Mapping global hotspots and trends of water quality (1992-2010): a data driven approach

Sebastien Desbureaux, Frederic Mortier, Esha Zaveri, Michelle van Vliet, Jason Russ, Sophie Aude, Richard Damania

CEE-M
CENTER FOR ENVIRONMENTAL ECONOMICS – MONTPELLIER

Mapping global hotspots and trends of water quality (1992-2010): a data driven approach

Sebastien Desbureaux
Frederic Mortier
Esha Zaveri
Michelle van Vliet
Jason Russ
Aude Sophie Rodella
&
Richard Damania

CEE-M Working Paper 2022-08

CEE-M

INRAE

cnrs

Montpellier SupAgro

Mapping global hotspots and trends of water quality (1992-2010): a data driven approach

Sebastien Desbureaux*[,1], Frederic Mortier[2,3,4], Esha Zaveri[5], Michelle van Vliet[6], Jason Russ[5], Aude Sophie Rodella[5], and Richard Damania[5]

[1]Center for Environmental Economics - Montpellier, Univ. Montpellier, CNRS, INRA, SupAgro, Montpellier, France
[2]CIRAD, Forêts et Sociétés, Montpellier, France
[3]Forêts et Sociétés, Univ. Montpellier, CIRAD, Montpellier, France
[4]Georgetown Environmental Justice Program, Georgetown University, Washington, DC, USA
[5]The World Bank, USA
[6]Department of Physical Geography, Utrecht University, Utrecht, The Netherlands

August 2022

Abstract

Clean water is key for sustainable development. However, large gaps in monitoring data limit our understanding of global hotspots of water quality and their evolution over time. We demonstrate the value added of a data-driven approach to provide accurate high-frequency estimates of surface water quality worldwide over the period 1992-2010. We assess water quality for six indicators (temperature, dissolved oxygen, pH, salinity, nitrate-nitrite, phosphorus) relevant for the Sustainable Development Goals (SDG). The performance of our modelling approach compares well to, or exceeds, the performance of recently published process-based models. The model's outputs indicate that poor water quality is a global problem that impacts low-, middle- and high-income countries but with different pollutants. When countries become richer, water pollution does not disappear but evolves.

*Corresponding author: sebastien.desbureaux@umontpellier.fr This research was undertaken as part of the *Quality Unknown: The Invisible Water Crisis* project within the World Bank's Water Global Practice. The authors are very grateful for comments from seminar participants at Eco-Publique, CEE-M, Espace-Dev and Georgetown EJP. The findings, interpretations, and conclusions are entirely those of the authors.

1 Introduction

Water quality deterioration is a global and growing problem for human development and ecosystem health. It negatively impacts health in the long term, and it decreases labor and agricultural productivity, which may result in lower incomes for people [1, 2]. As a consequence, targets of Sustainable Development Goal (SDG) 6 aim to ensure safely managed drinking water and sanitation services, improve ambient water quality, and protect water-related ecosystems. SDG indicator 6.3.2 tracks bodies of water with "good" ambient water quality, where "good" refers to a level of dissolved oxygen, salinity, nutrients (nitrogen and phosphorus) and acidity that does not damage ecosystem and human health. In addition, SDG 6.6 aims at protecting and restoring water-related ecosystems, for which these selected water quality indicators are highly relevant.

Although there are high ambitions in the SDGs to improve water quality, there is a paucity of data across much of the world. Furthermore, when data are available at a given location (primarily in the global north), time series are often incomplete, as illustrated by the GEMStat database (Fig. 1)– one of the largest databases of *in-situ* measurements of freshwater quality. In addition, a majority of data points are about thirty years old (Fig. 1), making them outdated and largely uninformative for policy purposes.

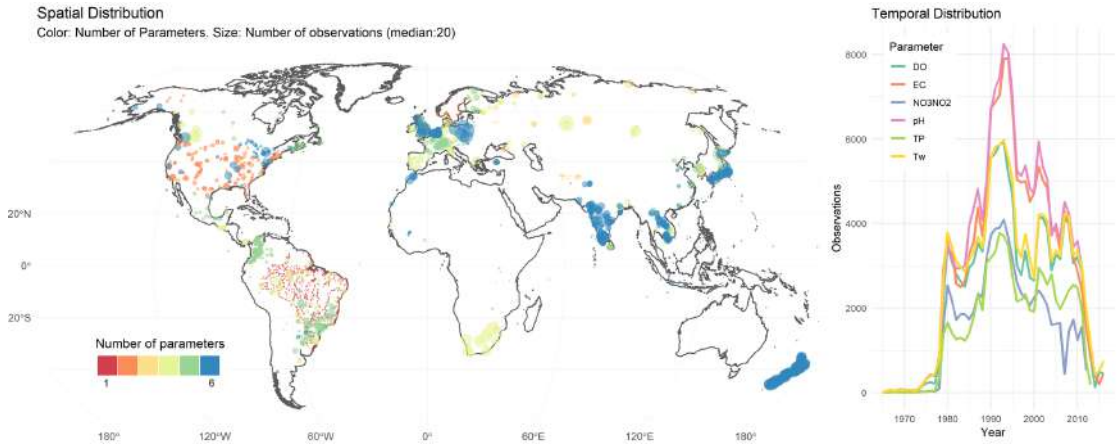


Figure 1: GEMStat data for dissolved oxygen (DO), electrical conductivity (EC), nitrate-nitrite (NOxN), pH, total phosphorus (TP), and temperature (Tw) between 1992 and 2010. The left panel shows the spatial distribution of the original observations per station. Dots size represents the number of observations per station. Dots color represents the number of indicators measured in a station. The right panel shows the temporal availability of data.

Process-based models are today the main modelling approach to fill data gaps in the water literature. Since 2010, there has been a rapid growth in the number of large-scale models for predicting indicators such as river water temperature [3–6], nutrients [7, 8], organic pollution [9–11], micro-organisms [12], chemicals [13], plastics [14], nanomaterials [15] and pesticides. Limited systems knowledge and parameter availability exist to mechanistically predict water quality variations at high temporal and spatial resolutions using process-based models [16]. In other fields, ranging from forest ecology [17] to development economics [18], machine learning models are increasingly used to flexibly predict missing data with high accuracy.

Our paper analyzes the value-added of predictive statistics models to traditional process based models in filling global data gaps. We use a fairly standard statistical model, Random Forests (RF), to predict six water quality indicators relevant for SDG 6 at a monthly temporal scale between 1992-2010 and globally a 0.5° resolution. These indicators are Dissolved Oxygen (DO) concentrations, Electrical Conductivity (EC) for salinity, Nitrate-Nitrite (NOxN) and Total Phosphorus (TP) concentration for nutrients, and pH for acidification. We compare our estimates to state-of-the-art process based models [19] to understand what can machine learning approaches can bring to the water literature when used at a large spatial scale. It completes recent results that have so far focused on

nutrient pollution only [20–22].

2 Methods

Random Forests (RF) are an ensemble, nonparametric modeling approach. It grows a "forest" of individual regression trees which improve upon bagging by using the best random set of predictors at each node in each tree.

2.1 Water Quality Data

We use water quality data from GEMStat which is a globally harmonized database on freshwater quality developed by UNEP-GEMS, maintained by the International Centre for Water Resources and Global Change (ICWRGC) and hosted by the Federal Institute of Hydrology in Koblenz. Raw data for the six water quality indicators are mapped in Supplementary Information 1. As many observations are not correctly encoded in GEMstat, we clean the raw data to exclude outliers, including observations flagged as "suspect" by GEMStat and observations with unrealistic values regarding the property of the pollutant and the long term distribution of the indicator in a given location.[1] Finally, some countries are overrepresented in our sample (Brazil for DO, TP and Tw; New Zealand for EC; South Africa for NOxN and pH). To limit spatial bias in the results, we randomly sample observations from the country with the highest number of observations and limit the number of observations to that of the second most represented country.

2.2 Predictors

We constructed a data set of 66 possible drivers to train and predict the model (Supplementary Information Table 1). Data come from 14 sources and include sanitation related variables, GDP per capita [23], population [24], urbanization rate, fertilizer use [25], croplands extent, livestock, precipitation and temperature [26], runoff [27], elevation, distance to shore, soil composition (soil pH and EC) [28], and river flows [29]. Squares, cubes, and interactions of variables were constructed to provide additional flexibility to the model.

2.3 Model

Model estimation, fitting, and prediction were done with the ranger and caret libraries in R 4.1 [30–32]. Model training for each water quality indicator was done as follows. First, covariates with a near zero variance were excluded. Second, we randomly split the sample into 10 folds and using Cross-Validation (CV) techniques — meaning that a given observation is used only to train or predict the model, but not for both. Third, we modelled water quality as a function of its drivers. We let the algorithm identify which variables to include for making accurate predictions. We estimate for each water quality indicator 1,000 trees. Fourth, we explored which drivers of water quality were selected by the model to ensure that they are coherent with the literature. Fifth, the final model is used to predict global values for all available grid cells.

[1] Thresholds are: DO $\in$ [0,18 mg/L], EC $\in$ [0,10 000 µg/l], NOxN $\in$ [0,90 mg/L], pH $\in$ [0, 14], TP $\in$ [0, 90 mg/L] and Tw $\in$ [0, 100°C]. Additionally, we drop observations that seem unusual in a given station when this value was outside an interval [Mean – 2.5 standard deviations, Mean + 2.5 standard deviations]. The final global predictions are not sensitive to the exclusion of these outliers.

We chose RF for its general prediction performance compared to other regression technics such as linear, partial least squares or support vector regressions. However, RF can present drawbacks, such as its sensitivity to time and/or spatial autocorrelations [33]. Such dependencies may lead to over-optimistic predictions. We test the accuracy of the predictions using station-blocs CV and water basin Leave One-Out CV (LOOCV). In the station-bloc CV, instead of randomly allocating water quality observations to folds, we attributed monitoring stations to 10 different folds and trained the model using CV. In the basin LOOCV all observations from a given basin were successively excluded from the training procedure and only used for testing. This was done to simulate the absence of a large geographical area. Finally, we conducted a temporal validation using an annual LOOCV approach (all observations of a given year are sequentially excluded from the training to be used for testing).

2.4 Area of Applicability

The validity and spatial transferability of RF predictions relies on the similarity that exists between the values of the predictors in the training and prediction samples. The spatial imbalances in our training dataset means that our model could not be able to predict trustworthy values for some part of the world. This is for example the case for Sub-Saharan Africa for which we have extensive observations from South-Africa, and more limited observations from Ghana, Lesotho, Mali, Senegal, Sudan and Tanzania. Recent advances allow to determine Area of Applicability (AOA) and Dissimilarity Index (DI) [34]. AOA is defined as the area, for which the cross-validation error of the model applies. It is based on DI, a metric based on the minimum distance to the training data in the predictor space. We determine the AOA for each water quality indicator [35].

3 Results

3.1 Accuracy

R^2 for random splits, basin-block, station-block cross validation, and temporal splits are synthetized in Table 1 and illustrated in Sup Information Figures 2, 3, 4, 5. A high correlation was found between observed and predicted water quality. With standard random splits validation techniques, the model explains 81% of the observed variability in the testing sample for pH, 70% for EC, 79% for DO, 71% for NOxN and 94% for Tw. This performance compares well to, or exceeds, the performance of other recently published process-based models. For example, the Root Mean Square Error (RMSE) of predictions for water temperature is half as large as reported for global process-based water temperature models [3–6]. A lower model performance was found for TP, where it predicts 37% of the observed variability in the testing sample. The prediction power of the models decreases, without collapsing, when using spatially structured cross-validation based on basins or stations. R^2 for Tw decreases only from 94% to 87%. For DO and pH, R^2 are also preserved, but at lower levels. Higher decreases are observed for EC, NOxN, and TP, particularly for basin-block cross-validation. However, further tests attest that this loss of predictive power is driven by a handful of basins (e.g., Schelbe basin for EC). The model preserves predictive power but cautions need to be taken to interpret local values. When station-block cross-validation is used, more predictive power is maintained compared to basin-block cross-validation. Finally, we follow previous assessments in process-based models by splitting water quality observations into three classes (good, medium or bad, thresholds displayed in Sup. Inf. Table 2). Our model accurately predicts the class of water quality and outperforms the process-based models in this task. As an illustration, for salinity, accuracy increases from 80% using a process-based model (Appendix B of [19]) to 96% in the data-driven approach described in this paper.

		Random CV		Station CV			Main Basin LOOCV			Year LOOCV	
	# Obs.	R2 (%)	Class Acc. (%)	# Stations	R2 (%)	Class Acc.(%)	# Basins	R2 (%)	Class Acc. (%)	R2 (%)	Class Acc. (%).
DO	81401	79	90	1724	69	81	173	62	80	9	85
EC	90993	70	96	1494	36	87	192	28	86	73	93
NOxN	111535	71	82	2154	35	57	163	35	43	73	71
pH	137471	81	98	2598	64	97	221	41	97	77	98
TP	78257	37	84	1610	19	61	135	1	50	38	75
Tw	81499	94	96	2089	9	91	189	87	89	93	93

Table 1: Synthesis of models' performance for each type of validation. For random and station cross-validation (CV), ten folds were constructed. For comparability with process-based models such as UNEP [19], we split water quality observations in three classes (bad, medium and good) and determine what percentage of the predicted values for water quality falls into the accurate class.

3.2 Predictions

Monthly time series data from 1992 to 2010 are generated for the six water quality indicators. Fig. 2 shows the predicted average value of water quality between 2000 and 2010. Fig. 3 displays the predicted change in annual average water quality between 1992 and 2010. Sup Inf. 6 shows global trends. Sup. Inf. 7 highlights that a combination of hydro-climatic and socio-economic variables best predict all pollutants. Sup. Inf. 8 maps DI. The results indicate that for EC, Temp, TP and to a lesser extent pH, we can confidently extrapolate predict water quality in continents like Sub-Saharan Africa despite the limited input water quality data. This is because the model can complements local Afircan data by data from other continents, possibly at different times (e.g.: Latin America or South Asia in the 1990s shared important similarities with lartge parts of Sub-Saharan Africa in later decades). For DO and NOxN, the procedure indicates that uncertainties exist in some areas to predict water quality, including in Sub-Saharan Africa.

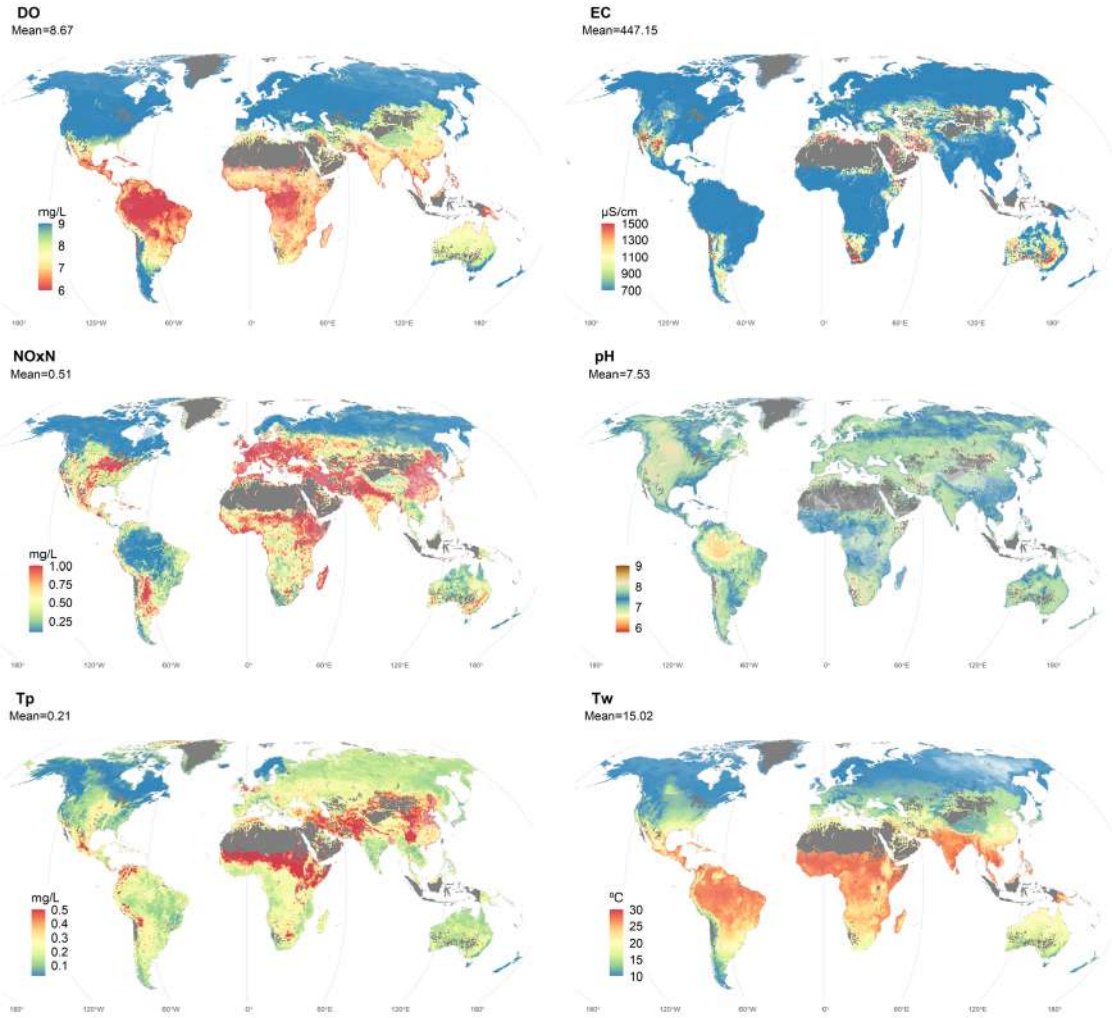


Figure 2: Global maps of river water quality risks for SDG pollutants dissolved oxygen (DO), electrical conductivity (EC), nitrate-nitrite (NOxN), pH, total phosphorus (TP), and temperature (Tw). Maps below present average values between 2000-2010. Regions with river discharge less than 1 m3/s or with missing data on covariates (e.g., Indonesia) are masked (grey). Blue and green represent good water quality, as defined by indicative thresholds provided in Sup Table 2. Yellow, orange and red represent moderate and poor water quality. For Tw, the color coding does not represent quality but the average level. Transparency was added to the cells in which a DI was too high between the training and predictions samples (Sup Inf. 8)

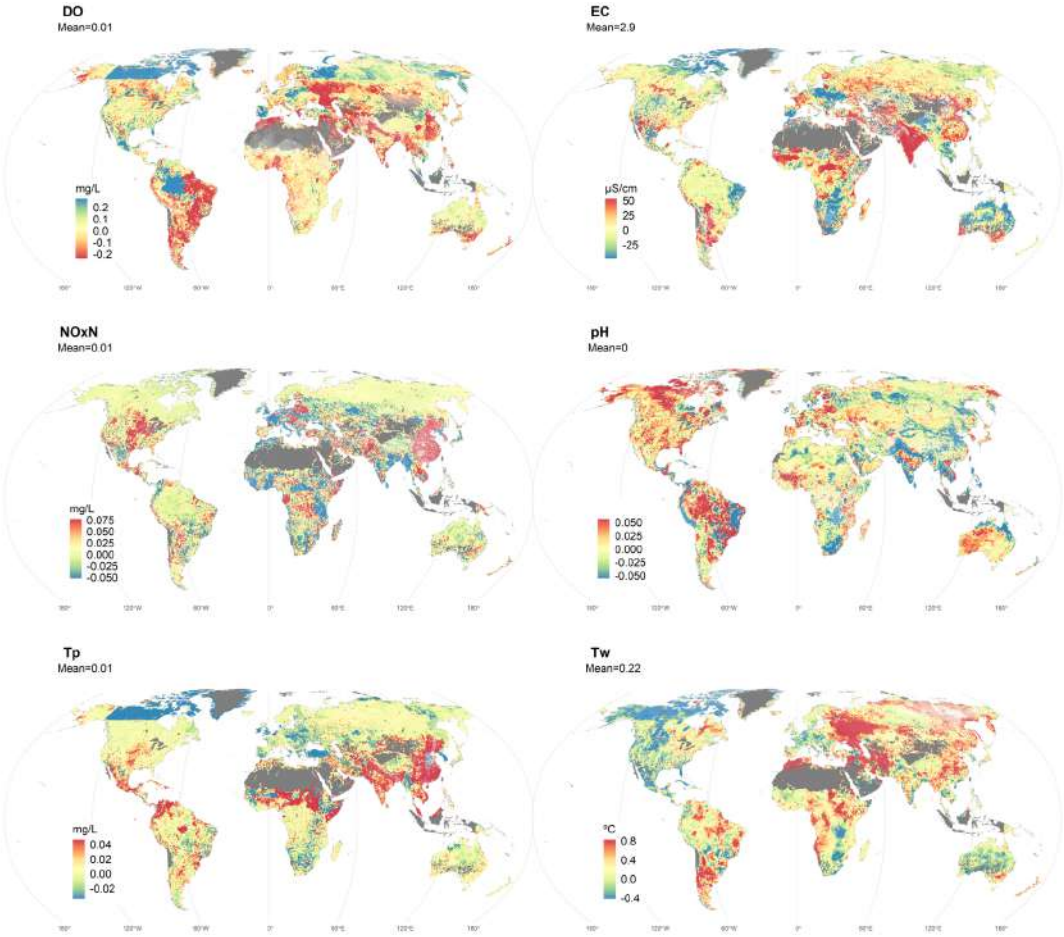


Figure 3: Evolution of water quality between 1992 and 2010 for SDG indicators dissolved oxygen (DO), electrical conductivity (EC), nitrate-nitrite (NOxN), pH, total phosphorus (TP), and temperature (Tw). Values for 1992 were estimated as the annual averages for the years 1992,93 and 94. Likewise, values for 2010 were estimated as the annual averages the years 2008,09 and 10. This was done to limit possible anomalies. Regions with river discharge less than 1 m3/s or with missing data on covariates (e.g., Indonesia) are masked (grey). Blue and green represent an improvement in water quality (e.g., decreases in DO, increases in NOxN). Orange and red represent a worsening of water quality. For pH, deviations with respect to pH=7 were calculated. For Tw, the water coding highlights cooling or warming. Transparency was added to the cells in which a DI was too high between the training and predictions samples (Sup Inf. 8)

Unsafe levels of water quality are widely found in most parts of the world, driven by both climate and anthropic pressures. Low-, middle- and high-income countries all face unsafe levels but for different types of pollutants. Our model uncovers water quality hotspots in data scarce regions. Low levels of DO – a sign of unsafe water when levels are below 5 – 6.5 mg/L – are widely predicted in large parts of Low- and Middle-Income Countries, such as most of Sub-Saharan Africa, Latin America, and South and Southeast Asia (Fig. 2). Along with hydro-climatic variables, the lack of access to basic sanitation is a key covariate associated with low levels of DO (Supplementary Inf. 7). The infrastructure gap that prevails in most low- and middle-income countries explains these low values of DO. In places where the infrastructure gap has widened because of high population growth and low investment in sanitation, DO has decreased during the study period. This is, for example, the case for coastal parts of China, India, Nepal, or in the northeast region of Brazil.

The concentration of NOxN in water is the highest in densely populated areas with intensive economic activities. England, Belgium, Germany, and some parts of France are the predicted global hotspots of nitrate-nitrite, notably because of intensive animal farming (poultry and pig) and agricultural activities. The challenge of NOxN in most high-income countries has persisted during the period studied and has worsened in fast-growing economies, such as in South Asia, East Asia (e.g. eastern China) and parts of Mexico (Figure 3). In these fast-growing areas, intensive animal farming,

combined with high population density, excessive fertilizer use, and infrastructure gaps contribute to high nutrient pollution levels (Supplementary Inf. 7). A certain degree of caution should be taken when interpreting data from some parts of East Asia, because of the dissimilarities between the training and testing samples for NoxN.

High levels of salinity, as reflected by EC, are driven by geological conditions, drier climates, and the use of fertilizers, which is in correspondence with the overview salinity drivers identified in various river basins across the world [36]. Thus, Australia, Mexico, the Southern USA and Central Asia are salinity hotspots because of their drier climates (Supplementary Inf. 7). Over the study period 1992-2010, EC is predicted to have increased the most in India. Turning to Tw and pH, we find that soil composition and air temperature are, respectively, strong determinants of observed levels (Supplementary Inf. 6). Large parts of the world have experienced increases in water temperature greater than 1°C in less than 20 years because of climate variations and change. Such increases in water temperature can have detrimental effects on aquatic life [37–39].

4 Discussion and conclusion

Filling data gaps for water quality will be key to better understanding where hotspots are, to determine trends, and to thus understand our progress in reaching SDG 6 targets. Our data-driven models, based on well-established statistical algorithms, can play a significant role in this endeavor and have shown suitable model performance. It flexibly identifies combinations of factors among a large set of possible drivers to provide accurate estimates of water quality that replicate intra- and inter-annual variations in water quality. They are robust to out of sample geographical predictions and perform at least as well as traditional measurement tools, thus offering a promising path forward for water quality mornitoring measurement. Because of their flexibility, high accuracy and ability to model uncertainties, machine learning approaches should be seen as highly complementary to existing process-based models.

Our results show that critical regions and hotspots of water pollution are found across low-, middle- and high-income countries, but for different water quality indicators. Fast growing middle-income countries tend to suffer from a combination of pollutants found in both low- and high-income countries. This is particularly salient wen synthetizing all pollutants in a synthetic water quality indicator (Sup. Inf 8). When the income levels of countries increase, our results illustrate that water quality does not automatically improve: economic development does not solve the problem of poor water quality, but transforms it. In low-income countries, the dominant concern is the *water pollutant of poverty* resulting from poor sanitation and litter that are mostly driven by infrastructure gaps in a fast-changing environment [40, 41]. Elsewhere there are concerns with *pollutants of prosperity* that result from more intensive economic activities, captured here by NOxN or in other studies by pesticide [42], plastic [14] and pharmaceutical pollutions [43]. Reaching SDG targets will require further investments in treatment, as well as emission control efforts to prevent the pollution happening in the first place.

Data driven models, such as the one presented here, are an accurate, low-cost and fast method to complement in-situ measurements collected in lakes and rivers. However, the performance of the models also strongly depends on the quality of the input datasets that are used. Although GEMStat is critically important for researchers, policy makers, and civil society, it suffers from important drawbacks. While some regions are well covered in the water quality monitoring database, such as North America, Brazil, and India, other regions such as Central and North Africa, Western and Central Asia, the South Pacific, and Australia are charactarized by large data gaps both in time and space. The absence of water quality monitoring data for large geographic areas such as sub-Saharan Africa, might introduce biases in the predictions if there are different drivers of pollution across regions. Our model serves as a starting point and future work could strengthen the results by expanding to more

relevant water quality indicators, using new algorithms (including hydrological grounded models), including more precise drivers, and employing richer water quality training data. We believe that the flexibility of the approach and its transparency can make these tools useful in the near real-time monitoring of water quality, notably in the context of the SDGs. To this end, an important conclusion of this research is the critical need to expand the spatial coverage of current databases to data-poor regions, notably Sub-Saharan-Africa, to enhance the spatial transferability of the results. The same spatial and temporal gaps are present in more rencently developed database, such as GRQA [44].

Our model serves as a starting point and future work could strengthen the results by expanding to more relevant water quality indicators, using new algorithms (including hydrological grounded models), including more precise drivers, and employing richer water quality training data. We believe that the flexibility of the approach and its transparency can make these tools useful in the near real-time monitoring of water quality, notably in the context of the SDGs. To this end, an important conclusion of this research is the critical need to expand the spatial coverage of current databases of water quality to data-poor regions, notably East Asia and Sub-Saharan-Africa, to enhance the spatial transferability of the results.

Authors contribution

SD, RD, JR, EZ, ASR conceived the study. SD analyzed the data, with the support of FM, and led the writing of the paper. MvV, RD, JR, EZ, and ASR commented on the analysis and provided critical inputs for the writing of the paper. RD led the overall team for the "Quality Unknown" project.

Data and Code Availability

All data analyzed in this study are publicly available. The input data on water quality need to be requested from GEMSTAT (https://gemstat.org/data/data-portal/custom-data-request/). All data and sources for covariates are presented in Supplementary Table 1. All predictions are made available through the platform xxx [Note: to be added after peer review process]

The analysis scripts for the random cross-validation, spatial cross-validation, and temporal cross-validation are available through figshare (https://doi.org/10.6084/m9.figshare.19486868.v1).

Competing interests

The authors declare no competing interest.

References

1. Vörösmarty, C. J. *et al.* Global threats to human water security and river biodiversity. *nature* **467,** 555–561 (2010).
2. Damania, R., Desbureaux, S., Rodella, A.-S., Russ, J. & Zaveri, E. *Quality unknown: the invisible water crisis* (World Bank Publications, 2019).

3. 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* **13,** 1052–1065 (2012).

4. Van Beek, L. P., Eikelboom, T., van Vliet, M. T. & Bierkens, M. F. A physically based model of global freshwater surface temperature. *Water Resources Research* **48** (2012).

5. Van Vliet, M. *et al.* Coupled daily streamflow and water temperature modelling in large river basins. *Hydrology and Earth System Sciences* **16,** 4303–4321 (2012).

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

7. Mayorga, E. *et al.* Global nutrient export from WaterSheds 2 (NEWS 2): model development and implementation. *Environmental Modelling & Software* **25,** 837–853 (2010).

8. Beusen, A., Van Beek, L., Bouwman, A., Mogollón, J. & Middelburg, J. 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,** 4045–4067 (2015).

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

10. Van Vliet, M. T. *et al.* Model inter-comparison design for large-scale water quality models. *Current opinion in environmental sustainability* **36,** 59–67 (2019).

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

12. Vermeulen, L. C. *et al.* Cryptosporidium concentrations in rivers worldwide. *Water Research* **149,** 202–214 (2019).

13. Van Wijnen, J., Ragas, A. M. & Kroeze, C. River export of triclosan from land to sea: a global modelling approach. *Science of The Total Environment* **621,** 1280–1288 (2018).

14. Jambeck, J. R. *et al.* Plastic waste inputs from land into the ocean. *Science* **347,** 768–771 (2015).

15. Dumont, E., Johnson, A. C., Keller, V. D. & Williams, R. J. Nano silver and nano zinc-oxide in surface waters–Exposure estimation for Europe at high spatial and temporal resolution. *Environmental pollution* **196,** 341–349 (2015).

16. Tang, T. *et al.* Bridging global, basin and local-scale water quality modeling towards enhancing water quality management worldwide. *Current opinion in environmental sustainability* **36,** 39–48 (2019).

17. Réjou-Méchain, M. *et al.* Unveiling African rainforest composition and vulnerability to global change. *Nature* **593,** 90–94 (2021).

18. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353,** 790–794 (2016).

19. UNEP, A. A snapshot of the world's water quality: towards a global assessment. *Nairobi, United Nations Environment Programme* (2016).

20. Marzadri, A. *et al.* Global riverine nitrous oxide emissions: The role of small streams and large rivers. *Science of The Total Environment* **776,** 145148 (2021).

21. Sheikholeslami, R. & Hall, J. W. A global assessment of nitrogen concentrations using spatiotemporal random forests. *Hydrology and Earth System Sciences Discussions,* 1–30 (2022).

22. Shen, L. Q., Amatulli, G., Sethi, T., Raymond, P. & Domisch, S. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Scientific data* **7,** 1–11 (2020).

23. Kummu, M., Taka, M. & Guillaume, J. H. Gridded global datasets for gross domestic product and Human Development Index over 1990–2015. *Scientific data* **5,** 1–15 (2018).

24. Klein Goldewijk, K., Beusen, A. & Janssen, P. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *The Holocene* **20,** 565–573 (2010).

25. Lu, C. & Tian, H. Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: shifted hot spots and nutrient imbalance. *Earth System Science Data* **9,** 181–192 (2017).

26. Willmott, C. J. Terrestrial air temperature and precipitation: Monthly and annual time series (1950-1996). *WWW url: http://climate. geog. udel. edu/˜ climate/html_pages/README. ghcn_ts. html* (2000).

27. Hejazi, M. I. *et al.* Integrated assessment of global water scarcity over the 21st century under multiple climate change mitigation policies. *Hydrology and Earth System Sciences* **18,** 2859–2883 (2014).

28. Batjes, N. H. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma* **269,** 61–68 (2016).

29. Van Vliet, M. T., Sheffield, J., Wiberg, D. & Wood, E. F. Impacts of recent drought and warm years on water resources and electricity supply worldwide. *Environmental Research Letters* **11,** 124021 (2016).

30. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). https://www.R-project.org/.

31. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77,** 1–17 (2017).

32. Kuhn, M. *caret: Classification and Regression Training* R package version 6.0-88 (2021). https://CRAN.R-project.org/package=caret.

33. Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications* **11,** 1–11 (2020).

34. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* **12,** 1620–1633 (2021).

35. Meyer, H. *CAST: 'caret' Applications for Spatial-Temporal Models* R package version 0.5.1 (2021). https://CRAN.R-project.org/package=CAST.

36. Thorslund, J., Bierkens, M. F., Oude Essink, G. H., Sutanudjaja, E. H. & van Vliet, M. T. Common irrigation drivers of freshwater salinisation in river basins worldwide. *Nature Communications* **12,** 1–13 (2021).

37. Verbrugge, L. N., Schipper, A. M., Huijbregts, M. A., Van der Velde, G. & Leuven, R. S. Sensitivity of native and non-native mollusc species to changing river water temperature and salinity. *Biological Invasions* **14,** 1187–1199 (2012).

38. Ficke, A. D., Myrick, C. A. & Hansen, L. J. Potential impacts of global climate change on freshwater fisheries. *Reviews in Fish Biology and Fisheries* **17,** 581–613 (2007).

39. Van Vliet, M. T., Ludwig, F. & Kabat, P. Global streamflow and thermal habitats of freshwater fishes under climate change. *Climatic change* **121,** 739–754 (2013).

40. Rozenberg, J. & Fay, M. *Beyond the gap: How countries can afford the infrastructure they need while protecting the planet* (World Bank Publications, 2019).

41. Thacker, S. *et al.* Infrastructure for sustainable development. *Nature Sustainability* **2,** 324–331 (2019).

42. Casado, J., Brigden, K., Santillo, D. & Johnston, P. Screening of pesticides and veterinary drugs in small streams in the European Union by liquid chromatography high resolution mass spectrometry. *Science of The Total Environment* **670,** 1204–1225 (2019).

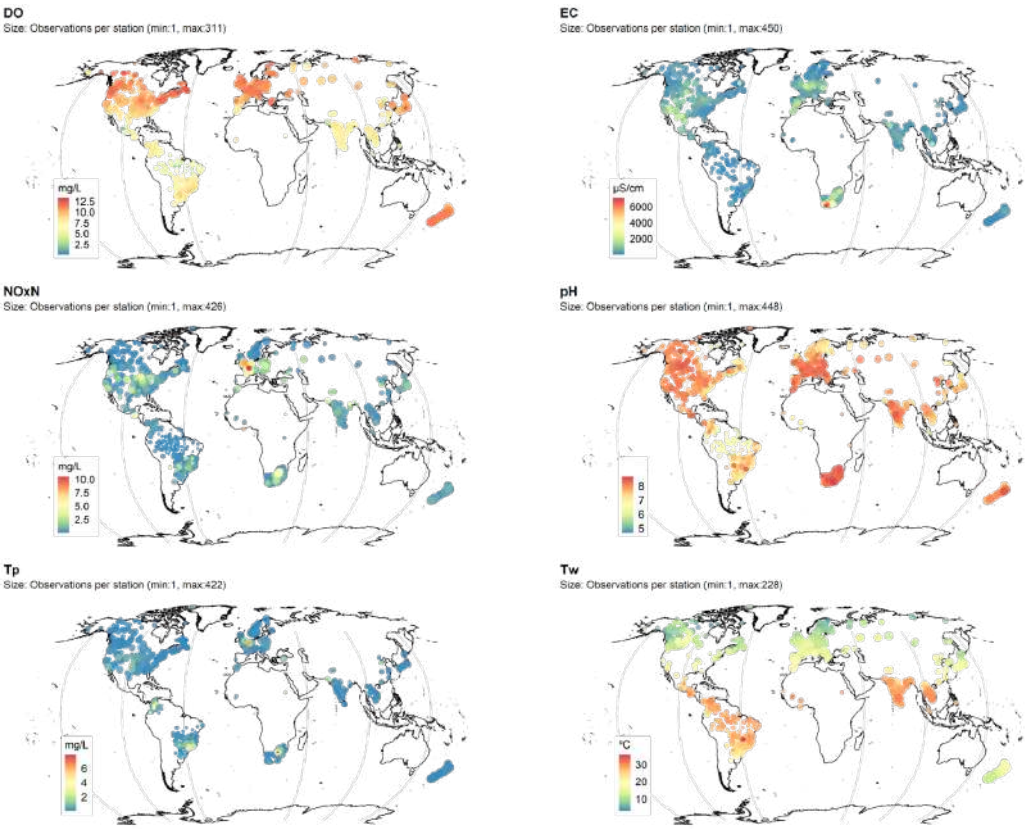
43. Johnson, A. C. *et al.* Do concentrations of ethinylestradiol, estradiol, and diclofenac in European rivers exceed proposed EU environmental quality standards? *Environmental science & technology* **47,** 12297–12304 (2013).

44. Virro, H., Amatulli, G., Kmoch, A., Shen, L. & Uuemaa, E. GRQA: Global River Water Quality Archive. *Earth System Science Data* **13,** 5483–5507 (2021).

Supplementary Information

Original UNGEMS observations by water quality indicator

Sup. Fig. 1: This figure plots the original GEMStat observation available by station. The color is a function of water quality. Dots size varies with the number of available observations in the station. We focus on six water quality indicators of interest for SDG 6: dissolved oxygen (DO, n= 81,401), salinity as reflected by electrical conductivity (EC, n= 90,993), sum of nitrate-nitrite concentrations (NOxN, n= 111,535), pH (n=137,471), total phosphorus (TP, n=78,257) and water temperature (Tw, n=81,499).



Covariates and data sources

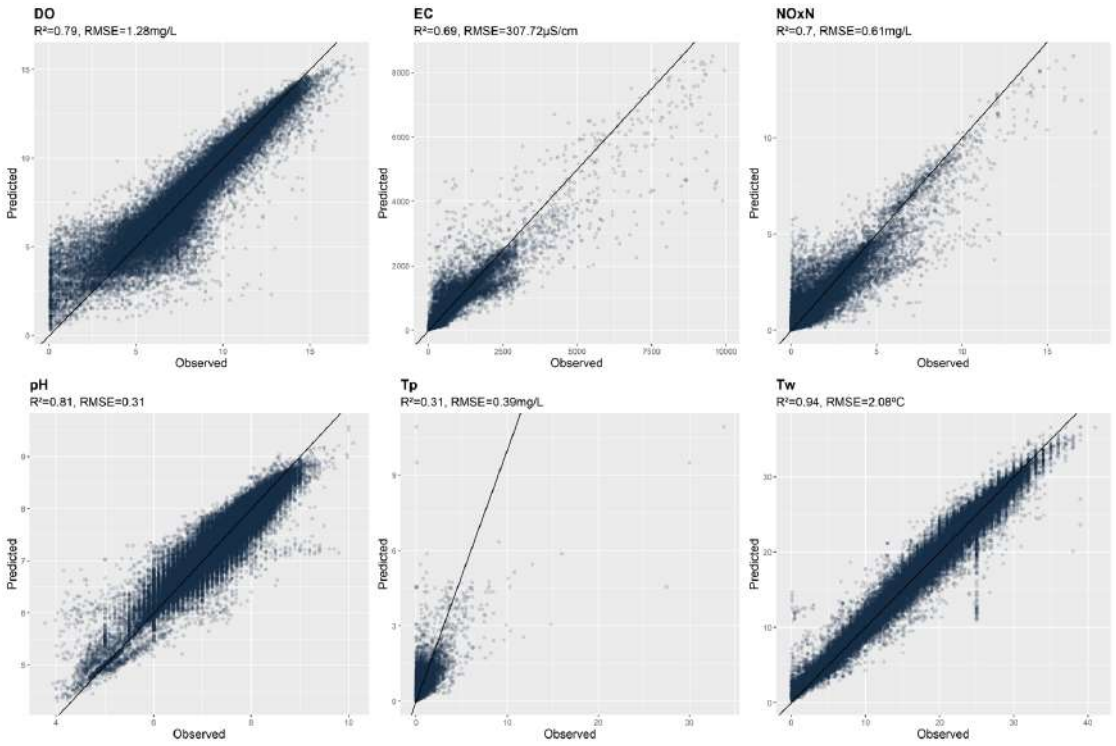
Name	Variables	Source	Spatial resolution	Note
Water Quality Data				
UNGEMS- GEMstat (version January 2022)	DO, EC, NOxN, pH, Temp, TP	gemstat.org	Monitoring station	
Drivers				
WHO/UNICEF (JMP)	Share and total population with At least basic access to sanitation, Limited service for sanitation, Unimproved sanitation, Open defecation, Sewer connections.	washdata.org	Country, disaggregated 0.5°x0.5°	Missing data linearly interpolated by country
Annual gridded GDP per capita 1990–2015	GDP per capita	Kummu, Taka, and Guillaume (2018)	0.5°x0.5°	
Annual population 1990–2015	Total population	HYDE 3.2 (Klein Goldewijk et al. 2010)	0.5°x0.5°	Derived from Kummu, Taka, and Guillaume (2018)
Urbanization	Share urban area 2010	SEDAC	Aggregated at 0.5°x0.5°	2010 value used for 1992-2010
Nitrogen and Phosphorus 1990-2015	Annual average used per ha and total cell	Lu and Tian (2017)	0.5°x0.5°	
Croplands 1992-2015	Share of croplands	European Space Agency Landcover project	Aggregated at 0.5°x0.5°	
Livestock 2010	Number and density of cattle, chickens, goats, pigs, sheeps	FAO (2012) Gridded Livestock of the World v3	Aggregated at 0.5°x0.5°	2010 value used for 1992-2010
Precipitation and Temperature 1900-2013	Monthly and annual precipitation and temperature	Willmott and Matsuura (2001)	0.5°x0.5°	
Runoff	Monthly and annual Runoff	GWAM – Hejazi et al. (2014)	0.5°x0.5°	
Elevation	Average elevation	SRTM	Aggregated 0.5°x0.5°	
Soil	Topsoil ECE, Topsoil PH, Subsoil ECE, Subsoil PH	Harmonized soil database – Wise v30s (Batjes 2016)	Aggregated 0.5°x0.5°	
Distance to coast	Distance to shore	Authors, based on GADM		
Flow	Monthly river flow	Van Vliet et al. (2016)		

Sup. Table 1: Additional interaction terms, squares and cubes were created to allow for more flexibility in the model: Fertilizer x Precipitation, Fertilizer x Temperature, GDP per capita x Fertilizer, Distance to shore x Elevation, Distance to shore x Topsoil ECE

Model Validation

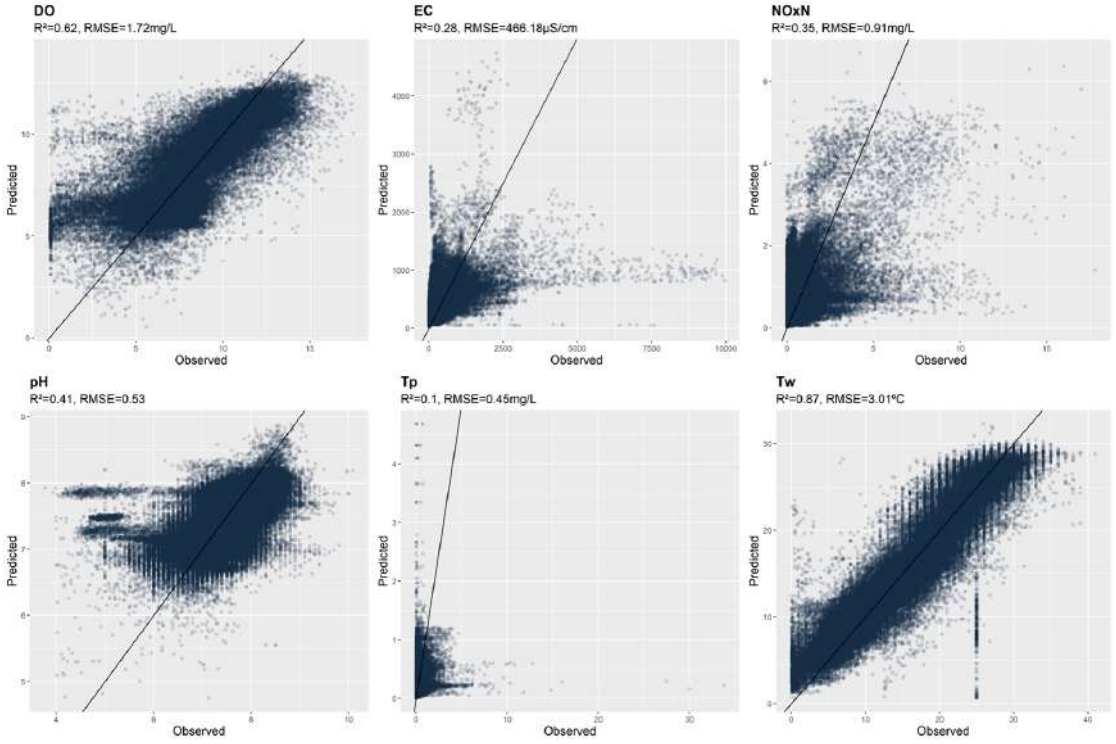
Random 10-folds cross validation

Sup. Fig. 2: Accuracy of predictions in the testing sample for the six water quality indicators using standard random splits. The model was trained and validated using traditional 10 folds cross validations. X-axis: observed values of water quality. Y-axis: predicted values of water quality.



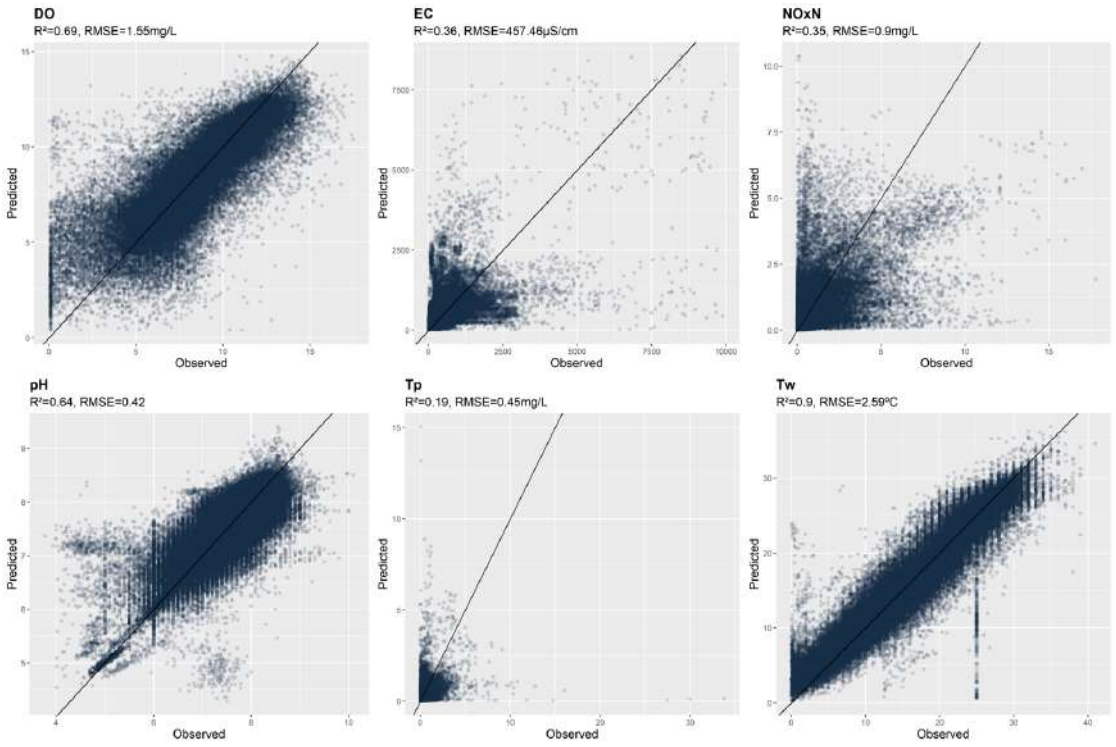
Basin block cross validation

Sup. Fig. 3: To assess the transferability of the predictions, we replace the random split between training and testing dataset by a geographical split. More precisely, we sequentially exclude one basin from the data, train the model and test it on the missing basin (river basin cross-validation). While the accuracy decreases, predictions remain globally valid – particularly for dissolved oxygen and water temperature.



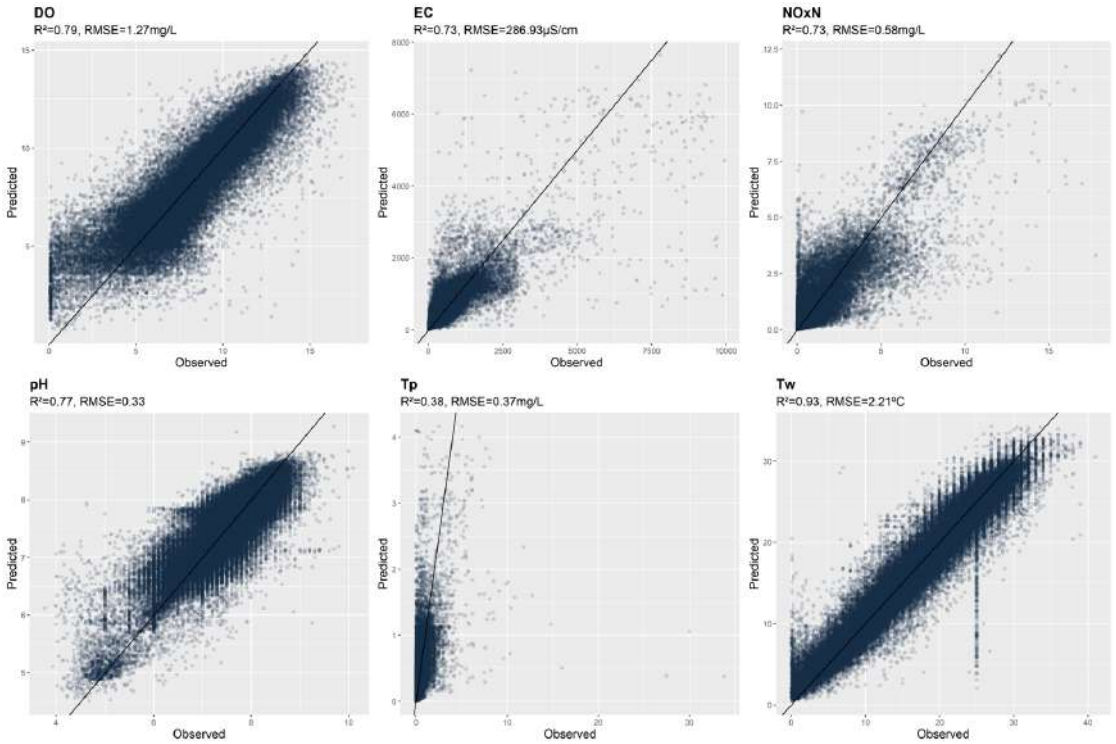
Station block cross validation

Sup. Fig. 4: We perform station-block cross validation to test for the spatial performance of RF for water quality. Stations are randomly attributed to 10 blocks. Each block of station is sequentially excluded from model training and is used for model testing. Accuracy decreases compared to random splits, at the exception of TP. This station-block cross validation provides however much higher accuracy than basin-block cross validation, notably for DO, EC and NOxN. Overall, it confirms that RF preserves predictive power when accounting for spatial dependencies for water quality.



Yearly block cross validation

Sup. Fig. 5: The temporal transferability of results is also key. To assess the transferability of the predictions, we replace the random split between training and testing dataset by a temporal split. We do that by successively excluding all observations from one year and use the rest of the years to predict the missing year (LOOCV). The model continues to replicate temporal variations observed in the data with almost no loss of R2 compared to standard random 10 folds validation.



Indicative thresholds for good, medium and bad water quality

Sup. Table 2: Good, medium and bad water quality: Organizations such as WHO have established safe concentration levels for many of the most common pollutants. Although these concentration levels are partially based on the latest science, there is great uncertainty about the true safe value for certain indicators of water quality. For instance, the WHO sets the limit for nitrate-nitrogen in drinking water at 10 milligrams per liter (mg/L). However, there is emerging evidence that this threshold may be too high (Ward et al., 2018; Zaveri et al., 2020). In addition, the safe thresholds can also differ depending on whether the guidelines relate to protecting human health, aquatic life or overall freshwater systems (Konmeng and Larsen, 2014) That such uncertainty exists speaks to how difficult it is to get the safe levels right. We therefore present indicative thresholds used for water quality. Note that the threshold for temperature does not reflect an idea of quality. Thresholds were however set to measure class accuracy as reported in table 1. Likewise, for pH, class accuracy was done distinguishing one the hand, <6 values and on the other hand >9 in order to keep three classes.

∗: The level of temperature is not representative of any water quality. The thresholds indicated here are simply for mapping and model validation purposes.

Indicative thresholds

Indicator	Unit	Good	Medium	Bad
hline DO	mg/L	9	5-9	5
EC	µS/cm	700	700 - 1500	1500
NOxN	mg/L	0.1	0.1 – 0.5	0.5
pH	-	6-9	6 or 9	
TP	mg/L	0.024	0.024 – 0.2	0.2
Tw*	ºC	10	10-20	20

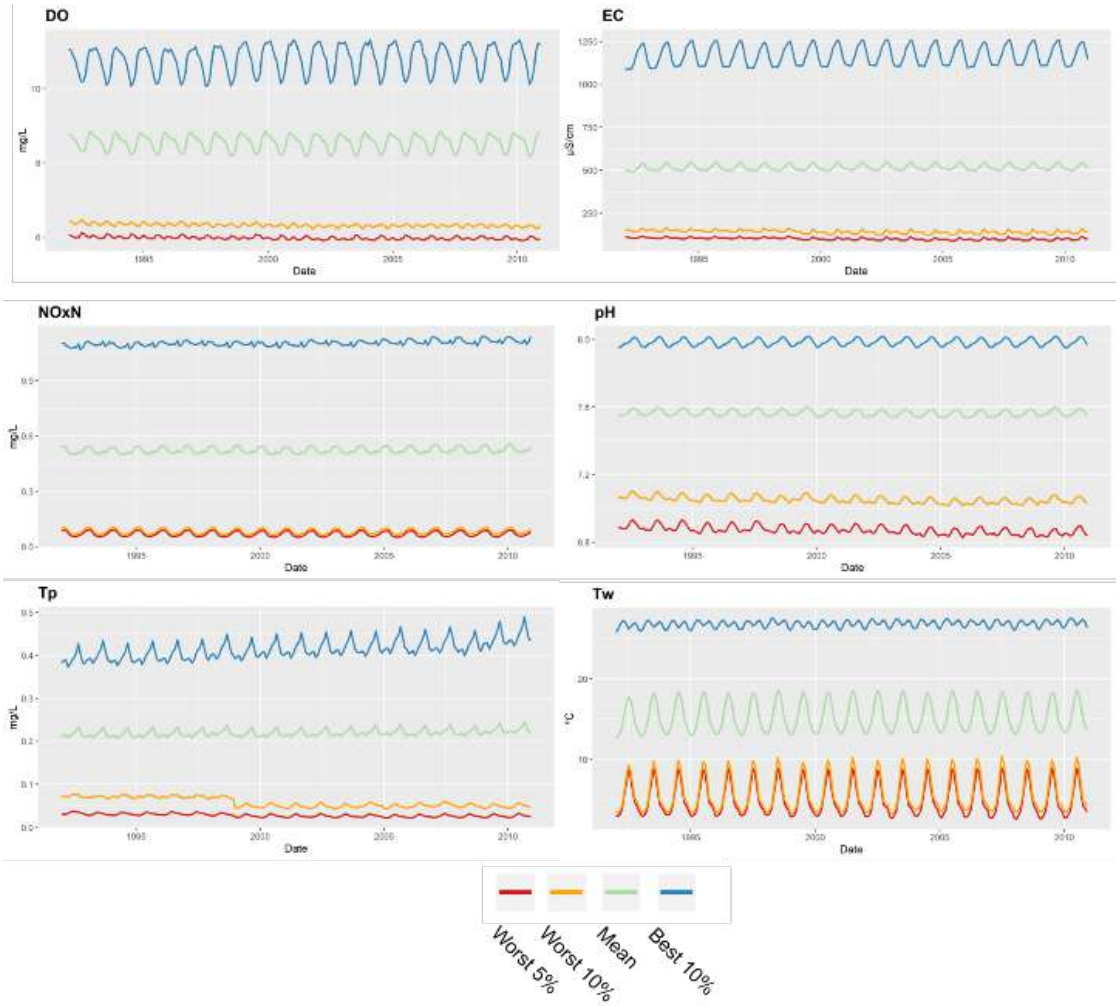
Additional references: Ly, K., Larsen, H., & Duyen, N. V. (2014). Lower Mekong regional water quality monitoring report. Technical paper, (60).

Ward, M., R. Jones, J. Brender, T. de Kok, P. Weyer, B. Nolan, C. M. Villanueva, and S. van Breda. 2018. "Drinking Water Nitrate and Human Health: An Updated Review." International Journal of Environmental Research and Public Health 15 (7): 1557.

Zaveri, E. D., Russ, J. D., Desbureaux, S. G., Damania, R., Rodella, A. S., & Ribeiro Paiva De Souza, G. (2020). The nitrogen legacy: the long-term effects of water pollution on human capital. World Bank Policy Research Working Paper, (9143).

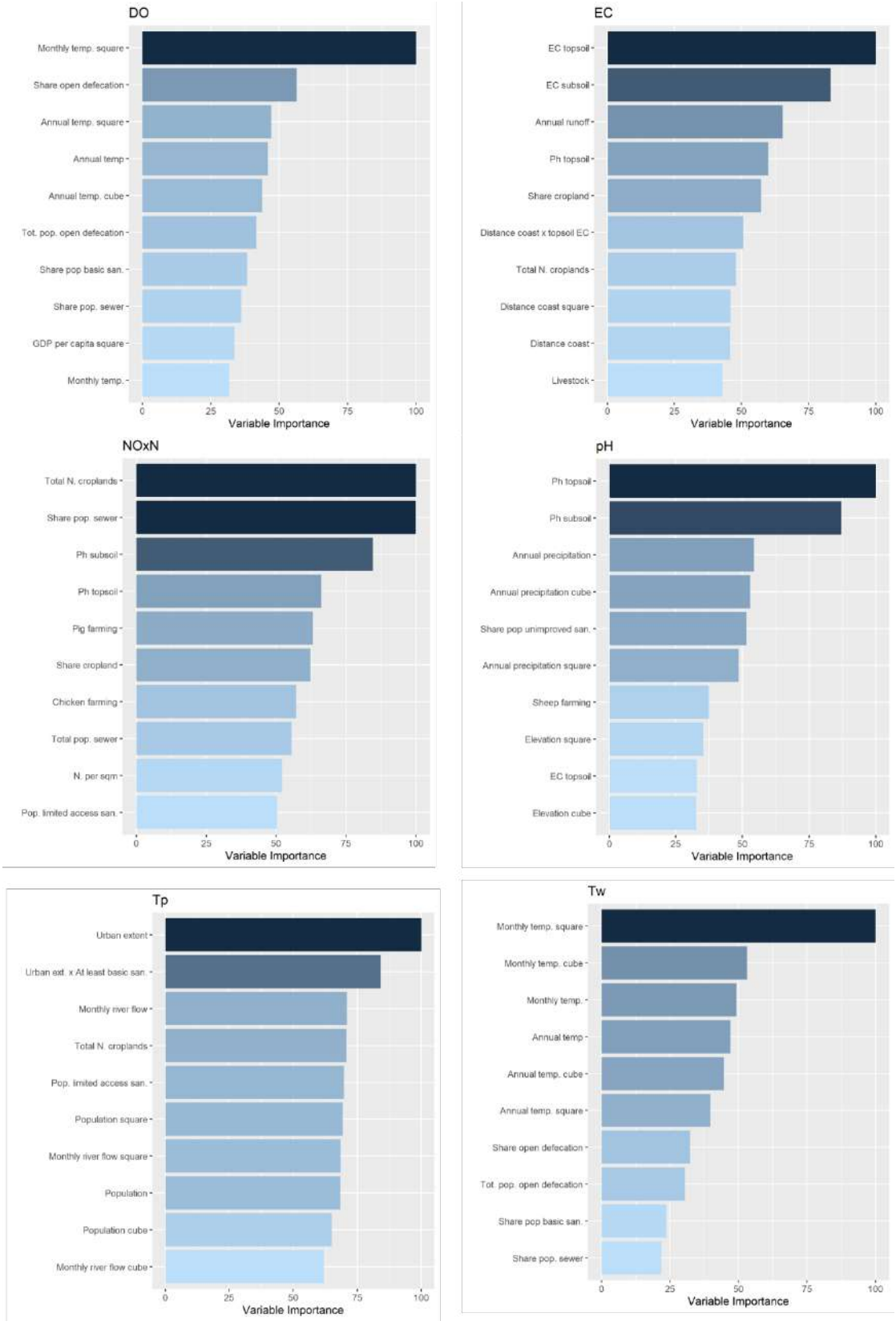
Global trends of water quality

Sup. Fig. 6: This figure provides the global monthly estimates from the model between 1992 and 2010. Mean, media worst 5%, first and tenth deciles are presented. Water quality presents important intra-annual variations mainly driven by seasons climate variations. Inter-annual variations were limited over 18 years. The legend worst and best does not apply to Tw for which values should be read as coldest. to hottest.



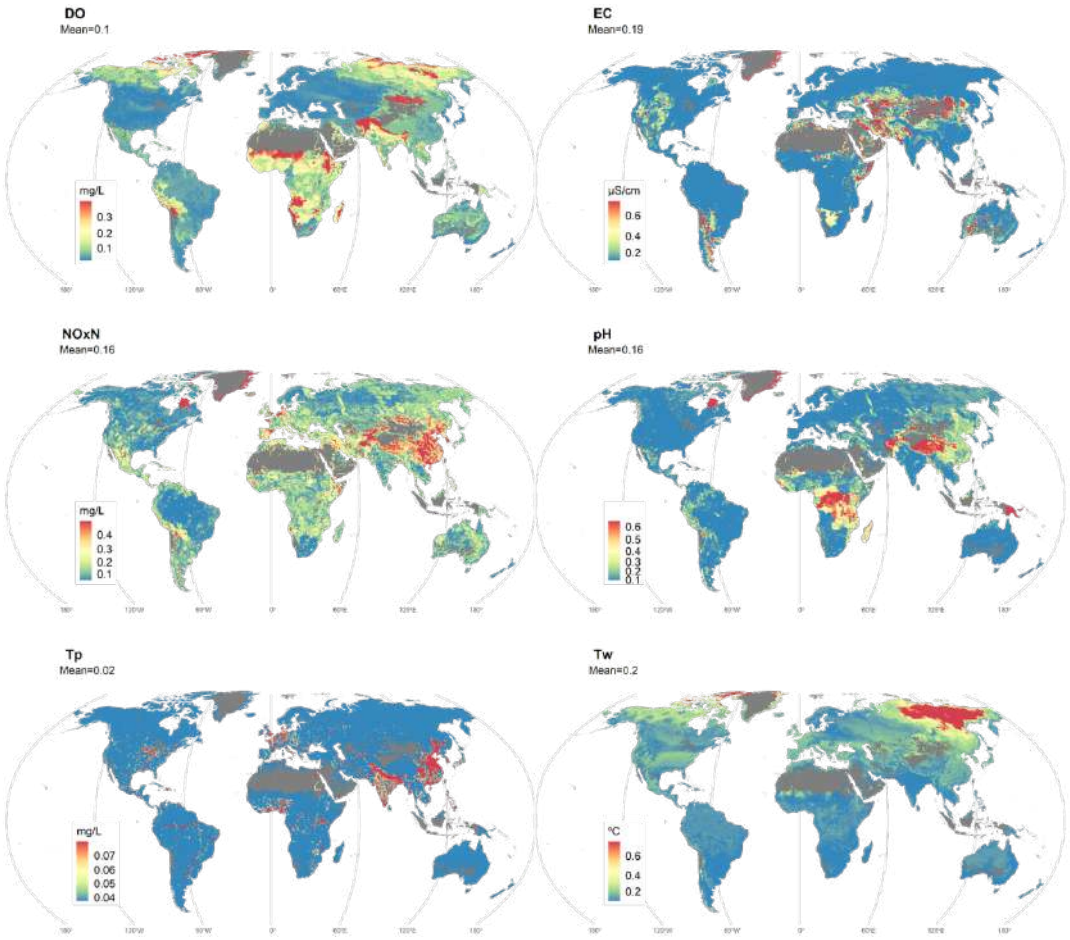
Variable Importance

Sup. Fig. 7: This figure displays for each water quality indicators the ten drivers that are the most important in explaining a given water quality indicator. Across the boards, sanitation and weather variables consistently explained water quality, at different degrees. Drivers of nitrate-nitrite varies significantly from the other indicators. Livestock farming and urbanization plays more for NO2NO3 than for other indicators.



Dissimilarity Index between training and prediction samples

Sup. Fig. 8: We follow Meyer and Pedesma (2021) to construct Area of Applicability (AOA) and Dissimilarity Index (DI). AOA is defined as the area, for which the cross-validation error of the model applies. It is based on DI, a metric based on the minimum distance to the training data in the predictor space. Cells are colored in blue when they fall within the AOA. Otherwise, a gradient color for green (small dissimilarity) to red (large dissimilarity) is used. The results indicate that for EC, Temp, TP and to a lesser extent pH, we can confidently extrapolate our results and predict water quality in continents like Africa despite the absence of input water quality data. This is because the predictors we use are covered in other continents. For DO and NOxN, the procedure indicates that uncertainties exist in some areas to predict water quality, including in Sub-Saharan Africa.



A synthesis index for SDG 6.3.2

Index Construction

We construct a water quality index that encompasses the water quality indicators tracked for SDG 6.3.2. The methodology for the construction of this index is inspired by well-established indices such as the Human Development Index and is in line with previous water quality indices. Each pollutant included in the index is provided with an equal weight since SDG does not give a priority or weighting preference to any water quality indicator. The steps for the construction of the index are as follow. First, DO and pH are transformed so that higher values of the transformed DO and pH mean worse water quality. For DO, DO_{bis} is such that $DO_{bis} = max(DO) - DO$. For pH, we create the variable $pH_bis = |7 - pH|$ to measures deviation from "pristine water". Second, each variable is scaled on a similar support to account for the fact that pollutants are initially measured different scale. Indeed, each indicator is originally distributed on a different support. EC ranges from 0 to 4000 µS/cm

while DO values in streams range between 0 and 18 mg/l. Therefore, scaling the indicators on a common support prevents one variable to influence more the index than the others. Third, we create the index variable WQI as described in equation 1:

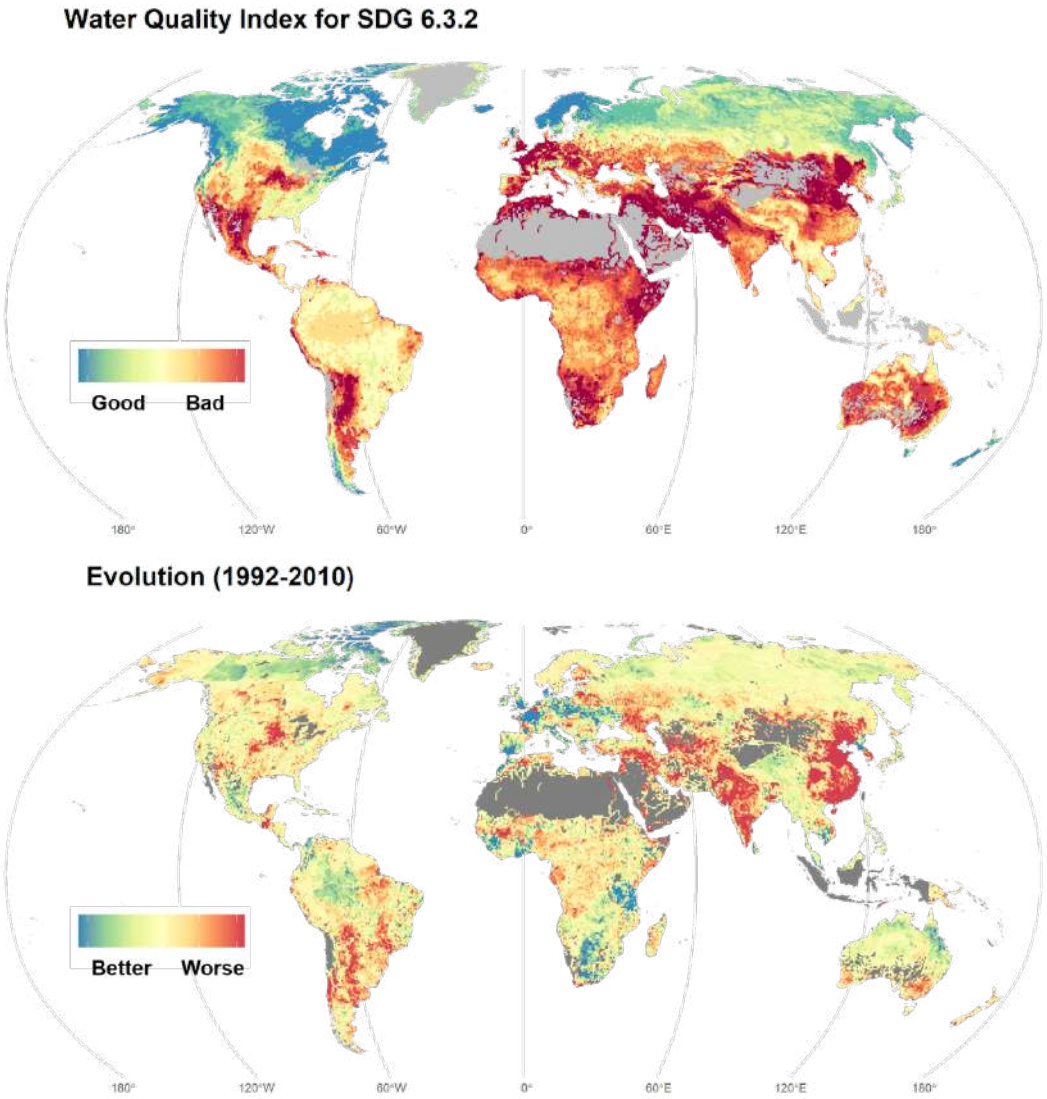
$$WQI_(SDG6.3.2) = DO_{bis} + EC + NOxN + pH_{bis} \quad (1)$$

The water quality index is by construction ordinal. It allows one to compare the values of water quality at different locations or its evolution across time. It is constructed such that higher values reflect poorer water quality.

Index

A global average of the index between 2000-2010 is shown in Sup Fig. 7. As for the other indicators, it is available monthly between 1992 and 2010. TP was not included in the main index because of the uncertain predictions provided by our model. Poor water quality hotspots occur in most populated places, importantly at all levels of development. The index shows that poor water quality status can be found in Africa, Asia, the Americas, and Europe, suggesting that income is not a key covariate associated with water quality as measured by this index. Between 1992 and 2010, water quality deteriorates the most in South Asia, China but also in high-income areas, such as the American Midwest.

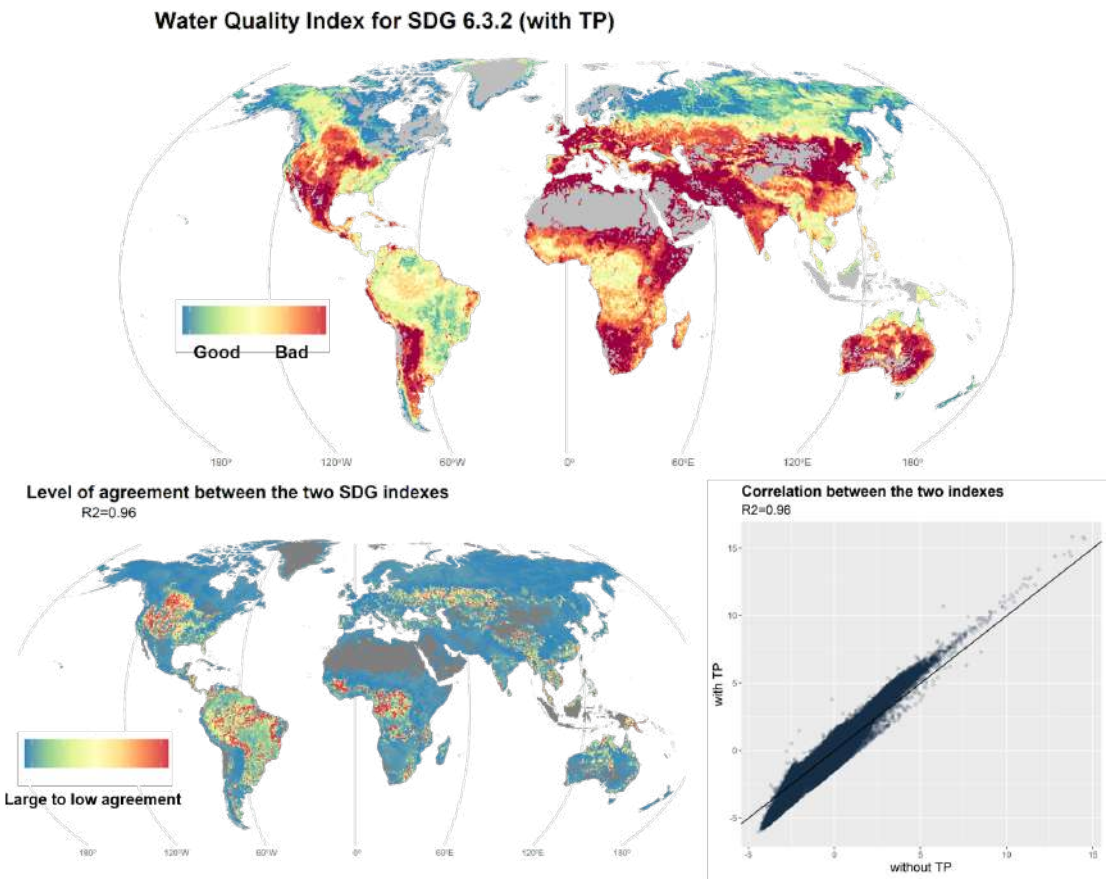
Sup. Fig. 9: Mapped water quality index for SDG 6.3.2. The top map represents the average value of the index for the period 2000-2010. The bottom map represents its change between 1992 and 2010. The index is the additional of the normalized values of each of four indicators: DO, EC, NOxN and pH. For DO, the scale was reversed so that higher value means lower water quality. For pH, the standardized difference between the observation and the ideal pH of 7 was calculated and incorporated in the index (see Methods). TP was not included in the main index because of the uncertainties associated to its prediction. Supplementary Figure 8 presents a map of the same index with TP and explore how both indexes correlate (R2 = 0.96). Finally, TW is not included as it is not directly relevant for SDG indicator 6.3.2.



4.1 Robustness of SDG 6.3.2. index to the inclusion of Total Phosphorus

We further checked that including TP does not change our conclusions. Both indexes (with and without TP) correlate well (R2 = 0.96). Dissimilarities however exist in areas predicted as hotspot for phosphorus (e.g., Central Africa but for which we had no observations to train the model).

Sup. Fig. 10: SDG indicator 6.3.2. including TP



CEE-M Working Papers[1] - 2022

WP 2022-01 Denis Claude & **Mabel tidball**
« Taking firms' margin targets seriously in a model of competition in supply functions »

WP 2022-02 Cauê D. Carrilhoa, **Gabriela Demarchi**, Amy E. Duchelle, Sven Wunder, & Marla Morsello
« Permanence of avoided deforestation in a Transamazon REDD+ initiative (Pará, Brazil)functions »

WP 2022-03 Francisco Cabo, Alain Jean-marie & **Mabel tidball**
« Positional effects in public good provision. Strategic interaction and inertia »

WP 2022-04 **Fabien Prieur**, Weihua Ruan & Benteng Zou
« Optimal lockdown and vaccination policies to contain the spread of a mutating infectious disease »

WP 2022-05 **Claudia Kelsall**, Martin F. Quaas & **Nicolas Quérou**
« Risk aversion in renewable resource harvesting »

WP 2022-06 **Raphaël Soubeyran, Nicolas Quérou** & Mamadou Gueye
« Social Preferences and the Distribution of Rewards »

WP 2022-07 **Antoine Pietri**
« Plus de voiture pour davantage de vélo ? Une étude sur la ville de Tours »

WP 2022-07 **Sebastien Desbureaux**, Frederic Mortier, Esha Zaveri, Michelle van Vliet, Jason Russ, Aude Sophie Rodella & Richard Damania
« Mapping global hotspots and trends of water quality (1992-2010): a data driven approach »

[1] **CEE-M Working Papers / Contact** : laurent.garnier@inrae.fr
- RePEc https://ideas.repec.org/s/hal/wpceem.html
- HAL https://halshs.archives-ouvertes.fr/CEE-M-WP/