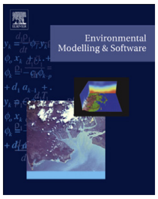
Environmental Modelling & Software 93 (2017) 255–267

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft



The impact of the objective function in multi-site and multi-variable calibration of the SWAT model

Eugenio Molina-Navarro*, Hans E. Andersen, Anders Nielsen, Hans Thodsen, Dennis Trolle

Department of Bioscience, Aarhus University, Silkeborg, Denmark

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form
19 December 2016
Accepted 22 March 2017
Available online 31 March 2017

Keywords:
Catchment model
Objective function
Multi-site calibration
Nutrients
SWAT
Calibration scheme

ABSTRACT

Automatic calibration of complex hydro-ecological models is an increasingly important issue which involves making decisions. One of the most relevant is the choice of the objective function, but its effects have been scarcely studied in complex models. We have used the SWAT model to assess the impact of the objective function for a multi-site (4 stations) and multi-variable (OrgP, OrgN, NO_3^-, PO_4^{3-}) calibration of the Odense catchment (Denmark). Six calibration schemes were tested, varying the objective function and the nutrient fractions targeted. The best performance metrics (R^2, NSE, PBIAS) were obtained when using NSE as objective function and targeting N-fractions and P-fractions separately. The scheme was validated in another SWAT set-up in northern Denmark. Although NSE is often questioned, we found it as an adequate objective function when addressing a multi-site and multi-variable calibration. Our findings may serve as guideline for hydro-ecological modellers, being useful to achieve watershed management goals.

Software and data availability

The Soil and Water Assessment Tool (SWAT) is a public domain model jointly developed by USDA Agricultural Research Service (USDA-ARS) and Texas A&M University. Source code, GUI, executables, documentation and test cases can be downloaded from http://swat.tamu.edu/. SWAT-CUP (SWAT Calibration and Uncertainty Programs, 270 Mb) is also a public domain program, developed by the Swiss Federal Institute of Aquatic Science and Technology (programming and maintenance by Neprash Technology Ltd). It is also available for download at the SWAT website. Exchange of ideas, reports of bugs and user support for both tools are driven by the user and developer community through Google sites, which are available through the website (http://swat.tamu.edu/support). Daily water discharges and nutrient concentrations were acquired from the Danish national environmental surface water database (Naturstyrelsen and Nationalt Center for Miljø og Energi, 2016).

1. Introduction

Hydro-ecological models have become very popular tools for assessing and managing water resources and nutrient transports at a catchment scale in the last decade. In recent years, improvements in integrated hydrological modelling, advances in calibration tools and availability of technology allow building more detailed and holistic models (Rouholahnejad et al., 2014). Their ability to simulate future scenarios and predict the changes in water availability and quality make them applicable and essential in development of water management and aquatic ecosystems policies.

However, setting up a reliable hydro-ecological model is not an easy task and a credible representation of the natural system presents a scientific challenge that requires a number of critical decisions (Ahmadi et al., 2014; Daggupati et al., 2015). Calibration is a key step in achieving a representative and reliable model. At a catchment scale, the calibration process can be addressed at different levels of complexity. Considering the calibration sites, it can go from a single-outlet calibration to a spatially differentiated (multi-site) calibration, which takes into account data from several locations. The latter approach has been recently recommended by several authors, since spatial variations within the catchment are likely to be better captured (e.g. Daggupati et al., 2015; Niraula et al., 2015; Zhang et al., 2008). Calibration complexity can vary also according to the variables subject for calibration: just one variable (e.g.

* Corresponding author. Department of Bioscience, Aarhus University, Vejlsøvej 25, 8600 Silkeborg, Denmark.
E-mail addresses: emna@bios.au.dk, eugenio.molinanavarro@gmail.com (E. Molina-Navarro).

http://dx.doi.org/10.1016/j.envsoft.2017.03.018
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

water flow) or a multi-variable calibration (e.g. different nutrients fractions). While focusing on a specific variable might yield a more accurate calibration result, often modelling applications require the simulation of several variables (e.g. nutrient fractions), and the calibration of multiple variables at the same time may be more convenient in terms of time effort needed. At the same time, this reduces the overall amount of parameters in need of calibration. Having spatially distributed data of different variables is becoming more frequent, so addressing multi-site and multi-variable calibrations is an increasingly important issue (White and Chaubey, 2005; Zhang et al., 2008). Nevertheless, a paradox arises here: the higher the number of calibration sites, the better reality is represented, but the more difficult it is to achieve an acceptable model calibration (e.g. in statistical terms) since more constraints are introduced in the calibration process (White and Chaubey, 2005).

During the calibration process, different values are iteratively assigned to several model parameters (preferably inside a realistic range) until the representation given by the model is considered acceptably accurate. Calibration can be done manually, but the use of automatic procedures have become frequent within hydrological sciences as computational power has become readily available, thus to some extent standardizing the calibration approach, and also accelerating the overall calibration process. Several decisions and assumptions have to be made when employing an automatic calibration procedure and the choice of objective function, being the performance metric that drives the automatic parameterization, is crucial here. In fact, the selection of objective function conditions the final values of the calibrated parameters (Abbaspour, 2015; Muleta, 2012), since different objective functions rely on different aspects of the variable targeted for calibration: some objective functions rely more on average deviations (e.g. the mean absolute error) while others tend to rely on the peaks (e.g. Nash-Sutcliffe efficiency, (Nash and Sutcliffe, 1970)). The objective function can also be scale-dependent (e.g. the summation form of the squared error) or can be ruled by the relative deviation (e.g. the percent-bias). Thus, the choice of the objective function is an important decision when setting-up an automatic calibration scheme for a catchment model, especially when several sites and/or different variables are being calibrated. However, identifying an objective function that meets all the requirements may be daunting (Muleta, 2012). Even though the final outcome of a model can be apparently good in statistical terms for one particular objective function, it can be very far from reality at the same time when doing an inappropriate choice of the objective function (Abbaspour, 2015; Muleta, 2012). Things become more complex in those models that additionally allow assigning different weights to different variables and/or sites in the calculation of the objective function during the calibration of catchment models with data from different variables (e.g. different nutrient fractions) and locations (multi-site).

The importance of the selection of different parameter calibration schemes and its ultimate implications for water resources management has been reported by many authors (e.g. Muleta, 2012; Zhang et al., 2008). However, studies assessing the impact of the choice of objective function in the final calibration results are relatively scarce and have mostly focused on hydrological calibrations alone. Already in 1991, (Servat and Dezetter, 1991) tested five objective functions, using three different lumped rainfall-runoff models and four catchments in Ivory Coast. Muleta (2012) examined the sensitivity of model performance to nine widely used objective functions during a single-output flow automated calibration (one variable). Other authors have evaluated the efficiency of single-objective versus multi-objective calibration methods in multi-site set-ups (Ahmadi et al., 2014; Zhang et al., 2008). However, assessment of the relevance of the objective function choice and calculation scheme in multi-site and multi-variable calibrations has to our best knowledge not been studied yet.

In this study, we used the Soil and Water Assessment Tool model (SWAT, Arnold et al., 2014). SWAT is one of the most widely used hydro-ecological catchment simulation models and it has been applied worldwide for hydrologic and water quality simulations. More than 2400 peer-reviewed articles using SWAT have been published since 1993 with 900 of them being published since 2014 (CARD, 2016), which reveals its increasing applicability and scientific acceptance under many different circumstances.

The main goal of this paper is to investigate the impact of the objective function (its choice and its calculation according to the weights assigned to different variables) when calibrating a multi-site and multi-variable SWAT catchment model (four gauges and four nutrients). Automatic calibration was performed with SWAT-CUP (SWAT Calibration and Uncertainty Programs) (Abbaspour, 2015), which allows the use of various types of objective functions, and also parameterization approaches. The assessment was based in a SWAT set-up for the Odense Fjord catchment (Fyn Island, Denmark) (Thodsen et al., 2015) and validated in a new and more complex set-up in northern Denmark.

Our findings may reveal new insights about the relevance of the objective function in complex model applications and are expected to serve as a guideline for hydro-ecological modellers (SWAT or other distributed models) when addressing automatic calibration procedures.

2. Material and methods

2.1. Study area

The Odense catchment is the major catchment in the island of Funen in Denmark (Fig. 1). It has an area of 1061 km^2 and flows into an estuary, the Odense Fjord. The climate is temperate with an annual mean temperature around 9 °C, being the warmest in July (17 °C average) and coldest in January (1 °C average). The mean annual precipitation is around 800 mm with no pronounced seasonality.

The altitude ranges from −8 to 125 m.a.s.l. and half of the territory has slopes below 2%. The catchment is settled over clayey moraines from the last glaciation (Weichsel glaciation) (Smed, 1982) and sandy-loam soils dominate the catchment.

Land use in the catchment is dominated by agriculture (68%), followed by urban areas (16%) and forest (10%). Odense, with 174,000 inhabitants, is the main city in the catchment. The dominant crops are winter wheat (42%), spring barley (21%) and oil seed rape (14%) (Thodsen et al., 2015). Several lakes are situated within the catchment, with Lake Arreskov, located in the upper part, being the largest (>3 km^2). The Odense Fjord estuary, with a surface of 62 km^2, is an important aquatic ecosystem in the catchment. Its ecological status is conditioned by the nutrient loads, which mainly originate from diffuse sources, i.e. agricultural areas. Thus, a reliable calibration of nutrient fractions for the applied catchment model is consequently an important step in achieving a valuable tool for aiding the decision making process of environmental managers.

The impact assessment of the objective function was validated in a new SWAT set-up covering 10,857 km^2 in northern Denmark (Fig. 1). Average precipitation and temperature are around 910 mm and 9 °C, respectively, and the altitude ranges from −8 to 131 m.a.s.l. Similarly to the Odense catchment, agriculture is the main land use in the catchment (65%), followed by forest and scrubland (17%) and urban areas (8%).

2.2. Model set-up and flow calibration

The Soil and Water Assessment Tool (SWAT) model, version 582,

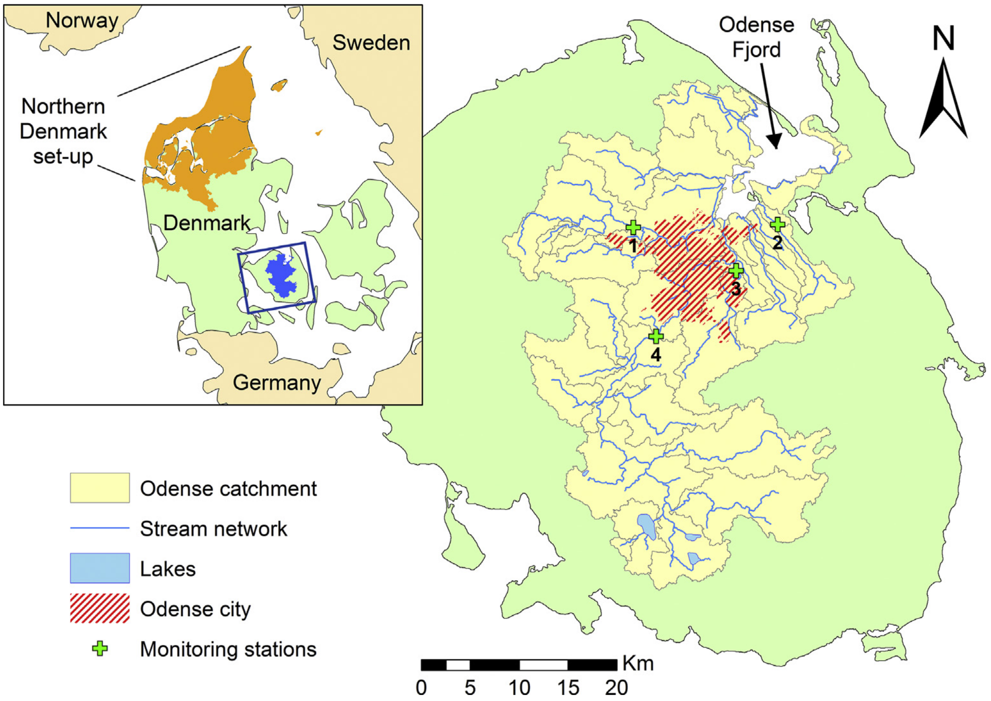


Fig. 1. Location of the Odense Fjord catchment, SWAT delineated sub-basin division and location of monitoring stations for calibration and validation. Extension of the SWAT set-up in northern Denmark used for validation is also depicted.

has been used in this study. SWAT is a physically based, basin-scale, continuous time, semi-distributed hydrologic model that uses spatially derived data on topography, land use, soil, and weather for hydrological and water quality modelling and operates on a daily time step. Based on the topography, SWAT divides the catchment into sub-basins and those are further divided into hydrological response units (HRUs), which are the computational units in the model and comprise a unique combination of land use, soil type and slope class. Nutrients are modelled by SWAT in the soil profile and in the shallow aquifer. They are introduced into the main channel through surface runoff, groundwater and lateral subsurface flow, and transported downstream with channel flow. In the shallow aquifer, nitrogen enters in recharge from the soil profile and SWAT allows for fluctuations in loadings over time, while the concentration of soluble phosphorus in the shallow aquifer may be specified. Detailed information about SWAT and nutrients cycling in the model can be found in (Arnold et al., 2014; Neitsch et al., 2011).

The Odense Fjord catchment model was previously set-up in (Thodsen et al., 2015) and a detailed description can be found there. In short, the catchment was divided into 31 sub-basins (Fig. 1) and 11 soil types, 6 land uses and 3 slope classes were considered. The agricultural land use was split in 14 different five-year crop rotation schemes, each with its corresponding management operations. It resulted in the creation of 3410 HRUs, which ensures a detailed representation of the reality of the catchment. Tile drainage was enabled in SWAT for agricultural areas with soils with a clay content over 8% and a slope lower than 2% (Thodsen et al., 2015), representing 39% of the catchment.

Four monitoring stations were used for calibration and validation (Fig. 1, Table 1). Relative to the study by (Thodsen et al., 2015), we updated the model to ArcSWAT 2012 rev. 582 and set-up a new calibration procedure. First, we added new parameters at sub-basin level thus allowing the relationships between streamflow and topography, land use, and precipitation to vary between sub-basins (Zhang et al., 2008). Second, we expanded the model set up to account for mineral and organic P exports. The Sequential Uncertainty Fitting Algorithm (SUFI2) (Abbaspour, 2015) was used to calibrate discharge and nutrients in respective order. SUFI2 is capable of analyzing a large number of parameters and measured data of different model output variables from various sites simultaneously, allowing the user to assign different weights for the different variables and/or stations calibrated when calculating the corresponding value of the selected objective function. SUFI2 is linked to SWAT in the SWAT-CUP software, and it has been recently seen to provide more reasonable and balanced predictive results than the other uncertainty analyses methods in SWAT-CUP (Wu and Chen, 2015). The latest SWAT-CUP version (5.1.6) has been used in this work.

Table 1
Monitoring stations used for model calibration and validation.

Monitoring station	1	2	3	4
Station code	450005	450058	450043	450003
SWAT sub-basin	14	20	21	22
Area upstream (Km^2)	80	26	62	487

Calibration was performed on a daily time step from 1 Jan. 2000 to 31 Dec. 2005, with a previous 10-year warm-up period to allow parameters to reach equilibrium after potential initialization biases. This relatively long warm-up period was chosen considering the subsequent use of the model for nutrient loads simulation, given that soil nutrient pools can take long time to stabilize. A set of sensitive parameters was selected based on expert judgement and assigned initial calibration value ranges based on previous SWAT applications in Denmark. For streamflow, five parameters at a basin-wide level and 17 at a sub-basin level were calibrated (Table 2). To compare simulated and observed stream flow data, the sum of squared errors was used as objective function to ensure that the model will rely more on the average values than on the high peaks. 3000 simulations were run through three stepwise iterations of 1000 simulations. SUFI2 creates a combination of values of

Table 2
Initial ranges and calibrated values (only discharge) for the selected basin-wide and sub-basin level parameters during discharge and nutrients calibration. The ranges of relative change from the initial value (IV) are showed in those parameters which values are changed through multiplication. For nutrient parameters, the nutrient fractions that are mainly expected to be influenced are indicated.

Discharge parameters

Parameter	Level	Initial range (relative change)	Calibrated values
SFTMP.bsn	Basin-wide	−1–1	0.94
SMFMN.bsn	Basin-wide	−1–2	1.12
SMFMX.bsn	Basin-wide	1.6–3.5	2.57
SMTMP.bsn	Basin-wide	−2.3–1	−0.14
SURLAG.bsn	Basin-wide	1–10	5.23
ALPHA_BF.gw	Sub-basin	0–1.011	0.07–0.97
ALPHA_BNK.rte	Sub-basin	0–1.1	0.003–0.94
CH_K2.rte	Sub-basin	0–75	9.68–74.31
CN2.mgt	Sub-basin	17.5–119.6 (IV·0.7 - IV·1.3)	17.5–67.16
DDRAIN.mgt	Sub-basin	770 - 1430 (IV·0.7 - IV·1.3)	792–1331
EPCO.hru	Sub-basin	0.01–1.103	0.11–0.66
ESCO.hru	Sub-basin	0–1	0.31–0.98
GDRAIN.mgt	Sub-basin	1.4–2.6 (IV·0.7 - IV·1.3)	1.72–2.26
GWQMN.gw	Sub-basin	0–2000	133.67–1735.02
GW_DELAY.gw	Sub-basin	0–600	28.19–339.88
GW_REVAP.gw	Sub-basin	0–0.2	0.06–0.13
OV_N.hru	Sub-basin	0.008–0.36 (IV·0.8 - IV·1.2)	0.008–0.32
REVAPMN.gw	Sub-basin	0–2000	368.71–1896.40
SOL_AWC.sol	Sub-basin	0.03–0.43 (IV·0.2 - IV·1.8)	0.08–0.36
SOL_BD.sol	Sub-basin	0.56–2.04 (IV·0.8 - IV·1.2)	0.62–2.02
SOL_K.sol	Sub-basin	1.28–129.03 (IV·0.2 - IV·3)	9.31–83.53
TDRAIN.mgt	Sub-basin	33.6–62.4 (IV·0.7 - IV·1.3)	46.56–60.48

Nutrient parameters

Parameter	Level	Initial range	Fraction(s) influenced
CMN.bsn	Basin-wide	0.002–0.003	All
PRF_BSN.bsn	Basin-wide	0.5–2	OrgN & OrgP
RSDCO.bsn	Basin-wide	0.02–0.1	All
SPCON.bsn	Basin-wide	0.001–0.01	OrgN & OrgP
SPEXP.bsn	Basin-wide	1–1.5	OrgN & OrgP
ADJ_PKR.bsn	Basin-wide	0.5–2	OrgN & OrgP
CDN.bsn	Basin-wide	0.02–0.3	NO_3^-
NPERCO.bsn	Basin-wide	0.21–0.95	NO_3^-
N_UPDIS.bsn	Basin-wide	20–100	NO_3^-
SDNCO.bsn	Basin-wide	0.8–0.88	NO_3^-
PHOSKD.bsn	Basin-wide	140–450	PO_4^{3-}
PPERCO.bsn	Basin-wide	11–16	PO_4^{3-}
PSP.bsn	Basin-wide	0.01–0.7	PO_4^{3-}
P_UPDIS.bsn	Basin-wide	28–95	PO_4^{3-}
ANION_EXCL.sol	Sub-basin	0.1–1	NO_3^-
BC1.swq	Sub-basin	0.1–0.99	NO_3^-
BC2.swq	Sub-basin	0.2–2	NO_3^-
CH_BED_KD.rte	Sub-basin	0.001–3.75	OrgN & OrgP
CH_BNK_KD.rte	Sub-basin	0.001–3.75	OrgN & OrgP
CH_D.rte	Sub-basin	0.58–3.57 (IV·0.7 - IV·1.5)	OrgN & OrgP
CH_L1.sub	Sub-basin	10.83–32.21 (IV·0.75 - IV·1.25)	OrgN & OrgP
CH_L2.rte	Sub-basin	6.10–21.51 (IV·0.75 - IV·1.25)	OrgN & OrgP
CH_N1.sub	Sub-basin	0–0.4	OrgN & OrgP
CH_N2.rte	Sub-basin	0–0.4	OrgN & OrgP
CH_ONCO.rte	Sub-basin	0–1500	OrgN
CH_OPCO.rte	Sub-basin	0–2500	OrgP
CH_S1.sub	Sub-basin	0.001–0.010 (IV·0.8 - IV·2.0)	OrgN & OrgP
CH_SIDE.rte	Sub-basin	0–5	OrgN & OrgP
CH_W1.sub	Sub-basin	1.72–22.62 (IV·0.75 - IV·1.25)	OrgN & OrgP
CH_W2.rte	Sub-basin	2.29–30.16 (IV·0.75 - IV·1.25)	OrgN & OrgP
GWSOLP.gw	Sub-basin	0.01–0.2	PO_4^{3-}
HLIFE_NGW.gw	Sub-basin	0–600	NO_3^-
USLE_P.mgt	Sub-basin	0–0.1	OrgN & OrgP

all the parameters calibrated for each simulation. Since the number of parameters in this set-up was large (73 in total because, for sub-basin level parameters, SUFI2 considers each parameter in each sub-basin as an independent one) a large number of simulations per iteration was chosen to ensure a representative number of parameter combinations.

Parameter calibration ranges were readjusted after each iteration following the suggestion given by the SUFI2 approach. New ranges for basin-wide parameters were obtained for all calibrated stations with a weight of the variable in the objective function equal to 1 in all the sub-basins. However, for sub-basin level parameters, new ranges were obtained by repeating the post-processing step separately for each sub-basin, giving a weight of the objective function equal to 1 to the sub-basin being processed, and equal to 0 for all others. Then, the optimized parameter ranges for each sub-basin were merged for the next model run, together with the basin-wide parameter ranges, thus taking all the sub-basins into consideration. Suggested new parameter calibration ranges were manually edited if going outside the initial wide (and presumed realistic) parameter ranges (Table 2). Unrealistic ranges can provide statistically satisfactory simulations (Muleta, 2012), and therefore the ranges should ideally be constrained to realistic values. However, equifinality of models can still be an issue even if the parameters are assigned within their permissible ranges. Nevertheless, limiting the ranges within realistic values aimed to ensure that real-world constrains were taken into account, favouring that intra-catchment processes produced realistic hydrologic responses even though it could result in somewhat weaker objective functions (Daggupati et al., 2015; Yen et al., 2014).

Once final parameter values were obtained, a split-sample validation (Refsgaard et al., 2014) was performed with an independent data set (1 Jan. 2006 - 31. Dec. 2009). Model accuracy during calibration and validation was evaluated using a range of performance metrics including Nash-Sutcliffe efficiency coefficient (NSE, (Nash and Sutcliffe, 1970)) coefficient of determination (R^2) and percent bias (PBIAS, positive values indicate model underestimation and negative values indicate overestimation). Visual inspection of simulated model output against observed data was also done, since using only performance metrics can be misleading and can produce good but unrealistic simulations (Daggupati et al., 2015). Finally, the values of the parameters providing the best model performance for discharge were identified, and then locked before continuing with the nutrient calibration.

2.3. *Nutrients calibration and objective function assessment*

A similar procedure was followed to perform the objective function assessment for calibration of daily nutrient loading; 14 and 19 parameters were selected at basin-wide and sub-basin levels respectively. Nevertheless, while discharge calibration involved just one output variable, here a multi-variable procedure was followed, calibrating nitrate (NO_3^-), organic nitrogen (orgN), phosphate (PO_4^{3-}, or mineral P “MinP”, as denoted by the SWAT model) and organic phosphorus (orgP) simultaneously. Again, 3000 simulations were run through three iterations of 1000 simulations, repeating the post-processing for basin-wide parameters and then for each sub-basin to optimize the parameter values and to obtain the final set of parameter values. New parameter ranges suggested by the automatic procedure were manually edited if out of range of realistic values to ensure real-world constrains.

This kind of calibration (several output variables at multiple sites) is a multi-criteria decision process (Ahmadi et al., 2014), and some of the decision needed to be taken concern the calculation of the objective function. Considering the different levels of complexity that can be addressed in the SUFI2 algorithm in SWAT-CUP, we have developed the following approach to assess the impact of the objective function in the multi-site calibration of four nutrient fractions:

- **First:** Several objective functions can be chosen to optimize the parameter ranges. Initially, we tested two objective functions, SUM and NSE, which were selected because of their different characteristics: SUM is addressing the mass balance and might be an appropriate objective function in those model applications focused on the average nutrient loads, while NSE is addressing the temporal dynamics and is by far one of the most widely statistics reported for hydrologic calibration (Gassman et al., 2007).
- **Second:** After each iteration, SUFI2 allows a user to repeat the post-processing, and to modify the weight of a variable and/or the sub-basin in the objective function. For example, the user can target all the variables at the same time, obtaining parameter ranges optimized for all the variables. But the user can also repeat the post-processing a number of times targeting each variable individually (i.e., assigning a weight equal to 1 for each variable one at a time -weight equal to 0 for the other variables-), obtaining different new ranges for the parameters optimized for each variable, so the user can choose the most appropriate range for a certain parameter depending on the variable influenced by that parameter. We targeted the four nutrient fractions individually, but also the two nitrogen and the two phosphorus fractions at the same time, for parameters that influence both fractions (weight equal to 1 for N fractions, equal to 0 for P fractions, and vice-versa). Thus, we obtained optimal ranges for each parameter depending on the nutrient(s) fraction(s) influenced by that parameter. This procedure was performed both basin-wide and at the sub-basin level. Then, we merged all the optimized parameter ranges obtained for the next model iteration, so all the nutrient fractions and sub-basins are taken into consideration. If a model parameter influenced several nutrient fractions, we chose the widest range possible before running the next iteration.
- **Third:** When the last run of the selected number of iterations is finished, the user may consider choosing another objective function to obtain the best value for each model parameter. In our case, after 3000 iterations, we needed a single value and not a range, so the “widest range possible” option was no longer valid. This means that we had to obtain a single final value for model parameters that might affect several nutrient fractions and/or monitoring stations that are post-processed simultaneously. Different fractions and/or sites imply different scales, so we considered that using a new objective function, not scale-dependent, could be more appropriate at this step. Thus, SUM was discarded and we decided to test PBIAS (ruled by the relative deviation), bR^2 (the coefficient of determination -R^2-multiplied by the coefficient of the regression line -b-, which accounts for the discrepancy of two signals as well as their dynamics) and again NSE (ruled by the temporal dynamics). SUM was discarded to obtain final parameter values.
- **Fourth:** During the final selection of best parameter values, post-processing in SUFI2 can be repeated as many times as desired changing the weight of the variables in the objective function, so several levels of detail can be implemented targeting different fractions individually, groups of fractions and/or sub-basins to get the final best parameter value depending on the properties of each parameter (same procedure as described for ranges in the second step). We targeted all the nutrient fractions, individually, or as groups of fractions (organic, N-fractions, P-fractions), both at a basin-wide and at a sub-basin level. Afterwards, the optimal values obtained for all the parameters were merged before the final run of the model.

Taking into account these four steps, six different calibration schemes were designed (Table 3).

The six calibration schemes in Table 3 were run from 1 Jan. 2000 to 31 Dec. 2005, with a 10-year warm up period, obtaining six different sets of final parameter values and their corresponding nutrient loads (NO_3^-, PO_4^{3-}, OrgN and OrgP). For each scheme, daily model performance was as for discharge described calculating three performance metrics, R^2, NSE and PBIAS, comparing the values yielded by the different schemes. These metrics are not only commonly recommended for catchment models (e.g. Ahmadi et al., 2014; Daggupati et al., 2015; Moriasi et al., 2007), but they also

Table 3
Calibration schemes designed to assess the impact of the objective function (OF).

Scheme	OF in iterations	Fractions targeted in iterations	OF for best parameters	Fractions targeted to obtain best parameters	
				Basin-wide	Sub-basin level
1	SUM	OrgN, OrgP, NO_3^-, PO_4^{3-}	bR^2	All, N, P	All
2	SUM	OrgN, OrgP, NO_3^-, PO_4^{3-}	PBIAS	All, Organic, OrgN, OrgP, NO_3^-, PO_4^{3-}	All, Organic, OrgN, OrgP, NO_3^-, PO_4^{3-}
3	SUM	OrgN, OrgP, NO_3^-, PO_4^{3-}	bR^2	All, Organic, OrgN, OrgP, NO_3^-, PO_4^{3-}	All, Organic, OrgN, OrgP, NO_3^-, PO_4^{3-}
4	NSE	N, P	NSE	All, N, P	All
5	NSE	N, P	NSE	All, N, P	All, N, P
6	SUM	N, P	bR^2	All, N, P	All

cover the major categories of quantitative statistics (Moriasi et al., 2007): standard regression (R^2), dimensionless (NSE) and error index (PBIAS). Good performance in all of them ensures a good calibration and helps to evaluate which scheme proves superior in terms of both magnitude and temporal variations across all the nutrients fractions and monitoring stations. The assessment of the influence of the objective function was evaluated not only considering the goodness of performance in the final calibrated model for each scheme, but also analysing the evolution of the performance metrics after each calibration step (raw uncalibrated model, flow calibrated, the three iterations steps of 1000 simulations and finally the best parameter selection). This analysis might also help to take decisions about the effort needed to obtain an adequate calibration.

2.4. Validation in a new model set-up

Once the six calibration schemes were evaluated, the one yielding the most satisfactory results was applied in another multisite and multi-variable SWAT set-up as a means of testing the approach and applicability to other lowland catchments. Here the SWAT model was applied for the northern part of Denmark (10,857 km^2, Fig. 1) with discharge and nutrient calibration performed at 17 and 12 monitoring stations, respectively.

3. Results and discussion

3.1. Discharge calibration

The initial and calibrated values of the selected parameters are given in Table 2. Fig. 2 shows the observed and simulated discharges in the monitoring stations during calibration and validation periods, and Table 4 shows the corresponding performance metrics values.

The observed discharges were well reproduced in all the stations during the calibration period (Fig. 2). Statistically, the model showed a better performance than in the previous calibration by Thodsen et al. (2015) (Table 4), which demonstrates the effect of allowing more parameters to vary at sub-basin level. Model performance at stations 2 and 4 was slightly better than at stations 1 and 3. Nevertheless, it can be considered a very good daily calibration overall, compared to the performance metrics for daily data in the literature (e.g. Gassman et al., 2007; Moriasi et al., 2007). Visual inspection of the model performance during the validation period also displayed a model capable of encompassing the observed heterogeneity in magnitudes (Fig. 2) and the statistical performances were also very good (Table 4), sometimes better than during calibration. This demonstrates the ability of the model to reproduce the discharge in the Odense catchment, which then serves as a good starting point when initiating the calibration of nutrient fractions.

3.2. Nutrients calibration and objective function assessment

Table 5 shows the values of the performance metrics for nutrients calibration at the end of the six different calibration schemes followed.

Calibration for mineral fractions showed higher R^2 and NSE values and lower biases (either positive or negative) than for organic fractions. None of the schemes were able to perform a satisfactory simulation of organic nutrients in station 2 according to the values of the three performance metrics used. It is the smallest and flattest sub-basin among those monitored and located at the border of the catchment (Fig. 1, Table 1). The discharge regime might be too low and slow for SWAT to produce a realistic organic nutrients load (which in a lowland catchment may be largely a result of riverbank erosion or collapses). For all the other variables and stations, statistical performance was good for one or more criteria (Table 5) compared to performance thresholds in e.g. (Moriasi et al., 2007), especially considering that our model was calibrated with daily values, which usually shows lower ratings (Gassman et al., 2007).

R^2 values remain quite consistent among the different schemes, while NSE and PBIAS values showed more variability. Scheme 4 and 5, which are those using NSE as objective function, gathered a larger number of best values of the different metrics than the other schemes, especially among the organic fractions. For the mineral fractions, best values were more widespread. However, scheme 4 and 5 were those providing the best combined result for the three performance metrics used (Table 5), which is always desirable since relying on one performance metric alone can be often misleading (Muleta, 2012). This first look at the final calibration results suggest that using NSE as objective function might be the best option to obtain a good calibration in a complex catchment where several variables are calibrated simultaneously for multiple sites. Scheme 4 and 5 not only yielded good values of NSE, as expected, but also R^2 and PBIAS, performing better than other schemes which used different objective functions.

Fig. 3 shows the evolution of the three statistical performance metrics after each calibration step. For all schemes R^2 showed similar behaviour with considerable increase in performance after the first iteration, especially in the organic fractions (Fig. 3a). Performance metrics for NO_3^- already showed high values in the raw model and the most significant PO_4^{3-} improvement was after flow calibration, probably because mineral fractions loads are mainly governed by runoff (Molina-Navarro et al., 2014). The representation of real-world constraints in the model (e.g. tile drains) might have allowed a good discharge simulation from the beginning (Daggupati et al., 2015; Yen et al., 2014). Increase in R^2 was very modest after each subsequent iteration step for all the schemes and nutrient fractions, except for scheme 4 and 5 and the mineral fractions, which yielded a higher increase of the statistics.

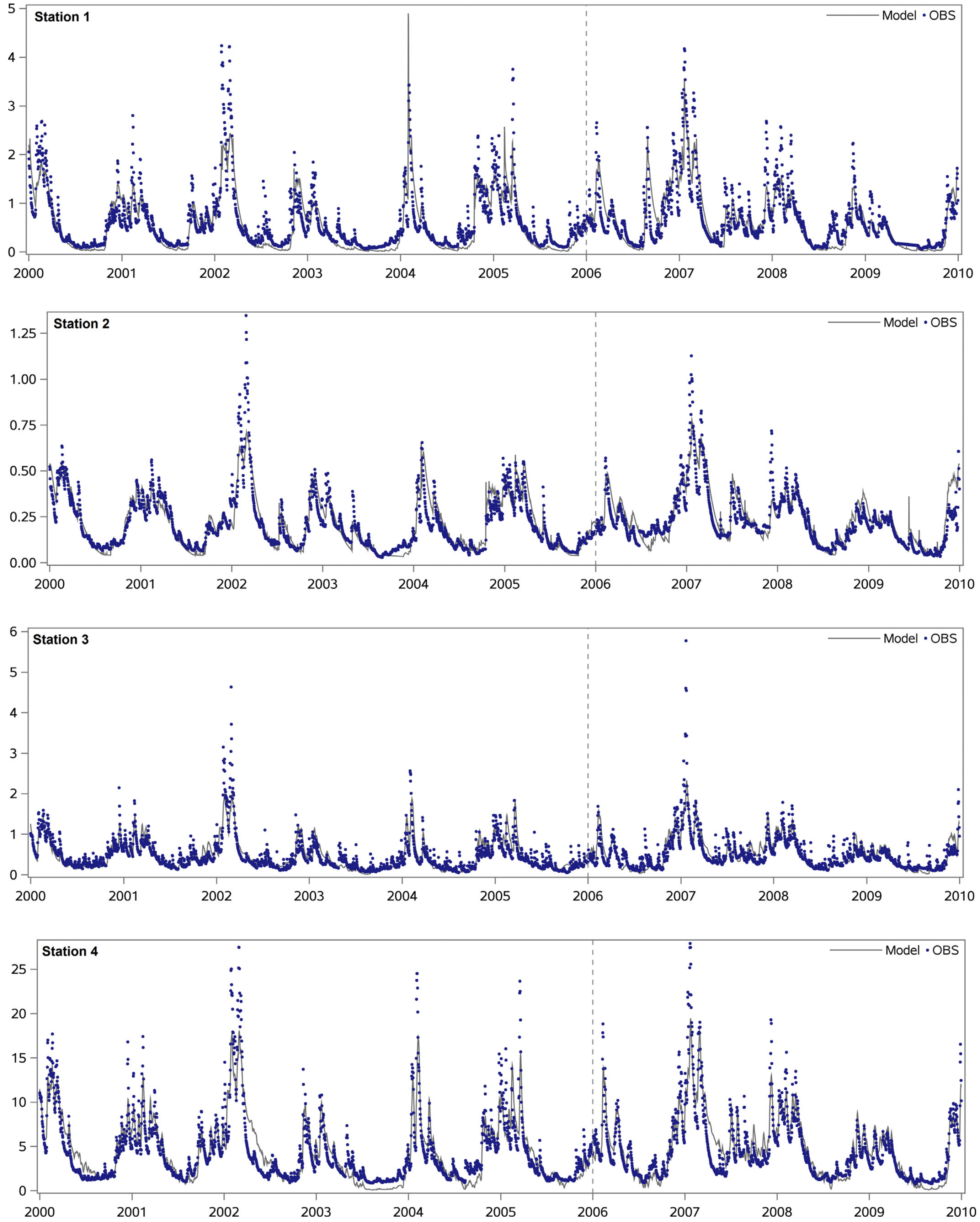


Fig. 2. Observed (blue dots) and simulated (grey line) daily discharge (m^3 s^{-1}) at the four gauging stations in the Odense Fjord catchment during calibration (2000–2005) and validation (2006–2009) periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Calibration (2000–2005) and validation (2006–2009, in brackets) performance metrics for daily runoff at the four monitoring points in the Odense Fjord catchment.

	Station 1	Station 2	Station 3	Station 4
NSE	0.73 (0.74)	0.82 (0.76)	0.71 (0.67)	0.79 (0.83)
R^2	0.74 (0.74)	0.83 (0.79)	0.71 (0.68)	0.79 (0.83)
PBIAS	9.5 (2.6)	−1.9 (−9.2)	0.8 (−1.9)	−0.6 (−2.4)

Regarding NSE (Fig. 3b), behaviour was not as uniform as for R^2. Flow calibration already improved NSE values noticeably for all the nutrient fractions and the highest values were again found for the mineral fractions, especially NO_3^-. Positive NSE values were not reached for organic fractions. A steady increase of NSE values in scheme 4 and 5 was found for every nutrient fraction, which was expected since NSE is also used as objective function in these schemes. OrgN, NO_3^- and PO_4^{3-} also showed a progressive increase of NSE in the remaining schemes, although lower values were reached for mineral fractions. Especially noticeable is the case of OrgP. After the first iteration of the nutrient calibration, NSE values only improved in scheme 4 and 5, noticeably decreasing again in scheme 1, 2, 3 and 6. Improvement arrived again after the second iteration of 1000 simulations, but in the third round and for the best parameter selection, NSE values dropped again. A similar pattern was found for PBIAS.

Contrary to the other performance metrics, PBIAS absolute values became worse (increased) after flow calibration for all the nutrient fractions except for OrgP (Fig. 3c). Then, a decrease of the absolute PBIAS followed for most of the schemes and fractions, especially for mineral fractions, again favoured by a good flow calibration. However, as seen also for NSE, OrgP only yielded an improvement in PBIAS for scheme 4 and 5, while in the other schemes PBIAS values were worse at the end of the calibration process (reaching values as high as 394% for scheme 2).

Results might be surprising since scheme 1, 2, 3 and 6 use SUM as the objective function in the three iterations of 1000 simulations, and SUM relies on the average deviation from a mean value, similar to PBIAS, while NSE tends to rely more on the peak events. SUM could be a natural choice of objective function in those applications where priority is a good simulation of the overall average nutrient loads and less so the temporal dynamics of the nutrient loads. However, SUM is a scale-dependent objective function, so when calibrating the different fractions at the same time and in different sites with SUM, parameters tend to be optimized towards those fractions and sites with a higher load. Consequently, the resulting model lack to simulate fractions with the lowest loads adequately (OrgP in this case). In other words, optimizing the objective function SUM at those sites with higher flows and for those variables with higher loads introduced in our case a serious bias for the other sites and variables, yielding a lower NSE and a higher absolute PBIAS. Thus, different parameter calibration schemes can lead to significantly different calibration results, which ultimately may have important implications if the models are used for water resource management purposes (Zhang et al., 2008). It must be acknowledged that this issue observed for SUM as a scale-dependent objective function could be counteracted by applying different weights to different sites and fractions in the calculation of the objective function (i.e. high weight to fractions and catchments with lower loads and vice versa). However, the purpose here was to assess the impact of the objective function in its original form, so a weight equal to 1 was always considered. Besides, the fact of selecting different weights in such a model set-up (four calibration sites and four nutrient fractions) would introduce such a complexity in the calibration procedure that it might become not operational.

The additional effort required in targeting all the different nutrient fractions when updating parameter value ranges after each iterations step in scheme 1, 2, 3, and also in the best parameter selection in scheme 2 and 3 was not fruitful despite being very time consuming (20 post-processings were needed after each iteration). There were always some parameters that influenced two or more nutrient fractions and/or several sub-basins (basin-wide parameters), and their existence did not allow a good calibration for all the nutrient fractions when using a scale-dependent objective function, yielding a comparatively worse calibration of the organic fractions and especially organic phosphorus. Using a new objective

Table 5
Calibration (2000–2005) performance metrics values for daily nutrient loads (organic nitrogen, organic phosphorus, nitrate and phosphate) at the four monitoring points (in brackets) in the Odense Fjord catchment. Bold numbers indicate the best statistical value obtained among the six calibration schemes.

	Scheme 1			Scheme 2			Scheme 3			Scheme 4			Scheme 5			Scheme 6		
	R2	*NSE*	*PBIAS*	*R2*	*NSE*	*PBIAS*	*R2*	*NSE*	*PBIAS*	*R2*	*NSE*	*PBIAS*	*R2*	*NSE*	*PBIAS*	*R2*	*NSE*	*PBIAS*
OrgN (1)	0.53	−1.78	−82.7	0.49	−0.34	89.9	0.50	0.36	45.3	0.51	0.44	24.4	0.51	0.35	49.7	0.54	0.09	−16.5
OrgN (2)	0.00	−1.34	98.0	0.17	−1.34	97.0	0.00	−1.34	98.1	0.03	−1.34	98.1	0.03	−1.34	98.1	0.02	−1.34	98.1
OrgN (3)	0.45	−5.00	−163.1	0.39	−0.58	90.7	0.45	0.06	−30.6	0.48	−0.61	90.7	0.47	−0.49	86.2	0.46	0.07	−21.2
OrgN (4)	0.50	0.05	71.8	0.49	0.06	66.9	0.49	0.38	38.0	0.51	−0.33	90.9	0.51	−0.33	91.1	0.44	0.14	67.8
OrgP (1)	0.54	−39.0	−471.3	0.51	−0.06	−16.8	0.52	−10.5	−198.9	0.53	**0.35**	**−2.7**	0.53	0.03	−25.9	**0.55**	−1.32	−78.7
OrgP (2)	0.01	**−1.04**	88.8	**0.16**	−1.13	**83.9**	0.00	**−1.04**	89.4	0.04	−1.05	88.6	0.04	−1.05	88.6	0.03	−1.05	89.1
OrgP (3)	0.43	−10.4	−357.1	0.44	0.36	43.1	0.43	−42.9	−708.7	**0.48**	**0.47**	**13.9**	**0.48**	−0.06	−61.4	0.47	−106.9	−1041.4
OrgP (4)	0.40	−1.14	−81.1	0.40	−8.18	−289.8	0.41	−36.5	−593.2	**0.42**	0.34	43.9	**0.42**	**0.35**	**41.8**	0.35	−16.5	−226.9
NO_3^- (1)	**0.72**	0.40	−58.8	0.70	0.69	−7.4	0.70	0.61	−33.6	0.70	**0.70**	**−2.3**	0.70	0.68	−17.6	**0.72**	0.57	−42.8
NO_3^- (2)	0.80	0.58	−56.8	0.01	−0.05	**25.3**	0.79	0.66	−41.9	**0.81**	0.69	−36.9	**0.81**	**0.71**	−31.3	0.64	−0.41	−116.2
NO_3^- (3)	0.69	−2.04	−139.1	0.77	0.76	−4.3	0.78	0.61	−32.6	0.74	0.69	−10.6	**0.79**	**0.78**	**1.9**	0.64	−0.70	−96.6
NO_3^- (4)	**0.79**	**0.79**	**4.4**	0.77	0.65	25.3	**0.79**	0.60	34.9	0.75	0.59	35.8	0.75	0.60	34.8	**0.79**	0.73	18.8
PO_4^{3-} (1)	0.36	0.27	−9.6	0.35	0.30	**−0.2**	0.35	0.19	−18.8	0.35	0.31	27.0	0.35	**0.33**	14.1	**0.38**	0.26	−15.2
PO_4^{3-} (2)	**0.51**	−0.01	−46.5	0.01	−0.12	−4.9	**0.51**	−0.62	−68.3	**0.51**	**0.51**	**2.2**	**0.51**	0.50	5.4	**0.51**	−0.02	−48.8
PO_4^{3-} (3)	0.00	−0.22	1.8	0.00	−0.24	**0.8**	**0.02**	−0.48	79.2	0.00	−0.23	**0.8**	0.00	**−0.16**	25.2	0.00	−0.37	−18.0
PO_4^{3-} (4)	**0.49**	**0.37**	**−0.4**	0.27	0.08	46.5	0.31	0.12	45.6	0.40	0.26	39.7	0.37	0.31	27.8	0.45	−1.43	−65.2
Average[a]	***0.45***	*−3.72*	*97.9*	*0.37*	*−0.57*	*55.8*	*0.44*	*−5.61*	*134.8*	***0.45***	***0.11***	***38.0***	***0.45***	*0.08*	*43.8*	*0.44*	*−8.01*	*126.8*
Av Org[a]	*0.36*	*−7.46*	*156.1*	***0.38***	*−1.40*	*97.3*	*0.35*	*−11.4*	*225.3*	***0.38***	***−0.22***	***56.7***	*0.37*	*−0.32*	*67.9*	*0.36*	*−15.9*	*200.8*
Av Min[a]	***0.55***	*0.02*	*25.0*	*0.36*	*0.26*	***12.5***	*0.53*	*0.21*	*36.0*	*0.53*	*0.44*	*18.8*	*0.54*	***0.47***	*15.4*	*0.52*	*−0.17*	*42.00*

[a] Average rows for PBIAS shows the average of the absolute values.

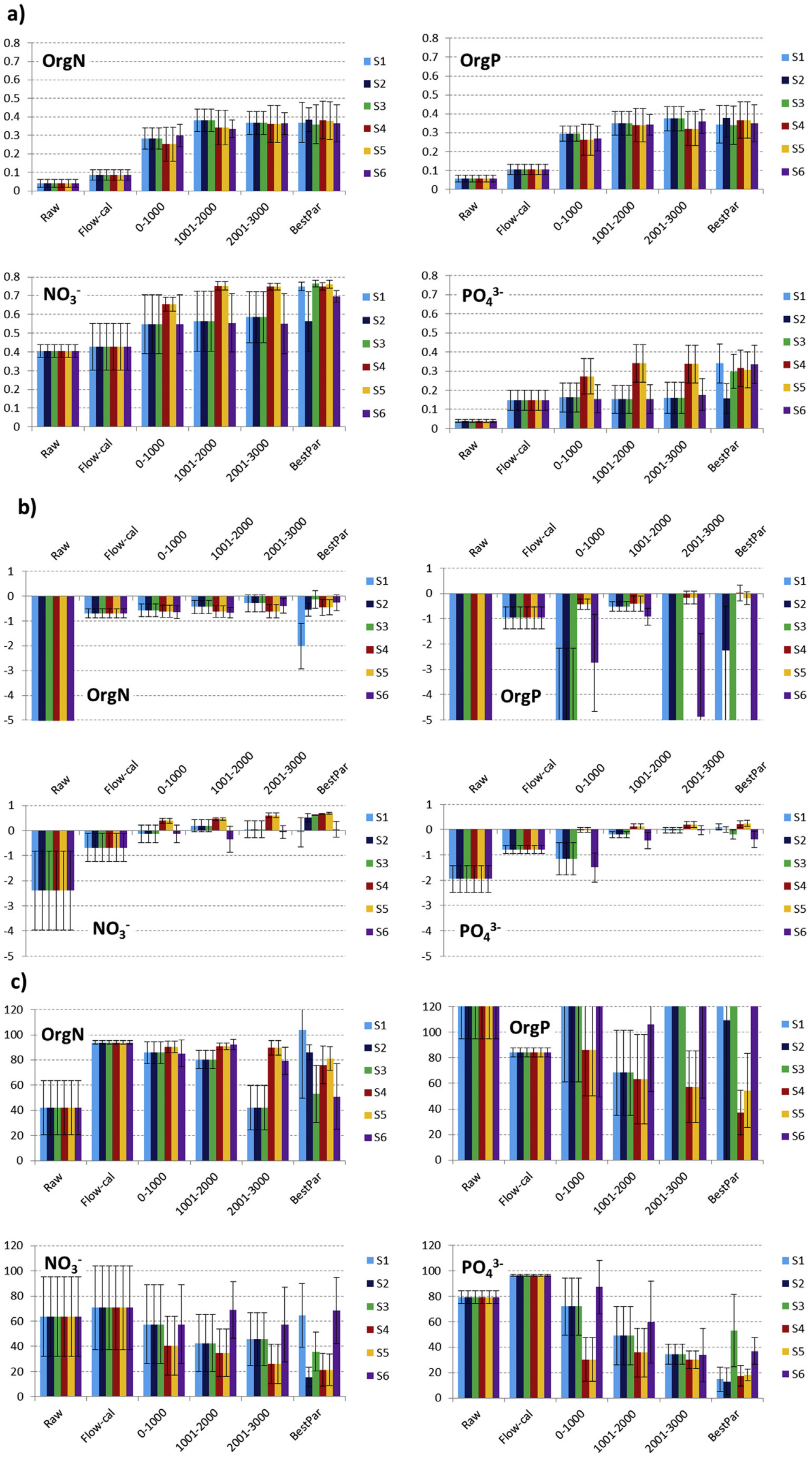


Fig. 3. Evolution of R^2 (a), NSE (b) and absolute PBIAS (c) values (four monitoring points average ± standard error) for each nutrient fraction in the six calibration schemes tested in the SWAT model set-up for the Odense Fjord catchment.

function to obtain the final parameter values (bR^2 -scheme 1, 3 and 6- or PBIAS -scheme 2-) did not improve the results.

The use of NSE as an objective function is often questioned in the modelling literature, especially in hydrological applications because it tends to rely more on the peaks of the calibrated variable (Zhang et al., 2015) or might be influenced by a few significant outliers in multi-site and multi-variable calibrations (Rouholahnejad et al., 2014). Several authors suggested that a model calibrated using NSE as objective function may yield high NSE values being in fact a poor model when evaluating other performance metrics, so they recommend using another objective function, using a multi-objective approach or evaluating model performance with multiple statistical metrics (e.g. Muleta, 2012). However, we did not find any reports in the literature assessing the impact of the choice of objective function on multi-site and multi-variable calibrations. Our results demonstrated that those schemes using NSE as an objective function (scheme 4 and 5) were generally the most efficient in achieving high performance metrics, also when looking at PBIAS and R^2. Some authors have pointed out that using only one objective function tends to calibrate the model parameters so that they fit that specific objective function, while ignoring important information contained in others (e.g. Zhang et al., 2008). However, scheme 4 and 5 not only improved NSE the most through the iterations of the nutrient calibration process, but they were also able to increase R^2 and decrease PBIAS, in many occasions more than the other schemes (Fig. 3). R^2 and NSE are correlation and model efficiency metrics (Bennett et al., 2013), which means that satisfactory values (over 0.5 for both criteria, Moriasi et al., 2007) preserve the data pattern. On the other hand, PBIAS is a key residual metric (Bennett et al., 2013), and a satisfactory value (below 75% for nutrients, Moriasi et al., 2007) suggests acceptable residuals and reduced under- or overestimation. It must be acknowledged that performance metric values for OrgN were generally poor, and not always the best when using scheme 4 and 5, showing specially higher positive bias (underestimation, Table 5 and Fig. 3c), but they achieved only minor deviation from the other schemes where satisfactory calibration for this particular fraction was never achieved. R^2 and NSE values were also unsatisfactory for PO_4^{3-} in station 3 (in all the schemes) because the timing of the peaks was not accurately simulated. Nevertheless, statistical performance for scheme 4 and 5 showed an overall good calibration, especially considering that it was carried out for a daily time step in a spatial (4 stations) and multi-variable (4 nutrients) model set-up.

So, despite the existence of many negative reports about using NSE as objective function, our results demonstrate that for a multi-site calibration of several nutrient fractions at the same time in a temperate lowland catchment model, NSE was a good choice. It does not mean, though, that NSE has no weaknesses as an objective function. The choice of the objective function plays an important role in determining the data points that are more critical in model calibration (Wright et al., 2016). Hence, NSE tends to rely more on the peaks of the calibrated variable and can be influenced by significant outliers, which, for example, might lead to inaccurate predictions of the nutrient loads during dry seasons. Besides, Fig. 3 also reveals that, despite being time consuming, it was worth running three iterations of 1000 simulations, because the values of the different performance metrics showed an improving trend through the whole process for most of the cases. However, the rate of improvement was minimal for the last 1000 runs (Fig. 3), so considering the computational time required, we would not recommend executing more than 3000 simulations.

NSE was the best choice among those tested and showed robust results. However, six plausible calibration schemes using mainly SUM and NSE as objective functions were selected to optimize the assessment of the objective function impact (Table 3), but we must acknowledge that other options could have been tested as well (11 objective functions are available in SUFI2). Since different objective functions rely on different aspects of the variable targeted, the choice of objective function also depends on the variable of interest and the expected application of the model. Thus, we also acknowledge that the most efficient calibration scheme described in this study might not be optimal in other applications. For example, a multi-site calibration procedure for discharge and nitrate was performed in the Black Sea Basin using bR^2 as objective function and setting the weights for all stations and variables to 1, obtaining calibration results ranging from very good to poor (Rouholahnejad et al., 2014).

Our results support the importance of checking the calibration performance with several statistical metrics of different types when using a single objective function for calibration (Moriasi et al., 2007; Muleta, 2012), besides considering additional information such as visual inspection (Daggupati et al., 2015) before making decisions based on model simulations. A representative example can be taken from the calibration of mineral fractions in station 2 with the calibration scheme 2, which uses SUM and PBIAS as objective functions (Table 3). PBIAS value was the best for NO_3^- (25.3) and the second best for PO_4^{3-} (−4.9) (Table 5). This is not surprising, since PBIAS was used as objective function, so the calibration procedure tends to optimize its value. However, relying solely on PBIAS would not have been appropriate, since the dynamics of both variables were not well captured (Fig. 4), yielding R^2 and NSE values of 0.01 and −0.05 for NO_3^-, 0.01 and −0.12 for PO_4^{3-} (Table 5). This demonstrates that obtaining satisfactory statistical performance metric values (PBIAS in this case) does not guarantee optimal parameter values representing the reality (Muleta, 2012). However, with scheme 4 and 5, absolute PBIAS deviations were similar (−36.9 and −31.3 for NO_3^-; 2.2 and 5.4 for PO_4^{3-}, respectively), but the calibration was much better for the other metrics (Fig. 4), yielding values of R^2 and NSE of 0.81 and 0.69 for NO_3^-, 0.51 and 0.51 for PO_4^{3-} in the scheme 4; 0.81 and 0.71 for NO_3^-, 0.51 and 0.50 for PO_4^{3-} in the scheme 5. This example shows again that using NSE as objective function was the best choice.

The difference between the two schemes using NSE as objective function is that in scheme 5 N and P fractions were targeted separately when obtaining the final best value of the sub-basin level parameters, while in scheme 4 just one post-processing per sub-basin was run, with a weight of the objective function equal to 1 for all the nutrient fractions. Fig. 3 and Table 5 show that the performance of the model was equally good in both schemes, even slightly better in scheme 4. Mann-Whitney W tests were carried out to compare each metric performance in scheme 4 vs. the other schemes, taking into account all the values (regardless of nutrient fraction and station). Although R^2 was never significantly higher, NSE and PBIAS were significantly better (higher NSE, lower absolute PBIAS) in scheme 4 than in schemes 1, 3 and 6 at the 90.0% confidence level. scheme 4 did not show statistically significant differences with scheme 2 and 5 for the entire data set. However, looking at individual nutrient fractions (Fig. 3), scheme 4 often performs better than scheme 2, especially for OrgP. No statistically significant differences found between scheme 4 and 5, which suggest that the additional effort done in scheme 5 did not yield superior calibration results. Thus, we recommend following calibration scheme 4 in the calibration of similar model set-ups.

3.3. Testing the optimal approach in a new model set-up

Following the results obtained, calibration scheme 4 proved superior, and was subsequently tested for a new SWAT set-up in northern Denmark (Fig. 1). Previously, discharge calibration was

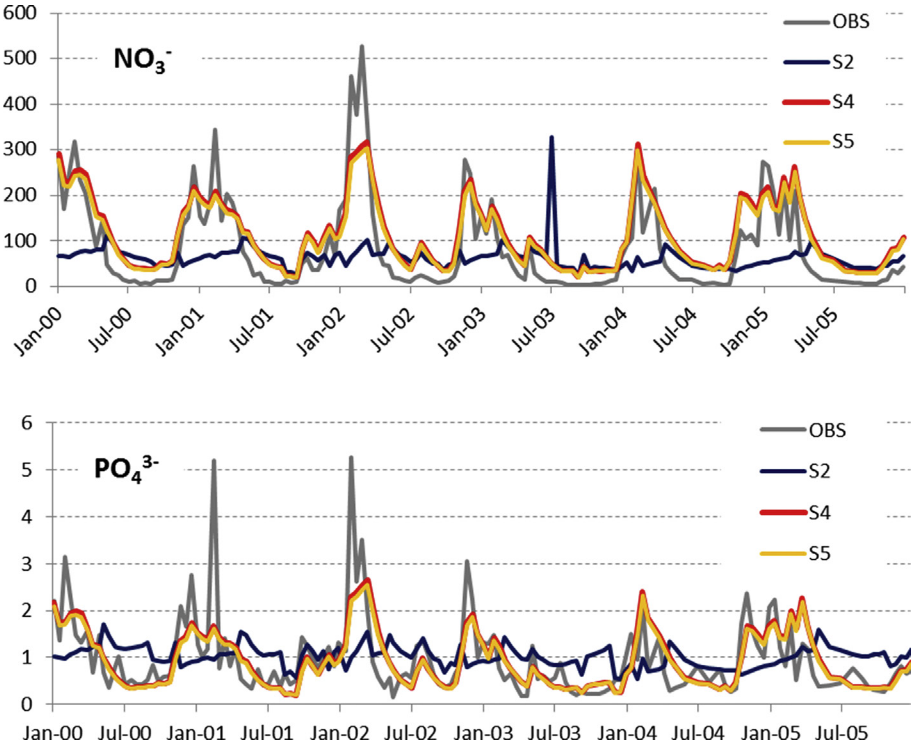


Fig. 4. Observed values and final result of NO_3^- and PO_4^{3-} loads calibration (Kg day^{-1}) in station 2 for scheme 2, 4 and 5.

undertaken using SUM as the objective function. R^2, NSE and PBIAS average values were 0.56, 0.48 and 4.9 respectively for daily calibration and 0.55, 0.39 and 13.4 for daily validation (average from 16 stations, absolute values for PBIAS). These results were satisfactory, especially taking into account that it is a very complex model, with 182 sub-basins and 8398 HRUs (31 and 3410 respectively in the Odense catchment). Fig. 5 shows the evolution of the three performance metrics evaluated along the nutrient calibration steps.

For the northern Denmark SWAT set-up many similarities with the results obtained in the Odense catchment could be found. Mineral nutrient fractions showed again the best calibration at the end of the procedure, especially NO_3^-, which showed a high (0.49) R^2 already in the early calibration stages favoured by a good hydrological simulation (Fig. 5a). The scheme was able to increase progressively the NSE value for all nutrient fractions (Fig. 5b). Similar to the Odense case, PBIAS increased after flow calibration in all fractions except from OrgP, but then the scheme was useful to reduce absolute PBIAS iteration by iteration (Fig. 5c). In summary, the calibration scheme was able to improve the value of all the performance metrics for all the nutrient fractions. Results revealed again that it was worthwhile to perform three iterations of 1000 simulations despite the additional cost in computation time (around five days for running each iteration with 1000 simulations and 1.5 h per post-processing on a Quad-core PC). The test of the calibration scheme for a new SWAT set-up confirmed that a calibration scheme using NSE as objective function and targeting N and P fractions separately in the weight of the objective function can be appropriate when calibrating several nutrient fractions simultaneously in a multi-site catchment model (using the SUFI2 calibration approach). Final performance metrics values were not as good as in the Odense calibration, but that was to be expected since the model covers a 10 times larger area (10,857 versus 1061 km^2) and includes more calibration stations (12 versus 4). Thus, the basin-wide parameters have to represent a much larger basin, and consequently with more imprecision.

The validation in another Danish lowland catchment confirms our recommendation of calibration scheme 4 in similar model set-ups. However, we acknowledge that there is incertitude about the suitability of this scheme under different hydro-climatic or geographical conditions. Nevertheless, nutrient pollution poses a big concern in lowland catchments. Results obtained in our study can be useful and serve as a guideline for modellers and water managers working in similar areas, as well as a starting point for similar multi-site and multi-variable calibrations in other catchments.

4. Conclusions

This paper assess for the first time the impact of the objective function (its choice and its calculation according to the weights assigned to different output variables) during the calibration of a multi-site and multi-variable catchment hydro-ecological model developed with SWAT. The model was set-up for the Odense catchment (Denmark) and calibrated for daily discharge. Then, six calibration schemes for daily nutrient load calibration were designed and tested, varying the objective function and the nutrient fractions targeted in its calculation. The evolution of three performance metrics during the calibration process and their final values in the different schemes were compared.

Our results showed that the best performance metrics were obtained with scheme 4, which is based on NSE as an objective function, running two post-processing steps for N and P fractions follow each iteration (in total three iterations of 1000 simulations), three post-processing steps for N-fractions, P-fractions and all fractions to obtain the single best value of basin-wide model parameters, and just one post-processing step targeting all fractions at a sub-basin level. This scheme was able to improve the value of all the performance metrics evaluated during the whole calibration

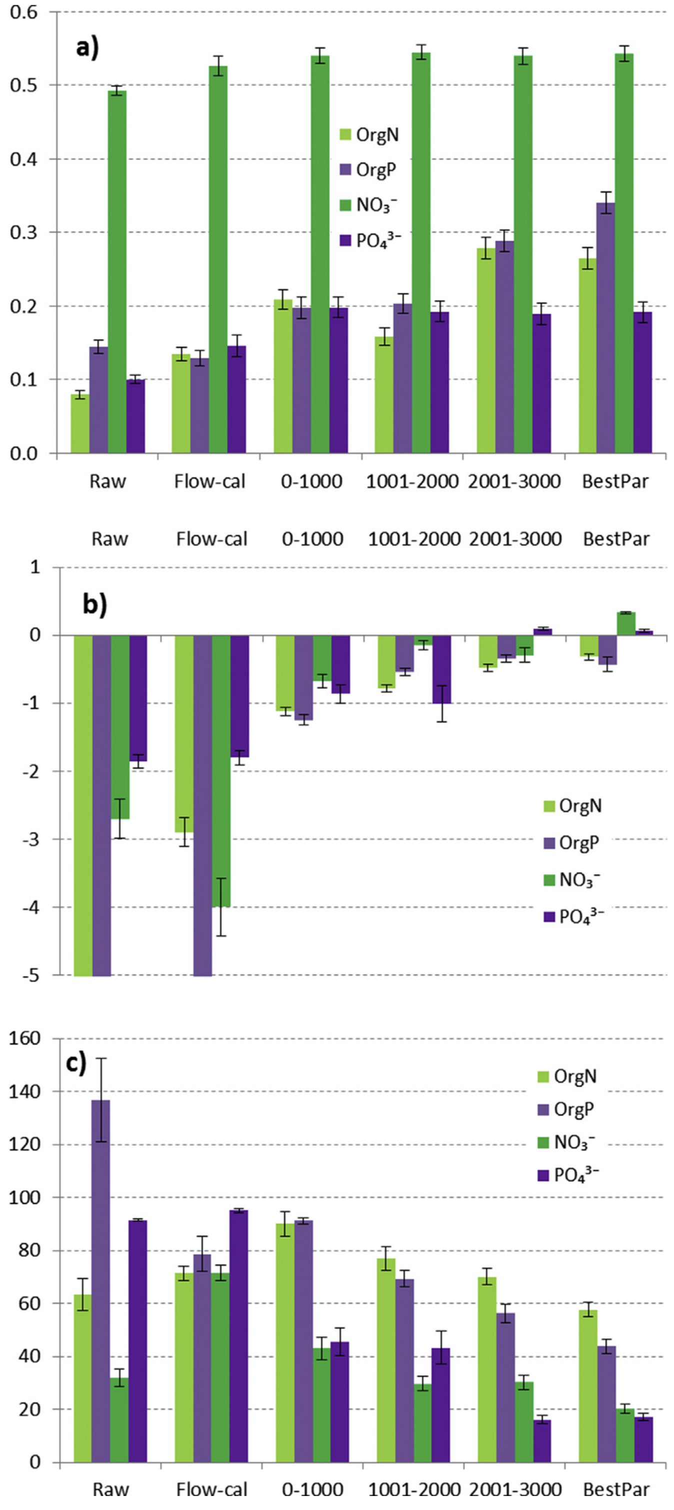


Fig. 5. Evolution of R^2 (a), NSE (b) and absolute PBIAS (c) values (average across 12 monitoring stations ± standard error) for each nutrient fraction following the calibration scheme 4 in a SWAT set-up in northern Denmark.

process, reaching satisfactory final values. The validation of the scheme in a new and more complex SWAT set-up for the northern part of Denmark indicate that the approach may be suitable and transferable to other sites.

In spite of the fact that the value in using NSE as an objective function has been questioned widely in the modelling literature, especially for hydrological applications, we found this particular metric a good choice when addressing the calibration of a multi-site and multi-variable (four nutrient fractions) catchment model. This article provides new insights about the choice of calibration schemes in complex catchment models and may serve as a guideline for hydro-ecological modellers facing similar automatic calibration procedures to achieve catchment management goals.

Acknowledgements

This study is carried out as part of the IMAGE research alliance (IMAGE 2014) funded by the Danish Strategic Research Council (09-067259), and as part of the MARS Project (Managing Aquatic ecosystems and water Resources under multiple Stress) funded under the 7th EU Framework Programme, Theme 6 (Environment including Climate Change), Contract No: 603378. Eugenio Molina-Navarro receives additional support from the Ramón Areces Foundation postdoctoral fellowship programme. This article is dedicated to the memory of Ángel Carpintero, mathematician and EMN's uncle, who passed away on 20/03/17.

References

Abbaspour, K.C., 2015. SWAT-CUP: SWAT Calibration and Uncertainty Programs - a User Manual (Eawag, Dübendorf).

Ahmadi, M., Arabi, M., Ascough II, J.C., Fontane, D.G., Engel, B.A., 2014. Toward improved calibration of watershed models: multisite multiobjective measures of information. Environ. Model. Softw. 59, 135–145.

Arnold, J.G., Kiniri, R., Srinivasan, R., Williams, J.R., Haney, E.B., Neitsch, S.L., 2014. Soil & Water Assessment Tool. Input/Output Documentation. Version 2012. Texas Water Resources Institute, TR-439, College Station.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Centre for Agricultural and Rural Development, 2016. SWAT Literature Database for Peer-reviewed Journal Articles. Available at: https://www.card.iastate.edu/swat_articles/. Last accessed: 10.05.2016.

Daggupati, P., Yen, H., White, M.J., Srinivasan, R., Arnold, J.G., Keitzer, C.S., Sowa, S.P., 2015. Impact of model development, calibration and validation decisions on hydrological simulations in West Lake Erie Basin. Hydrol. Process 29, 5307–5320.

Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil and water assessment tool: historical development, applications, and future research directions. Trans. ASABE 50, 1211–1250.

Molina-Navarro, E., Trolle, D., Martínez-Pérez, S., Sastre-Merlín, A., Jeppesen, E., 2014. Hydrological and water quality impact assessment of a Mediterranean limno-reservoir under climate change and land use management scenarios. J. Hydrol. 509, 354–366.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900.

Muleta, M., 2012. Model performance sensitivity to objective function during automated calibrations. J. Hydrol. Eng. 17, 756–767.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290.

Naturstyrelsen, Nationalt Center for Miljø og Energi, 2016. NOVANA. Det Nationale Program for Overvågning Af Vandmiljø Og Natur 2016 (Programbeskrivelse. Miljø- og Fødevareministeriet, Copenhague).

Neitsch, S.L., Arnold, J.G., Kiniri, R., Williams, J.R., 2011. Soil and Water Assessment Tool. Theoretical Documentation. Version 2009. Texas Water Resources Institute, TR-406, College Station.

Niraula, R., Meixner, T., Norman, L.M., 2015. Determining the importance of model calibration for forecasting absolute/relative changes in streamflow from LULC and climate changes. J. Hydrol. 522, 439–451.

Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., Hamilton, D.P., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, T.O., Trolle, D., Willems, P., Christensen, J.H., 2014. A framework for testing the ability of models to project climate change and its impacts. Clim. Change 122, 271–282.

Rouholahnejad, E., Abbaspour, K.C., Srinivasan, R., Bacu, V., Lehmann, A., 2014. Water resources of the Black Sea Basin at high spatial and temporal resolution. Water Resour. Res. 50, 5866–5885.

Servat, E., Dezetter, A., 1991. Selection of calibration objective functions in the context of rainfall-runoff modelling in a sudanese savannah area. Hydrol. Sci. J. 36, 307–330.

Smed, P., 1982. Landskabskort over Danmark. Sheet 2–4, Geografforlaget, Copenhague.

Thodsen, H., Andersen, H.E., Blicher-Mathiesen, G., Trolle, D., 2015. The combined effects of fertilizer reduction on high risk areas and increased fertilization on low risk areas, investigated using the SWAT model for a Danish catchment. Acta Agric. Scand. Sect. B — Soil Plant Sci. 65, 217–227.

White, K.L., Chaubey, I., 2005. Sensitivity analysis, calibration, and validations for a multisite and multivariable swat model. J. Am. Water Resour. Assoc. 41, 1077–1089.

Wright, D., Thyer, M., Westra, S., McInerney, D., 2016. The impact of objective function selection on the influence of individual data points. Geophys. Res. Abstr. 18, 11352.

Wu, H., Chen, B., 2015. Evaluating uncertainty estimates in distributed hydrological modeling for the Wenjing River watershed in China by GLUE, SUFI-2, and ParaSol methods. Ecol. Eng. 76, 110–121.

Yen, H., Bailey, R.T., Arabi, M., Ahmadi, M., White, M.J., Arnold, J.G., 2014. The role of interior watershed processes in improving parameter estimation and performance of watershed models. J. Environ. Qual. 43, 1601–1613.

Zhang, D., Chen, X., Yao, H., Lin, B., 2015. Improved calibration scheme of SWAT by separating wet and dry seasons. Ecol. Model 301, 54–61.

Zhang, X., Srinivasan, R., Van Liew, M., 2008. Multi-site calibration of the SWAT model for hydrologic modeling. Trans. ASABE 51, 2039–2049.