Model inter-comparison design for large-scale water quality models

Michelle T. H. van Vliet[1], Martina Flörke[2], John A.Harrison[3], Nynke Hofstra[1], Virginie Keller[4], Fulco Ludwig[1], J. Emiel Spanier[1], Maryna Strokal[1], Yoshihide Wada[5], Yingrong Wen[6], Richard Williams[4]

[1] *Water Systems and Global Change group, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands*

[2] *Center for Environmental Systems Research, University of Kassel, Wilhelmshöher Allee 47, 34109 Kassel, Germany*

[3] *School of the Environment, Washington State University, Vancouver Campus, Vancouver, WA 98686, USA*

[4] *NERC Centre for Ecology and Hydrology, Wallingford, Oxfordshire, OX10 9AU, United Kingdom*

[5] *International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria*

[6] *Department of Water Management, Delft University of Technology, Stevinweg 1, 2628CN, Delft, The Netherlands*

Abstract

Several model inter-comparison projects (MIPs) have been carried out recently by the climate, hydrological, agricultural and other modelling communities to quantify modelling uncertainties and improve modelling systems. Here we focus on MIP design for large-scale water quality models. Water quality MIPs can be useful to improve our understanding of pollution problems and facilitate the development of harmonized ~~data sets~~estimates of current and future water quality. This can provide new opportunities for assessing~~-~~ robustness in estimates of water quality hotspots and trends, improve understanding of processes, pollution sources, water quality model uncertainties, and to identify priorities for water quality data collection and monitoring. Water quality MIP design should harmonize relevant model input datasets, use consistent spatial/temporal domains and resolutions, and similar output variables to improve understanding of water quality modelling uncertainties and provide harmonized water quality data that suit the needs of decision makers and other users.

Highlights

- Model inter-comparison projects (MIPs) can identify robustness of ~~-~~water quality hotspots and trends
- Water quality MIPs can improve understanding of pollution causes and model uncertainties
- MIP design should focus on using consistent input datasets and harmonize output variables~~,~~ and spatial~~ and ~~/temporal resolutions,
- MIPs of lumped models should focus on pollutant loadings at river basin outlets
- MIPs of grid-based models can compare spatial water quality heterogeneity within basins.

1. Introduction

In the last decade, there has been a strong focus on global and regional model inter-comparison projects (MIPs), which ~~in various research fields, including climate, hydrology (water quantity) and agriculture (crop) modelling.~~have been used to contribute to a comprehensive and consistent picture of model-derived insights in several fields, including climate, hydrology (water quantity) and agriculture (crop) modelling. The concept of MIP offers a framework to consistently evaluate and compare models, and associated model input, structural, and parameter uncertainty under different objectives (e.g. climate variability and change, model performance, human impacts and developments). Some of the most representative global MIPs include the Coupled Model Inter-comparison Project (CMIP) [1], the Agricultural Model Inter-comparison Project (AgMIP) [2], WATCH Water Model Inter-comparison Project (WaterMIP) [3,4] and the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP) [5]. These MIPs were mainly designed to better understand past, present and future climate changes and associated impacts on respective sectors (e.g. hydrology, agriculture, biomes, energy). One of the important goals of MIPs is to make the multi-model output publically available in a standardized format (e.g. netCDF).

While there has been a significant amount of research and publications on MIPs and multi-model assessments for water availability, limited multi-model assessments for large-scale water quality studies exist [6,7]. Water quality problems exist in many parts of the world [8,9] and these issues may intensify due to climate change and socio-economic developments [10]. Robust estimates of current and future changes in water quality are needed to achieve sustainable management of clean accessible water for all, as required by the Sustainable Development Goal for clean water and sanitation (SDG 6) for 2030.

A large-scale water quality model is defined here as a model capable of simulating one or more water quality variables (pollutants) on a scale that exceeds the size of a single river basin~~.~~, which we define as the upstream land surface area contributing to the streamflow at the basin outlet (river mouth). Some examples of large-scale nutrient models are Global *NEWS*-2 [11,12], SPARROW [13], IMAGE-GNM [14,15], HYPE [16] and MARINA [17]. In addition, large-scale water quality models including nutrients, salinity (e.g. total dissolved solids (TDS)) and organic pollution (biochemical oxygen demand (BOD)) have been developed, such as WaterGAP-WorldQual [18,19] and GWAVA-WQ [20,21].

Development of large-scale nutrient models started in the 1990s, and since 2010 there has been a strong growth in the number of large-scale models for other pollutants too (Figure 1). For instance, global models have been recently developed for river water temperature [22-24], river water organic pollution [25], micro-organisms [26-28], chemicals [29], plastics [30-32], nanomaterials [33] and pesticides (insecticides) [34]. Most of the large-scale water quality models are spatially-explicit (commonly grid-based) and dynamic (i.e. account for temporal variability). The recent strong growth in the number of large-scale water quality models increases opportunities for comparing results from various models per water quality variable.

[Fig 1]

In this paper, we review work published on model inter-comparison of large-scale water quality models, discuss reasons to move forward on water quality MIPs and give suggestions for future directions on water quality MIP design. We first discuss the lessons learnt from previous MIPs in other sectors (climate, water) (Section 2.1) and from previous large-scale water quality model inter-comparison studies (Section 2.2). We then consider opportunities (Section 3.1), challenges and recommendations (Section 3.2) for design of water quality MIPs. We conclude by summarizing our main findings and examining how water quality MIPs could be designed to provide consistent, harmonized water quality model output datasets, which are more useful for policy makers and other users (Section 4).

2. Previous large-scale model inter-comparison studies

2.1 Lessons learnt from MIPs in other sectors

In ISIMIP, modelling protocols have been developed with an international network of climate-impact modellers to contribute to a comprehensive and consistent picture of the world's impacts under different climate-change scenarios across affected sectors (e.g. water, agriculture, energy, forestry, marine ecosystems) and spatial scales [35,36]. Overall, the focus of MIPs and associated concepts and modelling protocols is currently on understanding how model predictions vary across different sectors and different climate change scenarios. Within CMIP, the aim is to discover why different climate and earth system models provide different outputs despite receiving similar model input and identifying aspects of the simulations in which "consensus" in climate model projections or common problematic features exist [37]. To better understand the model spread and to reduce the associated uncertainties, a comparison of model performance and the sensitivity of the models to different warming rates may need to be studied further [38,39]. The consistent modelling framework of ISIMIP and CMIP using common input datasets and output variables has generated important datasets used by a broad research community and policy makers.

2.2 Previous water quality MIPs

Compared to other sectors (climate, water availability, agriculture) fewer MIP studies or multi-model assessments exist for water quality. Previous MIP studies for large-scale water quality have mainly focussed on nutrients. Comparisons of model results between different nitrogen (N) export models have been made, amongst others, at global scale [7], for Chinese basins [40], for the United States [6] and for selected sub-basins [e.g. 41,42]. These analyses have overall found fairly consistent loading predictions between similarly scaled models, despite varying levels of model complexity and differences in input data sources. The focus of most previous nutrient MIPs has been on comparing nutrient loads (e.g. kg N y^{-1}) with less attention on source apportionment. An exception is McCrackin et al. [6], where comparing results of SPARROW and Global *NEWS*-2 for the United States showed that for several regions similar N sources were identified by both models.

A model inter-comparison has also been published for global river water temperatures [43] using global grid-based (0.5°) simulations of the water temperature modules of the global hydrological models of PCR-GLOBWB [23], VIC-RBM [24,44], and WaterGAP-WorldQual [22]. All three models were run using consistent model input for climate forcing, land mask, basin delineation and river flow direction (routing network). The three river water temperature modules show similar spatial patterns of water temperature [43] and identified similar regions with highest water temperature increase under climate change. However, the magnitude of water temperature changes varied, and this was mainly attributed to different representations of impacts of hydrological change and snowmelt inputs/ice cover processes [43].

These previous nutrient/water temperature model inter-comparison studies have shown the importance of evaluating the performance of water quality models and highlighted the need of common input data to provide consistent water quality model output for comparison [41,43,45].

3. Opportunities, challenges and recommendations for design of water quality MIPs

3.1. Opportunities to move forward on water quality MIPs

Comparing water quality model results can lend credibility to water quality simulations and identify areas for future model improvement [6]. Water quality MIPs could facilitate the development of harmonised model output data sets of the current water quality status and future scenarios based on the water quality model ensemble. Overall, harmonized water quality model output datasets based on multiple models are more robust than results of a single water quality model, providing several new opportunities that are briefly discussed below.

1. *Identify robust water quality (pollution) hotspots*

Water quality MIPs can provide ~~more~~ better understanding of the robustness of ~~identification~~ identified ~~of~~ water pollution hotspots under present-day and under future climate and socio-economic conditions than are currently available. Limited knowledge in particular exist on how pollution hotspots will develop over the next decades. Using results from multiple water quality models will provide a more comprehensive picture and assessment of the robustness of identified pollution hotspots under certain future scenarios than results of a single water quality model. This information is needed by decision makers and water managers to assess what adaptive solutions should be implemented in specific regions to improve the quality of water resources for human water uses and ecosystem health.

2. *Assess robust trends in water quality*

Water quality model inter-comparison can be used to ~~identify~~ assess robustness of simulated trends in water quality. Various water quality models might show different responses and sensitivities to changes in climate, land use, and socio-economic development. Ensemble simulations of water quality models might therefore be more useful than stand-alone models by providing a more comprehensive projection and ~~increasing~~ understanding of ~~and anticipating~~possible future pollution changes.

3. Improve understanding of processes and sources of water pollution

Water quality MIPs can contribute to improved understanding of water quality processes and contribution of different pollution sources. Source apportionment across wide geographical domains can only be achieved through the use of large-scale water quality models, due to a lack of measurements at such scales [6,46]. Comparison of multi-water quality model outputs can provide a more comprehensive assessment ~~would allow more robust estimates~~ of sources and dominant pollution processes~~, ~~. MIPs can identify agreement on identified pollution sources apportioned by different water quality models, which is ~~which are~~ needed to inform and develop effective water quality solutions in certain regions.

4. Increase understanding of water quality model uncertainties

Ideally, observed water quality monitoring records are used to validate water quality model estimates and assess model uncertainties for regions worldwide. However, ~~In~~ in comparison to river discharge and meteorological data, there is a significant lack of water quality measurements for many regions worldwide (e.g. Africa) [8] to evaluate water quality model performances and uncertainties [47]~~–~~. A consistent comparison of the results of different water quality models contributes to lending credibility to water quality estimates. In addition, sensitivity analyses, perturbing water quality models with different input will enhance understanding of water quality model differences and uncertainties related to the structure and parameterization of different water quality models.

5. Identify and set priorities for water quality data collection and monitoring

Across many scientific domains, including water quality, monitoring and modelling are complementary approaches. The results of multi-model assessments of water quality could contribute to setting priorities and identifying regions for water quality data collection and monitoring [48].

3.2 Challenges and recommendations for water quality MIP design

A major challenge for water quality MIPs, so far, has been the limited number of large-scale water quality models per water quality variable (pollutant) available to compare and provide ensembles of water quality model results. However, several new large-scale water quality models have been developed over recent years (see Section 1; Supplementary Information Table S1) [47], providing new opportunities for water quality MIPs. Below we discuss the main challenges of designing a water quality model inter-comparison and propose recommendations to ensure useful harmonized water quality data are produced to suit the needs of decision makers and other users.

Challenge 1: Water quality models differ in spatial and temporal resolutions and domains

Water quality models differ both in terms of spatial and temporal domains (e.g. use of different basin delineations and model simulation periods), as well as temporal and spatial resolutions. Some models simulate daily or monthly water quality estimates whereas others simulate annual average values. Thus, when comparing models using different temporal resolutions, methods must be adopted to aggregate fine temporal scale estimates to compare with coarse-scale water quality estimates (e.g. select average year or use multiple years). In addition to temporal aspects, spatial resolution can also differ between models. Some water quality models are grid-based and spatially resolved at fine scales (e.g. WaterGAP-WorldQual). These are suitable to capture spatial heterogeneity of water quality, -while others are lumped at basins or sub-basins and are designed to compute basin-wide pollutant loadings or pollutant loadings of rivers to coastal zones (e.g. Global *NEWS*-2~~, SPARROW~~). Overall, the scale for comparison is generally limited to lowest temporal and spatial resolution and domain. MIPs including lumped water quality models (or a combination of lumped and grid-based model water quality models) should therefore focus on basin aggregated level, comparing loadings/concentrations at basin outlets (river mouths). MIPs that solely include spatially-explicit (grid-based) water quality models are more suitable to compare spatial heterogeneity of water quality and relate to acceptable water quality levels for different uses (e.g. domestic, irrigation, industrial) and ecosystem health within a basin.

We present ~~An~~ an illustrative example ~~is presented~~ for comparison of spatially-explicit organic pollution, focussing on simulated mean BOD concentrations derived from four large-scale grid-based water quality models, namely WaterGAP-WorldQual, GWAVA-WQ, VIC-QUAL and the global BOD model of Wen et al. [25] (Figure 2). We extracted ~~Simulated~~ simulated mean BOD concentrations from the model of Wen et al. [25] and global simulation of VIC-QUAL [49] at 0.5°x0.5° ~~were extracted~~ for Europe. ~~These~~ We compared the mean BOD data ~~were compared~~ with high-resolution simulations (5'x5') of GWAVA-WQ [20,21] and WaterGAP-WorldQual [18] for Europe, which were aggregated to 0.5°x0.5° using nearest neighbour resampling and averaged over the period 1990-2000 (Figure 2). ~~Overall,~~ These results show that organic pollution hotspots are roughly comparable but some differences exist due to differences in model structure, input datasets (e.g. hydrology) and pollution sources considered. For instance, lower BOD concentrations simulated by the model of Wen et al. [25], can be explained by the fact that this model focusses solely on BOD loadings from urban population and livestock, while the other models also consider organic pollution from manufacturing.

The importance of using similar temporal/spatial resolutions strongly depends on the purpose of the water quality model inter-comparison. For instance, full consistencies in temporal/spatial resolution amongst water quality models might be essential when aiming at understanding the water quality processes or quantifying model uncertainties, but possibly less so when the purpose of the inter-comparison is the identification (locations and intensity) of water quality hotspots (Table 1). Nevertheless, the use of similar spatial and temporal domains, and preferably also resolutions, of water quality models are overall recommended in water quality MIP design to provide consistent water quality model output.

- ➔ *Recommendation 1: Use similar spatial and temporal domains and, preferably, also resolutions of water quality models in MIP design. However, not all models can be compared for the same purpose. For instance, MIPs of lumped water quality models should focus on pollutant loadings at river basin outlets, while MIPs solely including grid-based models can compare spatial water quality heterogeneity within basins.*

[Fig 2]

Challenge 2: Water quality models differ in reported output variables

Water quality models show a high diversity in output variables, which complicates a direct comparison of model estimates. For instance, Some some water quality models focus on in-stream concentrations (e.g. in mg/l) while other models simulate loads (e.g. in kg/yr) or area specific yields (e.g. in kg/km^2 of basin/yr). In particular, nutrient models provide outputs for different nutrient forms. Several models focus on total nitrogen (TN) and total phosphorous (TP) (e.g. IMAGE-GNM, WaterGAP-WorldQual), whereas others (e.g. Global *NEWS*-2) simulate different forms of nitrogen, phosphorus, carbon and silica. We present An an illustrative example of comparison of river export of TN in loads (10^6 kg/yr) and yields (kg/km^2/yr) for Global *NEWS*-2 [11] and IMAGE-GNM [14] models for a single year, 2000, is presented (Figure 3). The Global *NEWS*-2 model simulates different forms of nitrogen, i.e. dissolved inorganic nitrogen (DIN), dissolved organic nitrogen (DON) and particulate nitrogen (PN). The individual loads for each form were summed in order to provide TN estimates, which were then compared to estimates of TN loads generated with IMAGE-GNM. We compared The the TN river export from the grid-based IMAGE-GNM (0.5°) at basin outlet gridcells was compared with TN river export from similar basin outlets of Global *NEWS*-2. Comparison of simulated TN loads (Figure 3a) and yields (Figure 3b) from both global nutrient models shows rather similar basins with high or low TN river export. Worldwide, lower values of TN river export were found for IMAGE-GNM (37 Tg N/yr) compared to Global *NEWS*-2 (45 Tg N/yr). This might be related to differences in model structure, process descriptions and input data. For instance, the approaches to simulate N retentions in the terrestrial and aquatic systems differ greatly between both models, as do the use of hydrological input data and basin delineations. The differences can also be explained by the different purposes of the models: e.g. Global *NEWS*-2 for scenario analyses and IMAGE-GNM for improved, spatial-explicit understanding of the processes controlling nutrient export. Overall,

~~it is~~we highly recommend~~ed~~ ~~to group~~grouping of water quality models per pollutant form and focus on similar output variables (e.g. total nitrogen concentrations, loads or yields) and units (e.g. mg/l, kg/km^2/yr)~~, in order to~~. This is needed to provide harmonized ensemble model outputs of water quality that can be used to identify in which regions models agree on simulated water quality changes, ~~that are useful for~~needed for water quality management and decision making, and to assess areas for model improvements. In line with model intercomparison projects within the climate community (e.g. CMIP6), a minimum ensemble size of three models is desired to assess the robustness of identified trends [50].

➔ *Recommendation 2: Use similar model output variables per pollutant form ~~for comparison of~~ to provide insights in the robustness ~~large-scale water quality models.~~of simulated pollution hotspots, trends and sources by large-scale water quality models.*

[Fig 3]

Challenge 3: Water quality models use different input datasets

Various water quality models use different climate forcing datasets, hydrological (discharge, runoff) input, reservoir, land use and waste-water treatment data and assumptions. This complicates direct comparison and understanding of differences in simulated water quality results between models. Therefore the use of similar model input datasets in water quality MIP design is strongly recommended to provide consistent water quality model results that are meaningful for water pollution management, decision-making and other possible uses. In global hydrological and land surface modelling, the development of the WATCH Meteorological Forcing Data [51], was a major accomplishment facilitating inter-comparison projects such as WaterMIP and ISIMIP. In a similar way, producing different input datasets for water quality can be an important step to provide harmonized water quality results. The level of harmonization on input data might differ, as certain water quality variables might have different driving forces and sensitivities to various input datasets. For example, river water temperature MIPs would prioritize the use of similar climate forcing data and hydrological datasets (reservoirs) into various water temperature models, while inter-comparison of organic pollution and nutrients models would ideally require harmonization also on land use and waste-water treatment input datasets. Furthermore, the main purpose for water quality model inter-comparison is important to consider. For instance, harmonization on all model input is preferred, but not absolutely trivial for the identification of present-day pollution hotspots. In contrast, strict harmonization on all model input ~~would be~~is essential when the focus of the MIP is on improved understanding of water quality processes and model uncertainties (Table 1).

➔ *Recommendation 3: Harmonize relevant input datasets to provide consistent output for water quality model inter-comparison.*

[Table 1]

4. Discussion, conclusions and future outlook

Large-scale MIPs such as CMIP, AgMIP and ISIMIP have contributed to a better understanding of important components of the Earth system and climate change impacts on various sectors, as well as the associated model uncertainties~~.~~ by bringing these modelling communities and together and consistently comparing model output. Given the recent proliferation of water quality models (Figure 1) and the fact that many people around the world are affected by water quality deterioration [8,9], pollution-driven water scarcity [52,53], and water security threats [54], there is now both an opportunity and a clear need to implement regional and global water quality MIPs.

Water quality MIPs can provide consistent, harmonized ensemble water quality model outputs, which is important for water policy and decision making [55]. Water quality MIPs can also contribute to improved understanding of pollution processes and pollution sources [6]. This is particularly important in world regions where observed water quality data are sparse (e.g. Africa, parts of southern America, Asia) [8]. In addition, water quality MIPs can be used to assess water quality trends and pollution hotspots, both for present-day and future scenarios. Such information is needed to assess potential strategies to provide clean water, both for human uses and ecosystems, and, to reduce pollution-driven water scarcity [52,53].

To further improve large-scale water quality modelling we believe a more coordinated effort for inter-comparisons is recommended. This paper has discussed some of the main challenges and recommendations for water quality MIPs. Harmonising model output by using similar spatial/temporal resolution and domains (recommendation 1) and by using similar water quality output variables (concentration, loadings) (recommendation 2) is of major importance to provide consistent results. In addition, previous water quality MIPs have shown the importance of evaluating the performance of water quality models [41,45]. An important next step is to further harmonize on model input data (recommendation 3) and perform sensitivity analyses to improve understanding of uncertainties related to differences in water quality model structure. The extent of harmonization between input datasets will depend on the aim and ambition of the MIP. We think t~~T~~here is a clear need for MIPs comparing model output for a single quality variable. However, MIPs comparing model output for multiple water quality variables may also be useful to identify hotspots for water pollution for selected pollutants with similar sources [47,56].

Several MIPs of climate models and integrated assessment models have not only been informative for the scientific community, they have also influenced policy, especially in relation to climate change [57,58]. We think a~~A~~ standardized set-up and input dataset on water quality observation and model outputs for both current conditions and for future scenarios will be helpful to address future water quality and scarcity problems, and identify where water quality improvement are needed. This could facilitate the development of harmonized water quality assessments that can contribute to sustainable management and solution(s) identification supporting the achievement of clean water for all (SDG6) in coming decades.

Acknowledgements

This paper evolved from a workshop titled “Water Quality: a new challenge for global scale modelling” held at Wageningen University 18-21 September 2017. Funding for this workshop came from the OECD-CRP, the support of which is gratefully acknowledged. Michelle van Vliet was financially supported by a Veni-grant (project no. 863.14.008) of NWO Earth and Life Sciences (ALW). We would like to think Xi Chen and Wietse Franssen (Wageningen University) for assisting with the IMAGE-GNM and technical support with VIC-QUAL. Dr. Arthur Beusen (PBL) is kindly acknowledged for checking the results for IMAGE-GNM and Global *NEWS*-2 comparison.

References

1. Covey C, AchutaRao KM, Cubasch U, Jones P, Lambert SJ, Mann ME, Phillips TJ, Taylor KE: **An overview of results from the Coupled Model Intercomparison Project**. *Global and Planetary Change* 2003, **37**:103-133.
2. Rosenzweig C, Jones JW, Hatfield JL, Ruane AC, Boote KJ, Thorburn P, Antle JM, Nelson GC, Porter C, Janssen S, et al.: **The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies**. *Agricultural and Forest Meteorology* 2013, **170**:166-182.
3. Haddeland I, D.B. Clark, W. Franssen, F. Ludwig, F. Voß, N.W. Arnell, N. Bertrand, M. Best, S. Folwell, D. Gerten, S. Gomes, S.N. Gosling, S. Hagemann, N. Hanasaki, R. Harding, J. Heinke, P. Kabat, S. Koirala, T. Oki, J. Polcher, T. Stacke, P. Viterbo, G.P. Weedon, and P. Yeh: **Multimodel Estimate of the Terrestrial Global Water Balance: Setup and First Results**. *J. Hydrometeor.* 2011, **12**:869–884.
4. Harding R, Best M, Blyth E, Hagemann S, Kabat P, Tallaksen LM, Warnaars T, Wiberg D, Weedon GP, Lanen Hv, et al.: **WATCH: Current Knowledge of the Terrestrial Global Water Cycle**. *Journal of Hydrometeorology* 2011, **12**:1149-1156.
5. Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O, Schewe J: **The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework**. *Proceedings of the National Academy of Sciences* 2014, **111**:3228-3232.

** 6. McCrackin ML, Harrison JA, Compton JE: **A comparison of NEWS and SPARROW models to understand sources of nitrogen delivered to US coastal areas**. *Biogeochemistry* 2013, **114**:281-297.

7. Boyer EW, Howarth RW, Galloway JN, Dentener FJ, Green PA, Vorosmarty CJ: **Riverine nitrogen export from the continents to the coasts**. *Global Biogeochemical Cycles* 2006, **20**.

** 8. UNEP: **A Snapshot of the World's Water Quality – Towards a global assessment.** . Edited by. Nairobi, Kenya: United Nations Environment Programme; 2016:162.

9. WWAP: **The United Nations World Water Development Report 2017. Wastewater: The Untapped Resource**. Edited by Programme UNWWA. Paris: UNESCO; 2017:180.
10. Xie H, Ringler C: **Agricultural nutrient loadings to the freshwater environment: the role of climate change and socioeconomic change**. *Environmental Research Letters* 2017, **12**:104008.
11. Mayorga E, Seitzinger SP, Harrison JA, Dumont E, Beusen AHW, Bouwman AF, Fekete BM, Kroeze C, Van Drecht G: **Global Nutrient Export from WaterSheds 2 (NEWS 2): Model development and implementation**. *Environmental Modelling & Software* 2010, **25**:837-853.
12. Seitzinger SP, Mayorga E, Bouwman AF, Kroeze C, Beusen AHW, Billen G, Van Drecht G, Dumont E, Fekete BM, Garnier J, et al.: **Global river nutrient export: A scenario analysis of past and future trends**. *Global Biogeochemical Cycles* 2010, **24**.
13. Schwarz G, Hoos A, B Alexander R, Smith R: *The SPARROW Surface Water-Quality Model: Theory, Application and User Documentation*; 2006.
14. Beusen AHW, Bouwman AF, Van Beek LPH, Mogollón JM, Middelburg JJ: **Global riverine N and P transport to ocean increased during the twentieth century despite increased retention along the aquatic continuum**. *Biogeosciences Discuss.* 2015, **2015**:20123-20148.
15. Beusen AHW, Van Beek LPH, Bouwman AF, Mogollón JM, Middelburg JJ: **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.* 2015, **8**:4045-4067.
16. Strömqvist J, Arheimer B, Dahné J, Donnelly C, Lindström G: **Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale**. *Hydrological Sciences Journal* 2012, **57**:229-247.
17. Strokal M, Kroeze C, Wang M, Bai Z, Ma L: **The MARINA model (Model to Assess River Inputs of Nutrients to seAs): Model description and results for China**. *Science of The Total Environment* 2016, **562**:869-888.

18. 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* 2012, **26**:2370-2384.
19. Fink G, Alcamo J, Flörke M, Reder K: **Phosphorus Loadings to the World's Largest Lakes: Sources and Trends**. *Global Biogeochemical Cycles* 2018.
20. Williams R, Keller V, Voß A, Bärlund I, Malve O, Riihimäki J, Tattari S, Alcamo J: **Assessment of current water pollution loads in Europe: estimation of gridded loads for use in global water quality models**. *Hydrological Processes* 2012, **26**:2395-2410.
21. Dumont E, Williams R, Keller V, Voß A, Tattari S: **Modelling indicators of water security, water pollution and aquatic biodiversity in Europe**. *Hydrological Sciences Journal* 2012, **57**:1378-1403.
22. Punzet M, Voß F, Voß A, Kynast E, Bärlund I: **A Global Approach to Assess the Potential Impact of Climate Change on Stream Water Temperatures and Related In-Stream First-Order Decay Rates**. *Journal of Hydrometeorology* 2012, **13**:1052-1065.
23. van Beek LPH, Eikelboom T, van Vliet MTH, Bierkens MFP: **A physically based model of global freshwater surface temperature**. *Water Resources Research* 2012, **48**.
24. van Vliet MTH, Yearsley JR, Franssen WHP, Ludwig F, Haddeland I, Lettenmaier DP, Kabat P: **Coupled daily streamflow and water temperature modelling in large river basins**. *Hydrology and Earth System Sciences* 2012, **16**:4303-4321.
25. Wen Y, Schoups G, van de Giesen N: **Organic pollution of rivers: Combined threats of urbanization, livestock farming and global climate change**. *Scientific Reports* 2017, **7**:43289.
26. Hofstra N, Vermeulen LC: **Impacts of population growth, urbanisation and sanitation changes on global human Cryptosporidium emissions to surface water**. *International Journal of Hygiene and Environmental Health* 2016.
27. Vermeulen LC, van Henger M, Kroeze C, Medema G, Spanier JE, van Vliet MTH, Hofstra N: **Cryptosporidium concentrations in rivers worldwide**. submitted.
28. Reder K, Flörke M, Alcamo J: **Modeling historical fecal coliform loadings to large European rivers and resulting in-stream concentrations**. *Environmental Modelling & Software* 2015, **63**:251-263.
29. van Wijnen J, Ragas AMJ, Kroeze C: **River export of triclosan from land to sea: A global modelling approach**. *Science of The Total Environment* 2018, **621**:1280-1288.
30. Siegfried M, Gabbert S, Koelmans A, Kroeze C, Löhr A, Verburg C: **River Export of Plastic from Land to Sea: A Global Modeling Approach.** . *EGU General Assembly Conference Abstracts. 2016.* 2016, **vol 18.**:p 11507.
31. Jambeck JR, Geyer R, Wilcox C, Siegler TR, Perryman M, Andrady A, Narayan R, Law KL: **Plastic waste inputs from land into the ocean**. *Science* 2015, **347**:768-771.
32. Lebreton LCM, van der Zwet J, Damsteeg J-W, Slat B, Andrady A, Reisser J: **River plastic emissions to the world's oceans**. *Nature Communications* 2017, **8**:15611.
33. Dumont E, Johnson AC, Keller VDJ, Williams RJ: **Nano silver and nano zinc-oxide in surface waters – Exposure estimation for Europe at high spatial and temporal resolution**. *Environmental Pollution (Barking, Essex : 1987)* 2015, **196**:341-349.
34. Ippolito A, Kattwinkel M, Rasmussen JJ, Schäfer RB, Fornaroli R, Liess M: **Modeling global distribution of agricultural insecticides in surface waters**. *Environmental Pollution* 2015, **198**:54-60.
35. Frieler K, Lange S, Piontek F, Reyer CPO, Schewe J, Warszawski L, Zhao F, Chini L, Denvil S, Emanuel K, et al.: **Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b)**. *Geosci. Model Dev.* 2017, **10**:4321-4345.
* 36. Schewe J, Heinke J, Gerten D, Haddeland I, Arnell NW, Clark DB, Dankers R, Eisner S, Fekete BM, Colón-González FJ, et al.: **Multimodel assessment of water scarcity under climate change**. *Proceedings of the National Academy of Sciences* 2013.
37. Taylor KE, Stouffer RJ, Meehl GA: **An Overview of CMIP5 and the Experiment Design**. *Bulletin of the American Meteorological Society* 2012, **93**:485-498.

38. Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE: **Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models**. *Water Resources Research* 2008, **44**.
39. Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA: **Challenges in Combining Projections from Multiple Climate Models**. *Journal of Climate* 2010, **23**:2739-2758.
40. Strokal M: *River export of nutrients to the coastal waters of China: The MARINA mdoel to assess sources, effects and solutions*; 2016.
41. Alexander RB, Johnes PJ, Boyer EW, Smith RA: **A comparison of models for estimating the riverine export of nitrogen from large watersheds**. *Biogeochemistry* 2002, **57**:295-339.
42. Valiela I, Bowen JL, Kroeger KD: **Assessment of models for estimation of land-derived nitrogen loads to shallow estuaries**. *Applied Geochemistry* 2002, **17**:935-953.
43. van Vliet MTH, van Beek LPH, Eisner S, Flörke M, Wada Y, Bierkens MFP: **Multi-model assessment of global hydropower and cooling water discharge potential under climate change**. *Global Environmental Change* 2016, **40**:156-170.
44. Yearsley JR: **A grid-based approach for simulating stream temperature**. *Water Resources Research* 2012, **48**.
45. Bowen JL, Ramstack JM, Mazzilli S, Valiela I: **Nload: an interactive, web-based modelling tool for nitrogen management in esturaties**. *Ecological Applications* 2007, **17**:S17-S30.
46. Kroeze C, Bouwman L, Seitzinger S: **Modeling global nutrient export from watersheds**. *Current Opinion in Environmental Sustainability* 2012, **4**:195-202.
** 47. Strokal M, Spanier JE, Kroeze C, Koelmans AA, Flörke M, Franssen W, Hofstra N, Langan S, Tang T, van Vliet MTH, et al.: **Global multi-pollutant modelling of water quality: scientific challenges and future directions**. *Current Opinion in Environmental Sustainability* submitted **special issue "Water Quality: a new challenge for global scale modelling"**.
48. Pérez CJ, Vega-Rodríguez MA, Reder K, Flörke M: **A Multi-Objective Artificial Bee Colony-based optimization approach to design water quality monitoring networks in river basins**. *Journal of Cleaner Production* 2017, **166**:579-589.
49. van Vliet MTH, Franssen WHP, Yearsley JR: **Global modelling of river water quality under climate change**. *Geophysical Research Abstracts, EGU General Assembly 2017* 2017, **Vol. 19,** :EGU2017-4713.
50. Gillett NP, Shiogama H, Funke B, Hegerl G, Knutti R, Matthes K, Santer BD, Stone D, Tebaldi C: **The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6**. *Geosci. Model Dev.* 2016, **9**:3685-3697.
51. Weedon GP, Gomes S, Viterbo P, Shuttleworth WJ, Blyth E, Österle H, Adam JC, Bellouin N, Boucher O, Best M: **Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century**. *J. Hydrometeor.* 2011, **12**:823–848.
* 52. van Vliet MTH, Flörke M, Wada Y: **Quality matters for water scarcity**. *Nature Geoscience* 2017, **10**:800 - 802.
53. Zeng Z, Liu J, Savenije HHG: **A simple approach to assess water scarcity integrating water quantity and quality**. *Ecological Indicators* 2013, **34**:441-449.
54. Vörösmarty CJ, McIntyre PB, Gessner MO, Dudgeon D, Prusevich A, Green P, Glidden S, Bunn SE, Sullivan CA, Reidy Liermann C, et al.: **Global threats to human water security and river biodiversity**. *nature* 2010, **09440**.
55. McIntyre NR, Wheater HS: **A tool for risk-based management of surface water quality**. *Environmental Modelling & Software* 2004, **19**:1131-1140.
** 56. Kroeze C, Gabbert S, Hofstra N, Koelmans AA, Li A, Löhr A, Ludwig F, Strokal M, Verburg C, Vermeulen L, et al.: **Global modelling of surface water quality: a multi-pollutant approach**. *Current Opinion in Environmental Sustainability* 2016, **23**:35-45.
57. Overpeck JT, Meehl GA, Bony S, Easterling DR: **Climate Data Challenges in the 21st Century**. *Science* 2011, **331**:700-702.
58. Ricke KL, Moreno-Cruz JB, Schewe J, Levermann A, Caldeira K: **Policy thresholds in mitigation**. *Nature Geoscience* 2015, **9**:5.

Paper of special interesent (*) or outstanding interest (**)

Figures and Tables

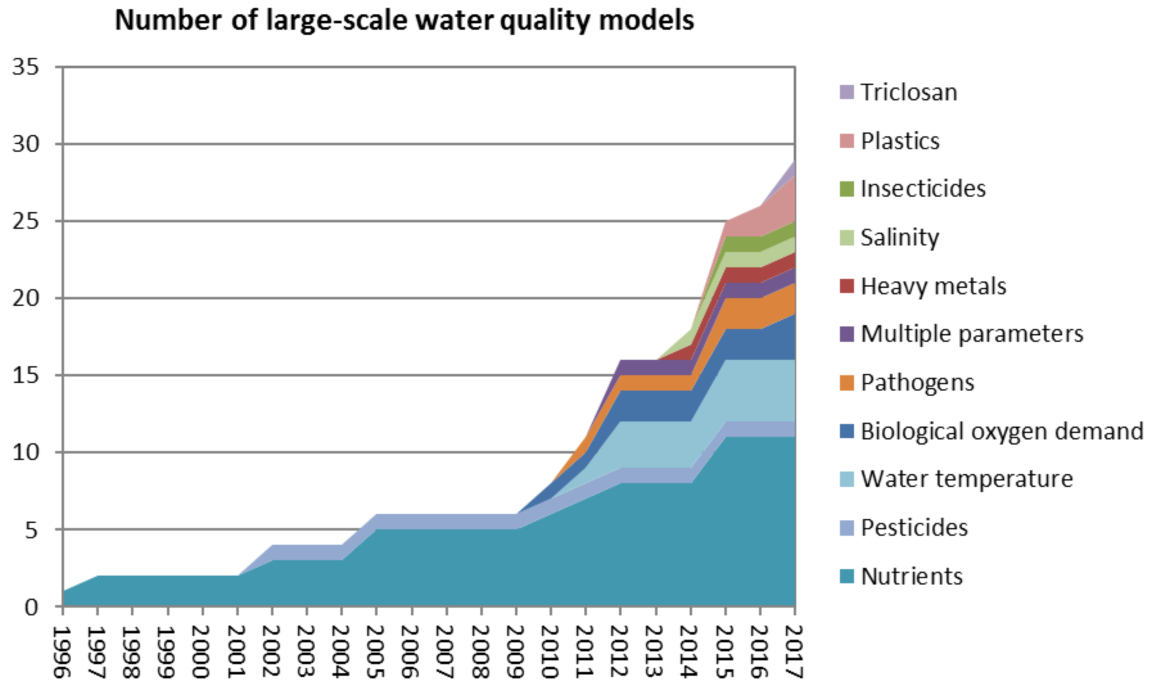

Figure 1: Increase in number of large-scale water quality models per water quality variable since the 1990s. A large-scale water quality model is defined here as a model capable of simulating one or more water quality variables on a scale that exceeds the size of one river basin. See Supplementary Information Table S1 for an overview of published studies per large-scale water quality model.

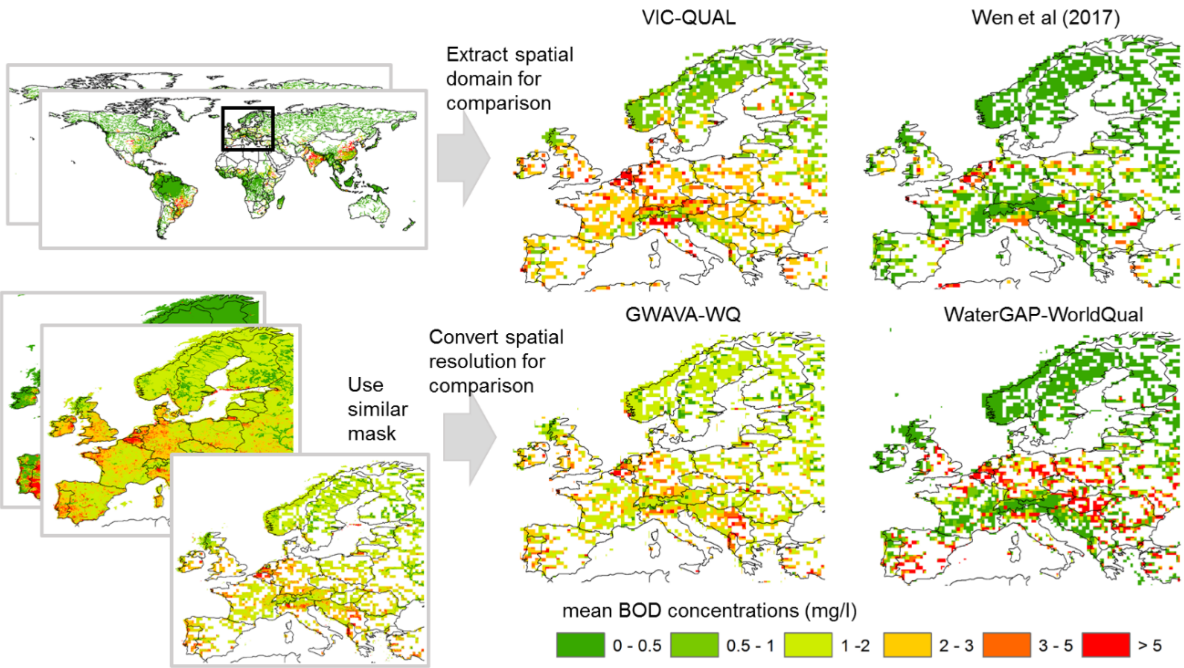

Figure 2: Model comparison of simulated mean BOD concentrations for Europe converting spatial domains and resolutions and aggregating to average values for the period of 1990-2000. Global gridded 0.5° simulations were extracted from the global models VIC-QUAL [49] and the global BOD model of Wen et al. [25] (upper panels), and BOD simulations from GWAVA-WQ [21] and WaterGAP-WorldQual [18] for Europe at 5’x5’ were aggregated to 0.5°x0.5° (lower panels). The BOD model of Wen et al. [25] excludes grid cells with very low water availability, and a similar mask to exclude grid cells with low water availability was therefore applied to the other BOD models to allow for a consistent comparison.

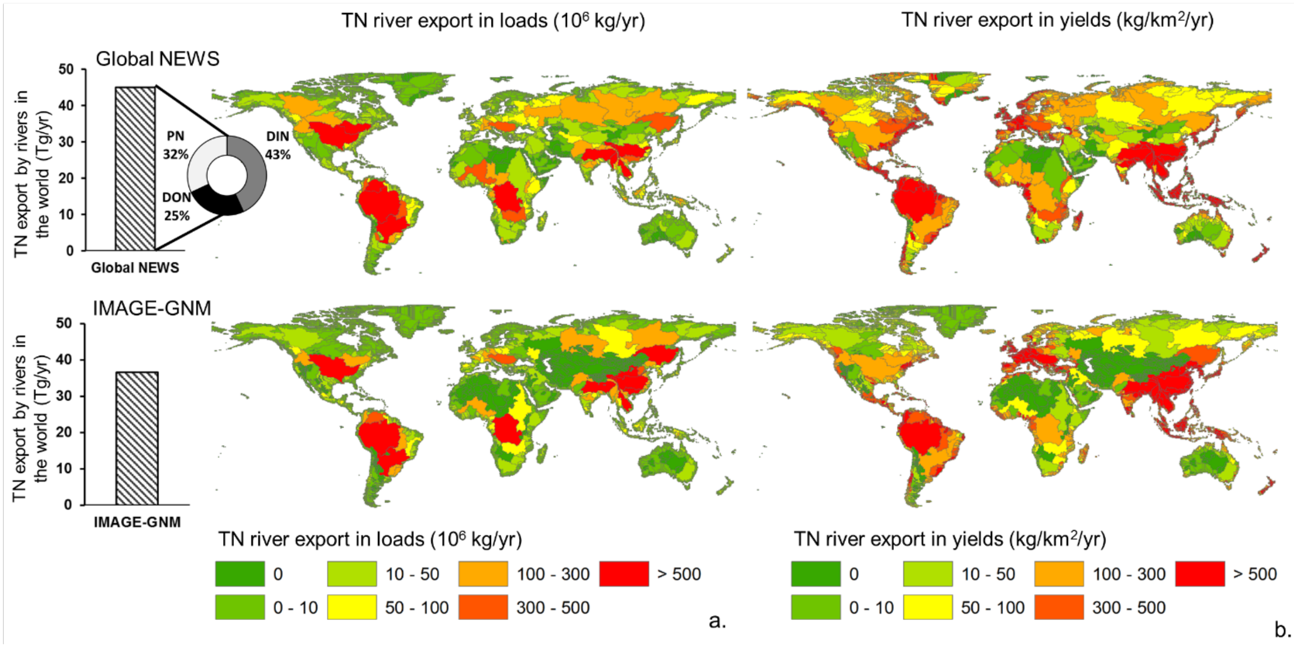

Figure 3. Use of similar model output variables and units for model inter-comparison of global total nitrogen (TN) river export in loads (a) and yields (b). Different nitrogen forms simulated by Global *NEWS*-2 [11] (upper panels) were aggregated to compare with total nitrogen (TN) river export from IMAGE-GNM [14] (lower panels). Different nitrogen forms are dissolved inorganic nitrogen (DIN), dissolved organic nitrogen (DON) and particulate nitrogen (PN). TN river export from the grid-based IMAGE-GNM (0.5°) at basin outlet gridcells were compared with TN river export from similar basin outlets of Global *NEWS*-2.

Table 1: Relative importance of proposed recommendations for the five main aims of water quality model inter-comparison. Greyscale indicates the relative importance (light grey = relevant; middle grey = important; dark grey = highly needed (compulsory) to include in water quality MIP design)

Aim / Recommen-dation	R1: Use similar spatial/temporal domains and resolutions (harmonize on model output)	R2: Use similar model output variables for comparison (harmonize on model output)	R3: Harmonize on main model input datasets
1. Identify robust water quality (pollution) hotspots			
2. Assess robust trends in water quality			
3. Improve understanding of processes and sources of water pollution			
4. Increase understanding of water quality model uncertainties			
5. Identify and set priorities for water quality data collection and monitoring			