Note: The comments from the community are in black while the responses from the authors are in blue.

Dear authors,

I would like to respectfully raise a number of concerns regarding our ability – as a hydrological modelling community – to predict water quality at the global scale. Some are generic; some are more specific to this manuscript.

Response: We sincerely thank Dr. Tobias Krueger for taking the time to read our manuscript and share his insightful concerns. We appreciate his thoughtful feedback and the opportunity to engage with the important questions he raises about global-scale water quality modelling.

Representing the relevant processes at the global scale and having data to parameterise them

Land management is known to be very important for modulating pollution transfers. But how to even define those sets of rules (decision tables) for scheduling management at the 2x2km grid scale or HRU scale (page 5)? How do decisions made at the farm scale or smallholder scale aggregate? How to represent heterogeneity?

Response: Thank you for the question. To explain this simply, consider an agricultural grid or hydrologic response unit (HRU), as shown in Figure R1. Above each of these grids, we have several management layers, including crop type, cropping calendar, and fertilizer or manure application. For each agricultural grid or HRU, we read the corresponding values from these management layers and apply them accordingly. This process is pre-processed using a Python script, which organizes the information into a decision table (an example is shown in Figure R2), following the workflow developed in Nkwasa et al. (2022). This allows us to automate the assignment of agricultural management practices across all agricultural grids or HRUs.

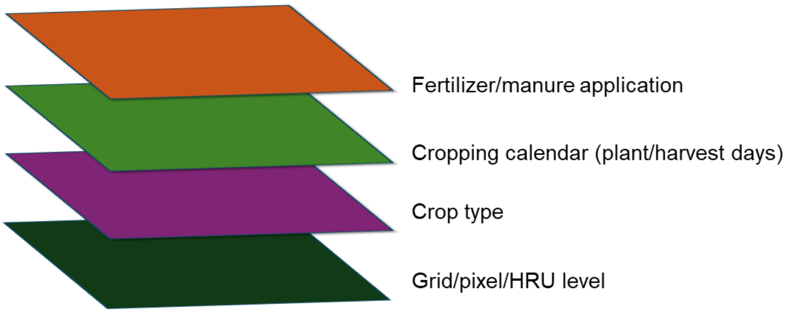
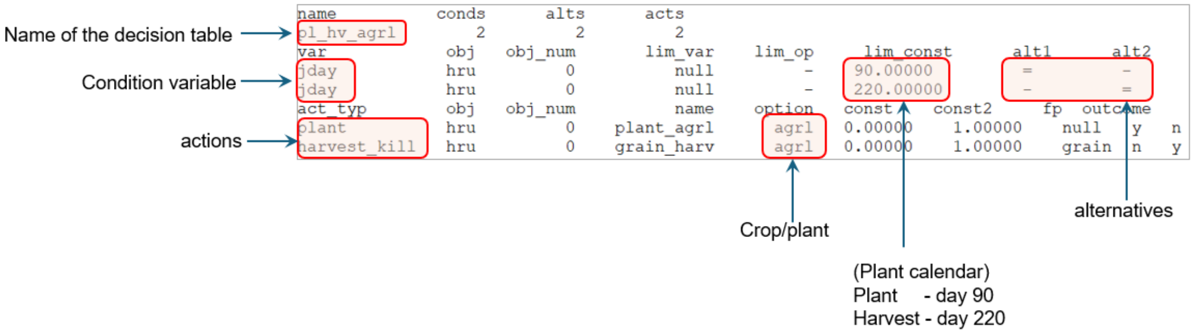

Fertilizer/manure application
Cropping calendar (plant/harvest days)
Crop type
Grid/pixel/HRU level

Fig. R1: Schematic representation of an agricultural grid or hydrologic response unit (HRU) and its associated management layers



```
name            conds     alts      acts
pl_hv_agrl          2        2         2
var              obj   obj_num       lim_var     lim_op     lim_const       alt1       alt2
jday             hru         0          null          -      90.00000          =          -
jday             hru         0          null          -     220.00000          -          =
act_typ          obj   obj_num          name     option     const      const2       fp  outcome
plant            hru         0    plant_agrl       agrl     0.00000    1.00000    null    y    n
harvest_kill     hru         0    grain_harv       agrl     0.00000    1.00000   grain    n    y
```

Fig. R2: Sample decision table showing the planting and harvesting of an agricultural crop

The question of farm or smallholder scale depends on the resolution of the datasets used for the management layers. In our study, we use freely globally available datasets, most of which are at a resolution of 0.5 degrees. Since our model operates at 2 km grid resolution, multiple 2 km grid cells fall within each 0.5-degree cell. This means that all agricultural grid cells within a single 0.5-degree grid share the same management practices.

In summary, the degree of heterogeneity we can capture depends on the resolution of the agricultural input datasets. To our knowledge, there are currently no global crop management datasets available at a higher resolution than those we have used. However, one strength of our workflow is its flexibility, if higher-resolution data become available in the future, this approach can incorporate and reflect that added heterogeneity.

We intend to add this short workflow description in the SI material.

What information exists to estimate tile drainage globally (figure 2)?

Response: We do not factor in tile drainage in our global study and we do not expect the model to perform so well in areas with a lot of artificially drained agricultural systems like in the Midwest United States and selected parts of Europe The only existing global dataset on drained areas, developed by Feick et al. (2005), presents a map depicting the fraction of drained land within 5 x 5 minute raster cells. While this map is useful when modelling global water resources, its still not suitable for use in model applications requiring area-differentiated delineation of drained land (Tetzlaff et al., 2009). This map also only show estimates of the area of land that is drained, or the area of land that is likely to be densely drained, but do not provide information on drain design characteristics or specific density metrics which are key in implementing tile drains. Additionally, there are no appropriately scaled estimates of nutrient delivery through tile drainage to compare with model results.

We intend to add a paragraph to the discussion which will highlight this limitation and bring it out as an area for both future model and data development under "section 4.2 Current missing links and future direction"

How do the uncertainties in global data (which are often generated themselves from models), e.g. on point sources, plant and harvest dates, and fertilizer and manure use rates, affect model predictions (page 6)?

Response: Thank you for this question. We agree that uncertainties in model structure and global input data such as point sources, planting and harvest dates, and fertilizer/manure application rates can significantly influence model outputs. These data are often derived from a combination of remote sensing, national statistics, and modelling approaches, each carrying inherent uncertainties due to spatial heterogeneity, spatial and temporal resolutions, reporting gaps, or even assumptions in upscaling methods. We acknowledge that these uncertainties can propagate through the model and affect both the magnitude and spatial patterns of predicted nutrient loadings and that such uncertainties are especially prevalent when modelling at large spatial extents.

We intend to add a paragraph to the discussion which will acknowledge the uncertainties and reflect on the implications for interpreting large-scale water quality modelling results.

Choosing an appropriate tool for what we want to predict

When it is said on page 13 that the primary objective of global water quality models is not to predict exact daily concentrations, then maybe a model that aims to represent processes at that scale is not appropriate. Monthly aggregated data are not predicted well by the daily model either, see below.

When the aim is instead to identify spatial and temporal pollution hotspots within global river networks (page 13), there may be other data-based methods that are better suited for this purpose.

Response: Thank you for this valuable comment. We agree that model selection should be driven by the intended purpose of the study. While this study does not aim to predict exact daily concentrations, but rather to initiate a global water quality model with the capacity to be applied at multiple temporal scales (from daily to annual), a process-based model like CoSWAT-WQ remains a suitable and effective choice given our objective. A key advantage of CoSWAT-WQ over more empirical or purely data-driven approaches is its ability to simulate cause-effect relationships between landscape processes and water quality outcomes. This allows us not only to identify where pollution occurs, but also to understand why it occurs by tracking the sources, pathways, and transformations of pollutants through the hydrological system. Furthermore, CoSWAT-WQ allows us to evaluate the effectiveness of different land and water management practices through scenario analysis, providing valuable insights for mitigating soil erosion and reducing nutrient loads from agricultural landscapes.

CoSWAT-WQ's semi-distributed structure captures spatial heterogeneity in land use, topography, and soil types, which is essential for modelling nutrient and sediment dynamics realistically across large and diverse catchments. Unlike black-box models, CoSWAT-WQ enables testing of land management scenarios by altering inputs or different cropping practices, making it a valuable tool for exploring mitigation/adaptation strategies and projecting future changes.

Additionally, we selected the CoSWAT-WQ model because SWAT is one of the most widely used water quality models across different spatial and temporal scales, supported by a large user community. This broad adoption means that the model is well-tested and familiar to many researchers, allowing others not only to contribute to its continued improvement but also to use parts of our global model such as individual subbasins for their own regional studies. We also anticipate that, with continued community involvement, simulations from this global water quality model will be further refined at higher temporal resolution (daily or even sub-daily).

We intend to rephrase our objective to show that we to initiate a global water quality model with the capacity to be applied at multiple temporal scales (from daily to annual).

Model calibration

The model in this study is said to be uncalibrated (page 6). How were parameter values determined instead?

Response: Thank you for your question. We used the default CoSWAT-WQ model parameters, which are based on literature and previous applications of the SWAT and SWAT+ models (Arnold et al., 2012; Arnold et al., 2013). As noted in the discussion (Line 270), future research will focus on model calibration and validation using observational data, especially from data-rich regions, to enhance the model's accuracy and reliability.

We intend to state clearly that we use the default CoSWAT-WQ model parameters.

The lack of sufficient observational data, strong spatial biases in available observations and uncertainties in input datasets (page 7) are put forward as obstacles to model calibration. What do these obstacles say about our ability to run a model at all?

Response: Thank you for your question. Indeed, these challenges are well-recognized and present significant obstacles to model calibration, particularly at the global scale. However, we believe these constraints should not preclude efforts to run global models. On the contrary, such modelling initiatives are essential for guiding future data collection, identifying priority areas for monitoring, and advancing methodological development. Our current approach focuses on identifying major spatial and temporal pollution hotspots, which can be cross-validated through model intercomparison efforts or supported by existing literature. While we acknowledge that uncertainties remain high, global models like this one, can still provide valuable insights especially for exploring scenarios and informing policy. Importantly, we see these modelling efforts as part of an iterative process. As more observational data and even remote sensing capabilities becomes available and modelling capacities continue to improve, we expect model performance and reliability to improve. We therefore view the current stage as an essential foundation for continuous development and improvement, rather than a limitation to modelling itself.

It is argued that the model is theoretical applicable in regions with no calibration data without a significant loss in performance (page 7). However, the hydrological modelling literature has repeatedly shown that even physically well-founded models require calibration to make them work in particular places. Among other things, this is because the processes that are represented are "effective" processes at the scale of application, either because physical theory is moved between scales or scale-bound abstractions are used in the first place (the MUSLE is an obvious example in this study).

Response: Thank you for this important observation. We fully agree that, as shown in a wide body of hydrological literature, including with SWAT, even physically based or semi-distributed models require calibration to effectively represent catchment-specific "effective" processes, especially when operating at larger spatial scales. Our statement on page 7 was not intended to suggest that calibration is unnecessary, but rather to reflect findings in the literature indicating that SWAT can still provide reasonable outputs in ungauged basins when default or parameters are used. For example; Chen et al. (2023) and Niraula et al. (2011) compared critical source areas (CSAs) for sediment and nutrients of a calibrated and uncalibrated SWAT model in southwest China and southeast U.S respectively. Both studies showed that CSAs locations had high similarity (81–93%) with and without calibration. These studies show that the SWAT model can be applied to identify pollution hotspots without calibration in watersheds especially data poor cases.

We intend to revise our statement on page 7 to clarify that it does not suggest calibration is unnecessary, but rather reflects findings in the literature indicating that SWAT can still produce reasonable outputs in ungauged basins when default parameters are used.

When it is said that uncalibrated physically-based models are often preferred for global change assessment on page 7, I am unclear whether this is because they are uncalibrated or because they are physically-based, and why exactly either of these features makes them preferrable. Keeping in mind the limits of physically-based models discussed above.

Response: Thank you for this question and sorry for the lack of clarity. Our intention was to highlight that physically-based models are preferred in global change assessments because they are physically-based, not because they are uncalibrated. Modelling approaches can broadly be categorized into statistical and physically-based models (Caissie, 2006). Statistical models estimate water quality variables using empirical relationships such as regression, stochastic methods, or machine learning based on available observations (Wanders et al., 2019). These methods can lead to satisfactory results; however, these statistical relationships need to be determined based on available observations. Physically-based models, on the other hand, use underlying physical relationships between water quality and drivers such as meteorological, hydrological, and socio-economic variables. Their advantage over statistical methods is most evident in ungauged basins and in assessing the impacts of global changes such as climate and socio-economic changes on water quality. Physically based models are particularly suited for these scenarios, as they rely on established physical processes to estimate nutrient levels from climatic, hydrological, and socio-economic inputs (Jones et al., 2023). This enables reliable predictions even under unobserved or future conditions, making them applicable in the context of global change. Of course, physically based models have limitations of large data requirements and larger computational requirements (Caissie, 2006).

We intend to revise our statement on page 7 to clarify that it the intention is to highlight that physically-based models are preferred in global change assessments because they are physically-based and not because they are uncalibrated.

Model intercomparison

If all models in an intercomparison are affected by the aforementioned shortcomings in process representation and data, can a model intercomparison really enhance confidence in large-scale water quality models as argued on page 7?

Response: Thank you for this question. While it's true that all models in an intercomparison may share some common limitations in process representation and data inputs, model intercomparison can still enhance confidence in large-scale water quality modelling. Model intercomparison is valuable not because it guarantees "truth" in estimates, but because it reveals robust patterns and structural uncertainties across models. When different models with different structural setups and assumptions converge on similar outputs, this consistency can enhance confidence in those specific results. Alternatively, when models diverge, the intercomparison highlights areas where uncertainty is greatest and where further research or data improvement is most needed. So while intercomparison cannot eliminate the fundamental limitations shared by models, it can still provide useful insights, highlighting consistencies across models, revealing the range of possible outcomes.

In this study, the log-scaling of the predictions in figure 5 make the discrepancies between the two models appear smaller than they are on the original scale, which are often an order of magnitude. When recalculating the R2 statistic on the back-transformed data, R2 decreases from 0.71 to about 0.41 for TP, for example. I am unclear why the same data appear twice with two different symbols in the graph.

I suggest that a model intercomparison would benefit from a deeper exploration of the mechanics of the individual models. What might explain the differences or similarities in predictions?

Response: Thank you for this question. We agree that the more detail is needed in the global model intercomparison between CoSWAT-WQ, IMAGE-DGNM (Beusen et al., 2015; Beusen et al., 2022) and MARINA-Multi (Micella et al., 2024). To expand more, model differences can be partially attributed to structural variations among the models, including differences in process representation and the estimation of diffuse and point source inputs under scenario assumptions. For instance; (i) MARINA-Multi operates at the sub-basin scale, IMAGE-DGNMIMAGE-GNM at a 0.5° grid resolution, and CoSWAT-WQ at the hydrologic response unit (HRU) level. (ii) CoSWAT-WQ does not simulate dynamic land-use change beyond 2010, considering only changes in fertilizer and manure application while holding land use constant. This assumption may lead to inconsistencies in nutrient estimates in regions experiencing substantial expansion of agricultural land or loss of natural land cover, which are captured by the other models. (iii) Both CoSWAT-WQ and MARINA-Multi do not account for nutrient legacies in groundwater, which are included in the IMAGE-DGN model. (iv) IMAGE-DGN model incorporates aquaculture as an anthropogenic nutrient, a factor not accounted for in the CoSWAT-WQ and MARINA-Multi models. Beyond the above highlighted differences, more distinctions in the main model assumptions and implementations of CoSWAT-WQ, IMAGE-DGN and MARINA-Multi are given in Table R1. Although these examples do not capture all model differences, such structural and process variations inevitably influence model outputs.

Table R1: An overview of the major assumptions and implementations of CoSWAT-WQ, IMAGE-DGNM (Beusen et al., 2015; Beusen et al., 2022) and MARINA-Multi (Micella et al., 2024) models

Parameter	CoSWAT-WQ	IMAGE-DGNM	MARINA-Multi
Landuse	Fixed on landuse of 2010 (ISIMIP)	Dynamic landuse from IMAGE-DGNM	Fixed land use of 2010
Point source	IMAGE-DGNM	IMAGE-DGNM	IMAGE-DGNM
Fertilizer	IMAGE-DGNM	IMAGE-DGNM	IMAGE-DGNM
Manure on crops	IMAGE-DGNM	IMAGE-DGNM	IMAGE-DGNM
Manure on grassland	No grassland	IMAGE-DGNM	IMAGE-DGNM
Atmospheric deposition (N)	ISIMIP	IMAGE-DGNM	IMAGE-DGNM
Uptake crops	CoSWAT-WQ	IMAGE-DGNM	IMAGE-DGNM
Crop types	Sugar cane, maize, soybean, wheat, banana	Upland crops, wetland rice and legumes	based on IMAGE-DGNM
Spatial calculation level	HRU	Grid cells of 0.5 * 0.5 degree	sub-basins
Hydrology	CoSWAT-WQ	IMAGE-DGNM - PCRGLOBWB	VIC (Variable Infiltration Capacity)

			model (van Vliet et al., 2016)
Legacy in groundwater (N)	No	Yes	No
Accumulation in soils (P)	yes	yes	no
N, P inputs from floodplains	no	yes	no
Aquaculture (N and P)	no	yes	no

We have summarized the differences between three global water quality models in a Table R1, and we will include this text and table discussing the model differences in the main text and supplementary material of the revised manuscript.

Model comparison with observations

The model does not fit the monthly data well (as seen by the KGE values and figure 8). Data uncertainties are mobilised to explain mismatches but not instances of better fits (page 10). However, if the data are uncertain then the better fits might equally well be due to chance.

Response: Thank you for this comment. We agree that while mismatches between model output and observations are partially attributed to data uncertainty, we also recognize that apparent agreements may likewise result from uncertainty rather than actual model skill. This possibility is consistent with findings in the hydrological modelling literature (Kavetski et al., 2003; Beven, 2006), where both poor and seemingly good fits can be driven by uncertain or noisy data. This also applies to calibrated models where it is significantly possible for calibrated parameter values to compensate for different types of errors (structural or observational) (Beven, 2006).

We intend to include this clarification in the discussion to provide a more balanced interpretation of model performance.

The recognition on page 13 that a comparison with observations and other global nutrient models can identify which parts of the model can be improved but that it is difficult to determine specific areas for improvement to me points to a promising research programme.

Response: We appreciate the reviewer's recognition of uncertainties in the input data (e.g., land use, climate, hydrology, wastewater flows) and the inherent complexity of surface, subsurface, and instream process interactions, to inform a promising direction for future research. These factors collectively complicate the identification of specific areas for model improvement. Disentangling the relative contributions of these uncertainties and process complexities represents a critical initial step towards improving model structure and parameterization. This is actually be a promising and necessary direction for future research, which will ultimately enhance the robustness of global nutrient models.

The recognition on page 13 that hydrology plays a particularly important role in model results whereas the discussion focuses on nutrient-related processes points into the same direction. The hydrological model setup is in a parallel discussion at EGUsphere.

Response: Thank you for the comment. Indeed, the hydrological model setup is discussed in a parallel manuscript at EGUsphere, which has received a positive initial evaluation and is currently under revision. To clarify, the revision does not involve any structural changes or re-runs of the hydrological model; rather, it focuses on improving the contextual presentation and clarity of the model description.

I agree with the assessment on page 14 that model validation and calibration to improve accuracy and reliability are needed.

Response: We are in agreement and this is the focus of ongoing and future work in our model development.

I hope these comments are helpful for improving global water quality assessments.

Response: We appreciate the reviewer's thoughtful comments and hope that the revised manuscript will contribute meaningfully to improving global water quality assessments.

References

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B., Neitsch, S.L., 2013. SWAT 2012 input/output documentation. Texas Water Resources Institute.

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R.D., Van Griensven, A., Van Liew, M.W., 2012. SWAT: Model use, calibration, and validation. Trans. ASABE 55, 1491–1508. https://doi.org/doi:10.13031/2013.42256

Beusen, A.H.W., Doelman, J.C., Van Beek, L.P.H., Van Puijenbroek, P.J.T.M., Mogollón, J.M., Van Grinsven, H.J.M., Stehfest, E., Van Vuuren, D.P., Bouwman, A.F., 2022. Exploring river nitrogen and phosphorus loading and export to global coastal waters in the Shared Socio-economic pathways. Glob. Environ. Change 72, 102426. https://doi.org/10.1016/j.gloenvcha.2021.102426

Beusen, A.H.W., Van Beek, L.P.H., Bouwman, L., Mogollón, J.M., Middelburg, J.B.M., 2015. Coupling global models for hydrology and nutrient loading to simulate nitrogen and phosphorus retention in surface water–description of IMAGE–GNM and analysis of performance. Geosci. Model Dev. 8, 4045–4067.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol., The model parameter estimation experiment 320, 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Caissie, D., 2006. The thermal regime of rivers: a review. Freshw. Biol. 51, 1389–1406. https://doi.org/10.1111/j.1365-2427.2006.01597.x

Chen, M., Janssen, A.B.G., de Klein, J.J.M., Du, X., Lei, Q., Li, Y., Zhang, T., Pei, W., Kroeze, C., Liu, H., 2023. Comparing critical source areas for the sediment and nutrients of calibrated and uncalibrated models in a plateau watershed in southwest China. J. Environ. Manage. 326, 116712. https://doi.org/10.1016/j.jenvman.2022.116712

Feick, S., Siebert, S., Döll, P., 2005. A digital global map of artificially drained agricultural areas.

Jones, E.R., Bierkens, M.F.P., Wanders, N., Sutanudjaja, E.H., van Beek, L.P.H., van Vliet, M.T.H., 2023. DynQual v1.0: a high-resolution global surface water quality model. Geosci. Model Dev. 16, 4481–4500. https://doi.org/10.5194/gmd-16-4481-2023

Kavetski, D., Franks, S.W., Kuczera, G., 2003. Confronting Input Uncertainty in Environmental Modelling, in: Calibration of Watershed Models. American Geophysical Union (AGU), pp. 49–68. https://doi.org/10.1029/WS006p0049

Micella, I., Kroeze, C., Bak, M.P., Strokal, M., 2024. Causes of coastal waters pollution with nutrients, chemicals and plastics worldwide. Mar. Pollut. Bull. 198, 115902. https://doi.org/10.1016/j.marpolbul.2023.115902

Niraula, R., Kalin, L., Wang, R., Srivastava, P., 2011. Determining nutrient and sediment critical source areas with SWAT: effect of lumped calibration. Trans. ASABE 55, 137–147.

Nkwasa, A., Chawanda, C.J., Jägermeyr, J., van Griensven, A., 2022. Improved representation of agricultural land use and crop management for large-scale hydrological impact simulation in Africa using SWAT+. Hydrol. Earth Syst. Sci. 26, 71–89. https://doi.org/10.5194/hess-26-71-2022

Tetzlaff, B., Kuhr, P., Wendland, F., 2009. A new method for creating maps of artificially drained areas in large river basins based on aerial photographs and geodata. Irrig. Drain. 58, 569–585. https://doi.org/10.1002/ird.426

van Vliet, M.T.H., van Beek, L.P.H., Eisner, S., Flörke, M., Wada, Y., Bierkens, M.F.P., 2016. Multi-model assessment of global hydropower and cooling water discharge potential under climate change. Glob. Environ. Change 40, 156–170. https://doi.org/10.1016/j.gloenvcha.2016.07.007

Wanders, N., van Vliet, M.T.H., Wada, Y., Bierkens, M.F.P., van Beek, L.P.H. (Rens), 2019. High-Resolution Global Water Temperature Modeling. Water Resour. Res. 55, 2760–2778. https://doi.org/10.1029/2018WR023250