Towards a data-effective calibration of a fully distributed catchment water quality model

Salman Ghaffar[a,c*], Xiangqian Zhou[a*], Seifeddine Jomaa[a], Xiaoqiang Yang[a,b], Günter Meon[c], Michael Rode[a,d]

[a]Department of Aquatic Ecosystem Analysis and Management, Helmholtz Centre for Environmental Research – UFZ, Magdeburg, Germany

[b]Yangtze Institute for Conservation and Development, Hohai University, Nanjing 210098, China

[c]Leichtweiß-Institute for Hydraulic Engineering and Water Resources, Technische Universität Braunschweig, Braunschweig, Germany

[d]Institute of Environmental Science and Geography, University of Potsdam, Potsdam-Golm, Germany

[*]Corresponding authors: Salman Ghaffar (salman.ghaffar@ufz.de), Xiangqian Zhou (xiangqian.zhou@ufz.de)

Abstract

Distributed hydrological water quality models are increasingly being used to manage natural resources at the catchment scale but there are no calibration guidelines for selecting the most useful gauging stations. In this study, we investigated the influence of calibration schemes on the spatiotemporal performance of a fully distributed process-based hydrological water quality model (mHM-Nitrate) for discharge and nitrate simulations at Bode catchment in central Germany. We used a single- and two multi-site calibration schemes where the two multi-site schemes varied in number of gauging stations but each subcatchment represented different dominant land uses of the catchment. To extract a set of behavioral parameters for each calibration scheme, we chose a sequential multi-criteria method with 300.000 iterations.

For discharge (Q), model performance was similar among the three schemes (NSE varied from 0.88 to 0.92). However, for nitrate concentration, the multi-site schemes performed better than the single site scheme. This improvement may be attributed to that multi-site schemes incorporated a broader range of data, including low Q and NO3- values, thus provided a better representation of within-catchment diversity. Conversely, adding more gauging stations in the multi-site approaches did not lead to further improvements in catchment representation but showed wider 95% uncertainty boundaries. Thus, adding observations that contained similar information on catchment characteristics did not seem to improve model performance and increased uncertainty. These results highlight the importance of strategically selecting gauging stations that reflect the full range of catchment heterogeneity rather than seeking to maximize station number, to optimize parameter calibration.

Keywords

Multi-Multi-site calibration, Spatiotemporal validation, Hydrological water quality model, Uncertainty, Parameter transferability

Highlights:

- Single- and multi-site calibration approaches generally led to similar model performance for discharge (Q) at the catchment outlet.
- Influence of calibration stations on the spatiotemporal performance of a fully distributed process-based hydrological water quality model.
- The quality of the nitrate model simulation depends less on the number of calibration stations than on their representativeness of the catchment characteristics.

1. Introduction

Distributed hydrological water quality models provide crucial support for water management decisions. The models include many parameters that represent spatial variability in hydrological and biogeochemical processes at the catchment scale that cannot be measured directly in the field (Li et al., 2010). Thus, parameters must be calibrated to optimize model performance (Engel et al., 2007; Moriasi et al., 2012; Saraswat et al., 2015).

Most commonly, hydrological water quality models are calibrated using measurements made at the catchment outlet and may thus poorly simulate dynamics at sites within catchments, given spatial variability in conditions (Cao et al., 2006; Refsgaard et al., 2016, Refsgaard et al. 2022). As spatially structured discharge and water quality data become increasingly available, researchers are calling for multi-objective calibration strategies that allow for the inclusion of multiple sites, variables, and criteria (Daggupati et al., 2015; Efstratiadis and Koutsoyiannis, 2010; Khu et al., 2008).

However, to date, findings are mixed regarding the performance of single- versus multi-site calibration techniques. Many studies have found that, for catchment outlets, multi-site calibration yields more accurate results than does single-site calibration (e.g., Ghaffar et al., 2021; Her and Chaubey, 2015; Jiang et al., 2015; Zhang et al., 2008). For example, Shrestha et al. (2016) found such to be the case for a SWAT model (Arnold et al., 2012; Arnold et al., 1998) simulating total nitrogen (TN) and total phosphorus (TP) loads. Ghaffar et al. (2021) reported the same for a HYPE model (Lindström et al., 2010) seeking to replicate nitrate ($\mathrm{NO_3^-}$) and TP concentrations across a suite of monitoring stations in central Germany’s Selke catchment.

In contrast, several other studies have found that performance was largely equivalent for multi-site and single-site calibration techniques (e.g., Franco et al., 2020; Lerat et al., 2012; Wu et al., 2022a). They explained the unimproved model performance with high degree of similarity between flow data used to evaluate the model performance (Lerat et al., 2012), errors in boundary conditions as well as in representations of spatially structured hydrogeological properties (Wang et al., 2012) and hydrological processes (Wu et al., 2022a). However, it is important to note that previous studies have largely utilized semi-distributed hydrological and water quality models (e.g., SWAT: (Leta et al., 2017; Zhang et al., 2008) and HYPE: (Ghaffar et al., 2021; Jiang et al., 2015) and that station choice has frequently been driven by availability. Guidance is lacking when it comes to selecting the most useful gauging stations when calibrating fully distributed hydrological water quality models.

Compared to their lumped and semi-distributed counterparts, fully distributed hydrological water quality models incorporate detailed spatial information for sites within catchments while also including a broader range of parameters (Khu et al., 2008; Refsgaard, 1997). The applicability of parameters across spatial and temporal scales (i.e., parameter transferability) presents a major challenge for the construction of distributed hydrological water quality models (Beven, 2001; Samaniego et al., 2010). Parameters defined using information from calibration locations can be applied to other locations using a process called regionalization, as per Bloschl and Sivapalan (1995). Regionalization can be based on spatial proximity (Oudin et al., 2008a; Parajka et al., 2005), similarity in climatic and catchment characteristics (Beck et al., 2016; Merz and Blöschl, 2004; Oudin et al., 2008b; Parajka et al., 2005), and non-linear transfer functions that relate the parameters to catchment characteristics (e.g., land use, soil type, and geological type) (Hundecha and Bárdossy, 2004; Pokhrel et al., 2008; Wagener and Wheater, 2006). Samaniego et al. (2010) specifically developed a multi-scale parameter regionalization (MPR) method, whose appeal stems from the fact that only the coefficients in the transfer functions (i.e., the global parameters) need calibration, and not the parameters for each grid, substantially reducing the dimensionality of the calibrated parameters (Parajka et al., 2013; Singh et al., 2014). When model parameters are tied to catchment characteristics, calibration data drawn from diverse gauging stations are assumed to better represent within-catchment heterogeneity and to enhance model performance at spatial scales. However, little is known about the impact of different calibration schemes on the spatial and temporal performance of fully distributed hydrological water quality models.

Hydrological water quality models are typically developed using current knowledge about the physical and chemical processes taking place in the focal catchment, an endeavor that inherently involves simplifications and assumptions (Beven, 2007; Gupta et al., 2005). Uncertainty in model simulations is rooted in uncertainty from the measurement data, used as input and for calibration, as well as from model structure and parameterization (Vrugt et al., 2005; Wagener and Gupta, 2005). Such is especially true for spatially distributed hydrological water quality models, which contain more parameters than those of a lumped or semi-distributed model. While the hydrological modelling community has spent considerable time and effort designing uncertainty analysis techniques, the latter are rarely applied to distributed process-based hydrological water quality models, perhaps due to model complexity (Wellen et al., 2015). In addition, contrasting estimates of model simulation uncertainty have been obtained with single- versus multi-site calibration techniques. Jiang et al. (2015) found that, compared to single-site calibration, multi-site calibration reduced the uncertainty around estimates of Q and $\mathrm{NO_3^-}$ concentrations in the HYPE model. In contrast, Her and Chaubey (2015) found the opposite effect for Q estimates from a SWAT model: better performance was obtained using single-site than multi-site calibration. Finally, Shrestha et al. (2016) reported mixed results: for a SWAT model, single-site calibration resulted in less uncertainty for simulated Q values, while multi-site calibration accomplished the same for simulated TN and TP loading values. Thus, there is a pressing need to explore the impact of multi-calibration techniques on the uncertainty associated with fully distributed models.

Recently, Yang et al. (2018) developed a fully distributed hydrological water quality model (mHM-Nitrate) that is based on both the mesoscale hydrological model (mHM) (Samaniego et al., 2010) and the HYPE model (Lindström et al., 2010). The mHM-Nitrate model appears to successfully handle different catchment characteristics (Wu et al., 2022b; Yang et al., 2019a), but it is unknown how well it deals with parameter transferability across space. Our study's overarching aim was to evaluate the effects of different calibration schemes on the spatiotemporal performance of the mHM-Nitrate model. The specific objectives were as follows: (i) to evaluate and compare three calibration schemes that differed in gauging station number and representation of within-catchment diversity (e.g., land use and stream order); (ii) to assess parameter transferability across space under the three calibration schemes using $\mathrm{NO_3^-}$ data from a large number of sampling locations; and (iii) to examine the effects of the three calibration schemes on the degree of uncertainty associated with simulated $\mathrm{NO_3^-}$ concentrations. Ideally, the study's results should help guide the choice of effective calibration schemes, depending on the availability of Q and water quality data.

2. Study area and methods

2.1 Study area

The Bode catchment has a area of 3,200 km^2 and is located in central Germany (Figure 1). It is part of the Harz/Central German Lowland Observatory, within the broader TERENO Earth observation network focused on integrated, multi-scale monitoring and intensive research (Wollschläger et al., 2016). There is dramatic spatial heterogeneity across the catchment, which extends from the Harz Mountains in the southwest to the lowlands of central Germany in the northeast. There is also a marked elevational gradient, ranging from 1,142 m above sea level (a.s.l.) at Brocken, the highest peak in the Harz Mountains, to 70 m a.s.l. in the central lowlands. These extremes are reflected in dramatic differences in mean annual precipitation at these two locations, equal to 1,500 mm and 500 mm, respectively (climatic data: 1990–2019). In the mountains, mean monthly temperature ranges from -0.4℃ in January to 16.6℃; for the lowlands, these figures are 1.3℃ and 18.9℃, respectively. In the mountains, land surfaces are dominated by forests, with some pastures (10%), agricultural fields (8%), and urban areas and lakes (7%). In the lowlands, land surfaces are largely dedicated to cultivating crops (81%), primarily winter wheat, winter barley, rapeseed, and sugar beet. There is much less representation of other land use categories: forests (7%), pastures (3%), and urban areas and small lakes (9%) (Figure 1a). The predominant soil types in the mountains and lowlands are cambisols and chernozems, respectively.

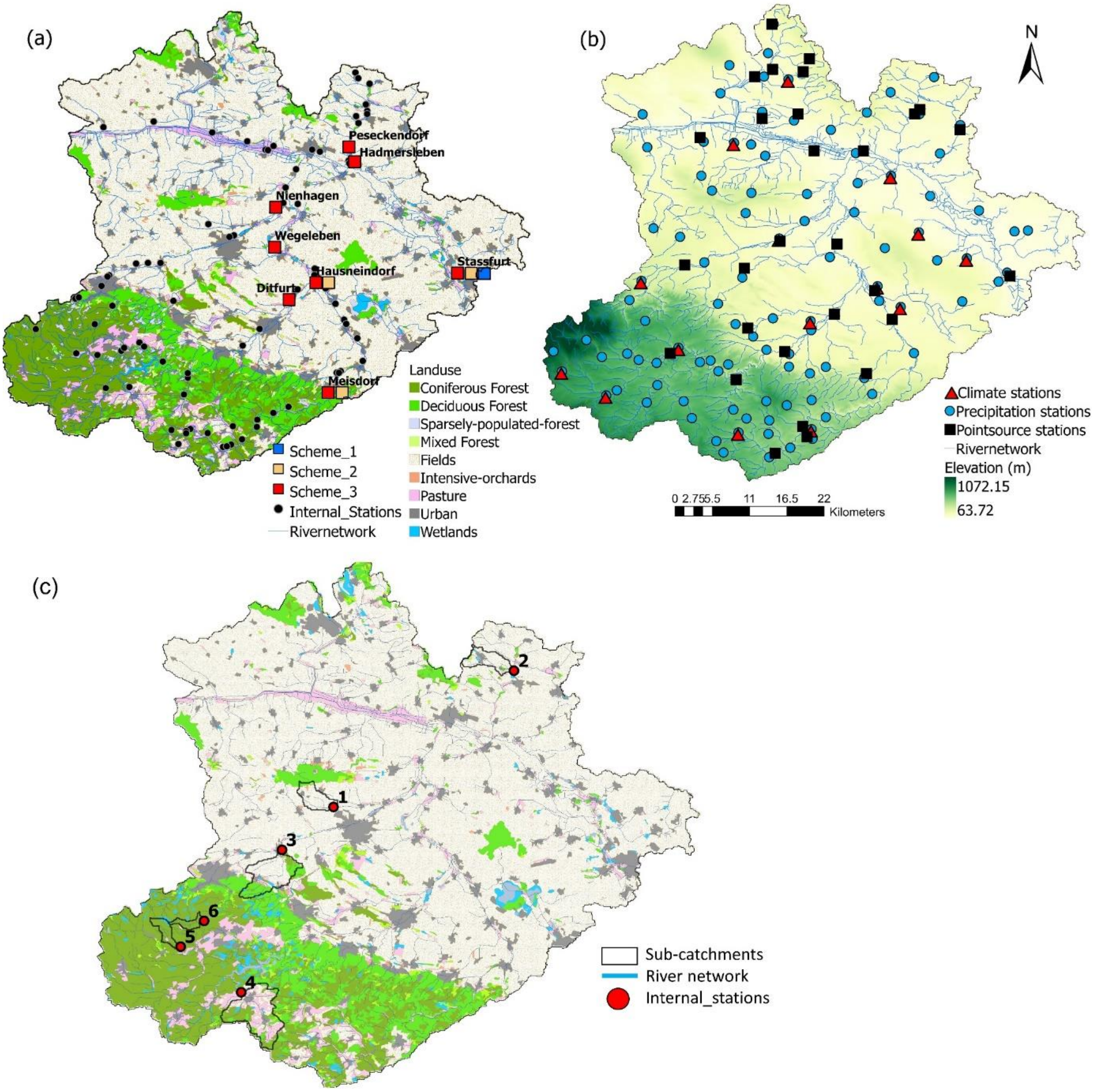

Figure 1. Maps of the Bode catchment showing (a) land use, the gauging stations, and the spatially distributed sampling locations as well as (b) elevation and the meteorological stations and (c) location of 6 internal stations presented in section 3.2.

We gathered observations of daily precipitation, daily temperature (maximum, mean, and minimum), and potential evapotranspiration to use as model input. These measurements spanned the period between 1993–2019 and were provided by the German Weather Service (DWD); they came from 78 rain gauges and 13 climate stations within the study area. To create the meteorological forcing dataset for the model, the daily precipitation and temperature data were spatially interpolated to 1 km × 1 km grid data using the External Drift Kriging method. This interpolation approach uses elevation, an external variable, to predict orographic effects on precipitation and temperature (Hundecha and Bárdossy, 2004). The daily potential evapotranspiration values were calculated using the Hargreaves and Samani (1985) method and interpolated at the same scale of spatial resolution.

To set up the mHM-Nitrate model, several sources of geographical data were used. Elevation measurements (spatial resolution: 90 m × 90 m) were obtained from the Shuttle Radar Topography Mission (SRTM) (Jarvis, 2008). The digitized geological map and the soil map (scale: 1:1,000,000) were provided by the German Federal Institute for Geosciences and Natural Resources (BGR) (https://produktcenter.bgr.de; last accessed 1 June 2020). The land cover data came from CORINE Land Cover 2012, which contains information on land cover/land use in the year 2012 (https://gdz.bkg.bund.de/index.php/default/open-data.html; last accessed 1 June 2020). These datasets were resampled to generate model input (spatial resolution: 100 m × 100 m).

For model calibration and validation, we used measurements of Q and $\mathrm{NO_3^-}$ concentrations from eight gauging stations. Daily measurements of Q at these stations were provided by the State Agency for Flood Protection and Water Management of Saxony-Anhalt (LHW) (http://gldweb.dhi-wasy.com/gld-portal/; last accessed 10 April 2020). High-frequency (15 minutes) $\mathrm{NO_3^-}$ concentrations for four stations (Meisdorf, Hausneindorf, Hadmersleben, and Stassfurt) between 2010 and 2019 were obtained from the Helmholtz Center for Environmental Research—UFZ; we aggregated these high-frequency measurements to daily values. For the other four stations (Ditfurt, Wegeleben, Nienhagen, and Peseckendorf), the $\mathrm{NO_3^-}$ data were low-frequency measurements collected every two weeks to every two months from 1994 to 2019 by LHW (http://gldweb.dhi-wasy.com/gld-portal/; last accessed 10 April 2020). Finally, we also gathered low-frequency $\mathrm{NO_3^-}$ measurements from 94 sampling locations to spatially validate the mHM-Nitrate model. The catchment characteristics at these sites are described in the Supplementary Materials (Table S1).

2.2 mHM-Nitrate model

The mHM-Nitrate model takes a grid-based approach and seeks to reliably represent complex processes (Yang et al., 2018). It includes the following hydrological processes: canopy interception, snow accumulation and melt, evapotranspiration, infiltration, soil moisture dynamics, runoff generation, percolation, and flood routing along the river network. The model incorporates nitrate processes described in the HYPE model (Lindström et al., 2010) as well as others: NO_3^- retention in deep groundwater, NO_3^- dynamics associated with spatially distributed crop rotations, and temporally variable point-source inputs of NO_3^-. These processes are fully integrated into hydrological cycling. Major N inputs include wet atmospheric deposition via precipitation, fertilizer and manure application, and plant/crop residues. For each soil layer, four N pools are defined—active solid organic N, inactive solid organic N, dissolved organic N, and dissolved inorganic N, along with soil N processes, namely denitrification, plant/crop uptake, and transformations among the four N pools. In-stream N transformations include denitrification, primary production, and mineralization. A more detailed description of the mHM-Nitrate model can be found in Yang et al. (2018), and the source code can be found in Yang and Rode (2020).

2.3 Model set-up

The mHM-Nitrate model was set up using available hydrometeorological and geographical data for 1993–2019 and was run at a daily time step (Table 1). To exclude the effects of a reservoir in the Harz Mountains, we used daily Q and NO_3^- concentrations measured at a downstream gauging station (Thale) as input.

Table 1. Description of the spatiotemporal data from the Bode catchment used as input for mHM-Nitrate model set-up.

General data type	Specific data type	Resolution	Source
Geographical	Digital elevation model	100 m × 100 m	SRTM
	Land use		CORINE Land Cover 2012
	Geological history		BGR

	Soil type		
Meteorological	Daily precipitation and mean air temperature	1km×1 km	DWD
Agricultural practices	Manure and inorganic fertiliser application, timing and amount of fertilization, sowing and harvesting	Land-use dependent	Field survey and scientific literature
Soil nitrogen content	Initial N storage		Scientific literature
Sewage treatment plants	N load	Daily time step	Operating reports from sewage treatment plants

2.4 Calibration schemes

The parameters of the mHM-nitrate model were related to catchment characteristics. Based on catchment characteristics, land use, mean NO_3^- concentration, and stream order, three calibration schemes were designed. In scheme. Scheme 1 used only data from the catchment outlet station (Stassfurt). Scheme 2 used data from Stassfurt and two gauging stations upstream (Meisdorf and Hausneindorf) (Table 1 and Figure 2). Scheme 3 used data from Stassfurt and seven gauging stations upstream (Figures 1a and 2).

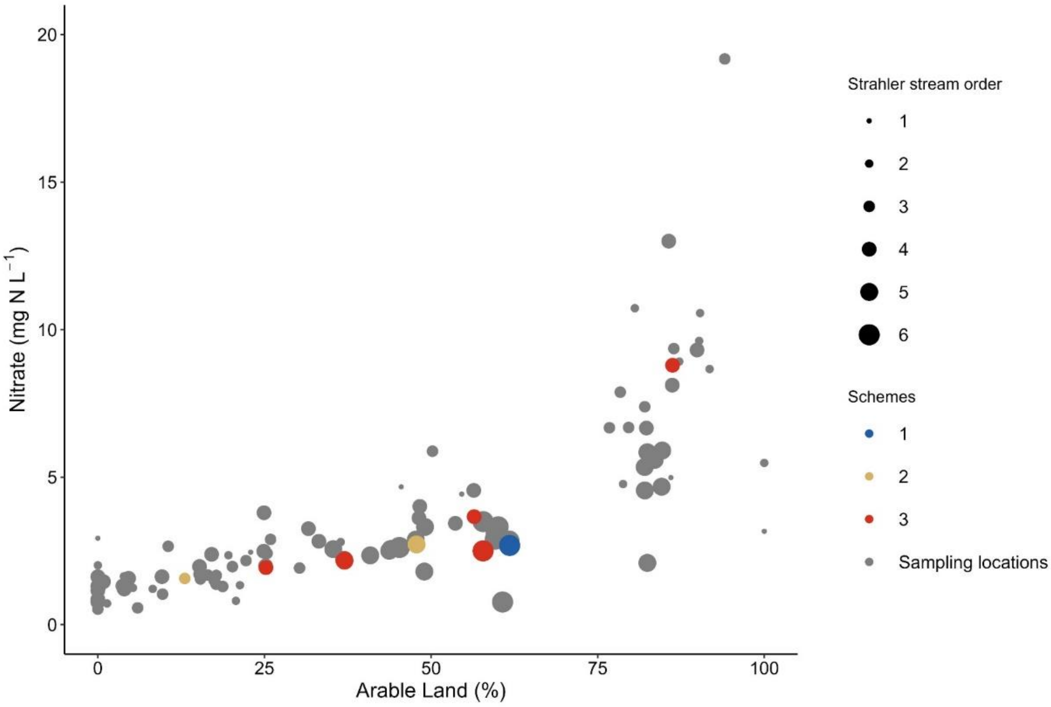

Figure 2. Relationship between nitrate concentration and share of arable land use with information on stream order of the sub-catchments represented by the eight gauging stations and the 94 spatially distributed sampling locations. Station inclusion within the calibration schemes is indicated (with higher-level schemes including the stations found in lower-level schemes).

The eight gauging stations used in scheme 3 reflect different combinations of land use and meteorological conditions found in the Bode catchment (Table 2). Compared to scheme 2, scheme 3 includes data from five additional gauging stations that are associated with larger streams (stream order: 4–6) (Krabbenhoft et al., 2022). There are four main gauging stations along the Bode River: Ditfurt (upstream), Wegeleben (intermediate stream), Hadmersleben (downstream), and Stassfurt (catchment outlet). Ditfurt and Wegeleben are in a forest-dominated subcatchment, while Hadmersleben and Stassfurt locate in an area dominated by farmlands. The headwaters of the Selke and Holtemme Rivers are located in the mountains, a region with extensive forests (71.9%) and low $\mathrm{NO_3^-}$ concentrations. In contrast, the lowlands are covered by agricultural fields, and $\mathrm{NO_3^-}$ concentrations are high. The Meisdorf station is located in the mountainous Upper Selke, while the Hausneindorf station is the Selke's outlet, an area with a mixture of forests and farms. The Nienhagen station is the Holtemme outlet, whose upstream and downstream areas are dominated by forest and agricultural surfaces (Ehrhardt et al., 2019), respectively. At Nienhagen, Q values are heavily affected by the presence of weirs (Kunz et al., 2017). The Peseckendorf station is the outlet of the Geesgraben stream, which merges into the Bode after Hadmersleben; the surrounding area is predominantly covered by crops (88.8%).

Table 2. Subcatchment characteristics for the eight gauging stations. Abbreviations: Subcatch = subcatchment; Precip = precipitation; Q = discharge; and $\mathrm{NO_3^-}$ = nitrate concentration range (mean).

Station	Subcatch	Area ($\mathrm{km^2}$)	Elevation (m)	Precip (mm $\mathrm{y^{-1}}$)	% Forest	% Farm land	Stream order	Q (mm $\mathrm{y^{-1}}$)	$\mathrm{NO_3^-}$ (mg N $\mathrm{L^{-1}}$)
Meisdorf	Selke	180	199–597	690	73.1	12.8	3	186	0.01–5.14 (1.57)
Hausneind.	Selke	458	106–597	590	37.8	48.5	5	99	0.44–8.55

									(2.73)
Ditfurt	Bode	714	107–1072	783	56.4	25.3	4	211	1.30–2.90 (1.93)
Wegeleben	Bode	1230	94–1072	698	46.9	36.9	5	166	1.10–4.75 (2.24)
Nienhagen	Holtemme	260	94–931	678	31.6	54.2	4	162	1.22–10.4 (4.59)
Peseckend.	Geesgraben	137	76–200	546	3.0	88.8	4	58	0.77–17.0 (8.80)
Hadmersl.	Bode	2620	76–1072	639	29.2	56.6	6	132	0.47–11.0 (2.51)
Stassfurt	Bode	3179	66–1072	617	24.7	61.6	6	114	0.46–8.10 (2.68)

2.5 Model calibration and validation

Parameter sensitivity analysis was performed using the Morris method (Morris, 1991). We calculated the elementary effect (EE) of each parameter using the Sensitivity Analysis For Everybody toolbox (SAFE; (Pianosi et al., 2015). We identified the eight most sensitive hydrological parameters and the six most sensitive water quality parameters (Table S1) based on the ranked values of the sensitivity indices (absolute mean and standard deviation of EE). This suite of parameters was then used in mHM-Nitrate model calibration. A more detailed description of the parameter sensitivity analysis is available in Zhou et al. (2022).

Instead of using an optimization algorithm, like a dynamically dimensioned search (DDS) (Tolson and Shoemaker, 2007), we opted for a sequential multi-criteria method (Wu et al., 2021) to filter out sets of behavioral parameters for each calibration scheme. This process involved two steps. During the first step, 300,000 parameter sets were created for the eight sensitive hydrological parameters. Next, the best 100 parameter sets were selected for each calibration scheme, a decision guided by the ranks of both the Nash-Sutcliffe coefficient (NSE) and percent bias (PBIAS) values for Q at the relevant gauging stations. During the second step, 300,000 parameter sets were generated for the six sensitive water quality parameters, which were combined with the 100 best Q parameter sets. For each calibration scheme, we selected the best 100 parameter sets from this second step based on the ranks of the NSE and PBIAS values for Q and $\mathrm{NO_3^-}$ concentrations for the relevant gauging stations. The preliminary calibration results revealed that 300,000 iterations allowed the objective function values to converge upon minimum values. This procedure made it possible to compare the three calibration schemes, as this allows each calibration scheme to achieve its own best performance from the same parameter space.

Following the split-sample test, this calibration procedure was applied to the mHM-Nitrate model incorporating Q and $\mathrm{NO_3^-}$ concentrations from 2011 to 2014. Each calibration scheme was validated (time period: 2015–2019) at all eight gauging stations for both Q and $\mathrm{NO_3^-}$ concentrations (Table 3). NSE and PBIAS were used as performance evaluation criteria. However, it is difficult to draw conclusions about the relative performance of calibration schemes when sample size is small. Therefore, we carried out spatiotemporal validation of the model using $\mathrm{NO_3^-}$ data from the 94 spatially distributed sampling locations (i.e., low-frequency measurements for 1994–2019). In this case, only PBIAS was used to evaluate model performance, which is satisfactory when values are less than 35%, according to Moriasi et al. (2015).

Table 3. Discharge (Q) and nitrate ($\mathrm{NO_3^-}$) concentration data used in model calibration and validation for the three calibration schemes.

Scheme	Calibration	Validation	
	2011–2014	Q and $\mathrm{NO_3^-}$ (2015–2019)	$\mathrm{NO_3^-}$ (1994–2019)
1	Q and $\mathrm{NO_3^-}$ at Stassfurt	Q and $\mathrm{NO_3^-}$ at Stassfurt, Meisdorf, Hausneindorf, Nienhagen, Peseckendorf, Ditfurt, Wegeleben, Hadmersleben	$\mathrm{NO_3^-}$ at 94 sampling locations
2	Q and $\mathrm{NO_3^-}$ at Stassfurt, Meisdorf, Hausneindorf		
3	Q and $\mathrm{NO_3^-}$ at Stassfurt, Meisdorf, Hausneindorf, Nienhagen, Peseckendorf, Ditfurt, Wegeleben, Hadmersleben		

2.6 The value of added calibration stations on parameter distributions and model performance

To assess the value of additional calibration stations on the identification of the model, the cumulative parameter distributions were computed for all calibration schemes utilizing the top 100 model runs from the second calibration phase of calibration schemes. To the extent that additional calibration stations change the cumulative distribution function of the individual model parameters defined due to model calibration. Significant differences in these cumulative distribution functions can be tested statistically and should allow an assessment of the added value of a modified data set for model identification. In this study, we determined the statistical significance of the differences in these cumulative distribution functions between calibration schemes using the two-sample Kolmogorov-Smirnov (Conover, 1999) test (D):

$$D = max|F(\theta_i) - G(\theta_i)| \quad (1)$$

where $F(\theta_i)$ and $G(\theta_i)$ are the empirical cumulative distribution functions of the parameter θ_i for calibration scheme 1(2) and 2(3). The null hypothesis is that the two samples are from the same continuous distribution. If D is closer to zero, it indicates that the probability of the two samples being drawn from the same population is higher. Moreover, the two-sample Kolmogorov-Smirnov test generates a p-value that corresponds to the calculated D statistic. A higher p-value (> 0.05) provides stronger support for the null hypothesis. The relative occurrences of certain, significant, KS statistics can be inspected by means of cumulative frequency plots. As different calibration stations result in varying levels of model parameters, distinct cumulative frequency curves of model performance will be observed.

2.7 Uncertainty analysis

To compare model uncertainty among the three calibration schemes, 95% uncertainty boundaries were calculated based on the 2.5th and 97.5th percentiles of the cumulative distributions for the best 100 model runs from the second calibration step. The R-factor quantifies differences between observed and simulated data and is calculated by dividing the average distance between the upper and lower 95% uncertainty boundaries by the standard deviation of the observed data (Abbaspour et al., 2007). The R-factor expresses the width of the 95% uncertainty and a value less than 1 is being desirable. The uncertainty analysis was performed for both Q and $\mathrm{NO_3^-}$ concentrations at all the gauging stations included in schemes 2 and 3. We compared model uncertainty for schemes 2 and 3 by comparing results for the stations shared by the schemes (Stassfurt, Hausneindorf, and Meisdorf).

3. Results

The mHM-Nitrate model was calibrated using the three schemes, resulting in different patterns of performance (parameter description: Table S1).

3.1 Model performance at gauging stations

The model performance of discharge (Q) for at the catchment outlet (Stassfurt station) was similar across the three calibration schemes (NSE—scheme 1: 0.82, scheme 2: 0.87, and scheme 3: 0.88; PBIAS—scheme 1: 0.30%, scheme 2: 0.0%, and scheme 3: -8.60%; Table 4). During the calibration period, at the Meisdorf and Hausneindorf stations, performance was lower for scheme 3 than for scheme 2 (NSE—scheme 2: 0.58 to 0.69 vs. scheme 3: 0.53 to 0.66; PBIAS—scheme 2: -7.80% to -23.5% vs. scheme 3: -20.2% to -32.0%). During the validation period, water balance was well captured across all the calibration schemes and gauging stations, with the exception of Nienhagen (PBIAS—scheme 1: -3.7% to 7.1%, scheme 2: -7.7% to 2.6%, and scheme 3: -12.7% to 1.4%). Performance was lowest at the Peseckendorf and Nienhagen stations across the three schemes, albeit lower for scheme 1 than for schemes 2 and 3 (NSE—scheme 1: -0.34 to 0.13 vs. scheme 2: 0.17 to 0.29 and scheme 3: 0.36 to 0.45; Table 4). It was also better at the Stassfurt, Meisdorf, and Hausneindorf stations during the validation period than during the calibration period across all calibration schemes (NSE—lower ranges: 0.53–0.88 and upper ranges: 0.71–0.92). The mean absolute PBIAS values for Q at all validation stations were 8.4%, 7.5%, and 9.2% for scheme1, scheme2, and scheme3, respectively.

Model performance of $\mathrm{NO_3^-}$ concentration at the catchment outlet Stassfurt station decreased from Scheme 1 to 2 and 3 during the calibration period (NSE—scheme 1: 0.67, scheme 2: 0.64, and scheme 3: 0.62; PBIAS—scheme 1: 0.40%, scheme 2: -6.90%, and scheme 3: 7.10%). Also during the calibration period, model performance at the Meisdorf station was better at scheme 2 (PBIAS: -

2.60%) than scheme 3 (PBIAS: -23.2%). At Hausneindorf, scheme 3 yielded better performance than did scheme 2 (PBIAS: -7.90% vs. 1.20%, respectively). During the validation period, performance was better at scheme 2 than at scheme 1 for all the gauging stations except for Nienhagen station, with PBIAS values in ranges —scheme 2: 1.8–33.9% and scheme 1: -10.1–23.3%, respectively. While NO_3^- concentration model performance decreased from Scheme 2 to 3 at all gauging stations except Nienhagen station, with larger absolute PBAIS values in Scheme 3 than Scheme 2. The mean absolute PBIAS values for NO3 were 15.7%, 9.5%, and 13.8% for scheme1, scheme2, and scheme3, respectively. These findings provide evidence that scheme 2 is the most promising option. Additionally, the results indicate that the model performance is categorized as good for Q and very good for NO_3^-.

Table 4. Model performance for discharge (Q) and nitrate (NO_3^-) concentrations during the calibration and validation periods across the three calibration schemes and their associated gauging stations.

Schemes	Stations	Q				NO_3^-			
		Calibration		Validation		Calibration		Validation	
		NSE	PBIAS (%)	NSE	PBIAS (%)	NSE	PBIAS (%)	NSE	PBIAS (%)
1	Stassfurt	0.82	0.30	0.92	4.20	0.67	0.40	0.33	12.5
	Meisdorf	-	-	0.71	7.10	-	-	0.32	33.9
	Hausneindorf	-	-	0.77	0.70	-	-	-0.08	8.10
	Wegeleben	-	-	0.92	-3.70	-	-	-1.19	16.4
	Hadmersleben	-	-	0.93	2.90	-	-	0.01	20.9
	Peseckendorf	-	-	-0.34	-3.50	-	-	-3.84	23.0
	Ditfurt	-	-	0.97	-0.20	-	-	-4.36	8.80
	Nienhagen	-	-	0.13	44.7	-	-	-0.66	1.80
2	Stassfurt	0.87	0.00	0.88	0.60	0.64	-6.90	0.23	9.30
	Meisdorf	0.58	-23.5	0.72	-1.70	0.66	-2.60	0.67	-10.1
	Hausneindorf	0.69	-7.80	0.76	-3.00	0.27	-7.90	0.31	-4.00
	Wegeleben	-	-	0.92	-3.10	-	-	-0.14	4.00
	Hadmersleben	-	-	0.92	2.60	-	-	0.26	14.3
	Peseckendorf	-	-	0.17	-7.70	-	-	-2.72	23.3
	Ditfurt	-	-	0.96	2.20	-	-	-2.18	1.60
	Nienhagen	-	-	0.29	38.8	-	-	-0.17	-9.10
3	Stassfurt	0.88	-8.60	0.90	1.40	0.62	7.10	-0.33	16.8
	Meisdorf	0.53	-32.0	0.71	-12.0	0.53	-23.2	0.71	-14.0
	Hausneindorf	0.66	-20.2	0.73	-12.7	0.31	1.20	0.20	-7.80

Wegeleben	0.87	-12.6	0.92	-5.60	0.37	-9.50	-1.39	11.0
Hadmersleben	0.87	-9.10	0.92	-0.90	0.21	14.0	-0.49	23.8
Peseckendorf	0.56	-21.6	0.45	-9.80	-0.44	-15.6	-1.70	24.7
Ditfurt	0.94	-3.40	0.96	1.00	0.35	-9.80	-3.56	9.20
Nienhagen	0.68	6.00	0.36	29.9	0.59	-14.2	0.39	-3.20

The seasonal dynamics of Q were captured by scheme 2 at its three gauging stations during both the calibration and validation periods as well as during low- and high-flow conditions (Figures 3a, 3c, and 3e). The same was true for the seasonal dynamics of $\mathrm{NO_3^-}$ concentrations (i.e., high values during high-flow periods and low values during low-flow periods; Figures 3b, 3d, and 3f). In addition, over the period from 2011 to 2019, $\mathrm{NO_3^-}$ concentrations followed a constant seasonal pattern at the Meisdorf station (Figure 3b) but tended to decline at the Hausneindorf and Stassfurt stations (Figures 3d and 3f), which were well captured by the model. Model performance for $\mathrm{NO_3^-}$ concentrations was greatest at the Meisdorf station (NSE—calibration: 0.66 and validation: 0.67; Table 4). It was lowest at the Hausneindorf station (NSE—calibration: 0.27 and validation: 0.31; Table 4). At Stassfurt, Meisdorf, and Hausneindorf, model performance for $\mathrm{NO_3^-}$ concentrations were satisfactory (PBIAS ranged between -7.9% and 9.3% during calibration and validation).

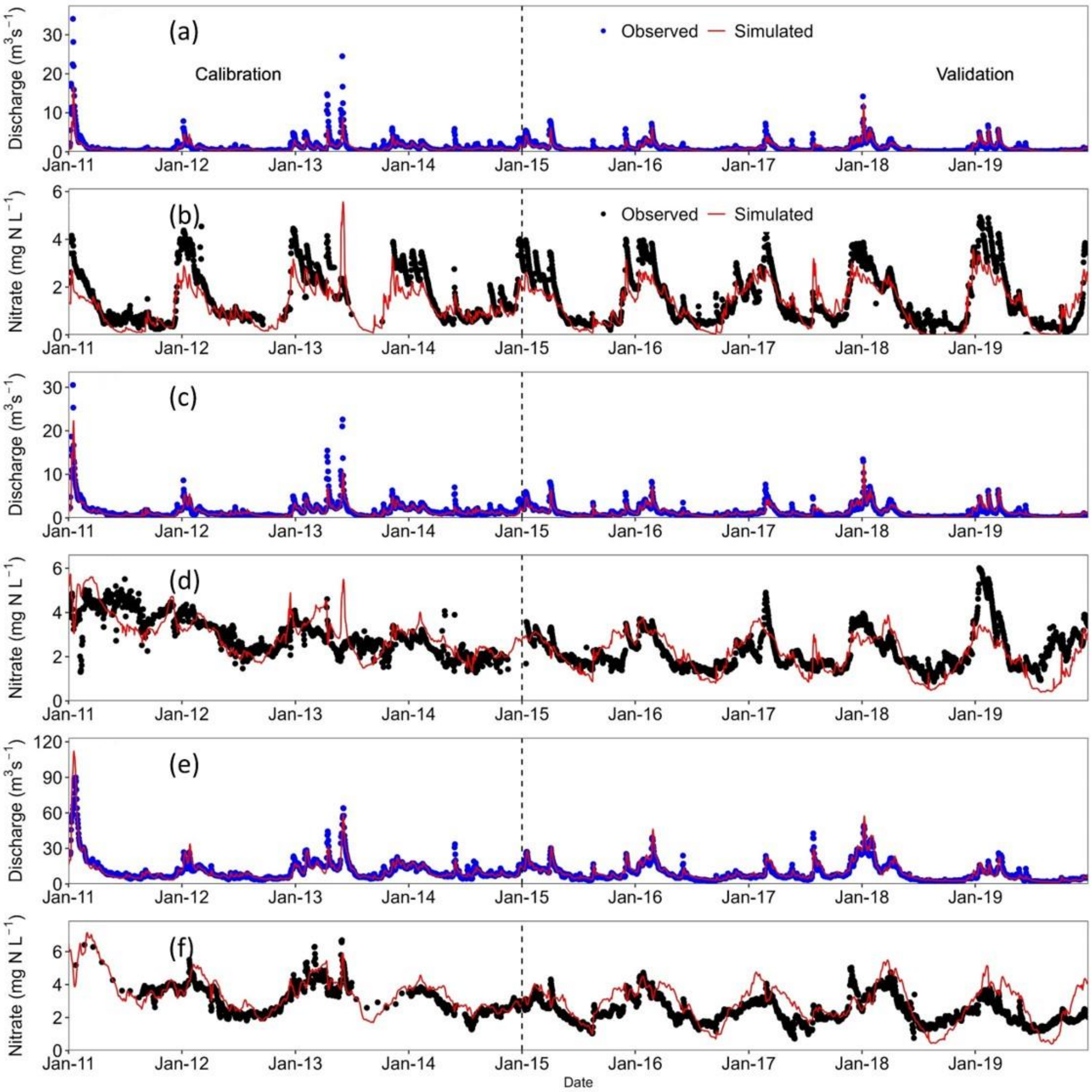

Figure 3. Observed and simulated Q and NO_3^- concentration at (a-b) Meisdorf, (c-d) Hausneindorf and (e-f) Stassfurt stations for calibration Scheme 2.

3.2 Model performance at spatially distributed sampling locations

We further tested how the calibration schemes affected model performance using NO_3^- data from the 94 spatially distributed sampling locations. Performance was generally better for scheme 2 than for scheme 1 (PBIAS ≤ 15.0%: 34 vs. 9 sampling stations, respectively, and PBIAS > 45%: 12 vs. 65 sampling stations, respectively) (Table 5). Performance was similar for schemes 2 and 3 (PBIAS ≤ 15.0%: 34 vs. 35 sampling locations, respectively).

Table 5. Frequency of sampling locations associated with different PBIAS ranges across the three calibration schemes.

PBIAS (%)	Scheme 1	Scheme 2	Scheme 3
0.00–15.0	9	34	35
15.1–25.0	9	19	16
25.1–35.0	8	17	20
35.1–45.0	3	12	10
> 45.1	65	12	13

We also examined how catchment characteristics might influence model performance by looking at the spatial distributions of the PBIAS values for all 94 sampling locations across the three calibration schemes (Figure 4). The model performance for $\mathrm{NO_3^-}$ concentration at each stream order and land use (farmland vs. forest) are shown in Figure S1. Overall, more locations showed a good level of performance (PBIAS $\leq$ 15.0%) at scheme 2 versus scheme 1; no such difference was seen between schemes 2 and 3. For example, in forested areas, scheme 2 demonstrated considerable improvement compared to scheme 1. By visually inspecting, there was no noticeable distinction between scheme 2 and 3 (Figure 4). More specifically, performance was better at scheme 2 than scheme 1 in areas dominated by farmlands for all stream orders (Figure S1). Additionally, performance was better for scheme 2 than scheme 3 except in the case of stream orders 2 and 4 in agricultural areas and stream order 5 in forested areas (Figure S1).

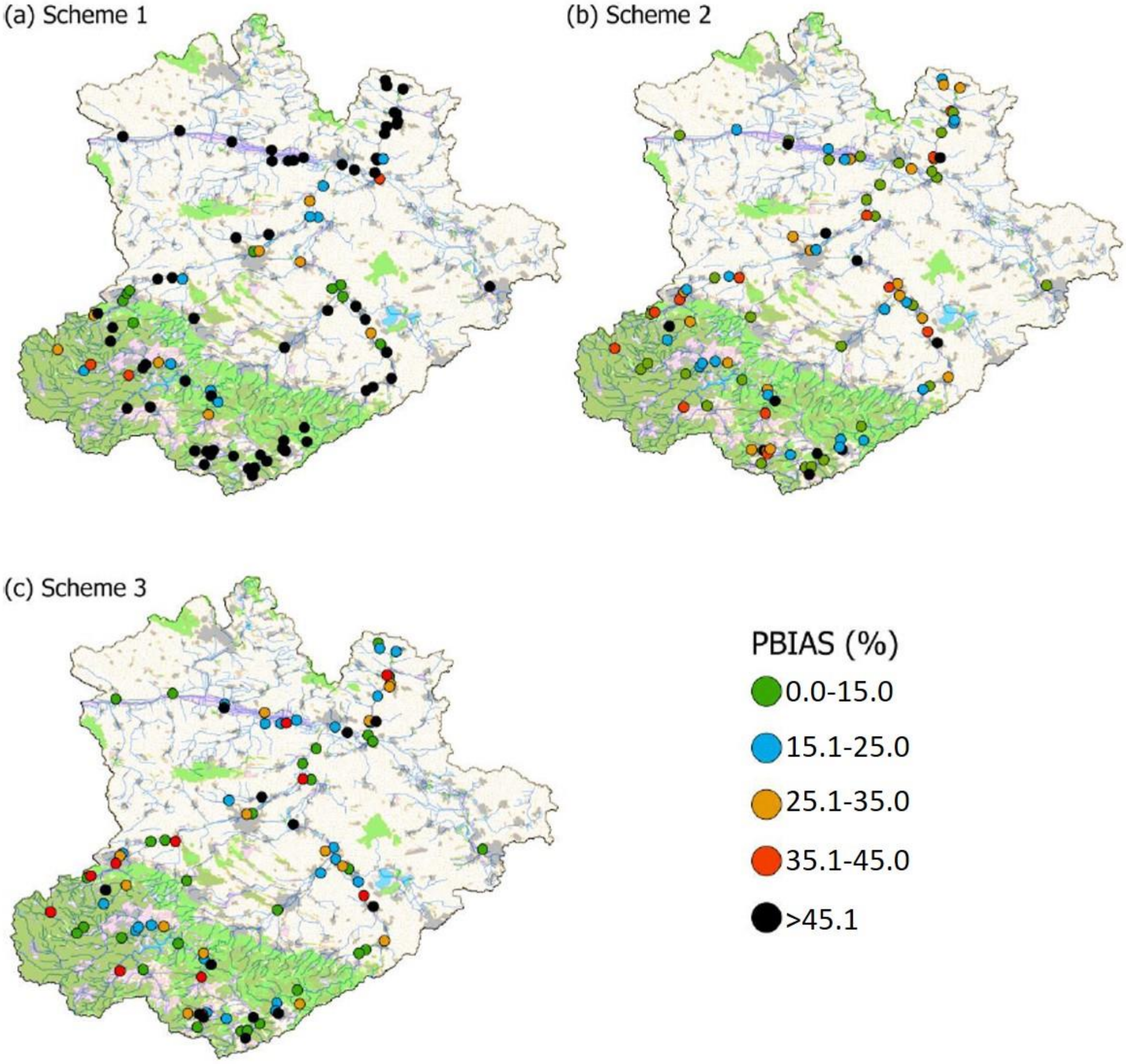

Figure 4. Performance of the mHM-Nitrate model for $\mathrm{NO_3^-}$ concentrations at the 94 spatially distributed sampling locations across the three calibration schemes.

We used the optimized parameter sets for scheme 2 to explore model performance in greater detail at six spatially distributed sampling locations that displayed distinct characteristics (map: Figure 1c; observed and simulated $\mathrm{NO_3^-}$ concentrations: Figure 6; PBIAS: Table 6). There was variation in the duration and frequency of the validation data for the six sampling locations. Seasonal patterns of $\mathrm{NO_3^-}$ concentrations were well captured by the model over different levels of $\mathrm{NO_3^-}$ (Figure 5), with PBIAS values ranging from -17.1% to 14.5% (Table 6). This result indicates that the mHM-Nitrate model was capable of representing $\mathrm{NO_3^-}$ dynamics within different subcatchments when scheme 2 was applied. The largest difference between mean observed and simulated $\mathrm{NO_3^-}$ concentrations occurred at $\mathrm{NO_3^-}$

sampling location 2 (Figure 5c) with PBIAS value of -17.1%, which represents an arable dominated sub-catchment. The best fit between mean observed and simulated NO_3^- concentration was found at NO_3^- sampling location 4 (Figure 5d; PBIAS = -9.3%), which is found in a mountainous sub-catchment that contains a mixture of farmland and pasture (Figure 1c).

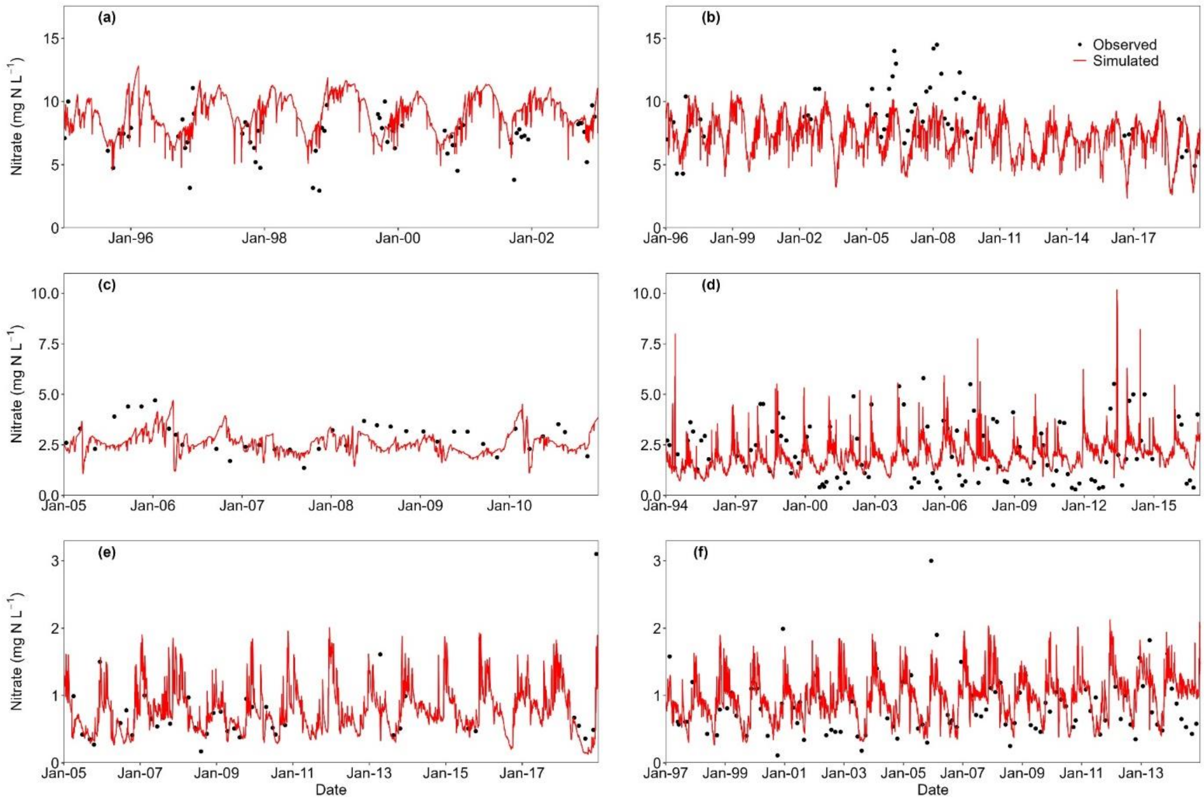

Figure 5. Observed and simulated nitrate (NO_3^-) concentrations (calibration scheme 2) for the six sampling locations displaying distinct characteristics.

Table 6. Summary of catchment characteristics represented by the six sampling locations, model performance for nitrate (NO_3^-) concentrations (PBIAS values), minimum and maximum values of simulated and observed NO_3^- concentrations at the sampling locations, and range (mean) of NO_3^- concentrations.

Sampling location	Sub-catchment area (km2)	Dominant land use	PBIAS (%)	Simulated NO_3^- concentration (mg N L^{-1})	Observed NO_3^- concentration (mg N L^{-1})
1=a	11.8	Arable (87.2%)	12.8	4.6-12.9 (9.1)	2.9-11.1 (7.4)
2=b	12.6	Arable (78.3%)	-17.1	2.4-10.8 (7.4)	4.3-14.5 (8.9)
3=c	26.4	Arable (53.6%) Forest (40.1%)	-12.6	1.1-4.7 (2.6)	1.4-4.7 (2.9)
4=d	37.1	Arable (22.2%) Pasture (29.0%)	-9.3	0.8-10.2 (2.0)	0.3-5.8 (2.2)
5=e	6.1	Forest (96.0%)	-11.7	0.1-2.0 (0.7)	0.2-3.1 (0.7)
6=f	3.9	Forest (100%)	14.5	0.3-2.1 (1.0)	0.1-3.0 (0.8)

3.3 Model parameter distributions

For the three calibration schemes, we constructed cumulative distribution functions for the most sensitive hydrological and water quality parameters using the best 100 model runs (Figure 6). From the results, it is clear that the hydrological parameters—infiltration shape factor (infil) and potential evapotranspiration (pet) differ significantly between schemes 1 and 2 as well as between schemes 2 and 3 ($p < 0.01$) (Figure 6 and Table 7). In the mHM-Nitrate model, soil infiltration is parameterized using the power function of soil saturation, whose exponent is determined by the infiltration shape factor (infil). Cuntz et al. (2015) reported that, as a parameter, infil is highly related to soil saturation, where higher infiltration occurs in mountain soils than in lowland soils. Because the Meisdorf station was included in scheme 2, a greater range of soil types were represented, allowing infil to be better defined. In contrast, scheme 1 averaged all the soil types present in the catchment, as reflected by the narrower ranges of infil for scheme 2 versus 1 (Figure 6). The cumulative distributions of four water quality parameters, namely in-stream denitrification rate (denitri), primary production rate (pprt), primary production coefficient in non-agriculture stream (pprt_na), and primary production coefficient in agriculture stream (pprt_agri), showed dissimilarities between scheme 1 and schemes 2 and 3. However, there were no differences in the cumulative parameter distributions between scheme 2 and scheme 3 ($p > 0.05$) (Figure 6 and Table 7). The four water quality parameters were better constrained for scheme 2 than scheme 1, as reflected by their narrower ranges in the former versus the latter (Figure 6). Yang et al. (2019b) found the control factors for denitri and pprt varied between the Meisdorf and Hausneindorf stations. At Meisdorf, both parameters have a strong correlation with stream discharge and benthic area, while at Hausneindorf they are highly correlated with terrestrial flows and fluxes. In summary, parameter distributions were dramatically affected by the increase in station number between scheme 1 and scheme 2. In contrast, the additional stations added in scheme 3 had little to no effect.

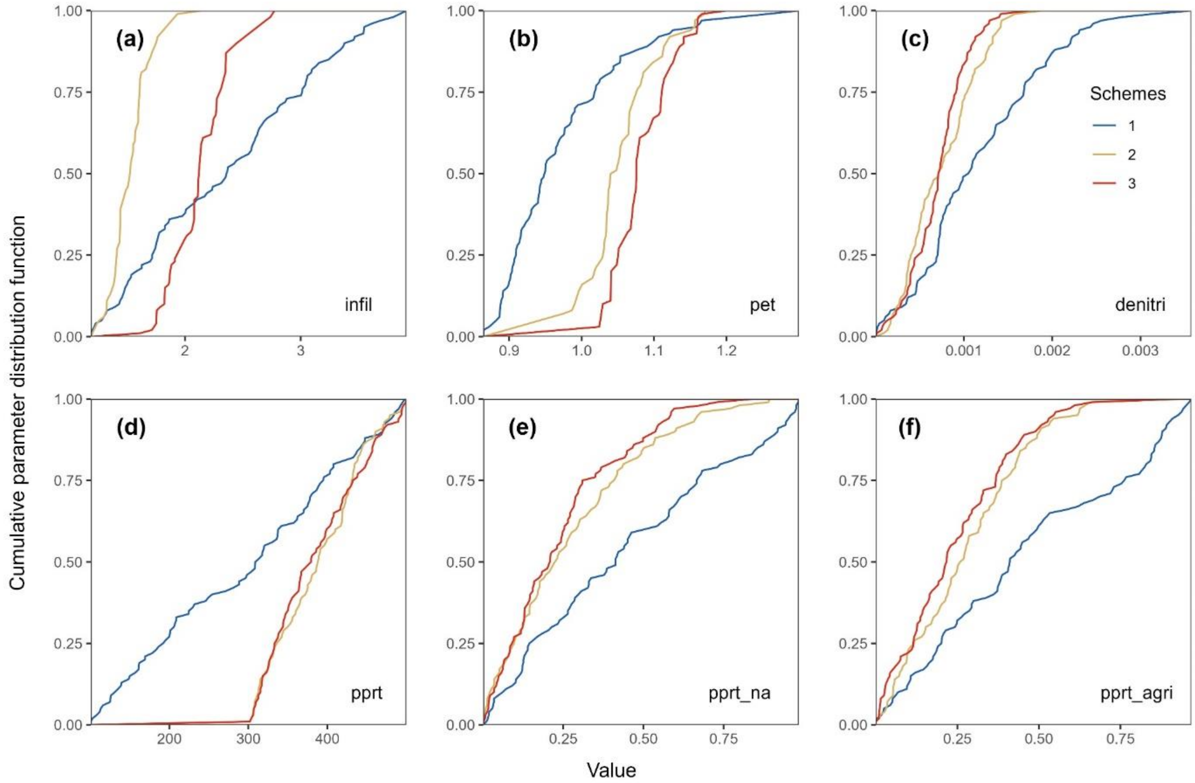

Figure 6. Cumulative distributions for the hydrological parameters infil (a) and pet (b) and four the water quality parameters, in-stream denitrification rate (denitri) (c), primary production rate (pprt) (d), primary production coefficient in non-agriculture stream (pprt_na) (e), and primary production coefficient in agriculture stream (pprt_agri) (f) across the three calibration schemes.

Table 7. Kolmogorov-Smirnov (KS) statistics and significance estimates for cumulative parameter distributions between calibration schemes.

Parameters	KS statistic and p-value		
	Scheme 1 and scheme 2	Scheme 1 and scheme 3	Scheme 2 and scheme 3
infil	0.64 (p<0.01)	0.38 (p<0.01)	0.88 (p<0.01)
pet	0.61 (p<0.01)	0.78 (p<0.01)	0.39 (p<0.01)
denitri	0.32 (p<0.01)	0.38 (p<0.01)	0.16 (p>0.05)
pprt	0.46 (p<0.01)	0.46 (p<0.01)	0.09 (p>0.05)
pprt_na	0.28 (p<0.01)	0.34 (p<0.01)	0.12 (p>0.05)
pprt_agri	0.31 (p<0.01)	0.37 (p<0.01)	0.14 (p>0.05)

3.4 Uncertainty analysis—nitrate concentrations

We calculated the 95% uncertainty boundaries for simulated daily NO_3^- concentrations at the Meisdorf, Hausneindorf, and Stassfurt stations for schemes 2 and 3 (Figure 7). The associated R-factors are given in Table 8. The 95% uncertainty boundaries for simulated daily Q associated with schemes 2 and 3 are available in the Supplementary Materials (Figure S3). Whether under low- or high-flow conditions, 95% uncertainty boundaries for daily NO_3^- concentrations were narrower for scheme 2 than for scheme 3 (Figure 7). For instance, they were nearly twice as wide for scheme 3 than scheme 2 at Hausneindorf (R-factor = 4.13 vs. 2.18, respectively) and Stassfurt (R-factor = 4.52 vs. 2.79, respectively) (Table 8). Furthermore, over 60% of the observed NO_3^- concentrations lay within the 95% uncertainty boundaries for scheme 2. When scheme 2 was used, the Meisdorf station, located in a forested subcatchment, displayed lower levels of uncertainty than did the Hausneindorf and Stassfurt stations, which are found in a subcatchment dominated by farmland. The same was also true for scheme 3. This finding was reflected in the narrower 95% uncertainty boundaries for Meisdorf versus Hausneindorf and Stassfurt (Figures 7a-b vs. 7c-f), as well as in the lower R-factor values for Meisdorf (scheme 2 = 0.92; scheme 3 = 1.08; Table 8).

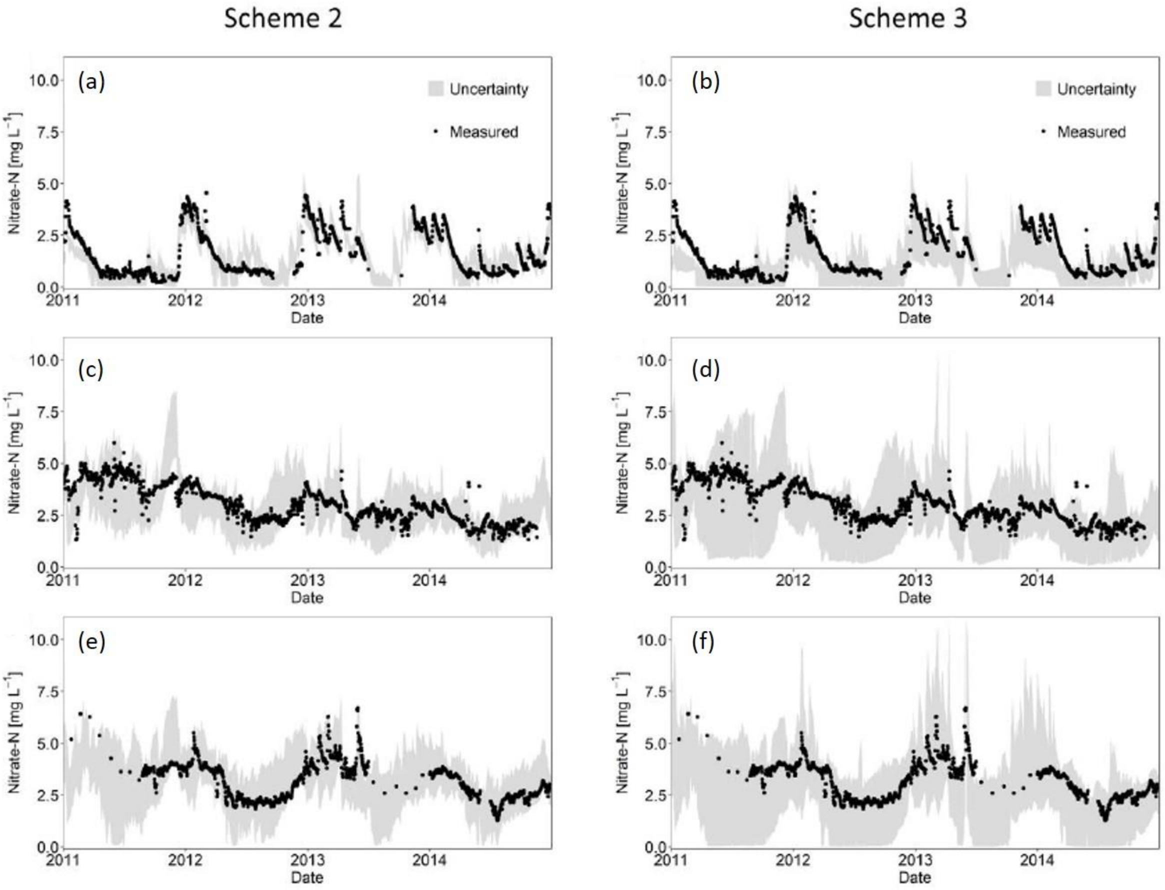

Figure 7. Comparison of 95% uncertainty boundaries for the simulated nitrate (NO_3^-) concentrations obtained with schemes 2 and 3 for three gauging stations: Meisdorf (a-b), Hausneindorf (c-d), and Stassfurt (e-f).

Table 8. R-factor values for nitrate (NO_3^-) concentrations at three gauging stations for schemes 2 and 3.

Stations	Scheme 2	Scheme 3
Meisdorf	0.92	1.08
Hausneindorf	2.18	4.13
Stassfurt	2.79	4.52

4. Discussion

4.1 Evaluation of model performance for different calibration schemes

We evaluated the ability of the mHM-Nitrate model to simulate discharge and nitrate concentrations at eight gauging stations. We specifically examined the transferability of hydrological and water quality parameters at spatial scales.

4.1.1 Model performance for discharge under three calibration schemes

During model validation, simulated discharge at the catchment outlet was similar whether the calibration data came from a single site (scheme 1: catchment outlet station) or multiple sites (scheme 2: 3 stations and scheme 3: 8 stations) (Table 4). This result suggests that, for discharge, the number of stations used during calibration did not affect model performance at the catchment outlet. Our finding is consistent with those of Chiang et al. (2014); Wang et al. (2012); Wu et al. (2022a).

That said, performance was better with scheme 2 than scheme 1 when discharge was simulated for all eight gauging stations, except in the case of Hausneindorf (Table 4). This result could have arisen because multi-site calibration better constrains model parameters by including information on catchment characteristics (e.g., land use and soil types) at upstream stations (here, Meisdorf and Hausneindorf); these characteristics are frequently heterogeneous in space and shape hydrological parameters (e.g., infil and pet, Figures 6a and 6b). Jiang et al. (2015) reported that, compared to single-site calibration, multi-site calibration may better capture dynamics in large, diverse catchments because it accounts for the effects of different hydrological processes (e.g., slow groundwater dynamics and quick interflows). For example, in the Bode catchment, interflow is the primary form of runoff in mountainous areas (Jiang et al., 2014), while the share of groundwater increases from the mountains to the lowlands (Zhou et al., 2022).

In contrast, model performance was similar for schemes 2 and 3 (NSE values for the eight gauging stations; Table 4), which suggests that adding more sites does not always improve simulations for upstream stations. This finding is consistent with those of previous studies (Her and Chaubey (2015);

Wang et al. (2012); Xie et al. (2021)) and could potentially be explained by station choice and the failure of scheme 3 to introduce any new catchment characteristics. As a result, schemes 2 and 3 displayed similar cumulative distributions for their hydrological parameters (Figures 6a-b). Therefore, during calibration, it may be challenging to optimize model parameters by relying on station number only.

4.1.2 Model performance for $\mathrm{NO_3^-}$ concentration under three calibration schemes

Simulated nitrate concentrations were significantly better for all gauging stations (with the exception of Nienhagen) when scheme 2 versus scheme 1 was used (Table 4). This could be due to the fact that the inclusion of Meisdorf in scheme 2 results in additional parameter constraint. The station is found in a forested subcatchment, which likely led to changes in the values of land-use-dependent parameters (e.g., pprt_na, pprt_agri; Figures 6e and 6f). These parameters were optimized in scheme 2 and, additionally, improved model performance at non-calibrated stations, such as Wegeleben and Ditfurt. Both stations are located in subcatchments with intermediate levels of forest cover (> 30% and > 56.4%, respectively). Similarly, the inclusion of Hausneindorf in scheme 2 improved model performance at Hadmersleben, which had not been part of the calibration process, because the two stations occur in regions with similar levels of farmlands (Table 2). This finding indicates that utilizing multi-site calibration schemes that capture diverse catchment characteristics can improve simulated nitrate concentrations even at locations that were not included in the calibration process. This result concurs with those of previous studies (Chiang et al., 2014; Jiang et al., 2015; Shrestha et al., 2016), which found that such improvements result from the fact that multi-site calibration schemes can account for dramatic variability in observed nitrate concentrations and hydrological regimes across catchments. These schemes can thus better constrain parameters associated with nitrate transport and transformation.

In contrast, model performance was slightly lower at all stations (except Nienhagen) for scheme 3 than scheme 2 (PBIAS values; Table 4), which suggests that adding more gauging stations to the calibration process cannot, by itself, result in further improvements to simulations of nitrate concentrations. This finding may have two explanations. First, the five additional gauging stations (Wegeleben, Hadmersleben, Peseckendorf, Ditfurt, and Nienhagen) included in scheme 3 did not introduce additional diversity in catchment characteristics, which was the case when Meisdorf and Hausneindorf were included in scheme 2 (Figure 2). For instance, except for Peseckendorf, four of the five additional stations have farmland surface areas and mean nitrate concentrations that are similar to that of Hausneindorf, which led to similar model parameter distributions for schemes 2 and 3 (Figures 6c-f). Second, three of the five additional stations have low-frequency measurements of nitrate concentrations (i.e., once or twice per month). Jiang et al. (2019) found that, when the HYPE model was applied to the Selke catchment, performance was better when calibration used nitrate concentrations that were collected daily versus every two weeks. The slight decline in performance from scheme 2 to scheme 3 could be affected by the model's attempt to satisfyingly balance the large number of additional observations resulting from site addition (Jiang et al., 2015). In other words, multi-site calibration approaches try to identify the parameter set that represents the best compromise given the presence of multiple subcatchments, which is a more intensive task than simply focusing on a single catchment outlet.

4.1.3 Comparison of hydrological and water quality model performance

In brief, the model's accuracy for predicting both discharge and $\mathrm{NO_3^-}$ concentration improved when using Scheme 2 compared to Scheme 1. However, while the model's accuracy for discharge remained consistent between Scheme 2 and 3, its accuracy for nitrate decreased in Scheme 3. On one hand, hydrology is a physical process that is well understood and can be easily quantified through measurements and modeling. On the other hand, nitrate dynamics are much more complex and can be influenced by a variety of specific factors that are unique to a particular location, such as the amount of fertilizer applied and the level of moisture in the soil. Nitrogen fertilizer application rates are often uncertain and can vary depending on crop type and management practices. Nitrate uptake by plants is also difficult to predict, as it is influenced by a range of factors such as soil moisture, temperature, and nutrient availability. Overall, nitrate simulations are likely to be more accurate in mountainous regions where quick flowing systems lead to less storage and transformation of nitrate (Table 4). In lowland agricultural systems, nitrate can persist in soils for several years and in groundwater for even longer time scales, leading to legacy effects that can complicate stream nitrate dynamics (Wriedt and Rode 2006, Ehrhardt et al., 2019; Hrachowitz et al., 2015).

4.2 Simulating nitrate concentrations across space

Scheme 1, which solely utilized data from the catchment outlet, was unable to accurately simulate nitrate dynamics at upstream sites within the large, heterogeneous Bode catchment. Indeed, PBIAS values were high (> 45%) for many of the 94 spatially distributed sampling locations when scheme 1 was used (Figure 4a and Table 5). The model performed much better when scheme 2 was employed. Its addition of two gauging stations to the calibration process thus appeared to greatly influence model performance at the catchment scale.

However, little to no further improvement was seen with scheme 3 and its five additional gauging stations. This assertion has two sources of support: schemes 2 and 3 had similar numbers of sampling locations within the different PBIAS ranges (Table 5) and displayed similar cumulative distributions for their parameters (Figure 6). Comparing cumulative parameter distributions can help identify informative calibration stations. It can also determine whether adding or removing calibration stations would improve.

Further results of the model performance of $\mathrm{NO_3^-}$ concentration at Scheme 2 shows varying performances among $\mathrm{NO_3^-}$ sampling locations that represent different catchment characteristics (e.g., precipitation, land use, and fertilizer inputs) (Figure 5). At sampling location 4, $\mathrm{NO_3^-}$ concentration was overestimated in summer, but the PBIAS value of the whole period was negative, it means that the model underestimated $\mathrm{NO_3^-}$ concentrations during other times of the year. This could be due to errors in the representation of hydrological processes, such as groundwater recharge, which can affect $\mathrm{NO_3^-}$ transport and concentration in the groundwater. This suggests that spatial representation of groundwater processes (such as groundwater $\mathrm{NO_3^-}$ concentration) are needed to be refined to obtain better model performance for small sub-catchments. Faramarzi et al. (2015) and Gao et al. (2016) concluded that the hydrological and water quality models that only rely on calibration without refining internal process representation (e.g., groundwater $\mathrm{NO_3^-}$ concentration) will often not result in further improvement. Nevertheless, the above analysis indicates that Scheme 2 is sufficient to ensure the satisfactory model performance at $\mathrm{NO_3^-}$ sampling locations, since 75% of the $\mathrm{NO_3^-}$ sampling locations showed absolute PBIAS ≤35% (Figure 4b and Table 5) and the mHM-Nitrate model was capable to present different magnitudes of $\mathrm{NO_3^-}$ levels for different sub-catchments which differ in their catchment characteristics (Figure 5). These findings are in line with Ghaffar et al. (2021), where they found that considering archetypal gauging stations in the calibration process leads better spatial validation of the model at internal locations that were not originally considered in calibration. These stations represent the maximum catchment characteristics in heterogeneous catchments in terms of dominant land-use and meteorological features. This highlights the need for multiple internal stations/locations to validate the model's capacity to accurately capture the complexity of natural processes and identify which process needs to be improved (Beven, 2001; Daggupati et al., 2015).

4.3 Impact of calibration approaches on model uncertainty

For the three gauging stations, there was more uncertainty around simulated nitrate concentrations for scheme 3 than for scheme 2 (Figure 7), likely because scheme 3 included stations with low-frequency measurements. This result highlights the effect of measurement frequency on simulation uncertainty. Indeed, low-frequency measurements may not capture the full range of variability in $\mathrm{NO_3^-}$ dynamics. Furthermore, multi-site calibration approaches that rely on low-frequency data may give rise to spatial representation issues, given that water quality can vary widely across heterogenous catchments and be influenced by local factors, such as land use and soil type. This finding is in line with those of previous studies (Jiang et al., 2019; Khorashadi Zadeh et al., 2019; Ullrich and Volk, 2010). For example, Jiang et al. (2019) found that, for the HYPE model, uncertainty was reduced when the calibration process used NO_3^- concentrations that had been collected daily versus every two weeks.

Once the catchment function is well captured by representative key stations (Scheme 2), additional measurements may not be cost-effective and could increase model simulation uncertainty (Scheme 3). Therefore, it is essential to consider the specific requirements of the study and the desired level of accuracy in model simulation. Depending on the goals and context, it may be necessary to find a balance between cost-effectiveness and model performance by considering the spatial distribution of measurements.

4.4 Implication of spatial evaluation of distributed hydrological water quality model

Improving the performance of hydrological water quality models has become a critical concern as these models grow more complex (Beven, 2001; Refsgaard et al., 2016; Refsgaard et al., 2022). Calibration using multiple sites is a crucial step in this process as it enables a better representation of the spatial variability of hydrological and water processes. It is also equally essential to extend the evaluation beyond calibration in order to gain insights into the spatial variability of hydrological and water processes and understand the underlying processes that govern the behavior of the system (Efstratiadis and Koutsoyiannis, 2010; Koch et al., 2015).

It is possible to use remote sensing data such as soil moisture (Mei et al., 2023; Rajib et al., 2016) and evapotranspiration (Rajib et al., 2018; Zhang et al., 2021) to evaluate spatial performance of distributed models for water quantity, but this approach cannot be applied to spatially evaluate models for NO_3^- and other chemicals. To effectively evaluate the spatial performance of distributed models for water quality, water quality monitoring or sampling is always necessary. This study highlights the significance of long-term and spatially distributed monitoring water quality data, which is readily accessible from authorities.

5. Conclusion

Using three different approaches, we calibrated a fully distributed process-based mHM-Nitrate model that was then validated spatially and temporally at 8 gauging stations (discharge and NO_3^- concentrations) and 94 spatially distributed sampling locations (NO_3^- concentrations) within the heterogeneous Bode catchment in central Germany. Scheme 1 used only data from the catchment outlet; scheme 2 used data from the catchment outlet and two upstream stations; and scheme 3 used data from the catchment outlet and seven additional upstream stations. Our study found that, for simulated discharge, model performance was similar at the catchment outlet for the three calibration schemes. Furthermore, model performance did not improve consistently across the upstream gauging stations.

In contrast, for NO_3^- concentrations, scheme 2 was better than scheme 1 when it came to simulating dynamics at sampling locations that had not been part of the calibration process. That said, model performance across the sampling locations was similar for schemes 2 and 3. Our results indicate that increasing the number of stations used in calibration does not necessarily improve simulations of NO_3^- concentrations. Additionally, we found that the use of low-frequency calibration data may increase the degree of model uncertainty.

In conclusion, this research provides guidance on selecting gauging stations for the purposes of model calibration: differences in cumulative parameter distributions should signal which stations can add helpful additional representation. Furthermore, our work highlights that this selection process must account for diversity in catchment characteristics, such as land use, meteorological patterns, and elevation. In this way, the calibration data will better represent spatial patterns, and the model will yield more accurate predictions. Overall, this study provides valuable insights into calibration-related decision-making when carrying out fully distributed hydrological water quality models to simulate dynamics within spatially heterogeneous catchments. This study also highlights the value of using readily available water authorities monitoring data with high spatial resolution but low temporal resolution for validating fully distributed models, even in the absence of discharge measurements.

Data Availability Statement

- The high-frequency monitoring data used are available at Zhang et al. (2022) via https://doi.org/10.48758/ufz.12911

- The discharge and low frequency monitoring data are available at the data portal (Datenportal) of the State Agency for Flood Protection and Water Management of Saxony Anhalt, Germany (LHW, 2022) https://gld.lhw-sachsen-anhalt.de/

- The high-frequency monitoring data is available at TERENO (TERrestrial ENvironmental Observatories) Data Discovery Portal https://ddp.tereno.net/ddp/ (TERENO, 2020).

References

Abbaspour, K.C. et al., 2007. Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. Journal of Hydrology, 333(2-4): 413-430. DOI:10.1016/j.jhydrol.2006.09.014

Arnold, J.G. et al., 2012. Swat: Model Use, Calibration, and Validation. Transactions of the Asabe, 55(4): 1491-1508.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part I: model development 1. JAWRA Journal of the American Water Resources Association, 34(1): 73-89.

Beck, H.E. et al., 2016. Global-scale regionalization of hydrologic model parameters. Water Resources Research, 52(5): 3599-3622. DOI:10.1002/2015wr018247

Beven, K., 2001. How far can we go in distributed hydrological modelling? Hydrology and Earth System Sciences, 5(1): 1-12. DOI:10.5194/hess-5-1-2001

Beven, K., 2007. Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process. Hydrology and Earth System Sciences, 11(1): 460-467. DOI:DOI 10.5194/hess-11-460-2007

Bloschl, G., Sivapalan, M., 1995. Scale Issues in Hydrological Modeling - a Review. Hydrological Processes, 9(3-4): 251-290.

Cao, W., Bowden, W.B., Davie, T., Fenemor, A., 2006. Multi-variable and multi-site calibration and validation of SWAT in a large mountainous catchment with high spatial variability. Hydrological Processes, 20(5): 1057-1073. DOI:10.1002/hyp.5933

Chiang, L.C., Yuan, Y.P., Mehaffey, M., Jackson, M., Chaubey, I., 2014. Assessing SWAT's performance in the Kaskaskia River watershed as influenced by the number of calibration stations used. Hydrological Processes, 28(3): 676-687. DOI:10.1002/hyp.9589

Conover, W.J., 1999. Practical nonparametric statistics, 350. john wiley & sons.

Cuntz, M. et al., 2015. Computationally inexpensive identification of noninformative model parameters by sequential screening. Water Resources Research, 51(8): 6417-6441. DOI:10.1002/2015wr016907

Daggupati, P. et al., 2015. A Recommended Calibration and Validation Strategy for Hydrologic and Water Quality Models. Transactions of the Asabe, 58(6): 1705-1719. DOI:10.13031/trans.58.10712

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal, 55(1): 58-78. DOI:10.1080/02626660903526292

Ehrhardt, S., Kumar, R., Fleckenstein, J.H., Attinger, S., Musolff, A., 2019. Trajectories of nitrate input and output in three nested catchments along a land use gradient. Hydrology and Earth System Sciences, 23(9): 3503-3524. DOI:10.5194/hess-23-3503-2019

Engel, B., Storm, D., White, M., Arnold, J., Arabi, M., 2007. A hydrologic/water quality model application protocol. Journal of the American Water Resources Association, 43(5): 1223-1236. DOI:10.1111/j.1752-1688.2007.00105.x

Faramarzi, M. et al., 2015. Setting up a hydrological model of Alberta: Data discrimination analyses prior to calibration. Environ Modell Softw, 74: 48-65. DOI:10.1016/j.envsoft.2015.09.006

Franco, A.C.L., Oliveira, D.Y.d., Bonumá, N.B., 2020. Comparison of single-site, multi-site and multi-variable SWAT calibration strategies. Hydrological Sciences Journal, 65(14): 2376-2389. DOI:10.1080/02626667.2020.1810252

Gao, H. et al., 2016. Accounting for the influence of vegetation and landscape improves model transferability in a tropical savannah region. Water Resources Research, 52(10): 7999-8022. DOI:10.1002/2016wr019574

Ghaffar, S., Jomaa, S., Meon, G., Rode, M., 2021. Spatial validation of a semi-distributed hydrological nutrient transport model. Journal of Hydrology, 593: 125818. DOI:https://doi.org/10.1016/j.jhydrol.2020.125818

Gupta, H.V., Beven, K.J., Wagener, T., 2005. Model Calibration and Uncertainty Estimation, Encyclopedia of Hydrological Sciences. DOI:https://doi.org/10.1002/0470848944.hsa138

Hargreaves, G.H., Samani, Z.A., 1985. Reference Crop Evapotranspiration from Temperature. Applied Engineering in Agriculture, 1(2): 96-99. DOI:10.13031/2013.26773

Her, Y., Chaubey, I., 2015. Impact of the numbers of observations and calibration parameters on equifinality, model performance, and output and parameter uncertainty. Hydrological Processes, 29(19): 4220-4237. DOI:10.1002/hyp.10487

Hrachowitz, M., Fovet, O., Ruiz, L., Savenije, H.H.G., 2015. Transit time distributions, legacy contamination and variability in biogeochemical 1/fαscaling: how are hydrological response dynamics linked to water quality at the catchment scale? Hydrological Processes, 29(25): 5241-5256. DOI:10.1002/hyp.10546

Hundecha, Y., Bárdossy, A., 2004. Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. Journal of Hydrology, 292(1-4): 281-295. DOI:10.1016/j.jhydrol.2004.01.002

Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database http://srtm.csi.cgiar.org/, International Centre for Tropical Agriculture (CIAT).

Jiang, S., Jomaa, S., Büttner, O., Meon, G., Rode, M., 2015. Multi-site identification of a distributed hydrological nitrogen model using Bayesian uncertainty analysis. Journal of Hydrology, 529: 940-950. DOI:10.1016/j.jhydrol.2015.09.009

Jiang, S.Y., Jomaa, S., Rode, M., 2014. Modelling inorganic nitrogen leaching in nested mesoscale catchments in central Germany. Ecohydrology, 7(5): 1345-1362. DOI:10.1002/eco.1462

Jiang, S.Y. et al., 2019. Effects of stream nitrate data frequency on watershed model performance and prediction uncertainty. Journal of Hydrology, 569: 22-36. DOI:10.1016/j.jhydrol.2018.11.049

Khorashadi Zadeh, F., Nossent, J., Woldegiorgis, B.T., Bauwens, W., van Griensven, A., 2019. Impact of measurement error and limited data frequency on parameter estimation and uncertainty quantification. Environ Modell Softw, 118: 35-47. DOI:https://doi.org/10.1016/j.envsoft.2019.03.022

Khu, S.-T., Madsen, H., di Pierro, F., 2008. Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering. Advances in Water Resources, 31(10): 1387-1398. DOI:10.1016/j.advwatres.2008.07.011

Krabbenhoft, C.A. et al., 2022. Assessing placement bias of the global river gauge network. Nat Sustain, 5(7): 586-592. DOI:10.1038/s41893-022-00873-0

Kunz, J.V., Annable, M.D., Rao, S., Rode, M., Borchardt, D., 2017. Hyporheic Passive Flux Meters Reveal Inverse Vertical Zonation and High Seasonality of Nitrogen Processing in an Anthropogenically Modified Stream (Holtemme, Germany). Water Resources Research, 53(12): 10155-10172. DOI:https://doi.org/10.1002/2017WR020709

Lerat, J. et al., 2012. Do internal flow measurements improve the calibration of rainfall-runoff models? Water Resources Research, 48(2). DOI:10.1029/2010wr010179

Leta, O.T., Griensven, A.v., Bauwens, W., 2017. Effect of Single and Multisite Calibration Techniques on the Parameter Estimation, Performance, and Output of a SWAT Model of a Spatially Heterogeneous Catchment. Journal of Hydrologic Engineering, 22(3): 05016036. DOI:doi:10.1061/(ASCE)HE.1943-5584.0001471

Li, X., Weller, D.E., Jordan, T.E., 2010. Watershed model calibration using multi-objective optimization and multi-site averaging. Journal of Hydrology, 380(3-4): 277-288. DOI:10.1016/j.jhydrol.2009.11.003

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., Arheimer, B., 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrology Research, 41(3-4): 295-319. DOI:10.2166/nh.2010.007

Merz, R., Blöschl, G., 2004. Regionalisation of catchment model parameters. Journal of Hydrology, 287(1-4): 95-123. DOI:10.1016/j.jhydrol.2003.09.028

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Transactions of the Asabe, 58(6): 1763-1785. DOI:10.13031/trans.58.10715

Moriasi, D.N., Wilson, B.N., Douglas-Mankin, K.R., Arnold, J.G., Gowda, P.H., 2012. Hydrologic and Water Quality Models: Use, Calibration, and Validation. Transactions of the ASABE, 55(4): 1241-1247. DOI:10.13031/2013.42265

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2): 161-174.

Oudin, L., Andreassian, V., Perrin, C., Michel, C., Le Moine, N., 2008a. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44(3). DOI:Artn W03413

10.1029/2007wr006240

Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008b. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44(3). DOI:10.1029/2007wr006240

Parajka, J., Merz, R., Bloschl, G., 2005. A comparison of regionalisation methods for catchment model parameters. Hydrology and Earth System Sciences, 9(3): 157-171. DOI:DOI 10.5194/hess-9-157-2005

Parajka, J. et al., 2013. Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies. Hydrology and Earth System Sciences, 17(5): 1783-1795. DOI:10.5194/hess-17-1783-2013

Pianosi, F., Sarrazin, F., Wagener, T., 2015. A Matlab toolbox for Global Sensitivity Analysis. Environ Modell Softw, 70: 80-85. DOI:10.1016/j.envsoft.2015.04.009

Pokhrel, P., Gupta, H.V., Wagener, T., 2008. A spatial regularization approach to parameter estimation for a distributed watershed model. Water Resources Research, 44(12). DOI:10.1029/2007wr006615

Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology, 198(1-4): 69-97. DOI:Doi 10.1016/S0022-1694(96)03329-X

Refsgaard, J.C. et al., 2016. Where are the limits of model predictive capabilities? Hydrological Processes, 30(26): 4956-4965. DOI:10.1002/hyp.11029

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research, 46(5). DOI:10.1029/2008wr007327

Saraswat, D. et al., 2015. Hydrologic and Water Quality Models: Documentation and Reporting Procedures for Calibration, Validation, and Use. Transactions of the Asabe, 58(6): 1787-1797. DOI:10.13031/trans.57.10707

Shrestha, M.K., Recknagel, F., Frizenschaf, J., Meyer, W., 2016. Assessing SWAT models based on single and multi-site calibration for the simulation of flow and nutrient loads in the semi-arid Onkaparinga catchment in South Australia. Agricultural Water Management, 175: 61-71. DOI:10.1016/j.agwat.2016.02.009

Singh, R., Archfield, S.A., Wagener, T., 2014. Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – A comparative hydrology approach. Journal of Hydrology, 517: 985-996. DOI:10.1016/j.jhydrol.2014.06.030

TERENO: Data Discovery Portal: https://ddp.tereno.net/ddp/, last access: 17 December 2020

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resources Research, 43(1). DOI:Artn W01413

10.1029/2005wr004723

Ullrich, A., Volk, M., 2010. Influence of different nitrate-N monitoring strategies on load estimation as a base for model calibration and evaluation. Environ Monit Assess, 171(1-4): 513-27. DOI:10.1007/s10661-009-1296-8

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resources Research, 41(1). DOI:10.1029/2004wr003059

Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stochastic Environmental Research and Risk Assessment, 19(6): 378-387. DOI:10.1007/s00477-005-0006-5

Wagener, T., Wheater, H.S., 2006. Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. Journal of Hydrology, 320(1-2): 132-154. DOI:10.1016/j.jhydrol.2005.07.015

Wang, S. et al., 2012. Multi-site calibration, validation, and sensitivity analysis of the MIKE SHE Model for a large watershed in northern China. Hydrology and Earth System Sciences, 16(12): 4621-4632. DOI:10.5194/hess-16-4621-2012

Wellen, C., Kamran-Disfani, A.R., Arhonditsis, G.B., 2015. Evaluation of the current state of distributed watershed nutrient water quality modeling. Environ Sci Technol, 49(6): 3278-90. DOI:10.1021/es5049557

Wollschläger, U. et al., 2016. The Bode hydrological observatory: a platform for integrated, interdisciplinary hydro-ecological research within the TERENO Harz/Central German Lowland Observatory. Environmental Earth Sciences, 76(1). DOI:10.1007/s12665-016-6327-5

Wu, H. et al., 2021. An improved calibration and uncertainty analysis approach using a multicriteria sequential algorithm for hydrological modeling. Sci Rep, 11(1): 16954. DOI:10.1038/s41598-021-96250-6

Wu, L., Liu, X., Yang, Z., Yu, Y., Ma, X., 2022a. Effects of single‐ and multi‐site calibration strategies on hydrological model performance and parameter sensitivity of large‐scale semi‐arid and semi‐humid watersheds. Hydrological Processes, 36(6). DOI:10.1002/hyp.14616

Wu, S., Tetzlaff, D., Yang, X., Soulsby, C., 2022b. Disentangling the Influence of Landscape Characteristics, Hydroclimatic Variability and Land Management on Surface Water NO3‐N Dynamics: Spatially Distributed Modeling Over 30 yr in a Lowland Mixed Land Use Catchment. Water Resources Research, 58(2). DOI:10.1029/2021wr030566

Xie, K. et al., 2021. Identification of spatially distributed parameters of hydrological models using the dimension-adaptive key grid calibration strategy. Journal of Hydrology, 598. DOI:10.1016/j.jhydrol.2020.125772

Yang, X., Jomaa, S., Buttner, O., Rode, M., 2019a. Autotrophic nitrate uptake in river networks: A modeling approach using continuous high-frequency data. Water Res, 157: 258-268. DOI:10.1016/j.watres.2019.02.059

Yang, X. et al., 2018. A New Fully Distributed Model of Nitrate Transport and Removal at Catchment Scale. Water Resources Research. DOI:10.1029/2017wr022380

Yang, X., Rode, M., 2020. A Fully Distributed Catchment Nitrate Model - mHM-Nitrate v2.0. DOI:10.5281/ZENODO.3891629

Yang, X.Q., Jomaa, S., Rode, M., 2019b. Sensitivity Analysis of Fully Distributed Parameterization Reveals Insights Into Heterogeneous Catchment Responses for Water Quality Modeling. Water Resources Research, 55(12): 10935-10953. DOI:10.1029/2019wr025575

Zhang, X., Srinivasan, R., Van Liew, M., 2008. Multi-Site Calibration of the SWAT Model for Hydrologic Modeling. Transactions of the ASABE, 51(6): 2039-2049. DOI:https://doi.org/10.13031/2013.25407

Zhou, X. et al., 2022. Exploring the relations between sequential droughts and stream nitrogen dynamics in central Germany through catchment-scale mechanistic modelling. Journal of Hydrology, 614: 128615. DOI:10.1016/j.jhydrol.2022.128615