Community comment: Jason Ke

This manuscript developed a global water quality model DynQual V1.0 and interpreted its results for TDS, BOD, and FC. Overall this manuscript is well-written with good-quality figures. Model results regarding the spatial patterns of concentration and temporal trends by region and economic development are interesting. However, there are some concerns about the model evaluation.

1) there seems no description of model calibration. How was the calibration done for the global water quality model? Is it a simultaneous calibration for both hydrology (discharge) and water quality (Tw, TDS, BOD, FC), or a two-step calibration strategy with discharge calibrated first followed by water quality calibration? Since the author mentioned that discharge was very important for model results (Supplement, Line 295), I would assume the discharge has to be well-calibrated before modeling water quality.

DynQual, in addition to the underlying model *PCR-GLOBWB2* (Sutanudjaja et al., 2018) and the original water temperature model *DynWat* (Wanders et al., 2019), are uncalibrated. This is an important point that is currently missing, and one that we will explicitly state and justify in the manuscript during revisions.

The process (physically) based nature and global scale of our model, combined with large data gaps in both space and time complicate meaningful calibration. For both the hydrology and water quality aspects, we want to avoid the creation of ‘calibration artefacts’ – whereby deficiencies in process descriptions are concealed by parameter estimation and there is a tendency to be biased towards areas/subsets where data availability is high, which could introduce a lack of spatial parameter consistency between different locations. This is especially problematic for calibrating large-scale water quality models, due to the strong spatial biases in the observations and the large number of parameters that need to be estimated. On the other hand, uncalibrated physical models can theoretically be applied in ungauged basins without loss of performance (Hrachowitz et al., 2013; Wanders et al., 2019).DynQual will also be used for global change assessments with different climatic and socio-economic scenarios, for which we preferably work with uncalibrated models.

All input parameter values for the global water quality modelling work are derived from previous (global) work (e.g. UNEP 2016, Reder et al., 2015, van Vliet et al., 2019). While PCR-GLOBWB2 is also parameterised on the basis of existing global datasets without further calibration, extensive model validation of hydrological simulations (e.g. discharge) have been performed using GRDC discharge data (5,363 stations) and GRACE total water storage thickness, as described in more detail in Sutanudjaja et al., (2018).

Please also note that the majority of global water quality models are currently uncalibrated, including QUAL (van Vliet et al., 2021), WorldQual (Voß et al., 2012) and IMAGE-GNM (Beusen et al., 2015) due to their process-based nature and aim of simulating water quality in data scare regions.

2) The model evaluation that is very important to the model development paper seems underdeveloped. It is essential to evaluate the model performance before the model result interpretation. For example, it is ideal to evaluate model performance whenever data are available. For example, there are 27,238 stations with TDS data in the Supplement. Perhaps the author could do the following evaluation regarding 1) spatial pattern of mean concentration (e.g., model mean vs. data mean from the station with high data availability); 2) temporal dynamics regarding seasonal fluctuations and long-term trends (e.g., Fig 11, add data points to the temporal trend plots to evaluate if the model could reproduce the long-term trends)

We agree that model evaluation is very important, and is somewhat underdeveloped in the current submission. Please note that some validation results are also presented in the Supplementary Information of Jones et al., (2022) (https://www.nature.com/articles/s43247-022-00554-y). We will endeavor to further improve this section, both in the manuscript and SI, e.g. by adding global maps of time-average model evaluation statistics at locations with water quality observations.

Overall, and as argued in other global water quality modelling work (e.g. Beusen et al., 2015; UNEP 2016) we believe that the overarching purpose and applications of the model must be considered when evaluating the performance. DynQual is ambitious in its aims to model surface water quality 1) using a consistent approach at the global scale; 2) dynamically; 3) considering multiple water quality constituents (multipollutant approach); and 4) at a high spatiotemporal resolution. We believe the presented approach is appropriate for investigating key research questions that only large-scale modelling efforts can help to answer, for example, those related to: global hotspot identification (Figure 4-6), the relative importance of different sectors across the globe (Figure 7) and meta-trends in water quality dynamics (e.g. Figure 10-11). For these analyses, we find it important to use globally consistent input data for DynQual runs made at the global scale (as per this paper), in order to facilitate meaningful comparisons across different world regions. Yet, this necessitates a simplified approach (see #8).

With these considerations, we choose to evaluate the global output from DynQual using metrics that focus on the residuals (i.e. prediction errors) using the normalized root mean squared error (nRMSE) (Figure 3, Figures S2a-5a); and by evaluating the ability of DynQual to simulate concentrations within a concentration range (Figure S6). This follows the evaluation approach adopted by global water quality models that are comparable to ours (e.g. Beusen et al., 2015; UNEP, 2016; van Vliet et al., 2021). Thanks for the good suggestion to add spatial patterns comparing the mean observed vs. modelled concentrations, these will supplement the existing evaluation approach nicely and so we will add these to the SI.

3) what is a good nRMSE value? It would be beneficiary to add the Nash–Sutcliffe model efficiency coefficient (NSE) which is a widely used dimensionless metric in hydrology and water quality literature.

Model evaluation spatial patterns presented in van Vliet et al. (2021) suggest acceptable performance below 100% (1.00) for total dissolved solids and biological oxygen demand. For large rivers (e.g. Rhine, Mississippi), Beusen et al., (2015) consider normalised RMSE (nRMSE) values of 50% (0.5) acceptable for nitrogen and phosphorus, in the view of the global scale of the model. Their model evaluation results including all rivers found a nRMSE of 124% (1.24) for N and 184% (1.84) for P, based on average annual concentrations, which they consider acceptable.

As also expressed in #4, the water quality parameter of interest is especially important to also consider here. For example, Reder et al., (2015) argue that as NSE and RMSE are sensitive to high extreme values (Moriasi et al., 2007), they need to be applied with caution in bacteria modelling where in-stream concentrations can vary across several order of magnitudes. The applicability of these metrics for model evaluation are also considered in the initial validation of DynQual in Jones et al., (2022), where it is shown for fecal coliform (FC) concentrations in some example stations (Figure 1) that while the model is generally capable of simulating concentrations within the correct concentration ranges, the magnitude of variability in the measured timeseries (also occurring over short time periods) is severely underestimated. High variability in observed FC concentrations is common across almost all monitoring stations, with 88% of stations reporting FC concentrations that range over three or more orders of magnitude (Jones et al., 2022).

We will further reflect on these points in the manuscript, supplemented with more comparisons to other global water quality models (as per #4).

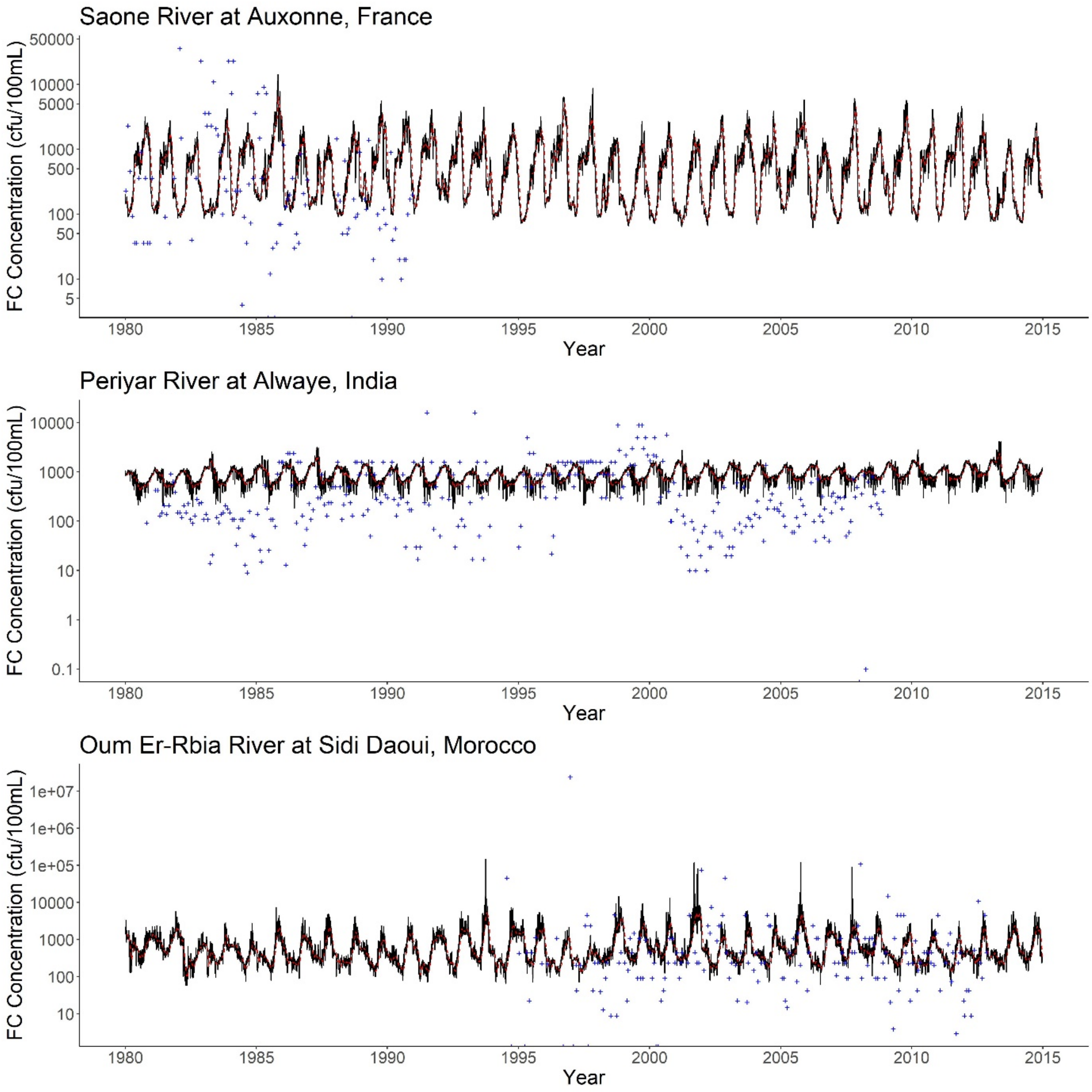

Figure 1. Selected time series of observed vs. simulated fecal coliform (FC) concentrations (cfu/ 100mL). Black (daily) and red (rolling 30-day average) lines indicated simulated FC concentrations; whereas blue crosses are observed concentrations. As displayed in Jones et al., (2022).

4) this manuscript in general lack literature discussion or comparison in terms of model performance (e.g., Figure 3), for example, what is other water quality model performance in terms of nRMSE? There might be few global scale water quality models. But I guess it could be useful to add a few comparisons with other watershed-scale water quality models.

Our comparisons to other global water quality models regarding performance are currently done in the discussion section (lines 508 – 526). We agree that more comparisons to other global water quality models regarding performance can be added – both of terms of the comparisons made to the other global water quality models we already discuss (e.g. van Vliet et al., 2021; UNEP 2016; Wen et al., 2017), and also to other global water quality models (e.g. Beusen et al., 2015). This will be added to the revised manuscript. When comparing global water quality models (including model performance), inherent differences between model set-ups must also be acknowledged. With regards to this, areas that are especially important to consider related to the model are the: 1) spatial extent (e.g. lumped vs. distributed models); 2) temporal resolution (e.g. daily vs. monthly vs. annual vs. decadal); and 3) water quality constituent of interest (also see #3). We will also reflect on these points in detail in the discussion section.

However, we find difficulties in comparing our global water quality model output directly to watershed-scale water quality models, which are typically applied to investigate vastly different research questions and have different purposes. These models can also incorporate locally relevant input data which are lacking in global approaches, can be parameterized for local conditions and typically have observation data of good quality and record length for calibration and validation. However, the flexible model does allow for applying DynQual at different spatial scales, and with different configurations. For example, pollutant loadings can be calculated independently of the DynQual emissions module and applied as a forcing – which could be particularly beneficial where knowledge of pollutant emissions exceeds the globally-applicable datasets. (Calibrated) hydrological output can also be forced directly to DynQual, thus making use only of the routing module of PCR-GLOBWB2/DynQual. Future work will seek to investigate the potential of DynQual to answer research questions at finer spatial scales (e.g. watershed, country-level). Here, comparisons to existing watershed-scale models will be important and relevant.

5) Line 200, can the decay coefficient be specified by the user?

Yes, as with all parameters/ coefficients in the model setup, decay coefficients could be specified by the user directly in the source code. We will clarify this in the manuscript. In the manuscript, we present the way this has currently been implemented within DynQual (based off existing global water quality modelling work). These are displayed in Equations 4 (BOD) and 5 (FC), with parameter values for FC in Table 1.

6) Line 220, is it a constant background concentration or a time-varying background concentration through each timestep?

We use a constant background concentration for TDS. While time-varying background concentrations could technically be implemented into DynQual easily, the availability of data (especially at the global scale) to determine time-varying background concentrations is a limitation. For other time-varying sources of pollutant loadings to the surface water network, for example those related to highly seasonal irrigation regimes, these are implemented at the daily timestep. We will clarify this in the manuscript.

7) what was the computational time to run for 1-year simulation?

In short, this will largely depend on: 1) model configuration (Figure 1); and 2) the target spatial extent.

Global 5 arc-min DynQual runs that are coupled with PCR-GLOBWB2, as done in this work, have a wall-clock time of approximately 6 hours when run with parallelization, due to the requirement to run with the kinematic wave routing option. This is a more computational demand routing equation than e.g. travel-travel time characteristic routing option (Sutanudjaja et al., 2018), but provides greater realism which is needed for higher accuracy discharge and water temperature simulations. Our calculation times are more or less equivalent to the PCR-GLOBWB2 run times given in Sutanudjaja et al., (2018) using kinematic wave routing. So called “offline” global DynQual runs, which use hydrological input (i.e. baseflow, interflow and direct runoff) as an external forcing, are somewhat (~20%) quicker.

The parallelization strategy for global PCR-GLOBWB2/DynQual involves dividing the model domain into 53 groups of river basins that run independently of each other as 53 separate processes (see Sutanudjaja et al., 2018 for more details). The length of time it takes for each individual process to run is dependent on the size of the land mask (i.e. the number of pixels). Thus, for global parallelization runs, the computational time to run for a 1-year simulation is equal to that of the largest land mask. Yet, PCR-GLOBWB2/DynQual does not necessarily need to be run at the global extent – users could alternatively select a particular river basin(s) or define their own land mask. The self-contained example for the Rhine basin that we provide (https://zenodo.org/record/7027242#.Y8fHOhfMJPY) takes approximately 45 minutes to run for 1 year.

8) Supplement Line 295, does it mean reaction is underestimated compared to discharge (dilution)?

The current model set-up for DynQual emissions module is to quantify loadings at the individual gridcells where the pollution activity is located (e.g. populations), with loadings entering the surface water network at the same location (which are subsequently routed through the surface water stream network based on the flow direction (routing) map; Figure 2). This approach is advantageous in that we can use globally available datasets for estimating pollutant emissions (e.g. populations). The ability to use globally consistent input data is important for DynQual runs made at the global scale (in line with the focal point of this paper), whereby we want to facilitate meaningful comparisons across different world regions.

This approach has the disadvantage of spatial mismatches between the generation of pollutant loadings and the actual locations where loadings enter the stream network occur. Simulated concentrations in gridcells with low water availability – i.e. headwater streams - are particularly sensitive to this, and concentrations in these gridcells are typically masked in global water quality studies. Similarly, point sources of pollution (e.g. wastewater treatment plants) typically discharge into higher-order streams, the location of which might not exactly coincide with population density. DynQual tracks the mass of pollutants (not concentrations) in discrete gridcells per timestep (Figure 1). Thus, these masses are transported downstream to higher-order streams whereby the pollutant mass will be more feasible with respect to the dilution capacity – hence the better model performance in gridcells with greater water availability.

9) Supplement Line 300, what is high data availability, and how many data points during 1980-2019?

“High data availability” is, of course, relative. In this paper, and as per Jones et al., (2022), high data availability refers to >30 measurements in the target time period. We will add this detail to the SI. In general, water quality observations are limited in both geographical space and fragmented across time . There are also issues related to data access, with data collected by both governmental organizations and private institutions not being made publicly accessible. These are key motivations for conducting large-scale water quality modelling efforts, such as DynQual, i.e. to provide insights on water quality status in data scare regions such as Africa and large parts of Asia.

10) Supplement Line 305, Figure S3 (b, c) what are the nRMSE and NSE values for these two rivers? It seems that the model overestimated a lot for peaks.

For Figure S3b, the nRMSE is 0.15, with a mean observed concentration of TDS 399 mg l^{-1} (349 observations) and a mean modelled concentration of TDS 388 mg l^{-1}. For Figure S3c, the nRMSE is 0.24, with a mean observed concentration of TDS 25 mg l^{-1} (468 observations) and a mean modelled concentration of TDS 24 mg l^{-1}. We wanted to explicitly include some time-series plots in our evaluation to demonstrate the dynamic nature of the model simulations, and selected stations based on the high data availability and the long-time series of observations.

However, comparing individual (instantaneous) observed concentrations vs. simulated concentrations comes with challenges at large scales – particularly when it comes to min and max concentrations. As demonstrated in ~2007 in the Drammenselva river, observed TDS concentrations fluctuate by a factor of 2 over short time periods (~25 mg l^{-1} to 50 mg l^{-1}), with the peak being entirely missed in our simulations. This is difficult to attribute (and is beyond the scope of DynQual), given the global scale of our model and the underlying assumptions – perhaps it is a missing emission process (e.g. delivery of road-salts to the stream network) or a temporal mismatch in the simulated vs. actual hydrological conditions (i.e. dilution effect). Yet, max TDS concentrations in the observed and simulated time-series for the Drammenselva river are very similar (~50 mg l^{-1}). Max simulated concentrations (~600 mg l^{-1}) are indeed somewhat higher than observed concentrations (~500 mg l^{-1}) in the Kalamazoo river. Again – difficult to attribute – this could similarly due to a bias in the dilution component of our model or due to an overestimation of TDS loadings originating from short-lived processes (e.g. those from urban surface runoff). As described in #2, the overarching purpose of the model must be considered, and please note that we are not using DynQual to draw conclusions regarding extremes (min-max concentrations).

For reasons described in #2&3, we refrain from calculating NSE values, which have not been used for evaluating the performance of any global water quality model to date.

References

Beusen, A. H. W., Van Beek, L. P. H., Bouwman, A. F., Mogollón, J. M., & Middelburg, J. J. (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. Geoscientific model development, 8(12), 4045-4067.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... & Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological sciences journal, 58(6), 1198-1255.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.

Reder, K., Flörke, M., & Alcamo, J. (2015). Modeling historical fecal coliform loadings to large European rivers and resulting in-stream concentrations. Environmental Modelling & Software, 63, 251-263.

Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., ... & Bierkens, M. F. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. Geoscientific Model Development, 11(6), 2429-2453.

UNEP: A Snapshot of the World's Water Quality: Towards a global assessment, United Nations Environment Programme, Nairobi, Kenya, 162pp, 2016.

van Vliet, M. T., Jones, E. R., Flörke, M., Franssen, W. H., Hanasaki, N., Wada, Y., & Yearsley, J. R. (2021). Global water scarcity including surface water quality and expansions of clean water technologies. Environmental Research Letters, 16(2), 024020.

Voß, A., Alcamo, J., Bärlund, I., Voß, F., Kynast, E., Williams, R., & Malve, O. Continental scale modelling of in-stream river water quality: a report on methodology, test runs, and scenario application. Hydrological Processes, 26(16), 2370-2384.

Wanders, N., van Vliet, M. T., Wada, Y., Bierkens, M. F., & van Beek, L. P. (2019). High-resolution global water temperature modeling. Water Resources Research, 55(4), 2760-2778.

Wen, Y., Schoups, G., & Van De Giesen, N. (2017). Organic pollution of rivers: Combined threats of urbanization, livestock farming and global climate change. Scientific reports, 7(1), 1-9.