://doi.org/10.1016/S1364-8152(03)00036-7)  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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 <sup>-1</sup> 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		

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 2. Annual precipitation rates (mm a<sup>-1</sup>) 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 <sup>-1</sup> )	883	839	892	893	1355	1013	739
Cumulative probability	44%	31%	48%	50%	98%	75%	10%

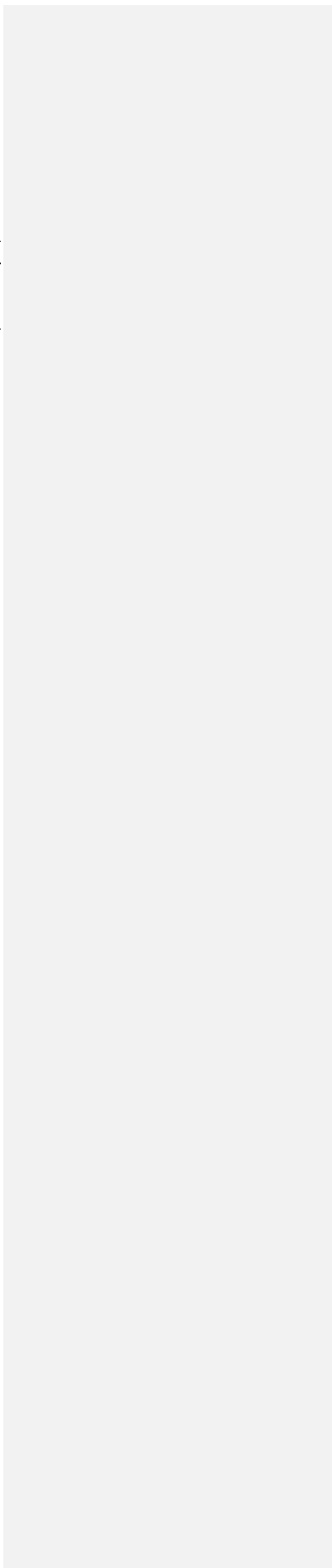


Table 3. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of volumetric water contents derived from calibrated soil moisture retention curves (Figure 2) and observed volumetric water contents at depths 0.35 m ( $n = 922$ ), 0.9 m ( $n = 1413$ ) and 1.8 m ( $n = 1456$ ) depth. **EPIC**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

Model	<i>MAE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>RMSE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>IoA</i>		
	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
<b>ARMOSA</b>	0.0064	0.0166	0.0308	0.0112	0.0176	0.0310	0.89	0.79	0.18
<b>COUP</b>	0.0341	0.0753	0.0391	0.0416	0.0775	0.0395	0.59	0.31	0.18
<b>DAISY</b>	0.0295	0.0340	0.0166	0.0326	0.0374	0.0178	0.63	0.62	0.38
<b>SIM-STO</b>	0.0212	0.0119	0.0064	0.0255	0.0130	0.0078	0.75	0.89	0.67
<b>SW-ANIM</b>	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	<i>MAE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>RMSE</i> (cm <sup>3</sup> cm <sup>-3</sup> )			<i>IoA</i>		
	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 ( $n = 1461$ )									
<b>ARMOSA</b>	0.0119	0.0247	0.0107	0.0168	0.0447	0.0123	0.79	0.75	0.46
<b>COUP</b>	0.0230	0.0104	0.0023	0.0288	0.0363	0.0031	0.74	0.84	0.85
<b>DAISY</b>	0.0956	0.0152	0.0105	0.1083	0.0630	0.0132	0.28	0.65	0.38
<b>EPIC</b>	0.0613	0.1563	0.0909	0.0662	0.0306	0.0925	0.49	0.90	0.07
<b>SIM-STO</b>	0.0180	0.0063	0.0028	0.0249	0.0271	0.0039	0.81	0.92	0.85
<b>SW-ANIM</b>	0.0101	0.0106	0.0072	0.0159	0.0285	0.0082	0.87	0.92	0.59
Validation 2009 – 2011 ( $n = 955$ )									
<b>ARMOSA</b>	x	0.0260	0.0130	x	0.0291	0.0149	x	0.52	0.47
<b>COUP</b>	x	0.0124	0.0030	x	0.0165	0.0041	x	0.74	0.84
<b>DAISY</b>	x	0.0152	0.0137	x	0.0193	0.0165	x	0.69	0.40
<b>EPIC</b>	x	0.1535	0.0924	x	0.1570	0.0939	x	0.19	0.09
<b>SIM-STO</b>	x	0.0093	0.0039	x	0.0134	0.0054	x	0.87	0.82
<b>SW-ANIM</b>	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			
	<i>MAE</i> (mm d <sup>-1</sup> )	<i>RMSE</i> (mm d <sup>-1</sup> )	<i>IoA</i>	<i>NSE</i>	<i>MAE</i> (mm d <sup>-1</sup> )	<i>RMSE</i> (mm d <sup>-1</sup> )	<i>IoA</i>	<i>NSE</i>
Calibration 2005 – 2008								
	<i>n</i> = 1461				<i>n</i> = 199			
<b>ARMOSA</b>	0.45	1.00	0.82	0.41	0.43	0.81	0.84	0.48
<b>COUP</b>	0.45	0.98	0.80	0.44	0.43	0.75	0.85	0.55
<b>DAISY</b>	0.57	1.16	0.68	0.21	0.54	0.90	0.74	0.35
<b>EPIC</b>	0.54	0.99	0.83	0.42	0.46	0.75	0.89	0.55
<b>SIM-STO</b>	0.34	0.87	0.86	0.55	0.30	0.62	0.91	0.69
<b>SW-ANIM</b>	0.38	0.91	0.86	0.51	0.37	0.72	0.88	0.58
Validation 2009 – 2011								
	<i>n</i> = 1084				<i>n</i> = 128			
<b>ARMOSA</b>	0.70	1.75	0.79	0.41	1.66	3.82	0.68	0.39
<b>COUP</b>	0.70	1.57	0.84	0.52	1.41	3.47	0.79	0.50
<b>DAISY</b>	0.73	1.77	0.77	0.39	1.74	4.34	0.56	0.21
<b>EPIC</b>	0.85	1.79	0.77	0.38	1.80	4.00	0.63	0.33
<b>SIM-STO</b>	0.51	1.43	0.90	0.61	1.69	3.94	0.76	0.35
<b>SW-ANIM</b>	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		
	<i>MAE</i> (mg L <sup>-1</sup> )	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i> (kg ha <sup>-1</sup> d <sup>-1</sup> )	<i>RMSE</i>	<i>IoA</i>
Calibration 2005 – 2008 ( <i>n</i> = 199)						
<b>ARMOSA</b>	15.71	20.37	0.78	0.043	0.085	0.77
<b>COUP</b>	6.74	9.60	0.97	0.041	0.085	0.78
<b>DAISY</b>	13.92	16.82	0.87	0.037	0.063	0.87
<b>EPIC</b>	19.55	25.63	0.76	0.049	0.084	0.82
<b>SIM-STO</b>	27.34	34.61	0.43	0.044	0.089	0.60
<b>SW-ANIM</b>	7.88	10.48	0.95	0.035	0.080	0.85
Validation 2009 – 2011 ( <i>n</i> = 128)						
<b>ARMOSA</b>	11.17	15.85	0.52	0.058	0.102	0.61
<b>COUP</b>	12.36	18.68	0.52	0.076	0.187	0.53
<b>DAISY</b>	8.54	11.40	0.78	0.045	0.095	0.54
<b>EPIC</b>	18.24	22.07	0.52	0.089	0.155	0.41
<b>SIM-STO</b>	8.88	10.44	0.78	0.058	0.138	0.56
<b>SW-ANIM</b>	19.97	29.37	0.43	0.205	0.800	0.12

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 7. Statistical parameters (*MAE*, *RMSE*, *IoA*) for the comparison of observed nitrate concentrations ( $\text{mg L}^{-1}$ ) 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; $n = 108$ )		
	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>	<i>MAE</i>	<i>RMSE</i>	<i>IoA</i>
<b>ARMOSA</b>	36.8	50.6	0.66	22.7	35.9	0.65	12.7	16.6	0.58
<b>COUP</b>	28.0	35.2	0.80	28.2	44.1	0.38	16.6	24.1	0.37
<b>DAISY</b>	32.2	43.9	0.68	29.1	50.9	0.46	12.9	21.5	0.55
<b>SIM-STO</b>	50.6	66.7	0.44	25.5	36.3	0.68	13.6	15.8	0.71
<b>SW-ANIM</b>	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 were preceded by catch crops (including leguminous crops).

Crop and period	Balance term <sup>†</sup> (kg ha <sup>-1</sup> )	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 <sub>3</sub> -N leaching <sup>§</sup> (-)	22.2	17.2	27.9	25.8	30.3	3.6	15.3
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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 <sub>3</sub> -N leaching <sup>§</sup> (-)	25.7	27.9	25.8	22.7	33.6	25.1	29.7
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-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 <sub>3</sub> -N leaching <sup>§</sup> (-)	5.9	4.4	7.1	6.3	5.4	8.8	5.8
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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 <sub>3</sub> -N leaching <sup>§</sup> (-)	18.9	13.5	18.5	11.7	12.3	5.7	22.2
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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 <sub>3</sub> -N leaching <sup>§</sup> (-)	33.1	44.2	61.5	26.4	16.0	32.5	72.1
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-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 <sub>3</sub> -N leaching <sup>§</sup> (-)	3.6	17.0	14.6	8.7	19.3	13.1	32.9



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		-0.2	72.7	-104.9	-39.7	58.3	102.8
	<hr/>							
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	Biological fixation (+)		0.0	12.8	0.0	0.0	0.0	18.2
-								
13.7.2011	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 <sub>3</sub> -N leaching <sup>§</sup> (-)	13.9	6.1	3.2	7.6	30.3	13.3	31.0
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		20.7	146.5	11.2	-128.8	-10.4	75.5
	<hr/>							
	Seven year totals 2005– 2011							
	<hr/>							
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	Biological fixation (+)		174.9	154.6	0.1	1.8	31.3	370.9
-								
13.7.2011	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 <sub>3</sub> -N leaching <sup>§</sup> (-)	123.3	130.3	158.6	109.1	147.3	102.2	209.0
	Other transport <sup>§</sup> (-)		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 <sup>#</sup>		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<sub>4</sub>-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	0.35 m	+	-	o	n.a.	o	+
		retention	0.9 m	o	-	o	n.a.	+	+
		relation	1.8 m	-	-	-	n.a.	o	+
		Simulated	0.35 m	o	o	-	-	+	+
		water	0.9 m	o	+	o	+	+	+
		contents	1.8 m	-	+	-	-	+	-
		Nitrate	0.9 m	o	+	o	n.a.	-	+
		concentration							
		Water flux, daily		o	o	-	o	o	o
		Water volumes per		o	o	-	o	+	o
Validation	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Nitrate concentration		o	+	o	o	-	+
		in water samples							
		Nitrate-N flux per		o	o	o	o	-	o
		sampling interval							
		Soil water	0.9 m	-	o	o	-	+	o
		contents	1.8 m	-	+	-	-	+	o
		Nitrate	0.35 m	o	-	-	n.a.	o	-
			0.9 m	-	-	-	n.a.	o	-
		Water flux, daily		o	o	o	o	+	o
		Water volume per		-	o	-	-	o	-
Validation	-: $IoA < 0.75$ o: $0.75 \leq IoA < 0.9$ +: $IoA \geq 0.9$	Nitrate concentration		-	-	o	-	-	
		in water samples							
		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<sub>7</sub>*) and for the best five seasons (*MAE<sub>5</sub>*).

Seasonal quantity	Indicators	ARMOSA	COUP	DAISY	EPIC	SIM-STO	SW-ANIM
Percolated water (mm)	<i>MAE<sub>7</sub></i>	21.3	24.2	63.9	48.6	14.6	40.3
	<i>MAE<sub>5</sub></i>	16.0	14.3	30.5	30.5	11.8	32.8
N crop off-take (kg ha <sup>-1</sup> )	<i>MAE<sub>7</sub></i>	36.5	32.7	47.7	31.0	33.0	21.5
	<i>MAE<sub>5</sub></i>	23.1	14.3	29.0	20.6	20.5	10.3
Leached NO <sub>3</sub> -N (kg ha <sup>-1</sup> )	<i>MAE<sub>7</sub></i>	6.6	8.2	4.6	10.3	6.6	14.2
	<i>MAE<sub>5</sub></i>	4.4	3.6	3.7	7.8	2.8	6.3

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## 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**

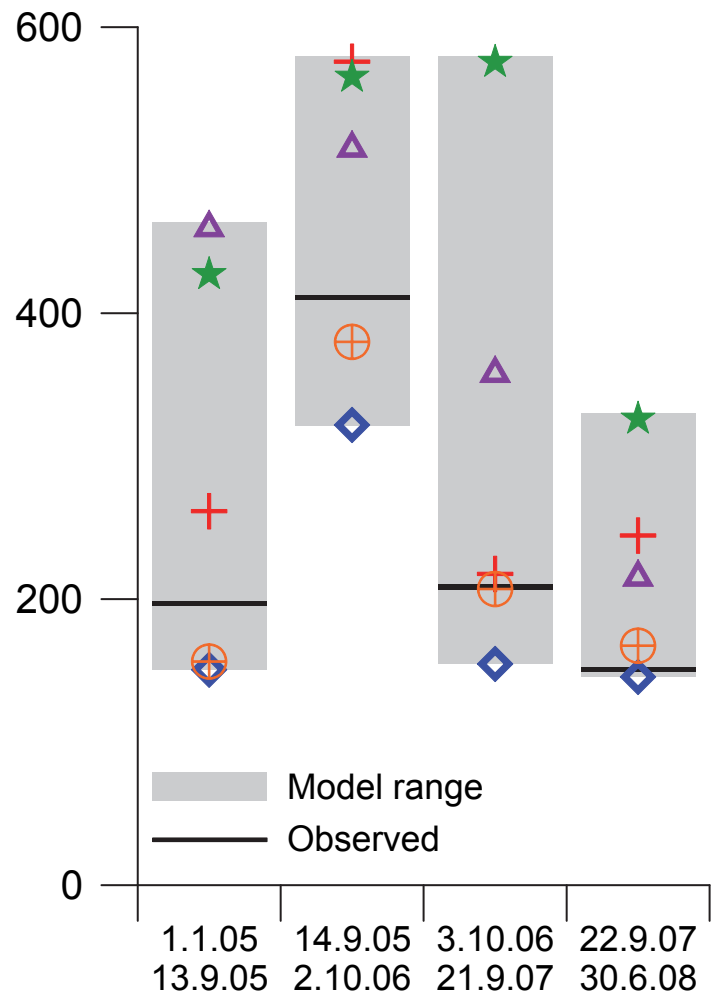
Figure 7 Seven years balances for fertilization minus crop off-take and nitrate-N leaching (all in  $\text{kg ha}^{-1}$ ), summed since the start of the calibration period

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).

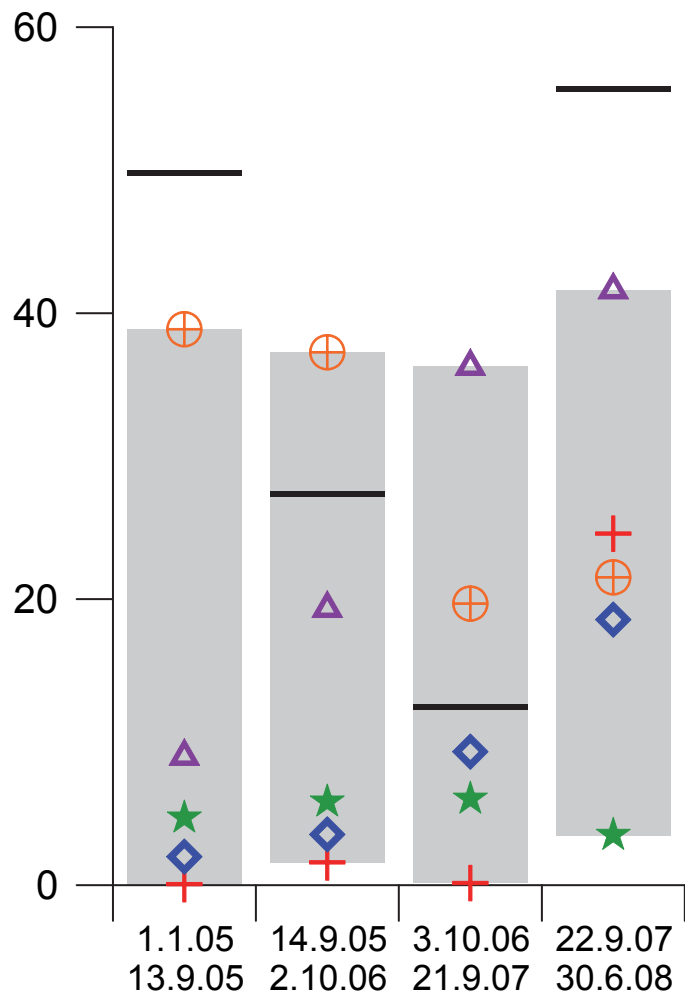
Figure 1

[Click here to download Figure: Fig\\_01\\_STOTEN-D-14-00814.pdf](#)

Seasonal percolation (mm)



Flow averaged NO<sub>3</sub> concentration (mg L<sup>-1</sup>)



Seasonal NO<sub>3</sub>-N leaching (kg ha<sup>-1</sup>)

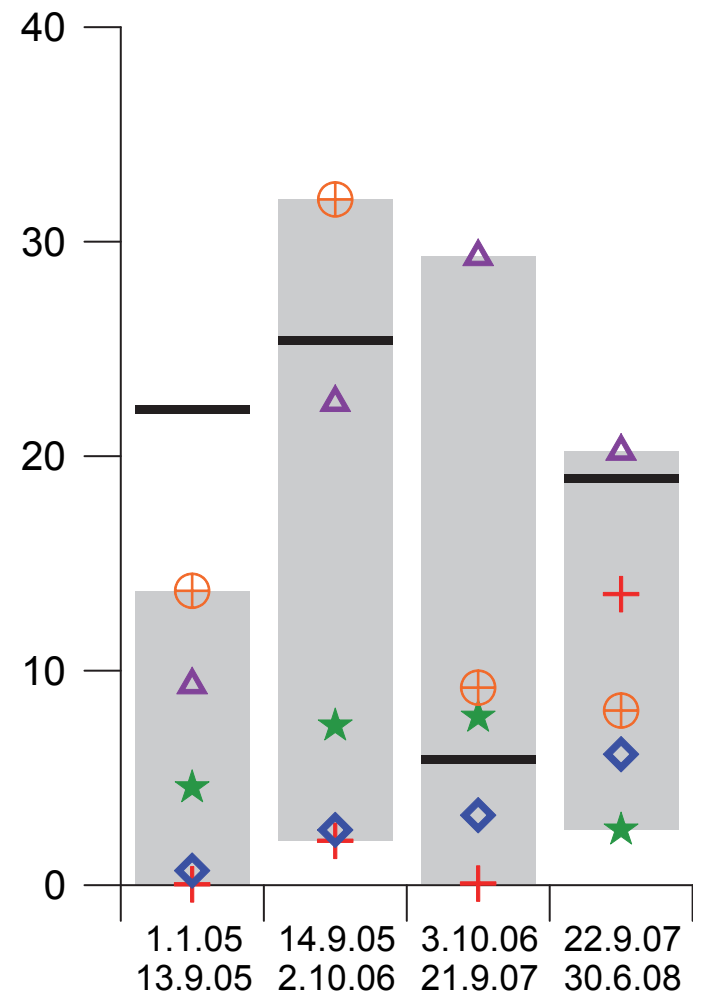


Figure 2

[Click here to download Figure: Fig\\_02\\_STOTEN-D-14-00814.pdf](#)

Absolute pressure head (cm)

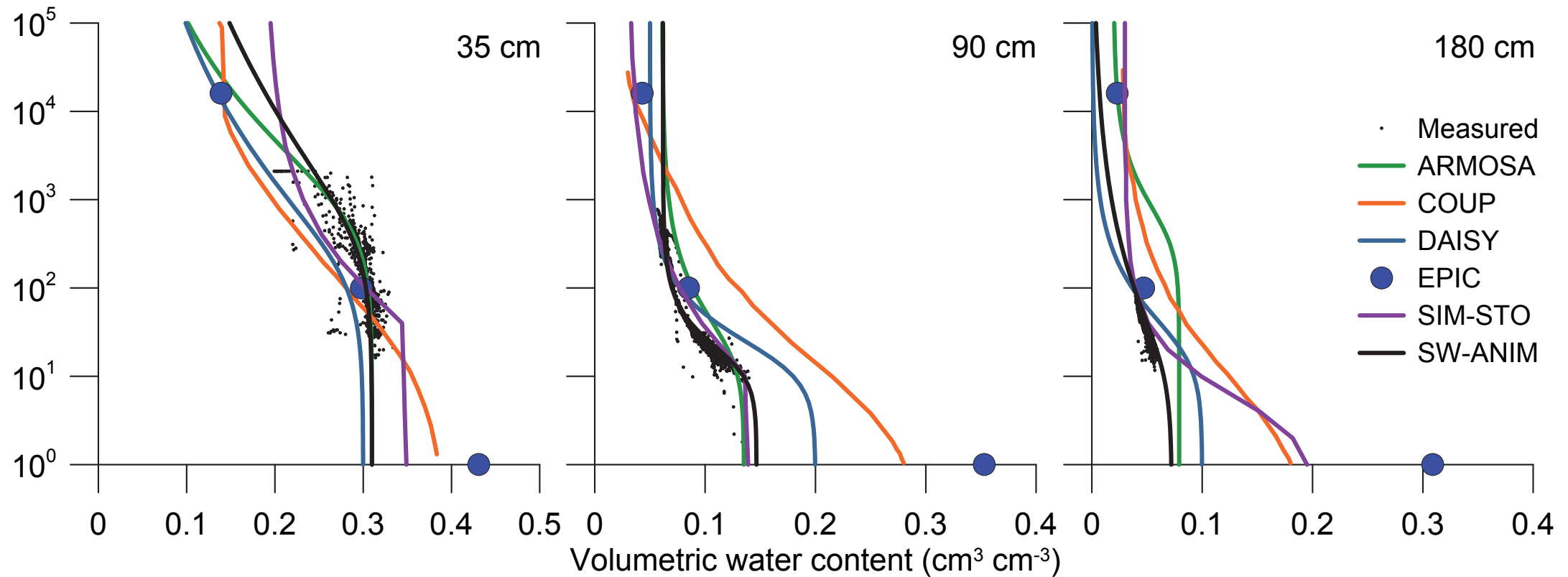


Figure 3

[Click here to download Figure: Fig\\_03\\_STOTEN.D.14.00814.pdf](#)

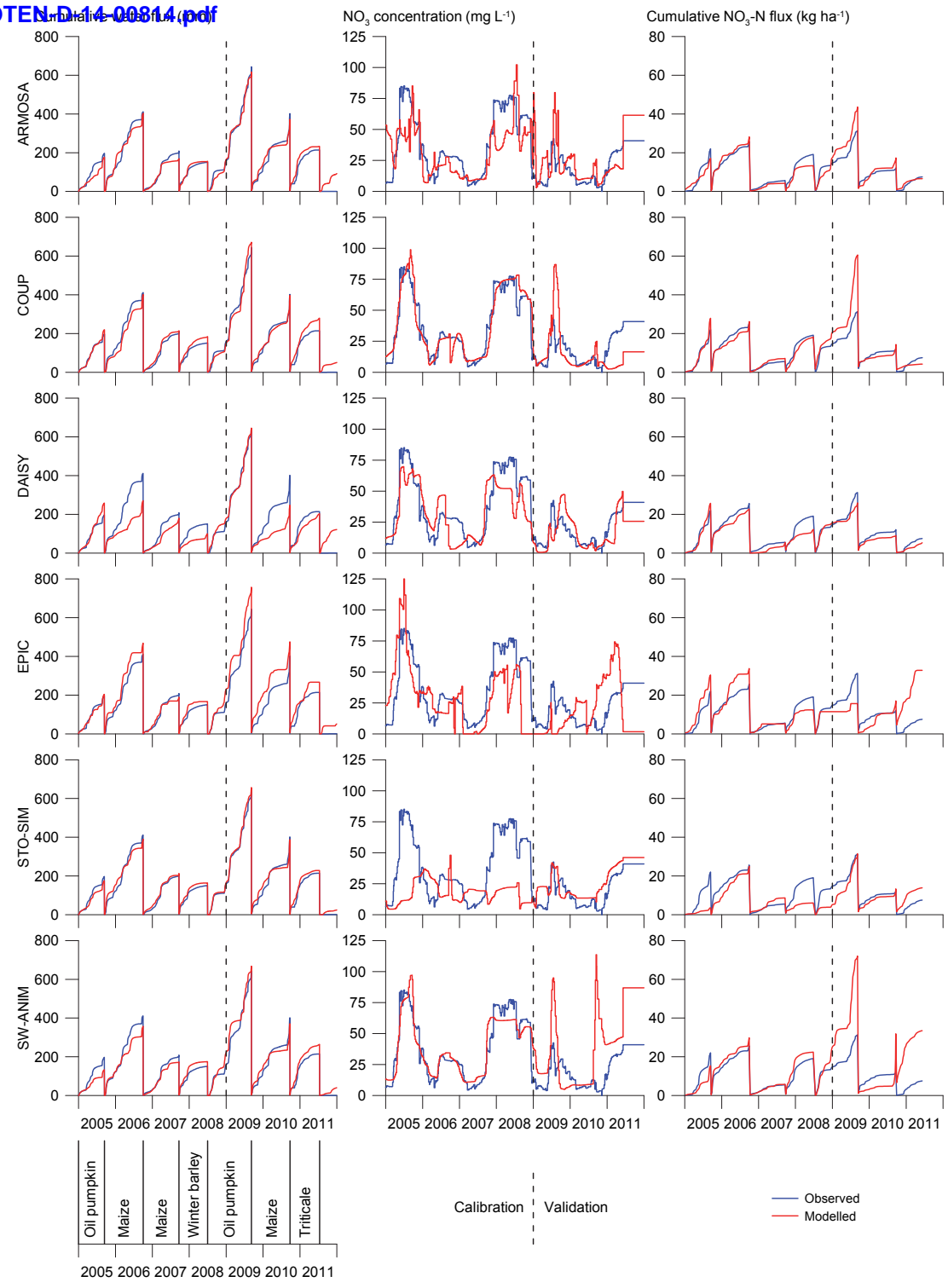


Figure 4  
[Click here to download Figure: Fig\\_04\\_STOTEN-D-14-00814.pdf](#)

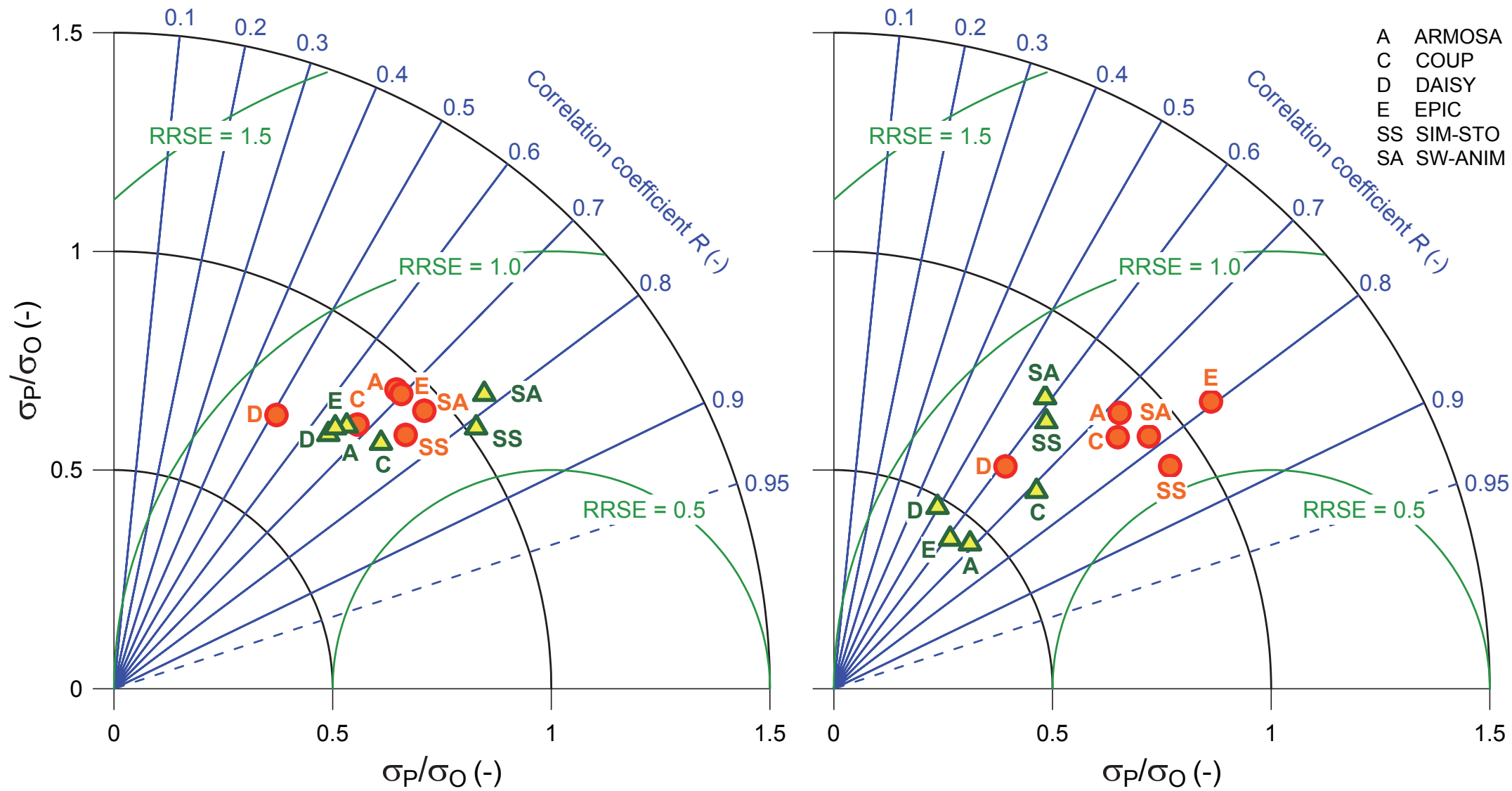


Figure 5

[Click here to download Figure: Fig\\_05\\_STOTEN-D-14-00814.pdf](#)

# Seasonal water flux at 180 cm depth (mm)

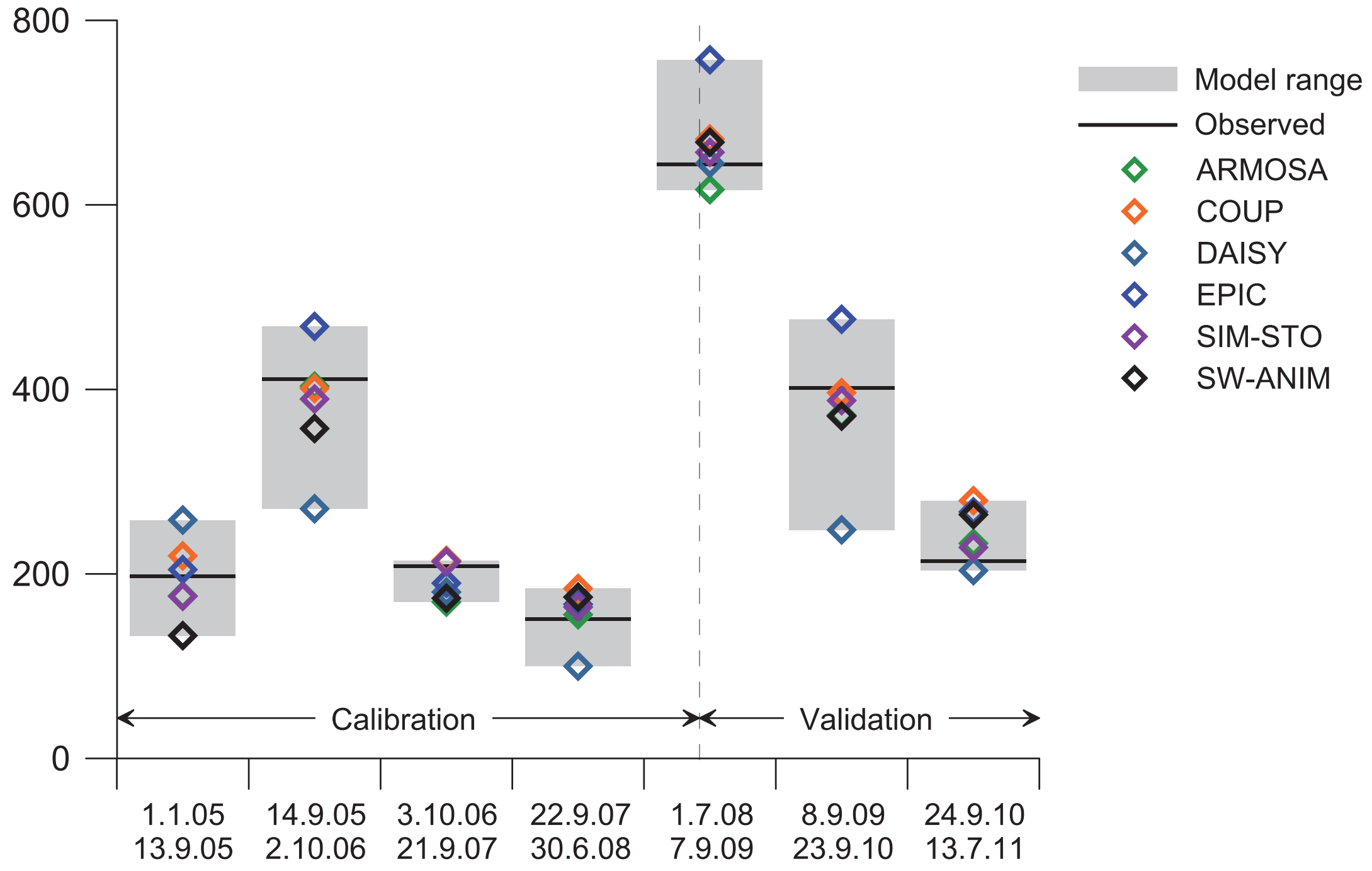




Figure 6

[Click here to download Figure: Fig\\_06\\_STOTEN-D-14-00814.pdf](#)

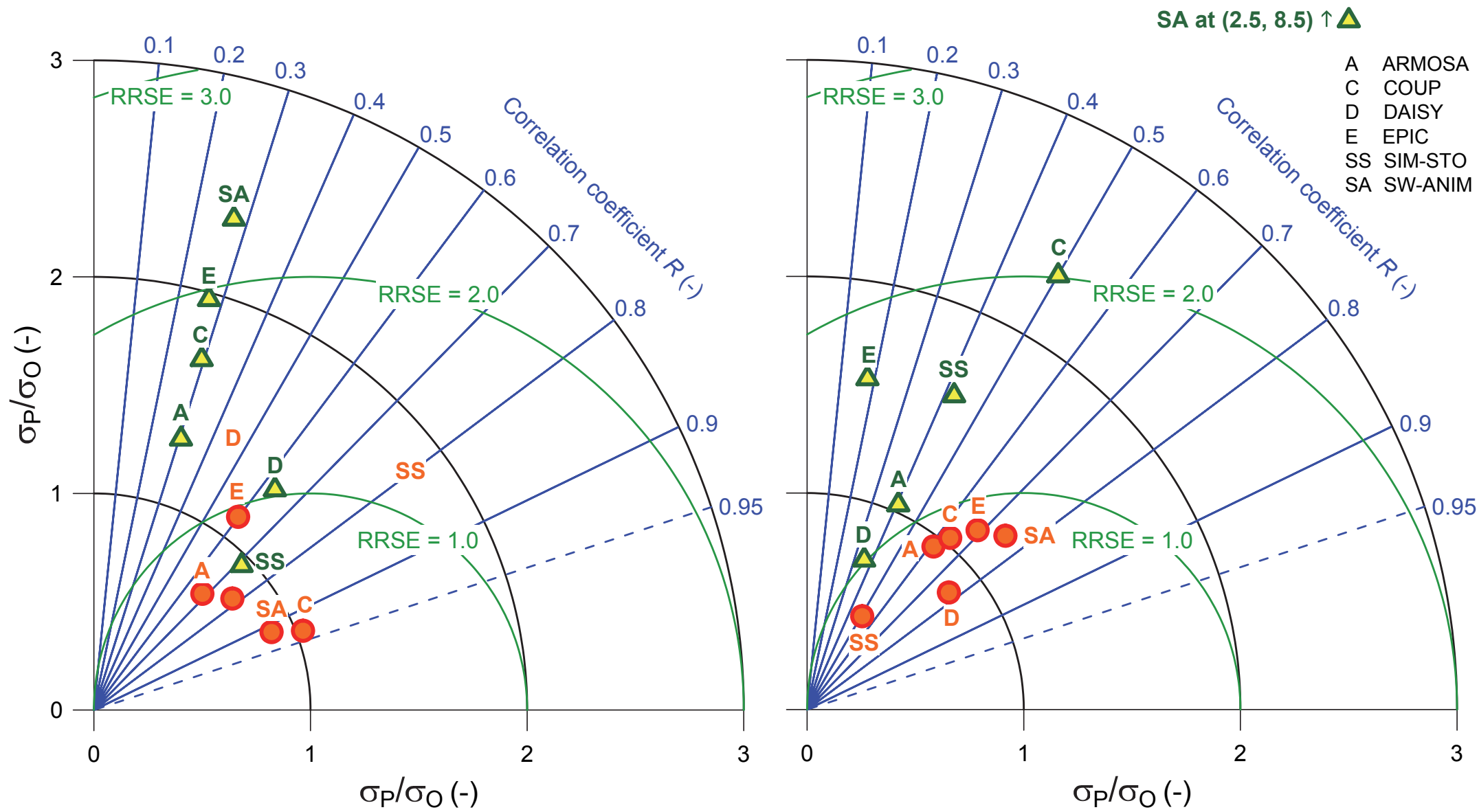


Figure 7

[Click here to download Figure: Fig\\_07\\_STOTEN-D-14-00814.pdf](#)

Seven years summed balance (kg ha<sup>-1</sup>)

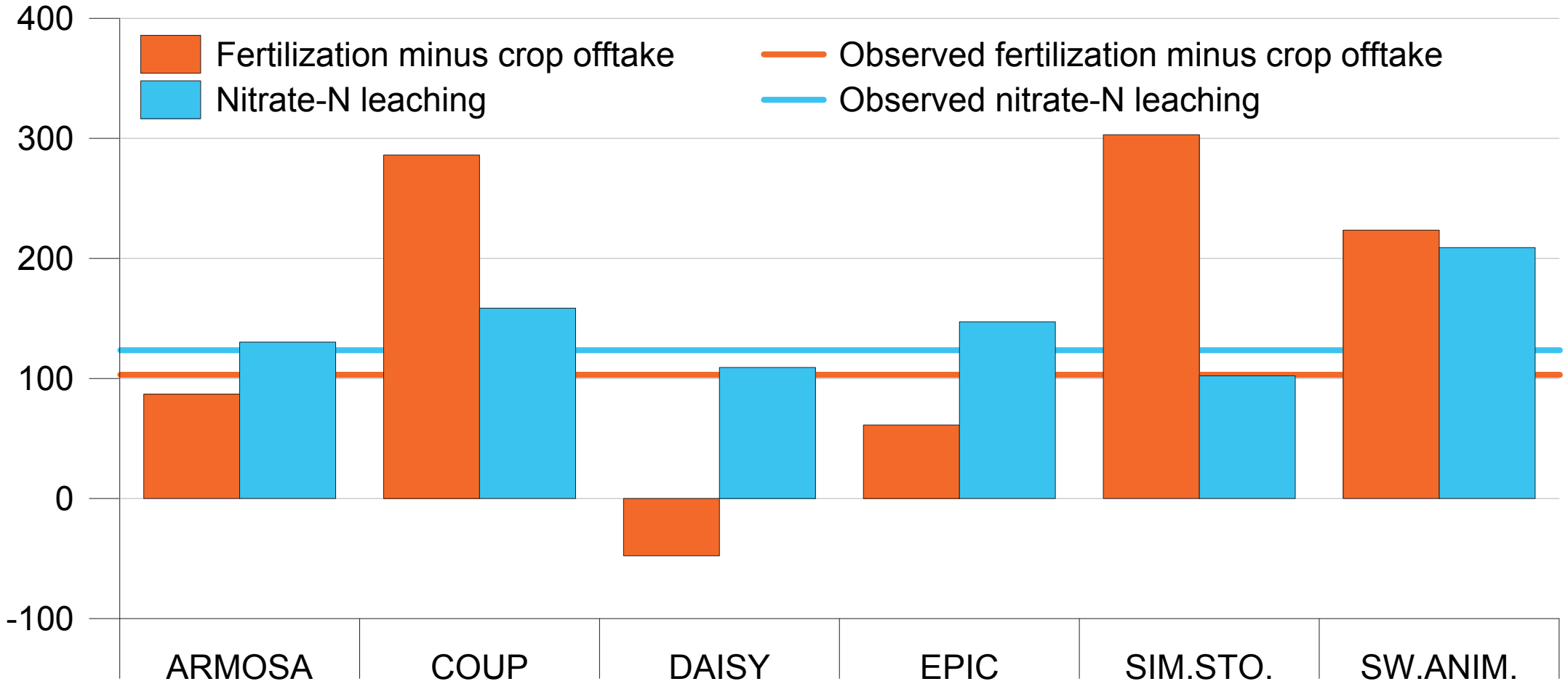
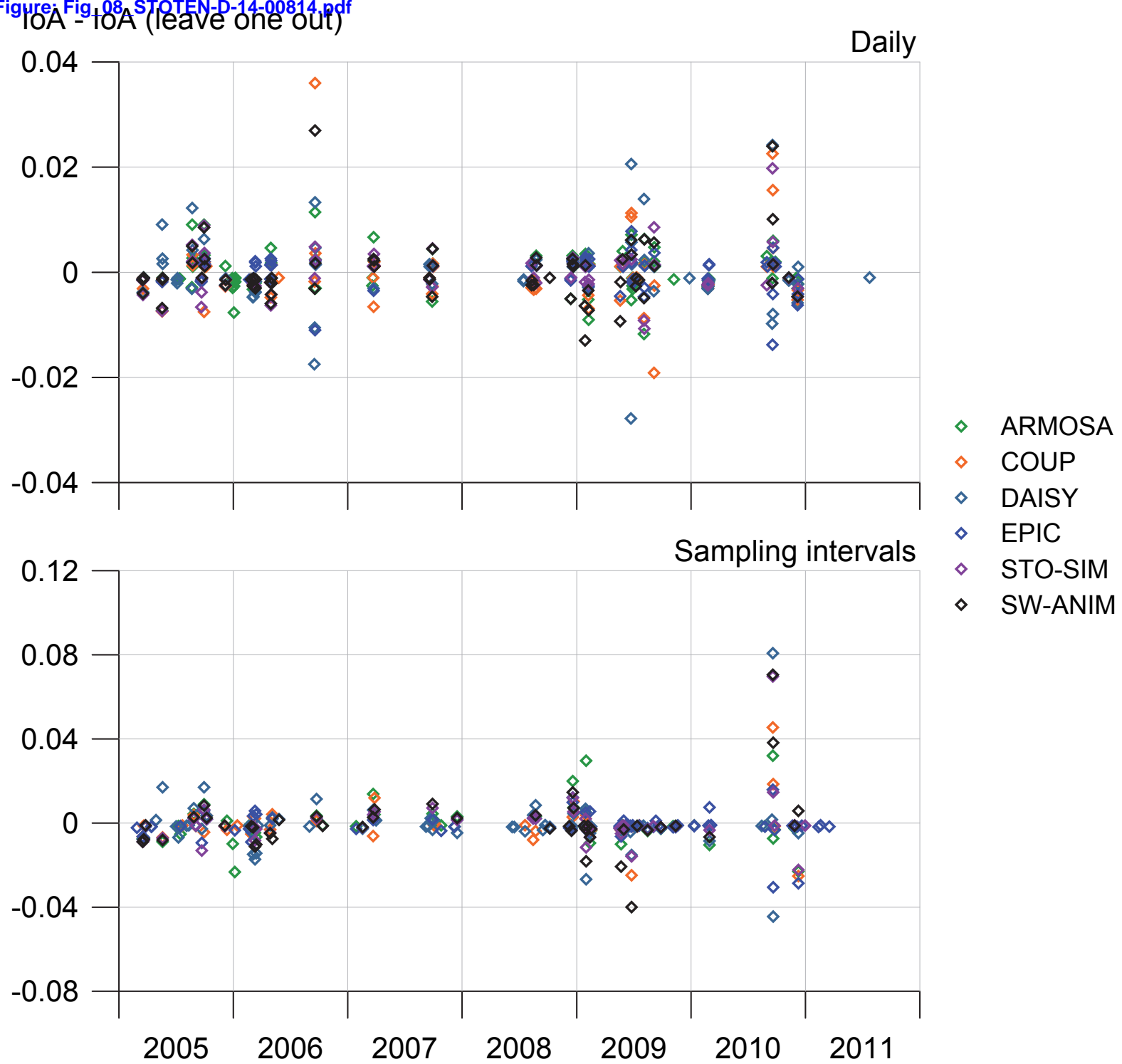


Figure 8

[Click here to download Figure: Fig\\_08\\_STOTEN-D-14-00814.pdf](#)



**Supplementary material for on-line publication only**

[Click here to download Supplementary material for on-line publication only: STOTEN-D-14-00814\\_Supplemental\\_Materials.docx](#)