Reviewer 2

General Comments

The manuscript entitled "Simulation of spatially distributed sources, transport, and transformation of nitrogen from fertilization and septic system in an exurban watershed" presents and uses an augmented version of the RHESSys Model to evaluate the hydrologic and biogeochemical N cycling, and transport in a mixed land use watershed characterized by anthropogenic N inputs from irrigation, fertilization, and on-site sanitary wastewater disposal in form of septic systems.

The study is motivated by enhancing the understanding of transport, cycling and subsequent export to streams of N in exurban watersheds. It declares the need to defer appropriate siting for effective best management practices (BMP) to reduce N export to downstream water bodies. To my perception, it fits within the thematic scope of HESS. It addresses a highly relevant research topic and could become a substantial contribution to scientific progress; however, the novelty of the presented approach remains unclear.

The manuscript promises to address certainly interesting aspects e.g., to evaluate how the spatial and temporal distribution of human nitrogen sources in exurban watershed controls N export to downstream water bodies and nitrification rates. However, the conclusions reached in the manuscript are rather general. At the current state, the concept of the study is limited to compare the simulated N loads and concentrations and nitrification rates between simulations without and with one or two human N input types and concludes that including fertilization and septic systems improves the simulation results when comparing to observations in the case study watershed. It remains unclear whether the augmented model can be transferred to other watersheds and how it can support to situate BMPs effectively.

The presentation of the results in the figures and tables does not keep up with the high quality in other articles published in HESS. Furthermore, to my understanding the methodology lacks significant steps: i) the model validation and ii) statistics that substantiate the results. Therefore, I suggest to reject the manuscript for publication at its current state and encourage the authors to improve it.

Response:

Thanks for your helpful comments to our manuscript. We addressed your concerns to our study in novelty and model uncertainty for N dynamics accordingly in responds to your General Comments:

We highlighted our revisions to the manuscript by using blue and calibri font with shading.

- Updates are highlighted in blue
- Responses in times new roman font

Thanks for all your insightful and helpful suggestions to improve the quality of our manuscript substantially. We addressed your three major concerns for 1) novelty of current approach, 2) too general conclusion, and 3) lack of model validation and statistics in methods accordingly.

1. **Unclear novelty**
   We highlighted the novelty of our study in the Abstract and Introduction. Our study addresses the **distribution and interaction of hillslope ecohydrological processes** in transporting natural and **human sources of water and nitrogen** in a long term monitored suburban watershed. Understanding processes and interactions at these scales promotes the design of retention features.

   To our knowledge, our model is the first fully distributed hydrologic model that includes i) spatial and temporal human-induced N and water sources at the household level, and ii) hillslope ecohydrological processes for routing and cycling water, carbon, and nitrogen. These processes are necessary to identify the space/time distribution of "hot spots" of N retention at scales amenable to restoration and Best Management Practices (BMPs) in the future.

   A significant aspect of the model is that it is calibrated for hydrologic processes **restricted to soil and subsurface hydraulic parameters**. It is **not calibrated for biogeochemical processes** which are subject to change with restoration activities. In contrast, the current set of ecohydrological models typically calibrate patch (grid cells, elements) to stream transfer, and biogeochemical cycling parameters. Therefore, our model could be generalized to other suburban watersheds with only discharge but no water chemistry observations.

   Abstract

   Line 21:

   We evaluated how the spatial and temporal distribution of nitrogen sources interacts with ecohydrological transport and transformation processes along surface/subsurface flowpaths to nitrogen cycling, and export. Embedding distributed household sources of nitrogen and water within hillslope hydrologic systems influences the development of both planned and unplanned "hot spots" of nitrogen flux and retention in suburban ecosystems.

   Line 29:

   With the model is calibrated for subsurface hydraulic parameters only and without calibrating ecosystem and biogeochemical processes, the model predicted mean […]

   In the Introduction, we thoroughly reorganized the order of paragraphs and firstly highlighted why understanding ecohydrological processes at "hillslope level" is required for planning Best Management Practices and promote N retention.

   Line 47:

BMPs can be both structural (e.g., constructed wetlands) and non-structural (e.g., changing fertilization and irrigation regimes). In addition to planned BMPs, spontaneously developed "hot spots" (Palta et al., 2017) may be responsible for a large share of nutrient retention, and therefore should be identified and protected. Both planned and unplanned retention features exist at very localized, sub-hillslope scales. Therefore, gaining a comprehensive understanding of the hillslope level ecohydrological behaviours and interactions between i) ecosystems and human derived nitrogen sources and ii) flowpath modification can lay the foundation for effectively mitigating these environmental issues through spatially well-conceived and sustainable management practices.

Then, we briefly reviewed how urban water quality is degraded by excessive human-induced N loads, emphasizing the widely used septic systems in suburban areas.

Line 60:

In the United States, about 20% of households (26.1 million) are reported to be served by septic systems in 2007 (U.S. EPA, 2008). Through our work in Baltimore Ecosystem Study, low density suburban areas have been shown to produce the highest NO_3^- load per unit developed land among different land uses, degrading local and downstream water quality (Groffman et al., 2004; Zhang et al., 2022).

We then discussed the research gap in current semi-distributed models in aspects of incapable of including i) household scale human water and N loads contributing the majority of N inputs in suburban watersheds in distinct landscape positions and ii) hillslope hydrologic flow paths to meet the planning purposes to design BMPs to reduce N export. We also discussed data-driven approaches which could include additional N inputs, but hillslope-level N transport and transformation is still missing.

Line 69:

With rapid suburban and exurban sprawl, decision makers are facing environmental challenges which requires detailed planning for siting BMPs effectively in watersheds to promote N retention, reduce N export in streams, and protect water quality. These include both constructed and "inadvertent" biogeochemical hot spots at specific hillslope locations (e.g., swales, wetlands, riparian areas) on N retention at resolutions required for landscape design. However, commonly used modelling frameworks could not couple distributions and interactions of hillslope ecohydrological processes in transporting and transforming natural and human-induced N sources to understand or predict local (neighbourhood or hillslope) scale N transport and retention. Semi-distributed. Semi […] lack(s) hillslope water and nutrient mixing along interacting surface/subsurface hydrologic flowpaths […]

Line 82:

Data-driven approaches, such as SPARROW (Ator & Garcia, 2016; Smith et al., 1997), are also developed to assess large scale water quality in streams by nonlinear regression from gauged discharge and solute concentrations. However, these models also do not investigate hillslope-scale transport and transformation processes. In addition, there does not exist the data at hillslope scales to develop sufficient data-based approaches to understand and predict retention processes (e.g., denitrification, uptake, immobilization).

Then, we emphasized, though fully distributed hydrologic models, such as MIKE-SHE, could simulate hillslope hydrology and biogeochemistry, they currently have no modules to include the household-level N inputs developed.

Line 87:
Fully distributed hydrology models, such as MIKE-SHE (Abbott et al., 1986a, 1986b) and RTM-PiHM (Bao et al., 2017; Zhi et al., 2022), ParFlow (Maxwell, 2013) and RHESSys (Tague & Band, 2004) could explicitly couple hillslope hydrologic and biogeochemical processes that are required to understand transport and transformation of these human-induced N loads along hydrologic flowpaths from upland to stream.

Lastly, we wanted to highlight that our model is designed to be generalized to watersheds without long-term water chemistry observations which are quite expensive to acquire. In other words, we do not calibrate our parameters for N inputs (e.g., fertilization and septic loads) or processes but only soil hydraulics against streamflow records. If the model could reasonably estimate NO_3^-, it compromises the generalization of the model.

Line 102:
Lastly, the framework should be capable to be extrapolated to watersheds without water chemistry data which are less available than discharge records worldwide. It would be a valuable feature of the framework to estimate nutrient dynamics reasonably if calibrating only hydrologic parameters could provide reasonable estimation of N dynamics. Calibrating nutrient dynamics may not allow generalization to watersheds without chemistry records or extrapolation to conditions in which water quality BMPs are implemented.

2. **Too general conclusion and model's ability to be transferred to other watersheds**
We emphasized the major results of reasonable NO_3^- concentration simulations, and the spatially explicit feature of our model allows assessments of BMPs' effects on promoting N retention when they are sited at areas **accumulating both high water and N loads** from upstream households in a watershed.

Also, by performing uncertainty analysis, our NO_3^- simulations include a reasonable range of biogeochemical outcomes while restricting calibration to subsurface hydraulic parameters. The model therefore can be applied to other sub/exurban watersheds which also use septic systems and fertilizers. In other words, with reasonable survey estimating human inputs and domestic water usage, our model could provide reasonable NO_3^- export of watershed by calibrating against streamflow records which are much more available than the water chemistry data. For numerous suburban watersheds, our model could reasonably help decision makers understand the current N levels and upland dynamics without water chemistry data.

We thoroughly revised our Conclusion section:

Line 579:

Our analysis provides important insights into how different sources of N input interact with ecohydrological processes to control N export from suburban and exurban watersheds where single-family households use individual groundwater wells for domestic water discharged to septic systems and lawn irrigation, and add additional nitrogen in the form of sanitary effluent and lawn fertilization. While atmospheric deposition is ubiquitous, the input of lawn fertilization and irrigation water, and septic effluent volume and N load are concentrated in limited areas of the watershed at much higher per unit area rates. These differences cascade through the watershed producing hot spots of N export and retention. Calibrating hydrologic parameters against streamflow observations only, our model yielded satisfactory simulations of in-stream NO_3^- concentration and upland N retention processes. Specifically, our model estimated the mean NO_3^- concentration as 1.43 mg L^{-1}, which is only less than 0.2 mg L^{-1} lower than the weekly observations from Baltimore Ecosystem Study for our study period. The simulated denitrification rates at fertilized lawns are also comparable to measurements in the study area and nearby watersheds in Baltimore, and rates at wetlands and riparian areas are similar to reported measurements in other studies. Our results strongly support the basis for small watershed-scale analysis and planning to address watershed N exports and are particularly relevant in areas such as the Chesapeake Bay that are highly sensitive to N-induced eutrophication. The spatially explicit, high-resolution simulations from our model could help local decision makers to identify existing and potential new hot spots of N retention processes (e.g., denitrification). Specifically, we found locations accumulating both high N loads and water from upstream are ideal locations for siting future BMPs (e.g., detention ponds, constructed wetlands) to promote N retention and improve water quality for local and downstream waterbodies. In summary, the improved RHESSys simulations with augmentations for more complete, spatially nested inputs of water and N and subsequent feedbacks between transport and retention highlight the importance of the structured spatial heterogeneity of human impacts to fully understand ecohydrological processes at hillslope level in developed watersheds. Existing models often miss the patterns and feedbacks water and N inputs at household levels and within hillslope hydrologic flowpaths. The spatially distributed inputs and our augmented RHESSys model structure may provide a reliable framework to comprehensively evaluate current coupled water, C and N cycles, and also understand and predict effectiveness of ecosystem restorations to improve water quality and ecosystem health in developed watersheds.

3. **Methods**

We appreciate your valuable suggestions to improve our Method section. We revised our approach to evaluate model outputs, expanded our discussion on the **model calibration and validation**, and **statistics** we used to quantify our model simulations in Results. Specifically, instead of showing the results from the simulation with highest NSE, we included more simulations from parameter sets yielding NSE greater than 0.5 for our calibrating period with gw_2 parameter less than 0.5. We also note that no parameters for N inputs and related processes were calibrated in the study, aiming to evaluate whether the model could reasonably estimate NO_3^- level by calibrating hydrologic parameters only. Please refer to **our response to Specific Comment #3 for details.**

From those behavioral simulations, we performed uncertainty analysis for streamflow and NO_3^- concentrations, which strengthened the argument that our model is capable of simulating NO_3^- dynamics without calibrating N related but only hydrologic parameters. With the updated method quantifying the uncertainty of our model, we updated our Results section substantially. Specifically, we composited simulations from 50 parameter sets with the **mean streamflow NSE from all simulations as 0.63 in the calibration period and 0.58 in the validation period**. For NO_3^- concentration and loads, we showed that we resampled the daily simulation to weekly means, as our sample were collected only once a week under conditions without large storm flows. Without calibrating N processes, our ensemble mean from 50 parameter sets estimates the daily mean concentration of 1.43 mg NO_3^--N L^{-1}, which is only 0.17 mg NO_3^--N L^{-1} lower than the observations in the study period.

We also thanks for your spotting of missing units in several equations, and we added units throughout all variables there.

Lastly, the detailed revisions are listed in our point-to-point responses to your suggestions in your attached PDF file (see Technical Corrections).

Specific comments

1. The motivation behind the study and the relevance of the research are well elaborated. However, the current state of the knowledge in regarding to the research questions is not elaborated. Are there no prior studies that have addressed similar research questions?

Response:

Thanks for the comment. To our knowledge, this is the first attempt to incorporate 1) spatial and temporal patterns of water and N inputs from irrigation, fertilizer, and septic systems in sub/exurban ecosystems and 2) evaluate their interactions with hillslope hydrologic and biogeochemical processes related to N retention. We reviewed several hydrologic and water quality models in the Introduction (in our response to 1. Unclear novelty), but found none include hillslope hydrology (i.e., explicit routing of water and nutrients within topography) and spatial and temporal patterns of N inputs from fertilizer and septic effluents simultaneously into one framework. Therefore, our augmented RHESSys model is by far the first fully distributed ecohydrological model that could meet the need to evaluate the current conditions of a watershed and designs of BMPs on forming hot spots for N retention.

2. The methodology should be written clearer and more structured. Given the fact that the study uses a rich base of data, for the reader it would be beneficial to have an overview of the data

used for setting up the model e.g., in the form of a Table providing specifications on each dataset and how it was employed in the study.

Response:
Thanks for the suggestion. We added a subsection to elaborate our calibration processes, and we also added a table in Appendix A to list all datasets we used for the study. We also changed our citation format thoroughly thanking to your suggestions.

Table A1. List of data in Baisman Run used to set up model and analyze water chemistry

Data	Detail	Source
Topography	Bare Earth DEM 2014	Baltimore County GIS, 2017
Land Use	Chesapeake Bay 1-m Land Use	Claggett et al., 2018
Discharge	United States Geological Survey	Gage ID: 01583580 (Baisman Run); 01583570 (Pond Branch)
Water Chemistry	Baltimore Ecosystem Study	Groffman et al., 2020; Castiblanco et al., 2023
Household Parcel	Baltimore County Parcels	Baltimore County GIS, 2019
Hydrologic Network	County Hydrolines	Baltimore County GIS, 2016

3. The model validation needs to be provided as well as the statistical methods for evaluating the simulation results.

Response:
Thanks for the suggestion. By summarizing your major comments to the Method section in the PDF file, we 1) reperformed model calibration and validation, 2) addressed why we chose the water year after 2010 to be evaluated in our study, 3) provided maps for initial values of soil properties from SSURGO in Supplementary (Fig. A2), which are further calibrated by multipliers to modify SSURGO properties' magnitudes but retain their spatial patterns.

Model validation:
We performed model calibration and validation again for our study, with the calibration period from water year 2013 to 2015 (Oct. 1, 2012 – Sep. 30, 2015) and validation period from water year 2016 to 2017 (Oct. 1, 2015 – Sep. 30, 2017). After calibration, we chose 50 behavioral parameter sets yielding highest NSE of streamflow to quantify uncertainty of model simulations using a 95% uncertainty boundary (i.e., 2.5th and 97.5th quantiles of simulations). The mean NSE of streamflow from the calibration period is 0.63 (range from 0.5 to 0.69), and 0.58 (range from 0.44 to 0.64) in the validation period. We noted that our calibration is only applied to hydrologic parameters of RHESSys, and no N-related parameters were calibrated.

Line 205:
We set the calibration period from water year 2013 to 2015 and validation period from water year 2016 to 2017. The original parameter values derived from SSURGO were further calibrated

by multipliers to vary their magnitudes but preserve the spatial patterns of soil hydraulic properties (Fig. A2). Specifically, the simulated streamflow was used to calibrate against the daily USGS discharge records (Gage ID: 01583580). From four thousands of parameter set realizations randomly chosen within specified limits, behavioural sets are chosen as yielding Nash-Sutcliffe efficiency (NSE; Nash & Sutcliffe, 1970) greater than 0.5 and fraction of groundwater loss to stream (i.e., gw2 in Table 1) less than 0.5 to estimate the ensemble means and uncertainties of model simulations. The latter condition was enforced to regulate the flashiness of groundwater dynamics, as BARN is found to have large saprolite storage to provide steady baseflow (Putnam, 2018). To assess uncertainty, we reported the 95% uncertainty boundaries for simulated streamflow and NO3- concentration and load from. Lastly, we noted that no calibration was performed for N inputs (e.g., fertilization rate and septic load) or N cycling/transport processes in the model, as an important aim of our methods is to evaluate the capacity of our model to regionalize to watersheds where no water chemistry but only streamflow observations were available.

Why resample simulations from daily to weekly means?

We compared the mean NO_3^- concentration between the observations and model's weekly means. We resampled our concentration simulations because the direct comparison of daily model simulation and observation is difficult. RHESSys simulates the daily mean NO_3^- under both low-flow and storm conditions, but our weekly grabbing samples were collected only in conditions with no large storm flows. Therefore, the weekly observations reflecting the average NO_3^- level of the week, which is better compared to our model's weekly means, though the bias would be unavoidable in this way. Also, since no calibration was performed for N-related parameters, we reported results for the whole study period.

Line 314:

To better compare our NO_3^- concentration results with the sampled weekly water chemistry from BES for BARN, we resampled the daily simulated concentration from RHESSys to weekly averages, expressed in unit of mg NO_3^--N L^{-1}. The weekly NO_3^- load was then estimated by the product of weekly mean NO_3^- concentration and streamflow, expressed in unit of kg N ha^{-1} $year^{-1}$. Note this approach may introduce bias for load as the once-a-week samples, typically not during major storms, and the observed daily mean discharges may not reflect the average load of the whole week.

Updated results for the validation period are shown in Table 3 with standard deviation reported from the means of NO_3^- concentration and load from 50 simulations for each scenario.

Table 1. Mean weekly NO_3^- concentration (mg N L^{-1}) and load (kg N ha^{-1} $year^{-1}$) and corresponding standard deviation from calibrated simulations for BES weekly observations (BARN and POBR) and RHESSys simulation scenarios in each season and the entire study period from water year 2013 to 2017

Variables	Season	Observation		RHESSys Scenarios			
		BARN	POBR	Both	Septic Only	Fertilizer Only	None

Concentration (mg N L^{-1})	Spring	1.5	0.02	1.4 (± 0.12)	0.76 (± 0.08)	0.77 (± 0.05)	0.27 (± 0.03)
	Summer	1.6	0.07	1.26 (± 0.13)	0.68 (± 0.1)	0.79 (± 0.1)	0.33 (± 0.06)
	Fall	1.57	0.06	1.41 (± 0.23)	0.77 (± 0.15)	0.94 (± 0.17)	0.41 (± 0.09)
	Winter	1.75	0.01	1.63 (± 0.18)	0.88 (± 0.12)	0.96 (± 0.1)	0.35 (± 0.05)
	Mean	1.6	0.04	1.43 (± 0.16)	0.77 (± 0.11)	0.87 (± 0.1)	0.34 (± 0.06)
Load (kg ha^{-1} $year^{-1}$)	Spring	10.93	0.01	8.86 (± 0.63)	4.84 (± 0.42)	4.77 (± 0.31)	1.62 (± 0.16)
	Summer	5.88	0.02	4.72 (± 0.36)	2.49 (± 0.25)	2.81 (± 0.23)	1.06 (± 0.16)
	Fall	4.72	0.01	4.72 (± 0.39)	2.57 (± 0.26)	3 (± 0.27)	1.23 (± 0.16)
	Winter	8.38	0.01	8.42 (± 0.68)	4.61 (± 0.46)	4.91 (± 0.38)	1.81 (± 0.18)
	Mean	7.44	0.01	6.68 (± 0.47)	3.63 (± 0.33)	3.87 (± 0.27)	1.44 (± 0.16)

Why water year after 2010?
The carbon and nitrogen cycling in RHESSys generally required a long spin-up period to stabilize. In BARN, the developed areas in headwaters of used to be farmlands before 2000. After about 10 years of slow transformation on the farmland, no major development was found, and the land cover has been stable as the form when the land cover data was collected in 2013. To reduce the uncertainty of N inputs due to changes of number of households, land cover, and fertilization practices, we chose water year 2012 to 2017 to assess our model in a stationary condition. We discussed this at line 175:

BARN had gradual suburban development in the headwater which converted from agricultural land over a few decades. New development was largely completed in the 1990s, with one last field developed in 2007-2009. Our study period could reduce the uncertainty of N inputs due to land cover change during urban development and allow for analysis of N dynamics in a stationary condition.

4. The results need to include estimates of errors and indication of deviation between the analyzed years.

Response:
Thanks for your suggestions. We thoroughly revised our Results section to update our simulation results from the 50 behavioral simulations.

For streamflow, we updated our multipliers values in Table 1, and showed the standard deviation from all our simulations at Line 343:

In the calibration period (i.e., water year 2013 to 2015, Fig. 3a), the ensemble of simulated mean (standard deviation) daily streamflow was 1.24 (□0.03) mm day^{-1}, with NSE of 0.63 (between 0.5 and 0.69) compared to the USGS observed 1.38 mm day^{-1}. In the validation period (Fig. 3b), the simulated ensemble mean (standard deviation) streamflow was 0.91 (□0.03) mm day^{-1}, with NSE of 0.58 (between 0.44 to 0.64) compared to the USGS's 0.86 mm day^{-1}.

For NO_3^- concentration/load, we reported the results of the ensembled mean value from 50 behavioral simulations in Table 3 (see above). Note we no longer reported the streamflow-weighted NO_3^- concentration in the revised version, as the reported ensemble results could 1) better assess the uncertainty of our model simulations and 2) be directly compared with results in Figure 3. We updated our contents in Sect. 3.2, Line 368:

We calculated weekly means of NO_3^- load and concentration of behavioural simulations. In our 5-year study period, the ensemble mean NO_3^- concentrations (Fig. 4a) for scenarios *none*, *septic only, fertilization only, and both* were 0.34, 0.77, 0.87, and 1.43 mg NO_3^--N L^{-1}, respectively (Table 4). The mean long-term observed concentration at the BARN USGS gauge was 1.6 mg NO_3^--N L^{-1}. Thus, the simulated bias of mean NO_3^- concentration considering both fertilization and septic loads decreased significantly from -1.26 mg NO_3^--N L^{-1} in the scenario *none* to 0.17 mg NO_3^--N L^{-1} in the scenario *both*. The 95% uncertainty boundary of weekly NO_3^- concentration in scenario *both* captured 67% of the weekly sampled observations. The seasonality of NO_3^- concentration is also well captured, except for the growing season (e.g., Jul. to Oct. in 2013 and 2016) when the model underestimated low flows (Sect. 3.1).

At line 394, we updated the ensemble results for water table depth, with a standard deviation of 1.1 m from 50 behavioral simulations. We also refined our results for the residential hillslopes (Fig. A6, hillslope 11 to 16) with urban development in BARN to see how human activities affect the ecohydrological behaviors.

The ensemble mean of water table depth (Fig. A4) from all behavioural simulations under scenario *none* was 4.52 m during the study period. Fertilization had overall negligible effects on watershed mean soil moisture or water table depth compared to the base (*none*) scenario (Fig. 6a – 6c), but minor increase of water table depth was detected in the residential areas, likely due to higher ET in lawns after fertilization. Septic processes decreased mean water table depth to 4.47 m by groundwater mounding, which increases shallow groundwater flow to surrounding patches along connected flowpaths. Specifically in septic drainage field patches, the mean water table depth decreased to 3.69 m (-0.66 m, -15%) in scenarios *both* and 3.72 m (-0.63 m, -14%) in *septic only* compared to the mean depth of 4.35 m, in scenarios *none* and *fertilization only*. With setting hillslope groundwater as the only source for septic process, we found groundwater withdrawal resulted in drier conditions (i.e., increase of water table depth) in riparian areas of these residential hillslopes (Fig. A6, hillslopes 11 to 16), where the mean water table depth increased by 5 (2%) and 8 (3.4%) mm in scenarios *septic only* and *both* compared to 219 mm depth in scenarios *none* and *fertilization only*. Though the **standard deviation** of each scenario from the 50 behavioural simulations was 1.1 m, the spatial distribution of soil moisture is consistent among all behavioural simulations.

We did the same for ET at line 406:

The watershed-scale mean ET was 43.9 mm month^{-1} in scenario *none* and 44.0 mm month^{-1} in scenario *fertilization only*. The standard deviation from 50 behavioural parameter sets was 0.8 mm month^{-1} for each scenario. With septic processes activated, mean ET increased to 44.1 and 44.2 mm month^{-1} in scenarios *septic only* and *both* in the residential hillslopes. […]. With septic processes activated, mean ET increased to 44.1 and 44.2 mm month^{-1} in scenarios *septic only* and *both* in the residential hillslopes, which could be contributed by the additional water extracted from groundwater to surface soil at the upland areas (in Fig. 6). When fertilization is activated in scenario *fertilization only*, ET in riparian areas of residential hillslopes decreased to (by) 54.7 (-0.1, -0.3%) mm month^{-1} compared to scenario *none*, while the upland of these hillslopes increased by 0.1 mm month^{-1}. This showed that fertilization in the upland residential lawns could support higher growth rate of vegetation but preventing water from draining towards downstream areas of a hillslope (in Fig. 6).

As a respond to the soil moisture condition, the ensemble watershed mean denitrification rate dropped compared to our previous simulation using only one parameter set. We reported the new results thoroughly at line 419:

Compared to scenario *none* (Fig. A5), the ensemble mean annual rates of denitrification at the watershed scale were 7.2, 7.8, and 9.1 kg N ha^{-1} year^{-1} in scenarios *fertilization only*, *septic only*, and *both*, respectively, increasing by 33%, 44%, and 68% (Fig. 6d – 6f & Table 4). The standard deviation from the 50 behavioural simulations was 1.5 kg N ha^{-1} year^{-1} for scenario *none* and *fertilization only* and 1.6 kg N ha^{-1} year^{-1} for scenario *septic only* and *both*. When fertilization and septic processes were activated, the denitrification rates increased at the residential hillslopes and their riparian areas. The only exception was found in scenario *septic only*, where 7 patches experiencing minor reduced denitrification (-1.4% in average). All these patches were found in riparian areas of residential hillslopes where the water table drops by 9 mm in average after the septic processes extracting groundwater in the upstream.

Lastly, according to your suggestions to improve our maps, we integrated the previous figures for water table depth and denitrification by only showing the differences from our scenarios (Fig. 6), and move the original maps as supplementary (Fig. A)

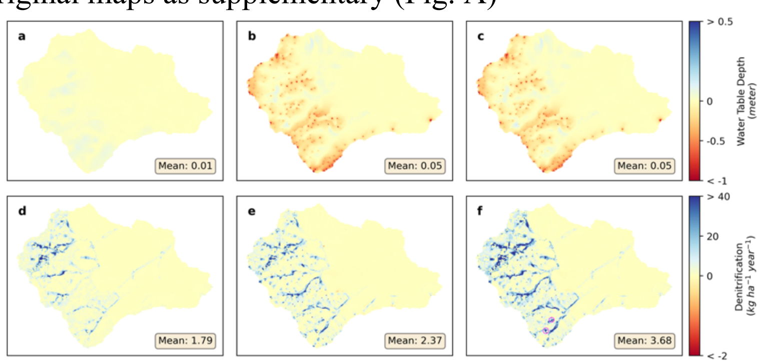

Figure 6. Ensemble mean differences of water table depth (top panel) and denitrification (lower panel) between scenario *none* and scenario fertilizer only (a & d), septic only (b & e), and both (c & f). The two hot spots of denitrification (i.e., wetlands in Fig. 1) were circled in (f).

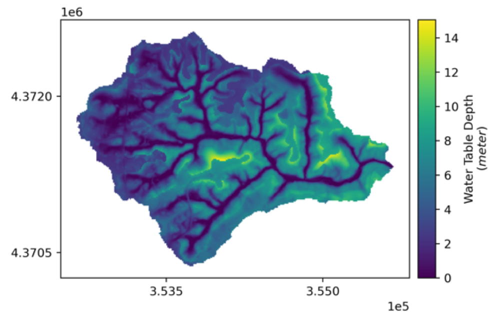

Figure A4. Spatial pattern of ensemble mean water table depth (meter) of Baisman Run during the entire study period (water year 2013 to 2017) from the 50 behavioral simulations. Map in projection NAD83 UTM 18N (EPSG: 26918).

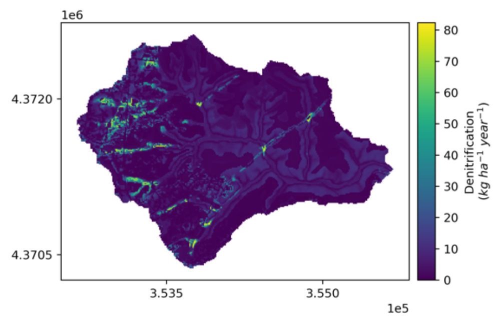

Figure A5. Spatial pattern of ensemble mean denitrification (kg N ha^{-1} year^{-1}) of Baisman Run during the entire study period (water year 2013 to 2017) from the 50 behavioral simulations. Map in projection NAD83 UTM 18N (EPSG: 26918).

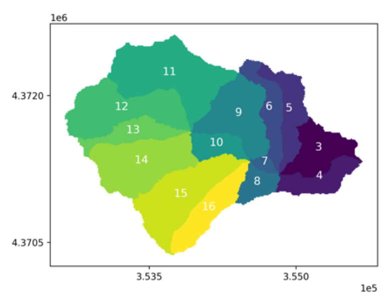

Figure A6. Hillslope indices of Baisman Run. Map in projection NAD83 UTM 18N (EPSG: 26918).

5. Some figures (maps) are obsolete as to my opinion the difference between them can not be spotted. Some figures have confusing axis labeling, lack titles for the legends and some have unclear captions (see technical corrections file for more details).

Response:

Thanks for your suggestions to improve our figures. According to your suggestions, the labels, legends, and captions of all figures are rephrased/fixed. Please refer to the end of this documents to see the updated figures. We made a new figure (Fig. 6, see above) to highlight the differences of water table depth and denitrification between scenario *none* and other three human N scenarios, since we agreed that the maps in the left panels showing absolute values are difficult to be differentiated. The spatial patterns of water table depth of scenario none were provided in Fig. A4 and A5 (see above). Other updated figures are shown here:

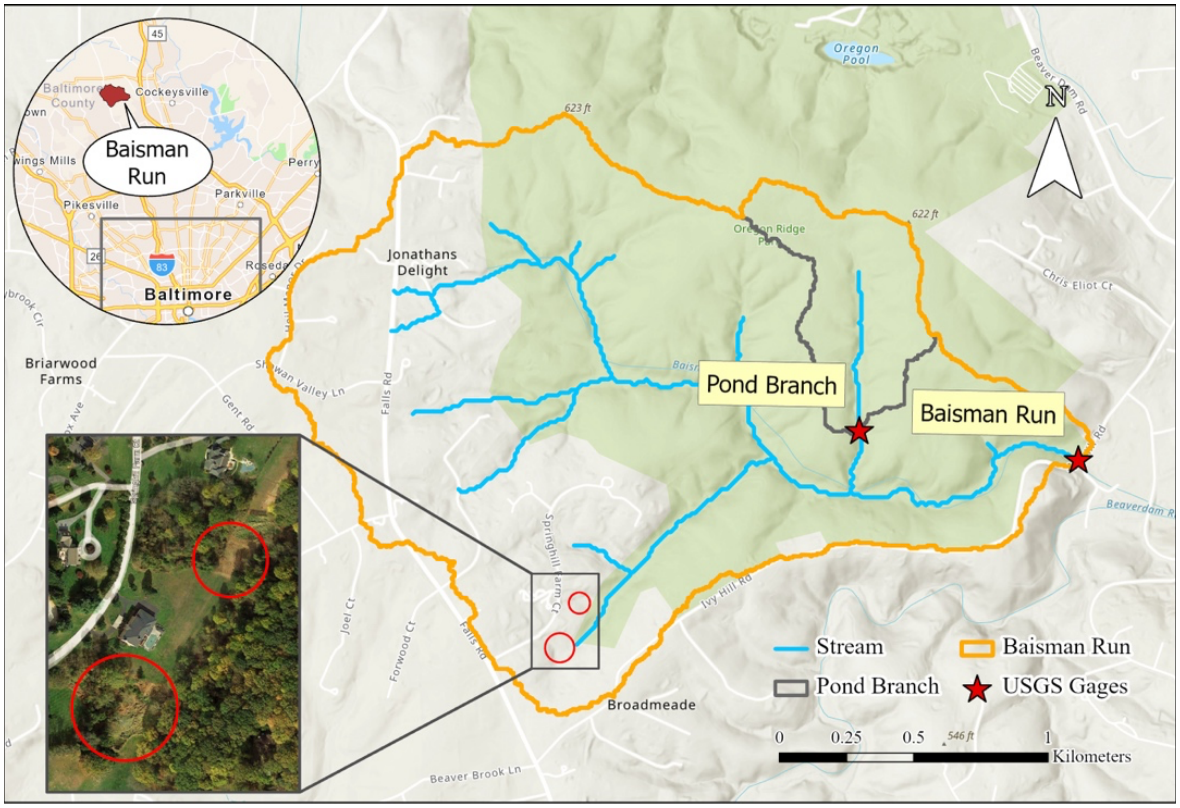

Figure 1. Study watershed Baisman Run (BARN) in suburban Baltimore County, Maryland (from ESRI). The black box highlights two N retention "hot spots": A sediment accumulation zone (upper circle) receiving drainage from roads and a constructed wetland (lower circle). These areas have a high capacity to prevent N from upland residential areas from being transported to streams.

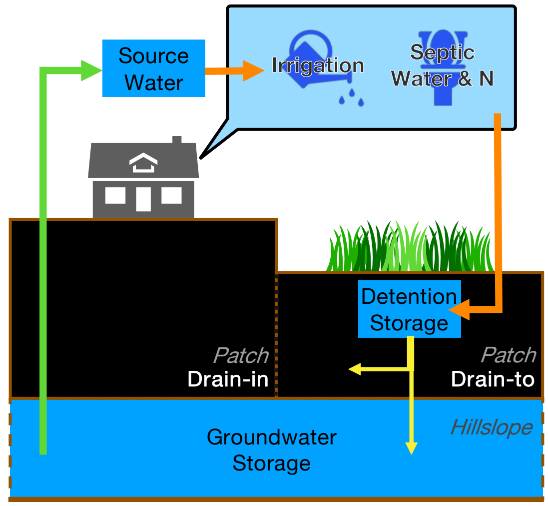

Figure 2. Groundwater extraction for irrigation and septic systems in the RHESSys model. The source water (green arrow) is extracted from groundwater storage of drain-in patches (i.e., house centroids) and redistributed (orange arrow) to surface detention in downstream lawn patches for septic effluents and irrigated lawn patches of a household. After redistribution of source water, infiltration to soil and percolation to hillslope groundwater (yellow arrows) would follow the original processing of RHESSys

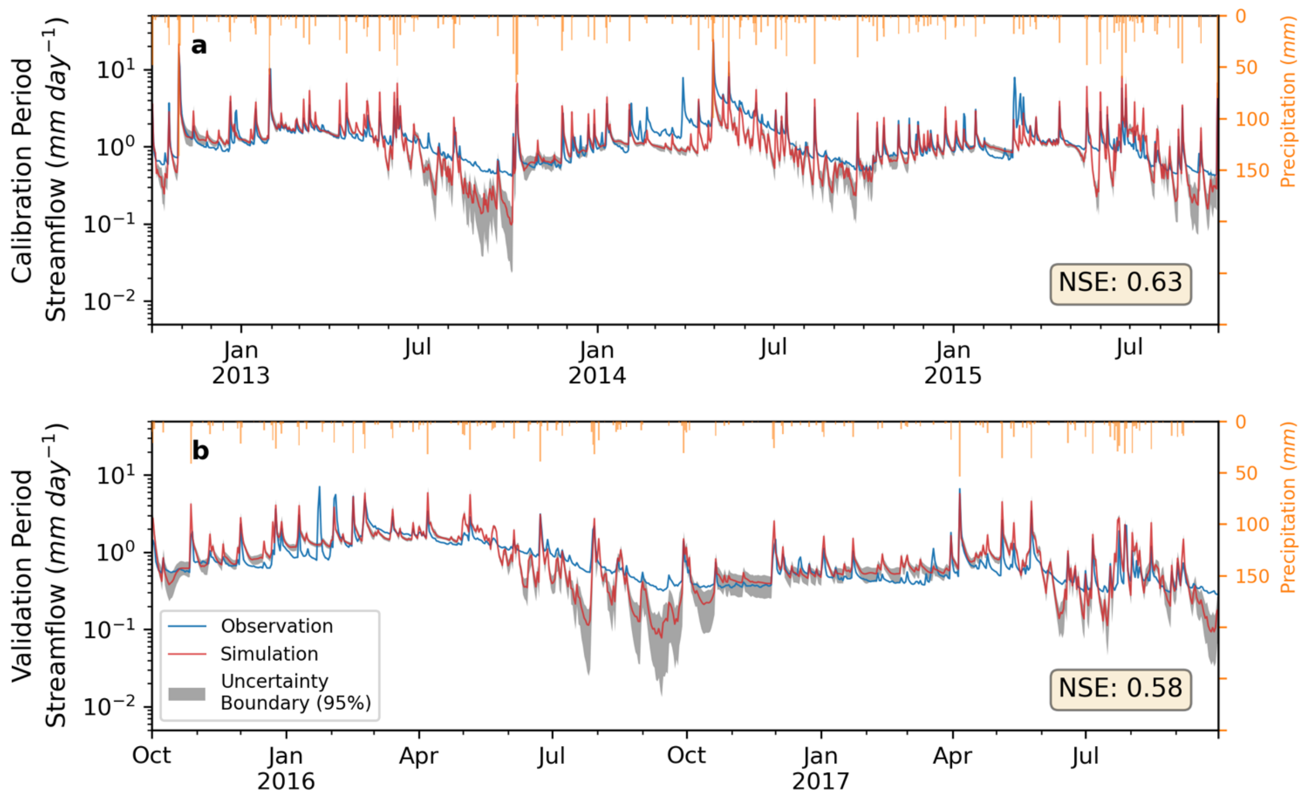

Figure 3. The ensemble mean of daily streamflow from simulations (red) with NSE greater than 0.5 and USGS observations (blue), with the daily 95% uncertainty range from 50 simulations in grey for the (a) calibration (Oct. 2012 – Sep. 2015) and (b) validation (Oct. 2015 – Sep. 2017) period. All simulations turned on irrigation, lawn fertilization, and septic processes

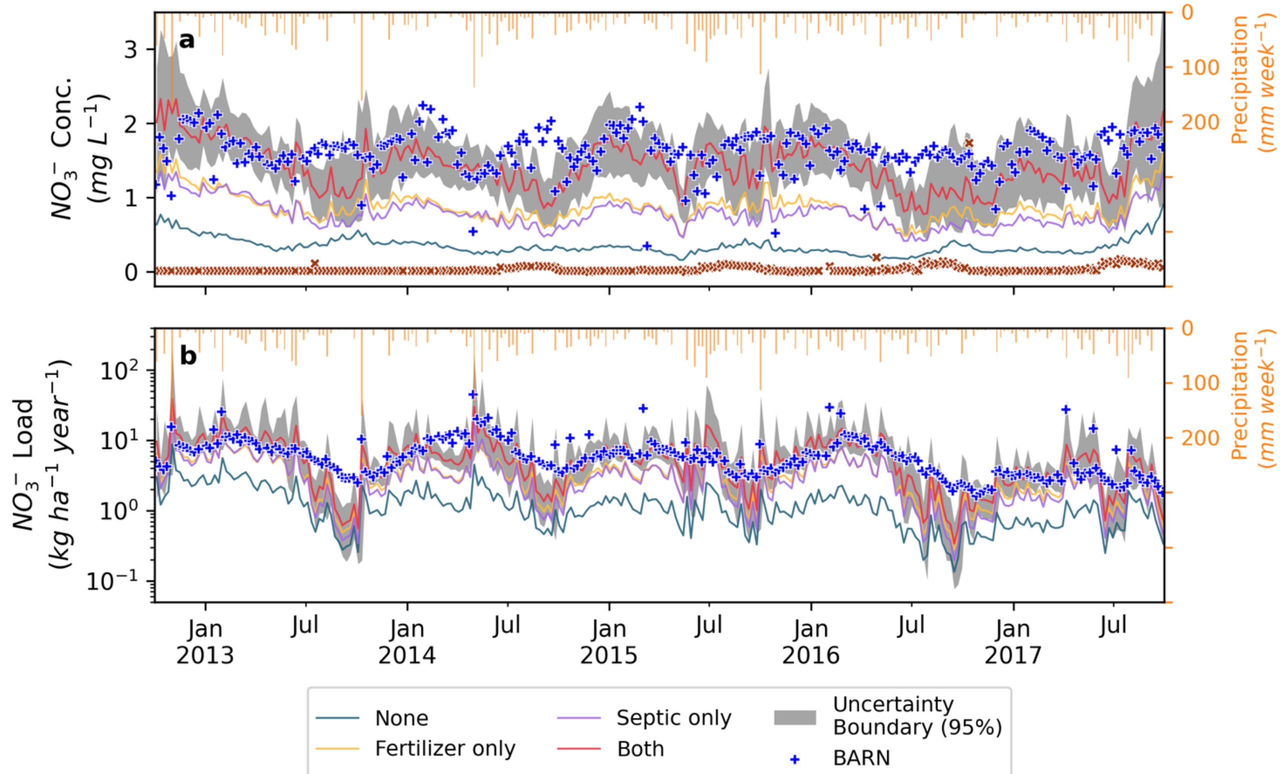

Figure 4. Ensemble weekly mean (a) NO_3^- concentration and (b) load at the outlet of Baisman Run over the entire study period (water year 2013 to 2017). The 95% uncertainty boundary for scenario *both* was shown in grey.

6. The manuscript needs to be reviewed i) to comply with the HESS requirements for manuscript composition, ii) for unit formatting as required by the submission guidelines, iii) for correct referencing of used data sets and software

Response:

Thanks for the comment. We have fixed our unit format. The hyperlinks for datasets have been fixed using the correct reference style according to HESS guidelines.

Technical corrections

Please find all technical corrections and suggestions for improvement of the figures as comments in the attached manuscript.

Response:

Thanks for all the corrections and suggestions your made to your manuscript. We listed major/important comments you have here for your reference. Corrections to typos are made in the manuscripts accordingly.

- Line 16: consistent terminology in abstract. Here you chose exurban and later in the abstract suburban.

  Thanks for the suggestion. We think suburban is the more general terminology for the study. Though BARN is not a typical suburban watershed, we noted it could be treated as a low-density suburban watershed, which is exchangeable with "exurban" for Baisman Run in the abstract.

  Excess export of reactive nitrogen in the form of nitrate (NO_3^-) from suburban watersheds is a major source of water quality […]. These processes in turn control the development of "hot spots" of nitrogen flux and retention in suburban ecosystems. We chose a well-monitored low-density suburban or exurban watershed, Baisman Run in Baltimore County […]

- Line 47: is this about the spatial distribution of them?

  Thanks for the question. Yes, this is about the spatial planning of the management practices. We rewrote here as:

  Line 54:
  Therefore, […] for effectively mitigating these environmental issues through spatially well-conceived and sustainable management practices.

- Line 93: because of the biogeochemical modules?
  Thanks for the question. RHESSys could simulate these detailed ecohydrological processes because it simulates fully distributed hillslope hydrology and coupled C and N dynamics in soil interacting with water and vegetations. We revised this sentence as at line 113:

  In this study, we augmented RHESSys to include household-level transfer of groundwater for lawn irrigation and domestic water use, with domestic water use routed to septic spreading fields. With coupling hillslope hydrology and biogeochemistry at spatially connected patches, RHESSys could estimate spatiotemporal patterns of soil moisture, lateral flow distribution, evapotranspiration, groundwater level, and N transportation, transformation, uptake, and immobilization in spatially explicit manners.

- Line 107: (The third research question) It's not very clear.

  Thanks for the comment. We agree the third research question could be further clarified as below at Line 130:

  What are the patterns of hot spots for N retention and associated implications to design future BMPs to promote N retention within suburban watersheds?

- Line 140: This sentence contains redundant information from L130, you can combine it.

Thanks for the suggestion. We have removed this sentence.

- Line 149: at which basis?

  Thanks for the comment. The atmospheric deposition of N was observation records from the National Atmospheric Deposition Program (NDAP) site MD99 (https://nadp.slh.wisc.edu/sites/ntn-MD99/). We added the reference at Line 170:

  Atmospheric N deposition was estimated as 11 kg N/ha/year from site MD99 of National Trends Network from National Atmospheric Deposition Program (NADP, 2022).

- Line 151: Could you elaborate why you chose this 5 year period out of the available data (that I understood to cover 2000-2018)?

  Thanks for the question. This is because there was continuous urban development before 2012 in BARN. The land use data were also acquired in 2013 and would not berepresentative to the conditions before it. We therefore chose study period after 2012 to make sure the stationarity of land over and excluded N loads uncertainties. We answered this in detail in the response to Specific Comment #3.

  Line 175:

  BARN had gradual suburban development in the headwater which converted from agricultural land over a few decades. New development was largely completed in the 1990s, with one last field developed in 2007-2009. Our study period could reduce the uncertainty of N inputs due to land cover change during urban development and allow for analysis of N dynamics in a stationary condition.

- Line 163: could you provide the number of houses this to quantify the stated uncertainty

  Thanks for the suggestion. We added this number and rewrote the sentence at Line 187:

  We identified 181 households, although 13 homes are located on the watershed divide, providing some uncertainty to the effective number of septic systems.

- Line 170: Did you perform this sensitivity analyses?

  Thanks for the question. We did test to set the starting date early and late (i.e., 35 days ahead and backward) for grass (the LAI would stay high for longer period), but found negligible changes in water and N dynamics for BARN. As this is beyond the scope of this study, we removed this sentence to avoid confusion for readers.

- Line 174: The initial estimated should be listed together with the calibrated multipliers in Table 1

Thanks for the suggestion. The maps of initial values of SSURGO soil properties are added as Fig. A2.

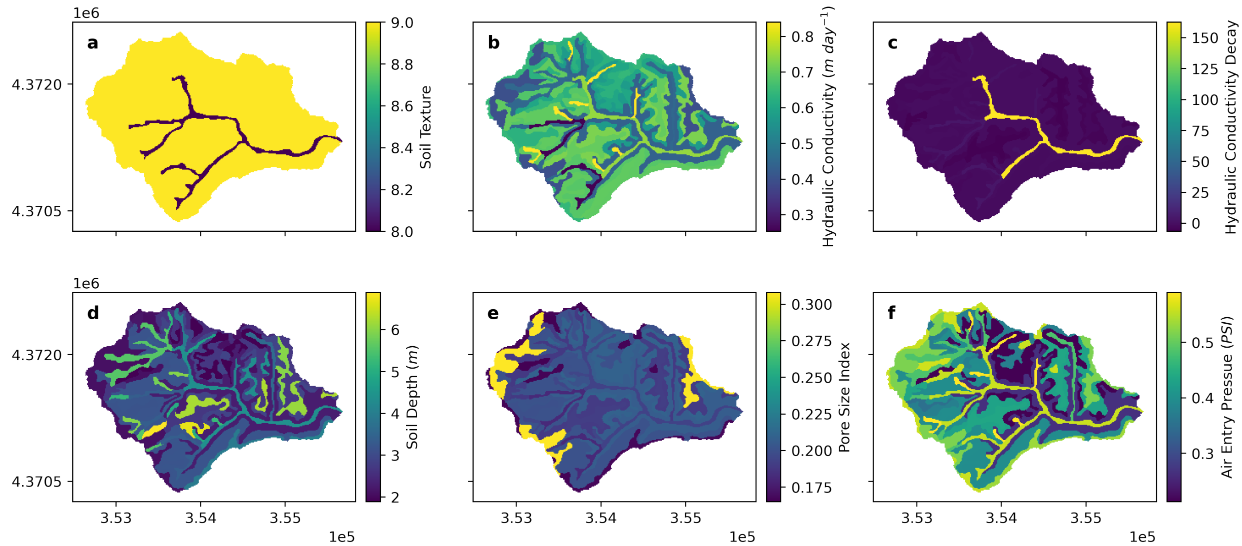

Figure A2. Soil types (a) based on SSURGO classification (USDA, 2019) and associated (b) lateral and vertical saturated hydraulic conductivities at surface (m day^{-1}), (c) lateral and vertical decay rates for lateral and vertical hydraulic conductivities, (d) soil depth (m), (e) pore size index, and (f) air entry pressure (pounds inch^{-2}).

We also updated our Table 1 to show ranges of multipliers applied on original SSURGO values.

Table 2. RHESSys parameters being calibrated and their physics (Tague and Band, 2004). Calibrated results shown as ranges of multipliers to original soil properties (Fig. A2 & A3) and groundwater component generating behavioural simulations with NSE greater than 0.5 for streamflow.

Parameter Groups	RHESSys Parameter Abbreviations		Detail	Source	Unit	Multiplier Range
Lateral soil hydraulics	s	m_l	Decay rate of lateral saturated hydraulic conductivity with depth	USDA SSURGO, 2019	-	0.31 – 2.91
		K_{sat0_l}	Lateral saturated hydraulic conductivity at the soil surface		m day^{-1}	0.38 – 2.93
		z	Soil depth		m	1.65 – 5.95
Vertical soil hydraulics	sv	m_v	Decay rate of vertical saturated hydraulic conductivity with depth	USDA SSURGO, 2019	-	0.51 – 1.98
		K_{sat0_v}	Vertical saturated hydraulic conductivity at the soil surface		m day^{-1}	0.52 – 1.98
Soil properties	svalt	b	Pore size index	USDA SSURGO, 2019	-	0.51 – 1.98
		φ_{ae}	Air entry pressure		pounds inch^{-2}	0.5 – 1.05
Groundwater dynamics	gw	gw_1	Fraction of bypass from the saturated zone to groundwater storage		-	0 – 0.13
		gw_2	Fraction of loss from groundwater storage to stream		-	0.03 – 0.5

gw_3	Fraction loss from surface to groundwater storage	-	0 – 0.07

- Line 179: How was the calibrated model validated?

  Thanks for the question. Please refer to our response to Specific Comment #3: Model calibration and validation.

- Line 180: The initial parameters and their units should be included in the table and the readers would profit from stating the Nash-Sutcliffe value here in the caption. Further, provide the meaning/names of the sensitivity parameters: s, sv, svalt, gw

  Thanks for your suggestions. We added a supplementary Fig. A2 (as above) to show the initial SSURGO values for each location of BARN, and improved our Table 1 as above.

- Line 212: Please add a (septic) reference if published.

  Thanks for the comment. We used data from Gold et al. (1990) and Lowe et al. (2009) to estimate the septic water and N load. The revised sentence at Line is:

  We estimated the N load from septic systems as 7.7 kg N capita^{-1} year^{-1} and water input as 110.5 m^3 capita^{-1} year^{-1} (~80 gal^{-1} capita^{-1} day^{-1}), resulting in a NO_3^- concentration of 70 mg N L^{-1} estimated from results of Gold et al. (1990), Lowe et al. (2009), and other sources for per capita water use and septic nitrogen concentrations.

- Line 248: Is this the local practice in the study area?

  Thanks for the question. This 4 mm day^{-1} threshold was set arbitrarily to constrain the groundwater extraction for septic or irrigation no more than this limit. Though we do not know the exact water extractions from each household, this limit allows abundant water usage that meets domestic water demand every day, and we did not see irrigation is beyond this limit during our study period assuming each house has 3.3 persons in average.

- Line 255: The survey by Law et al. (2004) and Fraser et al. (2013)?

  Thanks for the comment. We **removed** the sentence here to say it is consistent with survey results. We tried to include the maximal distance that people might irrigate their lawns, but there could be a quite large variations of this practice household by household.

- Line 257: Which method to you use to determine the differences between the scenarios?

  Thanks for the question. We compared the difference between the ensemble mean of NO_3^- concentrations from 50 simulations for each scenario (Fig. 4). As discussed above, the direct comparison from our simulated daily average NO3- concentration and weekly samples is difficult, we do not use traditional approaches (e.g., RMSE or R^2) in this study.

- Line 262: Could you elaborate on the method (resample daily to weekly)?

Thanks for the question. We answered this in our response to Specific Comment #3.

Section 2.4
To better compare our NO_3^- concentration results with the sampled weekly water chemistry from BES for BARN, we resampled the daily simulated concentration from RHESSys to weekly averages, expressed in unit of mg N L^{-1}. The weekly NO_3^- load was then estimated by the product of weekly mean NO_3^- concentration and streamflow, expressed in unit of kg N ha^{-1} $year^{-1}$. Note this approach may introduce bias for load as the once-a-week samples and the observed discharges at collecting days may not reflect the average load of the whole week.

- Line 273: I assume this subsection should be entitled Model calibration

Thanks for the suggestion. We changed this heading to "Model calibration and validation on streamflow". We also modified the axis titles in Figure 3.

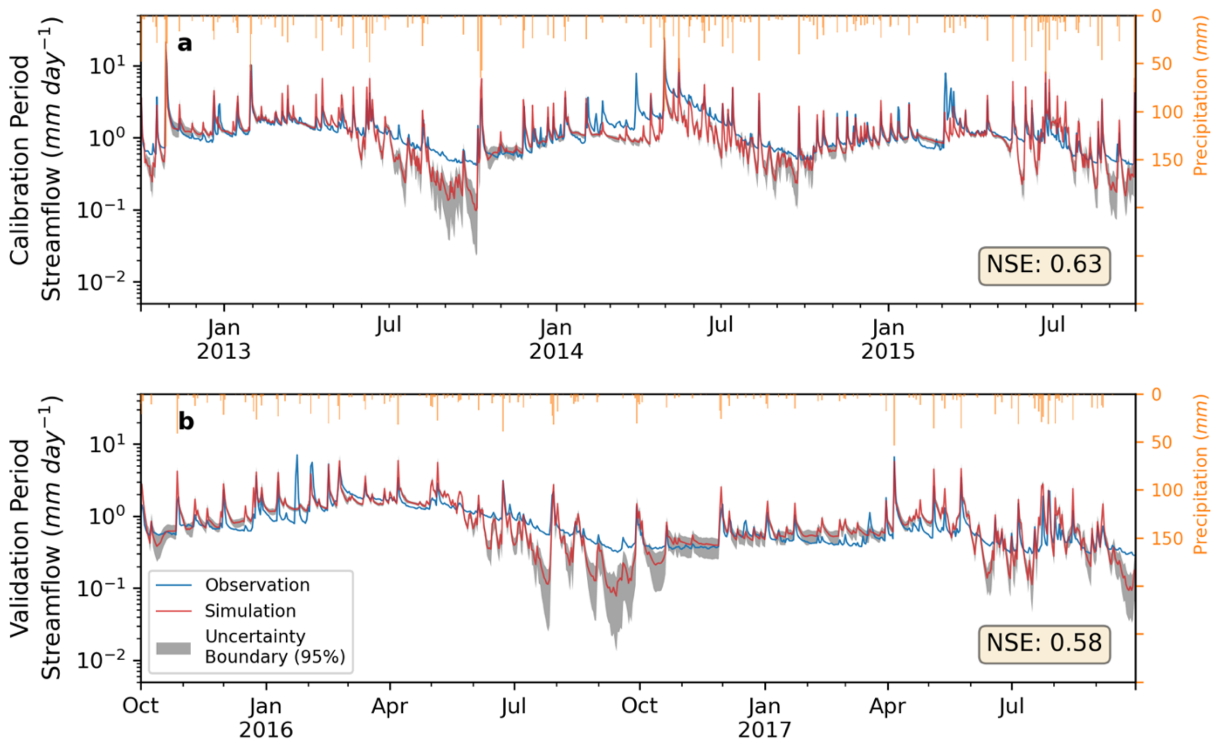

Figure 4. The ensemble mean of daily streamflow from simulations (red) with NSE greater than 0.5 and USGS observations (blue), with the daily 95% uncertainty range from 50 simulations in grey for the (a) calibration (Oct. 2012 – Sep. 2015) and (b) validation (Oct. 2015 – Sep. 2017) period. All simulations turned on irrigation, lawn fertilization, and septic processes.

- Line 282: Was this (underestimation of streamflow in growing season) the same in every modelled year? Please provide at least a standard deviation and consider elaborating on the results for every year.

  Thanks for the suggestion. With the uncertainty analysis, our 95% uncertainty range in Fig. 3 showed, most behavioral simulations would underestimate low flows in growing season (e.g., in 2013, 2016, and 2017). This is also found in previous studies in Baltimore (Miles, 2014). We discussed this at line 471:

  This may be due to local increases in septic water and nutrients increasing ET during the growing season, reducing groundwater recharge, lowering groundwater storage, and reducing watershed baseflow. We also noted that our model tended to underestimate the lowest streamflows during the growing season, which was also found in another suburban watershed, Dead Run, in Baltimore by Miles (2014).

  Reference

  Miles, B. C. (2014). *Small-scale residential stormwater management in urbanized watersheds: A geoinformatics-driven ecohydrology modeling approach* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

- Line 313: Please provide the standard deviations for NO3- concentration and load

  Thanks for the comment. We included the standard deviations of NO_3^- concentration and load in Table 3 (included in Specific Comment #3).

- Line 341: Please provide deviations with mean values

  Thanks for the suggestion. We added the standard deviation of annual mean denitrification among water years of our study period for each scenario.

  Compared to scenario *none* (Fig. A5), the ensemble mean annual rates of denitrification at the watershed scale were 7.2, 7.8, and 9.1 kg N ha^{-1} $year^{-1}$ in scenarios *fertilization only*, *septic only*, and *both*, respectively, increasing by 33%, 44%, and 68% (Fig. 6d – 6f & Table 4). The standard deviation from the 50 behavioral simulations was 1.5 kg N ha^{-1} $year^{-1}$ for scenario *none* and *fertilization only* and 1.6 kg N ha^{-1} $year^{-1}$ for scenario *septic only* and *both*.

- Line 381: Discussions and conclusion should be separate sections.

  Thanks for the suggestion. We have split out Discussion and Conclusion into two sections.

- Line 402: Please substantiate your discussion points with references. Do other studies using RHESSys encounter similar issues?

  Thanks for the question. The underestimation of low flows during growing season was also detected in previous RHESSys studies for another suburban watershed, Dead Run, in Baltimore by Miles (2014) at line 471.

  We also noted that our model tended to underestimate the lowest streamflows during the growing season, which was also found in another suburban watershed, Dead Run, in Baltimore by Miles (2014).

- Line 405: Please quantify the uncertainties (septic and fertilization)

  Thanks for the suggestions. We rephrased the sentence, emphasizing that each household has different fertilization or septic release rates and the spatial variation of N inputs could affect the N transport and transform and our model simulations. However, the actual rates of N inputs from fertilization and septic systems for all households is quite challenging to estimate at this point. We therefore, reported the range of surveyed fertilization rate from Law et al. (2004) to show the input variations.

  Line 490:
  In addition, we assumed identical N inputs acquired from Law et al. (2004) for all households in BARN, but the actual fertilization and septic effluents may have considerable spatial, and temporal variations which could impact the N cycling and transport significantly. Specifically, we used the annual fertilization rate on lawns as 84 kg N ha^{-1} from Law et al. (2004) in which the reported range of annual fertilization was from 10.5 to 369.7 kg N ha^{-1}.

- Line 408: Please substantiate with references. How many spin up years were used in other studies?

  Thanks for your suggestion. We added other studies for RHSSys, which used 500-year (Lin et al., 2015), 82-year (Son et al., 2019), or 47-years (Tague et al., 2013) spin-up periods to stabilize the model.

  Line 495:
  Compared to other RHESSys studies (e.g., Lin et al., 2015; Son et al., 2019; Tague et al., 2013), spinning up the model for 30 years may be insufficient to account for the export of this N from groundwater, which possibly caused the lower simulated mean NO_3^- concentration compared to BES measurements.

References

Lin, L., Webster, J. R., Hwang, T., & Band, L. E. (2015). Effects of lateral nitrate flux and instream processes on dissolved inorganic nitrogen export in a forested catchment: A model sensitivity analysis. *Water Resources Research*, *51*(4), 2680-2695.

Son, K., Lin, L., Band, L., & Owens, E. M. (2019). Modelling the interaction of climate, forest ecosystem, and hydrology to estimate catchment dissolved organic carbon export. *Hydrological Processes*, *33*(10), 1448-1464.

Tague, C. L., Choate, J. S., & Grant, G. (2013). Parameterizing sub-surface drainage with geology to improve modeling streamflow responses to climate in data limited environments. *Hydrology and Earth System Sciences*, *17*(1), 341-354.

- Line 423: Please compare the rates to the specific rates from the references like done for the hot spots in the following paragraph

Thanks for the suggestion. The denitrification rate at lawn was measured in lab with fixed environment settings (in Line 446, the Result section).

Assuming 210 days (~7 months) that denitrification would occur, Raciti et al. (2011) reported a denitrification rate of 204 kg N ha^{-1} $year^{-1}$ at 20 °C for saturated soil samples from fertilized lawns at the University of Maryland Baltimore County. At the same temperature, Suchy et al. (2023) reported a higher rate, 744 kg N ha^{-1} $year^{-1}$, when lawn soil samples collected from BARN lawns were saturated.

However, direct conversion of lab measured rates to the field measurements is impossible as the environment variables change all the time. We used Raciti et al.'s (2011) approach, the estimated denitrification rates were 13 and 40 kg/ha/year, respectively, using measurements from Raciti et al. and Suchy et al. These values were reported at Line 450. We added the cross reference to let readers to check the estimated rates.

Line 452:

The mean 25 and 85 percentiles of annual denitrification rate for lawns from all simulations in scenario *both* were 2.8 to 30.8 kg N ha^{-1} $year^{-1}$, respectively, which are quite comparable with the range of empirical measurements from low to high soil moisture conditions and various fertilization rates.

- Line 467: What does unaffected mean?

Thanks for your question. Our updated results suggested there was negligible change of water table depth at riparian areas at the whole watershed scale, but the drop of groundwater due to septic extraction is significant at hillslopes with dense residential development. We revised this sentence to explain this at line 562:

These results occur because while the septic effluent is depleted by evapotranspiration, the deeper groundwater that emerges in riparian areas is also affected at hillslopes with residential development. Thus, extraction of water for domestic use lowers riparian water tables even when this water is ultimately discharged back into the environment via a septic system.

- Line 478: Please specify where BMPs are sited effectively in a watershed. It would be interesting to run simulations with additional BMPs or BMPs in different locations throughout the watershed and compare those.

  Thanks for the comments. We mentioned that areas accumulating both upstream water and N inputs are ideal sites for BMPs. Running scenarios of siting BMPs in suitable areas would be the future research we will keep exploring.

  Line 575:
  These results suggest that effective siting of BMPs and a careful assessment of spontaneously existing (accidental) retention zones that accumulate both water and N loads from upstream can be used to achieve environmental goals for developed watersheds, by leveraging naturally occurring and built features providing ecosystem services.

- Line 481: The conclusion is very general. It needs to refer to your specific results presented before. Please elaborate whether the framework is applicable for other watersheds.

  Thanks for your suggestion. We elaborated our Conclusion with referring to our results of simulated NO_3^- concentration. We also specified our model can be applied to other suburban watersheds relying mainly on septic systems. Please refer to our response to your General Comment #2: Too general conclusion.