Universiteit
Leiden
The Netherlands

Assessing global regionalized impacts of eutrophication on freshwater fish biodiversity

Zhou, J.

Citation
Zhou, J. (2024, January 30). *Assessing global regionalized impacts of eutrophication on freshwater fish biodiversity*. Retrieved from https://hdl.handle.net/1887/3715136

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3715136

Note: To cite this publication please use the final published version (if applicable).

Chapter 3

A comparison between global nutrient retention models for freshwater systems

This chapter has been published as Zhou, J., Scherer, L., van Bodegom, P.M., Beusen, A.H.W., Mogollón, J.M., 2022. A Comparison Between Global Nutrient Retention Models for Freshwater Systems. *Frontiers in water* 4:894604.


Arthur Bensen

Abstract

Against the backdrop of increasing agricultural production, population, and freshwater/coastal eutrophication, studies are aiming to understand the behavior of nitrogen (N) and phosphorus (P) in the global freshwater system. Global nutrient models are typically used to quantify the nutrient amount and content in freshwater systems across different river orders and catchments. Such models typically use empirically derived nutrient retention equations for predicting nutrient fate, and these equations may be derived using data from a specific region or environment or for a specific context. Here we used IMAGE-GNM, a spatially explicit nutrient model at a half-degree resolution, to examine the performance of several well-known empirical equations by comparing the respective model outcomes with observed data on a global scale. The results show that 1) globally, the empirical retention equations work better for predicting N fate than P fate; 2) hydraulic drivers are the most important factor affecting the residual of total N and P concentrations, compared with the functional forms and the coefficients in the empirical equations. This study can aid in assessing the variability and accuracy of various retention equations from regional to global scales, and thus further strengthen our understanding of global eutrophication.

3.1 Introduction

During the 20th century, the global cycles of nitrogen (N) and phosphorus (P) have shown a rapid acceleration due to increasing nitrogen fixation and phosphate mining (Jenny et al., 2016). Over the 20th century, humans have almost doubled the global N and P delivery to freshwater systems from 34 to 64 Tg N yr^{-1}, and 5 to 9 Tg P yr^{-1}, respectively (Beusen et al., 2016). Due to a combination of N and P excessive nutrient loading, the global freshwater and coastal system has seen a major increase in eutrophication. Eutrophication can lead to the proliferation of algae blooms and hypoxia (Chislock et al., 2013; Müller et al., 2012), which consequently threatens the balance of environmental and ecological systems (Jenny et al., 2016; Vonlanthen et al., 2012; Schindler and Vallentyne, 2008). Toward the future, the rising trend of nutrient

accumulation in freshwater systems is set to continue due to the increase of fertilizer application and global population growth (Mogollón et al., 2018a). Moreover, warmer climates can lead to an acceleration of the hydrological cycle, which signifies both increasing evaporation and freshwater advection, and thus likely to exacerbate change in global nutrient cycles (Bouraoui et al., 2004; Statham, 2012). Thus, while global in-stream nutrient retention tends to vary slightly and stay stable under various future scenarios, N export to oceans is set to increase by up to 20% under future scenarios, unless human strictly takes sustainable practices in nutrient application and water use (i.e., Shared Socio-economic Pathway SSP1) (Beusen et al., 2022).

To better curb the increasing trend of eutrophication over the global aquatic system, the first step is to assess the fate of N and P, which requires regional to global nutrient models.

Despite the various modeling efforts, global estimates of nutrient exports are still highly variable. For instance, the estimated total phosphorus (TP) export of NEWS-2 (9 Tg yr^{-1}) is almost double the export of IMAGE-GNM (4 Tg yr^{-1}, Harrison et al. 2019), and total nitrogen (TN) of NEWS-2 (45 Tg yr^{-1}) is also higher than that of IMAGE-GNM (37 Tg yr^{-1}, van Vliet et al. 2019). van Vliet et al. (2019) and Grizzetti et al. (2015) reckoned that this issue results from the discrepancy in hydrological input data, spatial resolution, and the method used to calculate retention. Retention indicates the difference between nutrient input and output within a river segment or a lake. N retention includes the removal processes of denitrification, sedimentation, and uptake by aquatic vegetation (Saunders and Kalff, 2001), while P retention is affected by entrainment, sedimentation, sorption, as well as by uptake by plants and organisms (Reddy et al., 1999). Historically, retention is modeled through empirical equations based on regression analyses of localized nutrient input-output data (Behrendt and Opitz, 1999; Kelly et al., 1987). These regression analyses are based on localized studies (Kirchner and Dillon, 1975; Seitzinger et al., 2002). So far, current studies have never compared the performance of the various retention models globally**.** Identifying the best-performing retention models for global nutrient models can contribute to the

future knowledge of eutrophic impacts (e.g., nutrient loading/export to aquatic systems) (Jeppesen et al., 2009).

Kelly et al. (1987) proposed a mass balance model for N denitrification loss, and Howarth et al. (1997) employed this model to estimate N retention. Later on, Behrendt and Opitz (1999) found that this model can also be applied to P. They investigated 100 European rivers and developed a regression between retention and different hydraulic drivers, including hydraulic load and specific runoff. De Klein (2008) discovered large monthly variability in retention and the necessity to distinguish among drivers for N and P (e.g., P is highly related to temperature while N is not) after studying 13 catchments in the Netherlands and Germany. Furthermore, in contrast to N, P is susceptible to water body types due to its susceptibility to sedimentation and sorption (Reddy et al., 1999). Thus, the estimation of P retention should be based on different drivers for lakes vs. rivers. By analyzing 15 lakes in Canada, Kirchner and Dillon (1975) posited that the major driver of P retention was the areal water load (as opposed to the hydraulic load, the areal water load is related to specific runoff, Eq. 3.4), whereas Chapra (1975) argued that P retention could be better represented by apparent settling velocity in these lakes. Brett and Benjamin (2008) examined 305 input/output data of lakes and reservoirs in the USA and Canada and concluded that the main driver of lake P retention is residence time. In these studies, retention is dominated by hydrological drivers, i.e., hydraulic load and specific runoff. These drivers can only be converted from one into the other if the information of additional variables (i.e., water volume and depth) is provided. Such information is highly uncertain, which could potentially lead to biased estimates and increased uncertainties. Investigating this key feature was at the core of our study.

The aim of this study is to identify the best-performing retention model or set of regional retention models to assess the fate of global nutrients in freshwater systems. We adapted IMAGE-GNM to include a comprehensive set of retention equations. The retention models were examined by comparing the respective model outcomes with observed data. The model performance was also analyzed for different geographical

zones (“Geographical zone”, 2009), including the North Frigid Zone, the North Temperate Zone, the Torrid Zone, and the South Temperate Zone to discover the response of nutrient retention to hydrological conditions. The set of best-performing retention models can be applied to improve the accuracy of global nutrient models, which helps to better understand the global states of water quality.

3.2 Methods

3.2.1 Global nutrient model

In this study, we choose to use IMAGE-GNM (Beusen et al., 2015) as it is the best-fit nutrient model for our study among the most widely recognized nutrient models reviewed in MIPs (van Vliet et al., 2019). Of these, MARINA is a downscaled application of NEWS-2 to China and has not been employed for worldwide modeling (Strokal et al., 2016). HYPE has been used to estimate global hydrology (Arheimer et al., 2020), while for nutrients, this model was only developed at the regional scale, such as Europe (Strömbäck et al., 2019). Similarly, SPARROW was localized to the USA (McCrackin et al., 2013) and New Zealand (Alexander et al., 2002). Globally, NEWS-2 is differentiated at the watershed scale, while the resolution of IMAGE-GNM is gridded (0.5×0.5˚). Thus, compared to NEWS-2, IMAGE-GNM captures the inner-basin information, which is unneglectable since the geographical variation of nutrients within large watersheds is highly uneven. This spatial delineation allows validating nutrient data since the measurement stations are scattered over the watersheds and cannot reflect the estimation of the whole watershed.

IMAGE-GNM is a dynamic distributed model that depicts nutrient flow and delivery processes in soils, freshwater systems, and export to coasts. A detailed description and the code (written in Python 2.7) of IMAGE-GNM can be found in Beusen et al. (2015). In this study, different retention equations were implemented into IMAGE-GNM. The simulated concentration of TN/TP in rivers and TP in lakes were compared with respective observed data. We distinguished between lakes and rivers when validating P estimates to account for the strong links between P fate and lake ecology (Brett and

Benjamin, 2008). For N, we deemed this distinction unnecessary since N retention can be entirely represented by the water discharge difference between these water bodies (Saunders and Kalff, 2001). Due to a lack of TN observations in global lakes, the performance of simulated TN in lakes was not assessed in this study.

3.2.2 Retention models

Load-weighted nutrient water body retention (R_L, dimensionless) indicates the proportion of retention load ($R_{N,P}$, kg yr^{-1}) to the load of nutrients transported to the freshwater system ($L_{N,P}$, kg yr^{-1}).

$$R_{L_{N,P}} = \frac{R_{N,P}}{L_{N,P}} = \frac{L_{N,P} - O_{N,P}}{L_{N,P}} \quad (3.1)$$

where $O_{N,P}$ (kg yr^{-1}) denotes the export of nutrients at the outlet of the water body.

In our study, we only included strictly empirical models of mostly pure hydrological nature. Some empirical models also account for ecological nature, namely hydro-ecological retention models, whereas in this study the only two models that may be considered hydro-ecological models are the model of Wollheim et al. (2006) (section 3.2.2.1.1 (1)) and De Klein (2008) (for P, section 3.2.2.1.2 (2)). In this study, however, we only focused on the hydrological part and represented ecological impacts by temperature factors. The hydrological drivers in retention models are represented by the empirical function of hydraulic drivers, including hydraulic load (Eq. 3.2) and specific runoff (Eq. 3.3). We elaborate on these functions raised in literature in sections 3.2.2.1-3.2.2.2 and summarize all models in Table 3.1.

$$H_L = \frac{D}{t_r} \quad (3.2)$$

$$q = \frac{Q}{A} \quad (3.3)$$

where H_L (m yr^{-1}) is the hydraulic load represented by quotient between the depth (D, m) and residence time (t_r, yr) of the water body; q (L km^{-2} s^{-1}) is the catchment area-specific runoff, which equals the discharge (Q, L s^{-1}) divided by catchment area (A,

km^2).

The specific runoff can also be expressed as areal water load W_L(m yr^{-1}, Eq. 3.4), which denotes the annual value of the water column height per water surface area in the unit of specific flow:

$$W_L = \frac{q \times 8.64 \times 0.365}{W} \tag{3.4}$$

where q (L km^{-2} s^{-1}) is the specific runoff introduced in Eq. 3.3, W (%) is a ratio of the surface water area to the watershed area, and 8.64 × 0.365 is a coefficient to convert the unit from L km^{-2} s^{-1} to m yr^{-1}.

3.2.2.1 Riverine retention models for TN/TP

3.2.2.1.1 Hydraulic-load-driven models

(1) Wollheim et al. (2006, 2008)

Current IMAGE-GNM employs Wollheim et al. (2006, 2008)'s equation as the retention model. Here, the retention $R_{L_{N,P}}$ is defined as a first-order degradation process (Eq. 3.5).

$$R_{L_{N,P}} = 1 - exp\left(-\frac{v_f}{H_L}\right) \tag{3.5}$$

where v_f (m yr^{-1}) indicates the net uptake velocity expressing the biochemical features of a nutrient. v_f for P (Eq. 3.6) takes a basic value of 44.5 m yr^{-1} (Behrendt and Opitz, 1999) and is modified by the temperature factor $f(T)$ (Eq. 3.8):

$$v_{f_P} = 44.5 \cdot f(T) \tag{3.6}$$

For N, v_f (Eq. 3.7) is initialized to 35 m yr^{-1} (Wollheim et al. 2006, 2008) and modified by the temperature factor and concentration factor $f(C_N)$, which is proposed by Beusen et al. (2015):

$$v_{f_N} = 35 \cdot f(T) \cdot f(C_N) \tag{3.7}$$

where $f(C_N)$ represents the effect of concentration on denitrification resulting from electron donor limitation if excessive N is transported into the water (Mulholland et al.,

2008). $f(C_N)$ was calculated as an approximation of a hyperbolic function which contains the following points:7.2 at C_N = 0.0001 mg L^{-1} and 1 at a turning point C_N = 1 mg L^{-1}, and continues to decline mildly to 0.37 at C_N = 100 mg L^{-1} and keep constant for a higher concentration (Marcé and Armengol, 2009).

$$f(T) = \alpha^{T-20} \tag{3.8}$$

where α is 1.06 for P (Marcé and Armengol, 2009) and 1.0717 for N (Mulholland et al., 2008); T is average annual temperature (˚C).

(2) Kelly et al. (1987)

Kelly et al. (1987) proposed a simple mass balance model for the N denitrification losses in lakes and Howarth et al. (1997) used this mass transfer model to estimate the N retention of rivers. Behrendt and Opitz (1999) found this equation can be used to estimate phosphorus retention. Their studies have shown that this function form can be applied to both river systems and lakes.

$$R_{L_{N,P}} = \frac{S_{N,P}}{S_{N,P} + H_L} \tag{3.9}$$

where $S_{N,P}$ is an average mass transfer coefficient given in m yr^{-1}. Behrendt and Opitz (1999) estimated the mass transfer coefficient S_N for nitrogen (N) as 11.9 and S_P for phosphorus (P) as 16.1.

(3) Seitzinger et al. (2002)

Seitzinger et al. (2002) combined N observations from 10 rivers and 23 lakes in the USA. This study provided the equation of N retention as Eq. 3.10 and proved it applies to rivers, lakes, and reservoirs:

$$R_{L_N} = 88.45 \cdot {H_L}^{-0.3677} \tag{3.10}$$

3.2.2.1.2 Specific-runoff-driven models

(1) Behrendt and Opitz (1999)

Behrendt and Opitz (1999) investigated Dissolved Inorganic N (DIN) measurements

and provided two correlation equations for nutrients. While IMAGE-GNM calculates TN, DIN is the major component of TN. We, therefore, included these two equations in our research. Note that they defined "emission" as the inflow flux of nutrients to the aquatic system, which is equivalent to "load ($L_{N,P}$)" in IMAGE-GNM, while the term "load" used in their study indicated the nutrient exported at the outlet of the river, which equals the "output ($O_{N,P}$)" defined in IMAGE-GNM. Therefore, it necessitates a conversion from the output-weighted retention $R_{O_{N,P}}$ to load-weighted $R_{L_{N,P}}$ (Eq. 3.11).

$$R_{L_{N,P}} = \frac{R_{O_{N,P}}}{1 + R_{O_{N,P}}} \tag{3.11}$$

The first statistical equation is expressed by a power function of areal water load W_L:

$$R_{O_{N,P}} = a \times {W_L}^b \tag{3.12}$$

where *a* and *b* are statistical coefficients. For N, *a* equals 5.9 and *b* equals -0.75; for P, *a* and *b* are 13.3 and -0.93, respectively.

The second retention equation, in which the driving force is the catchment area-specific runoff q, can be expressed as:

$$R_{O_{N,P}} = c \times q^d \tag{3.13}$$

where *c* and *d* are statistical coefficients. For N, *c* is 6.9 and *d* is -1.10; for P, *c* and *d* are 26.6 and -1.71, respectively.

Behrendt and Opitz (1999) (W_L) and Behrendt and Opitz (1999) (q) were used to identify the retention equations driven by areal water load W_L and the catchment area-specific runoff q, respectively, in the following sections.

(2) De Klein (2008)

De Klein (2008) studied monthly TN retention for catchments whose areas ranged from 20.8 km^2 to 486 km^2. The results of this study showed that load-weighted nitrogen retention R_L is inversely related to surface water area-specific runoff (SR, m^3 ha^{-1} s^{-1}). The *SR* can be expressed as a ratio of specific runoff to the surface water area.

De Klein (2008) gave a retention equation based on the monthly time step. It was then aggregated to an annual scale by summing the monthly inputs and the estimation of

monthly exports. De Klein (2008) argued that the difference between monthly retention and annual retention of N was negligible, whereas, for P, the status remains uncertain. However, we assume that the equation still works for P at an annual time step.

Herein, the retention equation of N can be expressed as:

$$R_{L_N} = 0.0246(SR)^{-0.57} = 0.0246\left(\frac{e \cdot q}{W}\right)^{-0.57} \quad (3.14)$$

where e is a unit conversion coefficient of 10^7, W (%) is the percentage of surface water area to watershed area (including land area and water area).

Besides SR, P retention is also determined by temperature:

$$R_{L_P} = 0.253(SR)^{-0.20} \times 1.01^{(T_i-22)} \quad (3.15)$$

where T_i is the average water temperature (ºC).

(3) Venohr et al. (2005)

Venohr et al. (2005) provided another group of statistical coefficients for TN retention based on the same dataset as Behrendt and Opitz (1999). Venohr et al. (2005) distinguished water bodies by assigning different coefficients for lakes, rivers, and reservoirs (Eq. 3.16):

$$R_{L_P} = \frac{f \times {W_L}^g}{1 + f \times {W_L}^g} \quad (3.16)$$

where *f* and *g* are statistical coefficients. *F* is 1.9 and *g* is -0.49 for rivers; *f* is 7.279 and *g* is -1 for lakes and reservoirs.

3.2.2.2 Lake retention models for P

(1) Kirchner and Dillon (1975)

By analyzing nutrient budget information from 15 Canadian lakes, Kirchner and Dillon (1975) developed an empirical equation for the retention of phosphorus in lakes:

$$R_{L_P} = 0.426\exp(-0.271W_L) + 0.574\exp(-0.00949W_L) \quad (3.17)$$

(2) Chapra (1975)

In contrast to Kirchner and Dillon (1975), Chapra (1975) argued that the retention of P can be more precisely related to both the areal water load W_L and the settling velocity

of P-contained particles (υ), assuming the lake is at a steady state:

$$R_{L_P} = \frac{\upsilon}{W_L + \upsilon} \tag{3.18}$$

where υ (m yr^{-1}) is the apparent settling velocity of TP, which was estimated as 16 m yr^{-1}.

(3) Brett and Benjamin (2008)

Brett and Benjamin (2008) conducted a statistical reassessment of total phosphorus (TP) input/output data to determine which hydraulic driver is most strongly associated with lake phosphorus concentration and retention. They provided the best-fit equation as Eq. 3.19:

$$R_{L_P} = 1 - \frac{1}{1 + 1.12{t_r}^{0.53}} \tag{3.19}$$

where t_r (yr) denotes the water residence time of lakes and reservoirs.

3.2.3 Sample data for validation

Water quality sample data, including TN and TP concentrations, were obtained from the Global Freshwater Quality Database (GEMStats, UNEP GEMS/Water Programme (2007)), Global River Chemistry Database (GLORICH, Hartmann et al. (2019)), and United States Geological Survey (USGS, Aulenbach et al. (2007)). We downloaded the datasets on September 17, 2021. The sample data from literature covers the main rivers of Africa and Asia, including the Nile River (El-Sadek, 2011; Sinada and Yousif, 2013), the Pearl River (Liu et al., 2009), the Yangtze River (Liu et al., 2018; Maotianet al., 2014; Sun et al., 2013a; Sun et al., 2013b), and the Yellow River (Chen et al., 2004; Tao et al., 2010). We used a DIN/TN ratio of 50% to transform dissolved inorganic nitrogen (DIN) into TN for the Yangtze River (Liu et al., 2018; Yan et al., 2001) and took a DIN/TN (the same as NO_3/TN, since nitrite NO_2 occupies less than 1% of DIN and the

Table 3.1 Summary of retention models proposed by previous studies

Approach	Driving force	Applicability	Nutrient	Original scale	Function form
Wollheim et al. (2006)	H_L, v_f (C_i, T_i)	River and lake	N, P	Global	$R_{L_{N,P}} = 1 - exp\left(-\frac{v_f}{H_L}\right)$
Kelly et al. (1987)	H_L	River and lake	N, P	North America and Norway	$R_{L_{N,P}} = \frac{S_{N,P}}{S_{N,P} + H_L}$
Seitzinger et al. (2002)	H_L	River and lake	N	Northeastern U.S.A.	$R_{L_N} = 88.45(H_L)^{-0.3677}$
Behrendt and Opitz (1999) (W_L)	W_L	River and lake	N, P	Europe	$R_{L_{N,P}} = \frac{a \times {W_L}^b}{1 + a \times {W_L}^b}$
Behrendt and Opitz (1999) (q)	q	River and lake	N, P	Europe	$R_{L_{N,P}} = \frac{c \times q^d}{1 + c \times q^d}$
	q	River and lake	N	The Netherlands	$R_{L_N} = 0.0246\left(\frac{e \cdot q}{W}\right)^{-0.57}$
De Klein (2008)	q, T_i	River and lake	P	The Netherlands	$R_{L_P} = 0.253\left(\frac{e \cdot q}{W}\right)^{-0.20} \times 1.01^{(T_i - 22)}$

Venohr et al. (2005)	W_L	River and lake	N	Europe	$R_{L_P} = \frac{f \times {W_L}^g}{1 + f \times {W_L}^g}$
Kirchner and Dillon (1975)	W_L	Lake	P	Canada	$R_{L_P} = 0.426\exp(-0.271W_L) + 0.574\exp(-0.00949W_L)$
Chapra (1975)	W_L	Lake	P	Canada	$R_{L_P} = \frac{\upsilon}{W_L + \upsilon}$
Brett and Benjamin (2008)	t_r	Lake	P	North America	$R_{L_P} = 1 - \frac{1}{1 + 1.12{t_r}^{0.53}}$

Note: Driving forces are site-related variables to be determined by the observed or simulated data, whereas the non-driving-force parameters in the retention equation are constant coefficients provided by literature. Definitions of the variables as the driving force of retention: H_L (m yr^{-1}) is hydraulic load; W_L (m yr^{-1}) is areal water load; q (L km^{-2} s^{-1}) is specific runoff; C_i (mg L^{-1}) is the nutrient concentration; T_i (ºC) is average annual temperature; t_r denotes the water residence time for lakes and reservoirs; υ (m yr^{-1}) is the apparent settling velocity of total phosphorus.

ammonium concentration is low in rivers) ratio of 77% for the Nile River, the Yellow River, and the Pearl River (Turner et al., 2003). For computing TP, we used a ratio of 62% to transfer PO_4 into TP data (Turner et al., 2003).

We selected the data reported in the year 2000 since it is the last representative (most recent) year of IMAGE-GNM (Beusen et al., 2015). The samples include 9770 items of TN data from 1199 river stations, 19701 items of TP data from 2261 river stations, and 141 items of TP data from 23 stations of 7 lakes. The depth and residence time of lakes were derived from the World Lake Database (Herschy, 2012) except for Ashkui at narrows in Seal Lake and Wuchusk Lake, which lack measured data. For these two lakes, we applied the prediction of PCR-GLOBWB, the global hydrological model running on a grid cell level that has been integrated into IMAGE-GNM. Note that in the validation of lake retention equations, including Kirchner and Dillon (1975), Chapra (1975), and Brett and Benjamin (2008), we apply Wollheim et al. (2006)'s equation to calculate river retention in the cells that contain no lakes or reservoirs.

For validation, the cells with an invalid hydrological parameter (i.e., zero discharge and zero volume) were removed. To avoid errors raised by inadequate spatial data representation, basins with fewer than 10 grid cells were also excluded (Beusen et al., 2015). Consequently, 82% of the river sample items from 1157 river stations for TN and 91% of the river sample items from 2185 river stations for TP were included in the analysis (Figure 3.1). The validation was conducted based on a 0.5×0.5˚ grid-cell scale based on the resolution of predicted results of IMAGE-GNM. When stations were located within the same cell, the average of the samples was taken as observed data.

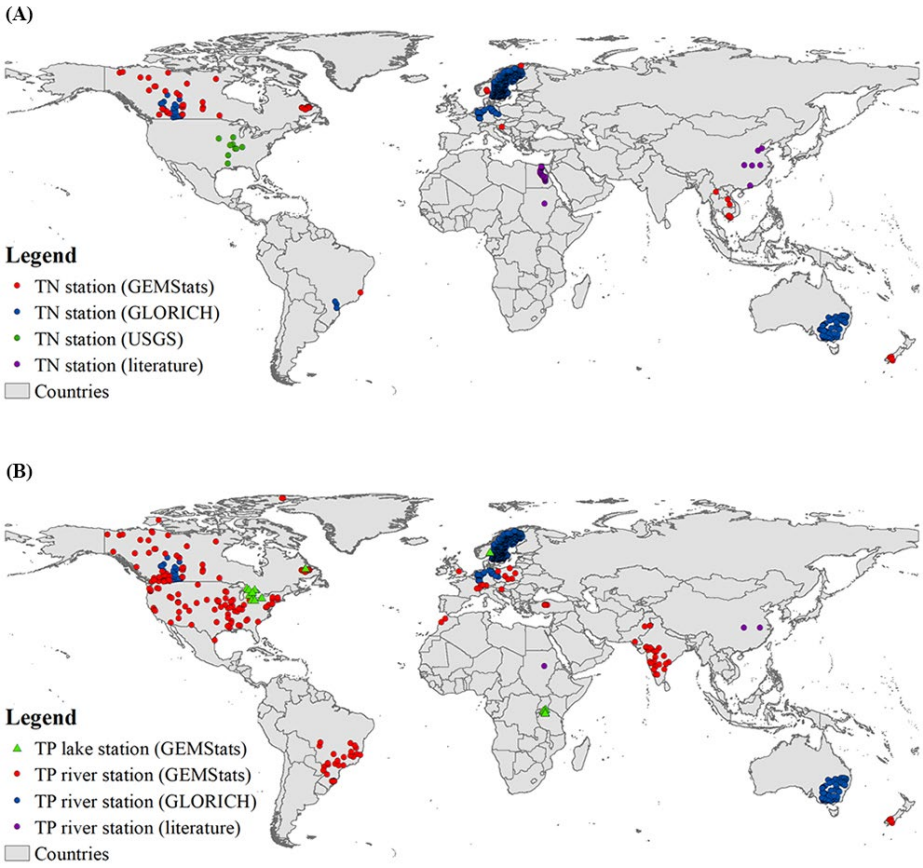

Figure 3.1 Sampling stations of (A) TN and (B) TP concentration over the globe. For N, all the stations are river stations. Note that (1) due to a lack of TN observations in global lakes, lake TN was not assessed in this study; (2) for TP, GEMStats has included USGS data. (3) in total, there are 1157 TN river stations including 63, 823, 261, and 10 stations distributed in North Frigid Zone, North Temperate Zone, South Temperate Zone, and Torrid Zone, respectively; for TP, the respective geographical zone contains 68, 1535, 493, and 89 river stations (2185 TP river stations globally).

3.2.4 Assessment of the model performance

We used the interquartile range ($IQR = Q_3$-Q_1) to describe the dispersion and employed

Yule's coefficient (Yule's coefficient = $\frac{(Q_3+Q_1-2Q_2)}{Q_3-Q_1}$) to depict the skewness of simulated retention through non-parametric coefficient (Yule and Kendall, 1968); where Q_1, Q_2, Q_3 denote the 25th percentile, 50th percentile, and 75th percentile respectively.

We employed the mean-Normalized Root Mean Square Error (NRMSE) to evaluate the error between predicted and observed nutrient concentrations of each retention model.

$$NRMSE = \frac{1}{\bar{O}}\sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \tag{3.20}$$

where $\bar{O}$ is the average of observations; n is the number of pairs of predicted-observed data; O_i and P_i are the observed value and predicted value of the *i*th cell, respectively.

The retention model that has a minimal NRMSE generates the lowest discrepancy between predicted values and observed values. NRMSE is a widely used criterion for the validation of nutrient concentrations (e.g., Beusen et al. 2015; Liu et al. 2018). However, NRMSE is quite sensitive to extremes, in particular to extremely high values. The Pearson correlation coefficient (r) is complementary to it and assesses the dynamic behavior of the model rather than the bias.

We used a logarithmic transformation to linearize the pairwise data and use r to evaluate the correlation between predictions and observations. Meanwhile, taking r of logarithmic data into account also lessens the likelihood risk of misjudging the performance of right-skewed residuals.

$$r = \frac{\sum_{i=1}^{n}\left(\log O_i - \overline{\log O_i}\right)\left(\log P_i - \overline{\log P_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(\log O_i - \overline{\log O_i}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\log P_i - \overline{\log P_i}\right)^2}} \tag{3.21}$$

Ideally, NRMSE is close to zero (on a range from 0 to unlimited) and r close to 1 (on a range from -1 to 1).

3.2.5 Significance of difference

We applied one-way Analysis of variance (ANOVA) to evaluate the significance of differences in performance among retention models. Here, as a measure of performance, the difference in simulated and observed concentration in a sampled grid cell was taken. The mean difference (i.e. whether a model consistently over- or underestimated retention and corresponding concentration) was evaluated.

To verify normality, the distribution of residuals of each model was judged based on probability plots. Then, we examined the homoscedasticity with the Brown–Forsythe test (Brown and Forsythe, 1974) due to its robustness and its maintenance of good statistical power (Derrick et al., 2018). TP showed heteroscedasticity and was analyzed with Welch's ANOVA instead. To evaluate the differences in retention between specific pairs, we conducted a pairwise comparison using Tukey's honestly significant difference (HSD) for homoscedastic data and a Games-Howell post hoc test for heteroscedastic data between pairs of samples.

The analysis was accomplished using Python 3.7. Details of packages/versions/functions are listed in Supporting Information Table S3.1.

3.3 Results

3.3.1 Validation

The plots of riverine simulation against observations show that the empirical equations perform better for TN than for TP (Figure 3.2 and Figure 3.3). Furthermore, the NRMSE of TN outcomes ranges from 1.62 to 2.31, which is much smaller than the NRMSE of TP whose interval is between 4.97 and 13.84. The Pearson's r of TN is higher than that of TP (Table 3.2 and Table 3.3).

The retention models of Behrendt and Opitz (1999) (q) generated the lowest NRMSE and a satisfactory Pearson's r (>0.5) for both N and P, being the best option for estimating riverine retention of TN/TP.

Among TN retention models, with the exception of Behrendt and Opitz (1999) (q) and

Seitzinger et al. (2002), the models' NRMSEs are higher than 2. The largest NRMSE (2.31) was generated by the retention model of Kelly et al. (1987) despite having the largest r value of 0.71. Behrendt and Opitz (1999) (q)'s r is 0.62, which shows an acceptable correlation between the simulated and observed concentrations. Hence, the retention model of Behrendt and Opitz (1999) (q) performs best for TN according to our analyses and validation dataset. Compared with Wollheim et al. (2006), which is the currently used retention equation in IMAGE-GNM, Behrendt and Opitz (1999) (q) can reduce the NRMSE by 41% for estimating riverine TN concentration globally.

The retention model of Behrendt and Opitz (1999) (q) also simulated the lowest NRMSE (4.97) for P retention, followed by that of De Klein (2008) (6.40), whose Pearson's r is the lowest (0.26). Excepting the retention model of De Klein (2008), the difference in Pearson's r among the models is quite minor, ranging from 0.42 to 0.54. However, aside from Behrendt and Opitz (q) and De Klein (2008), the NRMSEs of the models exceed 10. The best-performing model, Behrendt and Opitz (1999) (q), can reduce the NRMSE of Wollheim et al. (2006) by 107%.

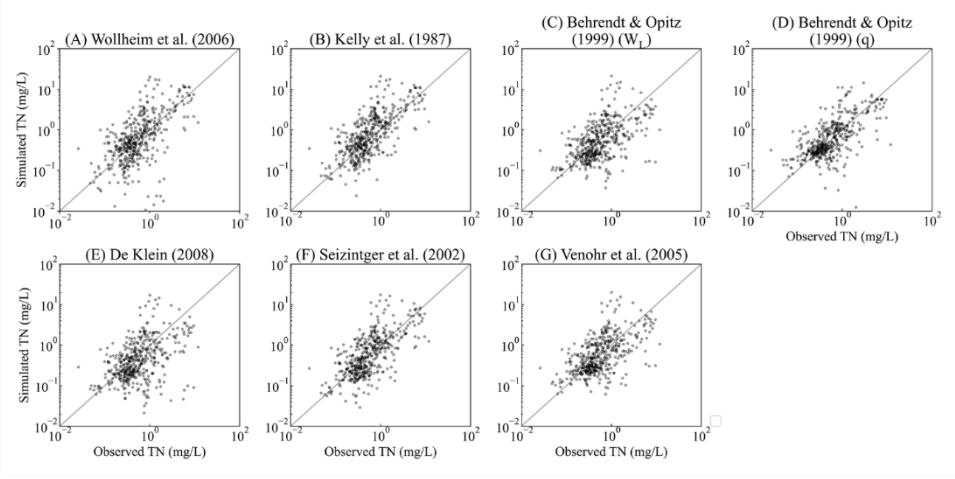

Figure 3.2 Validation of predicted values against observations of annual average concentration for riverine N (each dot represents the predicted values against average observed N concentration of the measurement stations within the same cell). The sample size is 449, the number of grid cells covered by measurement stations.

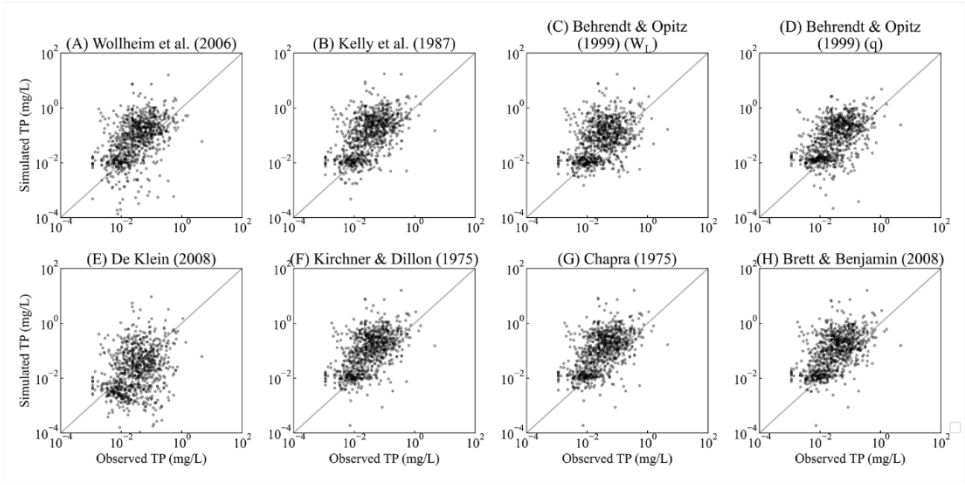

Figure 3.3 Validation of predicted values against observations of annual average concentration for riverine P (each dot represents the predicted values against average observed P concentration of the measurement stations within the same cell). The sample size is 849, the number of grid cells covered by measurement stations.

The comparison between simulated and observed TP concentrations in lakes is shown in Figure 3.4. Since the measurements vary considerably across the locations of stations within a lake, we plotted measurements as boxplots to show the variation. In Mjøsa and Wuchusk Lake, the simulations of all the models are higher than the observed TP concentration, while in other lakes, simulated TP is closer to the observations. De Klein (2008)'s residuals (i.e. the difference between simulated and average observed concentration in a lake) in Mjøsa and Wuchusk Lake are the smallest among empirical equations. Besides, De Klein (2008)'s simulations of other lakes do not deviate from the observed measurement intervals, yielding the best performing empirical equation. The NRMSE and Pearson's r of De Klein (2008) are 1.09 and 0.77 (Table 3.3). De Klein (2008) has the second-lowest NRMSE following Kelly et al. (1987) (0.89), but Kelly et al. (1987)'s r shows the second-worst performance (0.59). Behrendt and Opitz (1999) (q) has the highest Pearson's of 0.92 as well as the highest NRMSE (8.18). NRMSE and r of Wollheim et al. (2006) are 1.81 and -0.47, respectively, both of which perform worse than Kelly et al. (1987), Behrendt and Opitz (1999) (W_L), De Klein (2008), and

Brett and Benjamin (2008). Replacing the retention equation of De Klein (2008) with Wollheim et al. (2006) in IMAGE-GNM can reduce the NRMSE in lakes by 66%.

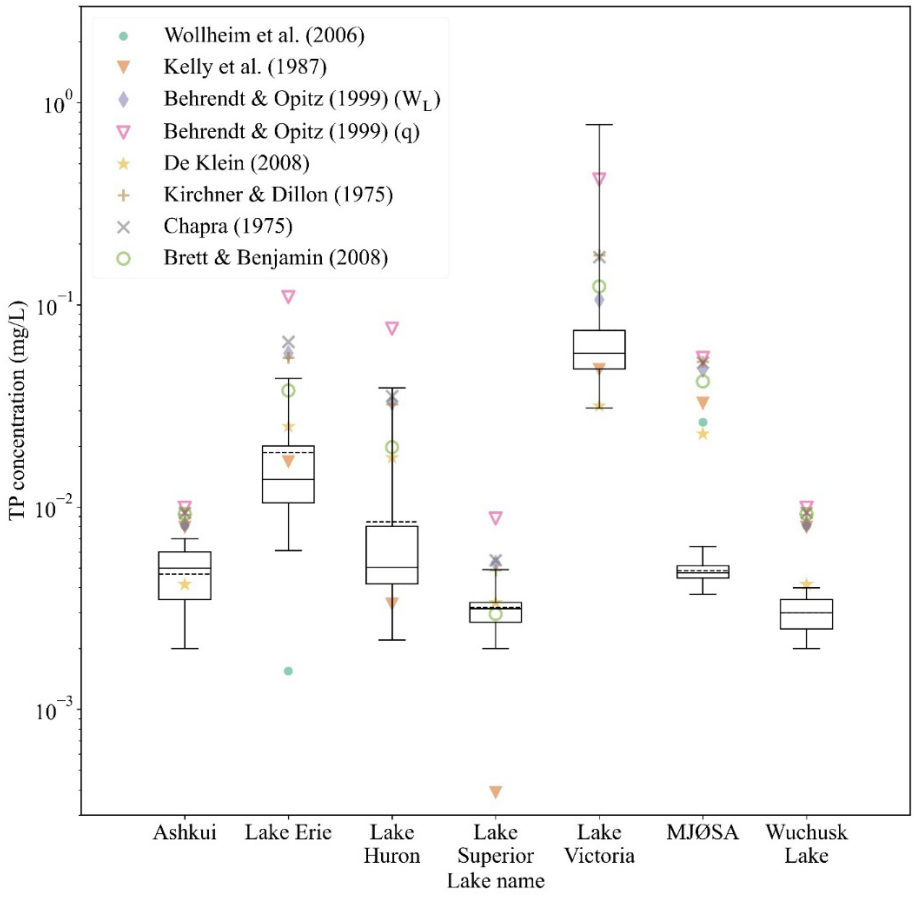

Figure 3.4 Simulated TP concentration of lakes compared with observed values. The boxplot shows the range of observed TP concentrations of each lake: the dark solid lines in the boxes are the median value of observations; the dark dash lines are the average of observations; the upper and lower borders of the boxes indicate 75th percentile and 25th percentile of observations; the whiskers denote upper and lower extremes of observations. IMAGE-GNM simulates lake concentration at the outlet cell of the lake, providing one value (average concentration) for each lake. Note that Wollheim et al.

(2006)'s simulation of Lake Huron, Lake Superior, and Lake Victoria is not shown in the log-scaled figure, as it predicted an extremely low concentration (< 0.0001 mg/L) for these lakes.

The performance of empirical equations differs by geographical zone (Table 3.2 and Table 3.3). For the N retention models, Behrendt and Opitz (1999) (q) obtained the lowest NRMSE in the North Frigid Zone, North Temperate Zone, and South Temperate Zone, which can lower the NRMSE by Wollheim et al. (2006) by 63%, 45%, and 32% in the respective regions. In the Torrid Zone, Venohr et al. (2005) performed the best, as it reduces the NRMSE by 88% compared with the currently used retention equation. For P, Behrendt and Opitz (1999) (q) is the best-performing retention model in North Temperate Zone and Torrid zone, reducing the NRMSE of Wollheim et al. (2006) by 132% and 146%, whereas the riverine retention model Wollheim et al. (2006) combined with the lake retention models of Kirchner and Dillon (1975) or Chapra (1975) provides the best fit of retention in the North Frigid Zone. Wollheim et al. (2006) is also recommended in the South Temperate Zone, as it has both the second-lowest NRMSE and the second-highest r. The best retention models of different geographical zones are presented in bold in Table 3.2 and Table 3.3 and listed in Supporting Information Table S3.2.

Table 3.2 Assessment of the performance of N retention models for rivers. The values of the best-performing models are shown in bold on a global or regional (geographical zone) scale. Note that only river samples were included due to a lack of lake sample data.

Region and Observation Type	Criteria	Wollheim et al. (2006)	Kelly et al. (1987)	Behrendt and Opitz (1999) (W_L)	Behrendt and Opitz (1999) (q)	De Klein (2008)	Seitzinger et al. (2002)	Venohr et al. (2005)
Global	NRMSE	2.29	2.31	2.12	**1.62**	2.02	1.93	2.04
	r	0.58	0.71	0.55	**0.62**	0.45	0.68	0.59
North Frigid Zone	NRMSE	0.57	0.51	0.42	**0.35**	0.49	0.48	0.43
	r	0.14	0.18	0.32	**0.25**	0.24	0.09	0.18
North Temperate Zone	NRMSE	2.35	2.33	2.10	**1.62**	1.99	1.85	1.97
	r	0.59	0.73	0.65	**0.68**	0.57	0.71	0.66
Torrid Zone	NRMSE	1.71	2.15	1.13	2.18	**0.96**	0.92	0.91

	r	0.05	-0.22	0.20	-0.31	**0.16**	-0.12	0.15
South	NRMSE	1.91	2.06	2.02	**1.45**	1.97	2.03	2.10
Temperate Zone	r	0.46	0.55	0.09	**0.40**	-0.02	0.49	0.19

Table 3.3 Assessment of the performance of P retention models for rivers and lakes. The values of the best-performing models are shown in bold on a global or regional (geographical zone) scale.

Region and Observation Type	Criteria	Wollheim et al. (2006)	Kelly et al. (1987)	Behrendt and Opitz (1999) (W_L)	Behrendt and Opitz (1999) (q)	De Klein (2008)	Kirchner and Dillon (1975)	Chapra (1975)	Brett and Benjamin (2008)
Lakes	NRMSE	1.81	0.89	1.59	8.18	**1.09**	2.70	2.73	1.47
Global	r	-0.47	0.59	0.83	0.92	**0.77**	0.87	0.87	0.84
Rivers	NRMSE	10.29	13.84	10.96	**4.97**	6.40	10.60	10.91	10.61

Global	r	0.42	0.54	0.42	**0.52**	0.26	0.54	0.54	0.54
North Frigid Zone	NRMSE	2.36	2.37	2.49	2.51	2.75	**2.33**	**2.33**	2.37
	r	0.32	0.28	0.05	-0.03	-0.67	**0.35**	**0.35**	0.23
North Temperate Zone	NRMSE	10.94	12.41	8.27	**4.72**	5.40	11.24	11.60	11.25
	r	0.39	0.56	0.48	**0.55**	0.33	0.56	0.55	0.55
Torrid Zone	NRMSE	11.40	27.35	26.48	**4.64**	14.22	11.52	11.54	11.50
	r	0.22	0.24	0.15	**0.27**	0.09	0.22	0.23	0.22
South Temperate Zone	NRMSE	**4.90**	7.99	6.96	5.67	3.03	5.56	5.81	5.57
	r	**0.32**	0.24	-0.05	0.31	-0.08	0.32	0.32	0.33

3.3.2 Retention model comparison

Figure 3.5 (TN) and Figure 3.6 (TP) show that different retention models generate similar hotspot distributions. The hydraulic-load-driven models (i.e., retention models of Kelly et al. (1987), Wollheim et al. (2006), and Seitzinger et al. (2002)) predicted relatively lower retention than specific-runoff-driven models (i.e., retention models of Behrendt and Opitz (1999), De Klein (2008), and Venohr et al. (2005)).

Despite different hydraulic driving forces among retention models, the hotspots of all the models are located in arid zones, South Africa, West Argentina, Mississippi River Basin, and Colorado River Basin. However, low retention values are quite distinct. For N, retention values under or equal to 0.1 cover over 50% of the global area in hydraulic-load-driven models (i.e., the retention models of Kelly et al. (1987), Wollheim et al. (2006), and Seitzinger et al. (2002)). In contrast, in specific-runoff driven models (i.e., the retention models of Behrendt and Opitz (1999), De Klein (2008), and Venohr et al. (2005)), low retention (≤0.1) occurs in only 24% to 30% of the global area. For P, regions with retention under or equal to 0.1 calculated by hydraulic-load-driven models (i.e., retention models of Kelly et al. (1987) and Wollheim et al. (2006)) occupy 58% and 66% of the global area, respectively. In contrast, low retention values (≤0.1) in specific-runoff driven models (i.e., the retention models of Behrendt and Opitz (1999)(WL), Behrendt and Opitz (1999)(q), and De Klein (2008)), occur in <36%. In particular De Klein (2008)'s model only generated 5% low-value retention globally.

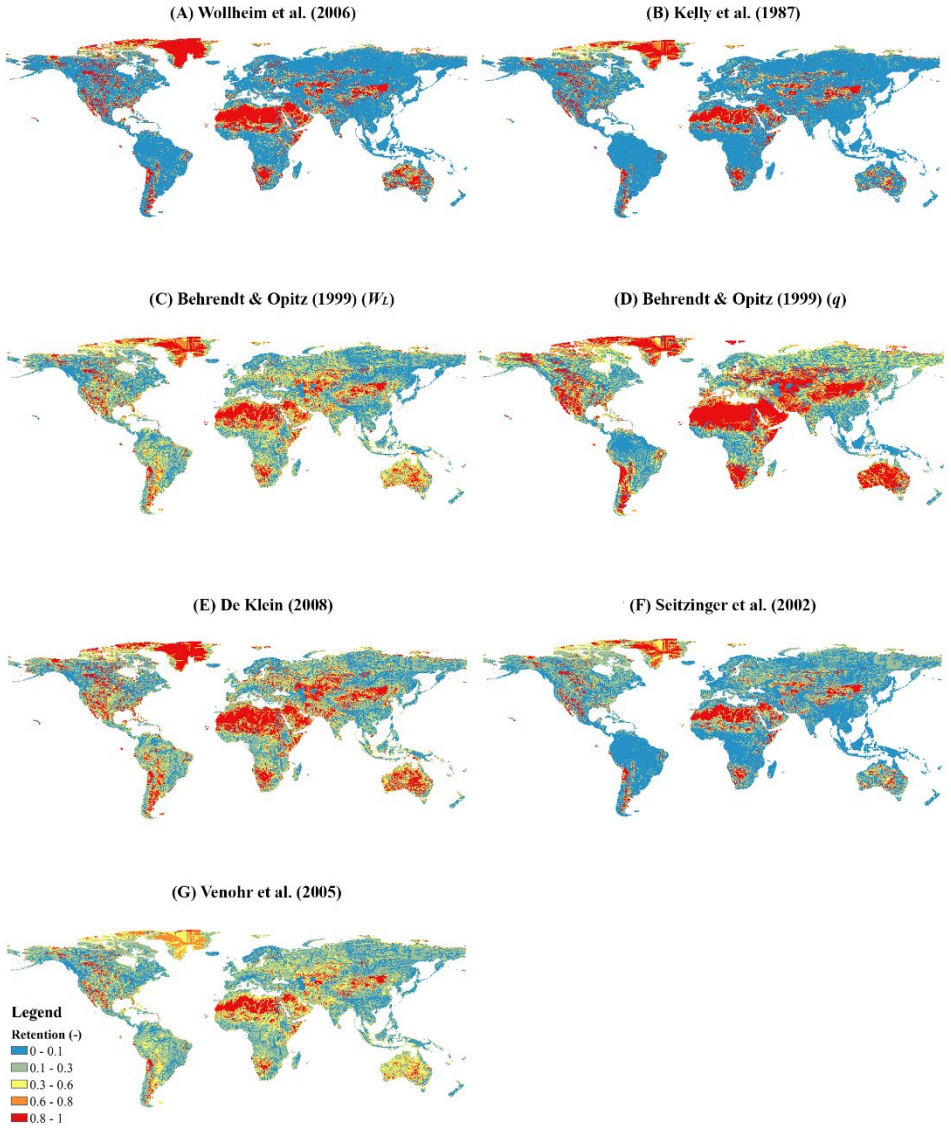

Figure 3.5 N retention for different models at a half-degree resolution (Retention is dimensionless, and the unit was labeled as "-")

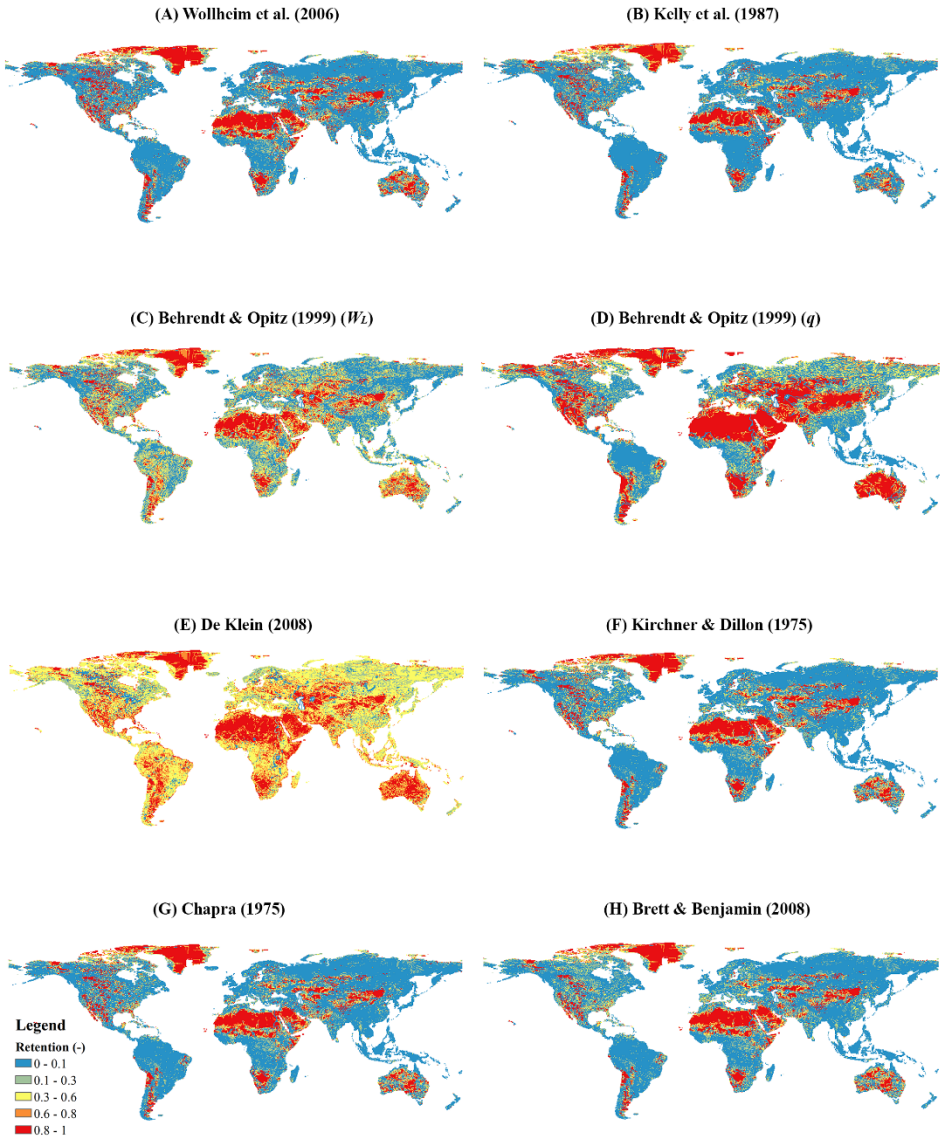

Figure 3.6 P retention maps for different models at a half-degree resolution (Retention is dimensionless, and the unit was labeled as "-")

The dispersion (represented by IQR) and skewness (represented by Yule's coefficient) between N and P retention showed only minor differences except for De Klein (2008)

(Table 3.4 and Table 3.5), which predicted a much smaller dispersion and skewness for P when compared with N.

For both N and P, the retention from hydraulic-load-driven models displayed larger skewness than specific-runoff-driven models. Yule's coefficient of retention predicted by hydraulic-load-driven models is larger than 0.5, while Yule's coefficients of specific-runoff-driven models range from 0 to 0.5. Thus, the retention simulated by models with a driving force of hydraulic load is more asymmetrically distributed than that of runoff-driven models. Nevertheless, the retention from all the models is positively skewed.

For N, the retention model of Behrendt and Opitz (1999) (q) predicted the largest average retention globally, followed by the simulation of De Klein (2008) and Behrendt and Opitz (1999) (W_L), while those models with a driving force of hydraulic load predicted relatively smaller average retention. The IQR of the simulation following Behrendt and Opitz (1999) (q) is the highest, revealing that the model simulates more dispersed retention than other models.

For P, the retention model of De Klein (2008) predicted the largest average retention globally, with the second and third largest average retention modeled by Behrendt and Opitz (1999) (q) and Behrendt and Opitz (1999) (W_L), respectively. In contrast, the retention models of Wollheim et al. (2006) and Kelly et al. (1987) with a hydraulic load driver simulated smaller average retention. Lake retention models including Kirchner and Dillon (1975), Chapra (1975), and Brett and Benjamin (2008) cause little impact on global riverine retention. Thus, the prediction of these models is close to that of Wollheim et al. (2006) on a global scale. Larger difference in IQR was found between different specific-runoff-driven models, as IQRs of modeled retention following Behrendt and Opitz (1999) (q) and De Klein (2008) are 0.893 and 0.382 respectively, while IQRs of hydraulic-load-driven models range from 0.234 to 0.398.

Table 3.4 Descriptive statistics of N retention of different retention models

Retention models		Wollheim et al. (2006)	Kelly et al. (1987)	Behrendt and Opitz (1999) (W_L)*	Behrendt and Opitz (1999) (q)*	De Klein (2008)	Seitzinger et al. (2002)	Venohr et al. (2005)
Average		0.273	0.203	0.328	0.430	0.386	0.228	0.285
	5%	0	0	0	0	0	0	0
	25%	0.024	0.011	0.081	0.081	0.099	0.068	0.102
Quartiles	50%	0.061	0.031	0.222	0.335	0.243	0.101	0.203
	75%	0.413	0.201	0.505	0.804	0.639	0.214	0.378
	95%	1.0	0.998	0.991	1.0	1.0	1.0	0.936
Dispersion (IQR)		0.389	0.190	0.424	0.723	0.540	0.146	0.276
Skewness (Yule's coefficient)		0.814	0.790	0.332	0.297	0.466	0.560	0.265

* W_L and q were used to identify different retention equations of Behrendt and Opitz (1999) as driven by areal water load W_L and the catchment area-specific runoff q, respectively.

Table 3.5 Descriptive statistics of P retention of different retention models

Retention models		Wollheim et al. (2006)	Kelly et al. (1987)	Behrendt and Opitz (1999) (W_L)	Behrendt and Opitz (1999) (q)	De Klein (2008)	Kirchner and Dillon (1975)*	Chapra (1975)*	Brett and Benjamin (2008)*
Average		0.278	0.219	0.354	0.437	0.553	0.263	0.263	0.257
	5%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	25%	0.028	0.015	0.067	0.029	0.358	0.027	0.027	0.029
Quartiles	50%	0.068	0.042	0.237	0.311	0.517	0.065	0.065	0.069
	75%	0.426	0.249	0.600	0.922	0.740	0.384	0.358	0.327
	95%	1.0	0.999	0.998	1.0	1.0	1.0	1.0	1.0
Dispersion (IQR)		0.398	0.234	0.533	0.893	0.382	0.357	0.331	0.298

Skewness (Yule's coefficient)	0.802	0.774	0.363	0.370	0.166	0.787	0.774	0.730

* Kirchner and Dillon (1975), Chapra (1975), and Brett and Benjamin (2008) are lake retention models; for those cells without lake cells, Wollheim et al. (2006)'s equation is used to calculate river retention.

3.3.3 Difference score performance of retention models

Both TN and TP showed significant differences in their mean subtraction between simulated and observed concentration among the retention models. For TN, Tukey's HSD showed a clear distinction between hydraulic load-driven models on the one hand and specific-runoff-driven models on the other hand (Figure 3.7 (A), Supporting Information Table S3.3). The Games-Howell post hoc tests showed similar differences in model groups for TP (Figure 3.7 (B), Supporting Information Table S3.4). Particularly, the retention models of De Klein (2008) deviated strongly in performance, which may relate to the difference of their coefficients and the consideration of temperature in De Klein (2008).

Generally, the average difference between observed and simulated concentration is lower for specific-runoff-driven models than for hydraulic-load-driven models. Note that concentration is inversely proportional to the estimation of retention. A positive average difference between simulated and observed concentrations signifies an overestimation of concentration and thus an underestimation of retention. For both TN and TP, retention models, except for the TN equations of De Klein (2008), tended to underestimate retention, particularly in low-retention regions (retention ≤ 0.1).

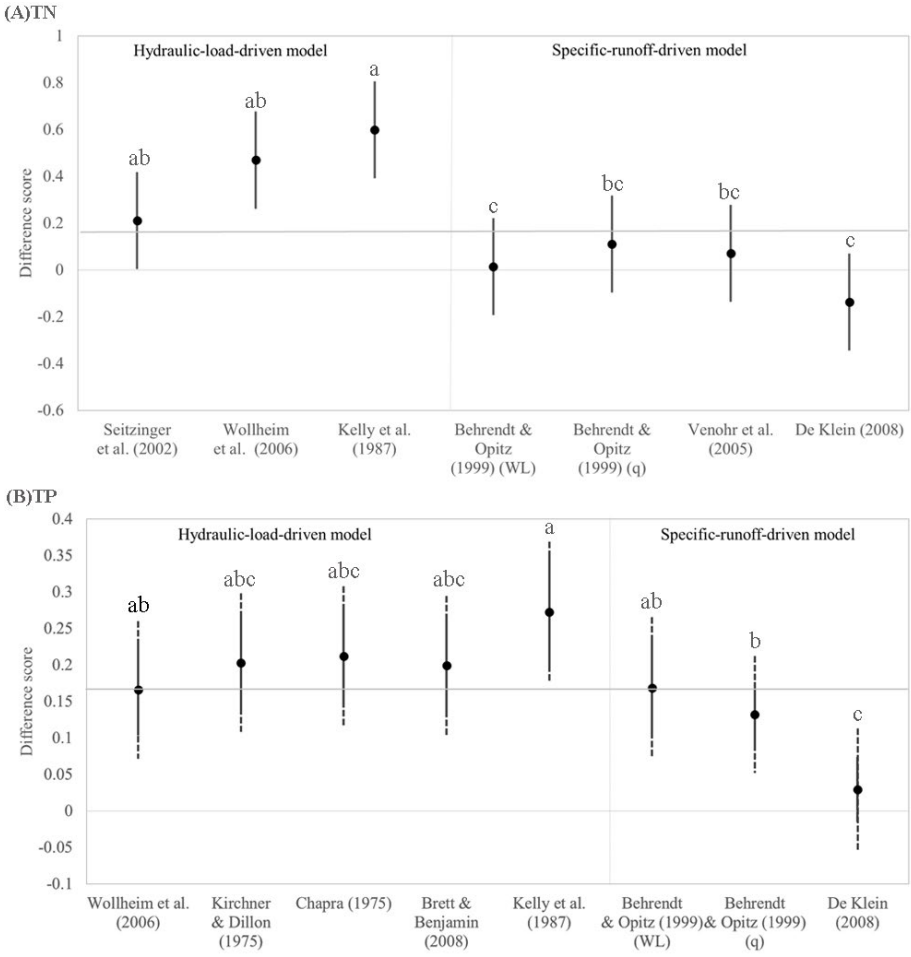

Figure 3.7 Difference score between simulated and observed concentration for (a) TN and (b) TP among different retention models. Black points are the average difference. The length of the "wings" (shown as black lines around the average) equals $SE \cdot q_{\alpha,k,\nu}$ with critical value $q_{\alpha,k,\nu}$ and standard error SE determined by Tukey's HSD or Games-Howell post hoc test. The solid lines show the minimal $SE_{i,j} \cdot q_{\alpha,k,\nu}$ and the dashed lines of the wings indicate the range of pairwise $SE_{i,j} \cdot q_{\alpha,k,\nu}$ from the minimum to the maximum. Different letters a, b, c on top of the wings identify significant differences (p-value < 0.05) in concentration among the retention models.

3.4 Discussion

As far as we are aware, this is the first study to assess the performance of empirical retention equations for global nutrient models and to investigate the role of driving forces, function form, and equation coefficients. The strengths of this study include its analyses on the global and regional performance of retention equations using multiple criteria (NRMSE, Pearson's r, and relative bias (i.e., average difference score shown in Figure 3.7) and the comparison of model residuals for different water bodies based on abundant samples from diverse sources.

We applied NRMSE and Pearson's r as the performance criteria and used an ANOVA and a post hoc test to investigate the performance of (and differences between) retention models. The results revealed that the impact of function form and coefficients are inferior to the hydraulic driver. Since most of these models were developed based on a local dataset (Table 3.1), the coefficients and function forms were expected to represent the corresponding local systems better than the globe. However, our results show that some of these local studies perform better globally than those developed at global scales (i.e., Wollheim et al., 2008). Particularly, empirical retention equations whose driving force is hydraulic load predicted relatively lower retention than specific-runoff driven models. Hydraulic-load-driven models tended to underestimate retention and overestimate concentration, particularly for TN. The specific-runoff driven equations of Behrendt and Opitz (1999) (q) and De Klein (2008) provided the best fit for the simulation of riverine nutrient retention and P retention in lakes, respectively.

Our study reinforced the importance of temperature as a secondary driving force of P retention (D'angelo et al., 1991; Jensen and Andersen, 1992; Kim et al., 2003), since the retention models of De Klein (2008) lowered the difference between simulations and observations and is the only model that considered temperature. Our results were also in line with the discovery that riverine N and P retention depends on the specific runoff rather than hydraulic load (Behrendt and Opitz, 1999) and predicted P values disperse more than predicted N values using empirical equations to estimate retention (Hejzlar et al., 2009).

Using the best combination of retention models for geographical zones (Table S3.2), we simulated the global export to coastal waters of N and P are 30.5 Tg N yr^{-1} and 5.8 Tg P yr^{-1}. For the global N export, our estimation is lower than those of NEWS-2 (45 Tg N yr^{-1}, Mayorga et al. 2010) and IMAGE-GNM (37 Tg N yr^{-1}, Beusen et al., 2016). The combination of retention models for various zones can better represent the realistic retention and results in a lower global export that is closer to observations. For P, our estimation falls between the global export of NEWS-2 (9 Tg P yr^{-1}, Mayorga et al., 2010) and IMAGE-GNM (4 Tg P yr^{-1}, Beusen et al., 2016). Moreover, the best combination of P retention models avoids the bias caused by Wollheim et al. (2006) to predict zero P loads in the high-retention regions.

Our assessment of lake P retention differs from the multiple comparison results of Brett and Benjamin (2008), who compared and optimized the retention equation for lake TP in the USA and Canada and regarded Brett and Benjamin (2008)'s equation, a residence-time-driven equation, as the best fit. We identify the reason as the difference of performance criteria and spatial coverage of sample data. Brett and Benjamin (2008) used the logarithm coefficient of determination r^2 as the performance criterion, which is equivalent to the square of Pearson's r of linearized log-transformed data. Indeed, the model of Brett and Benjamin (2008) got high r scores among all the models in our research, but their model performs worse than De Klein (2008) and Kelly et al. (1987) if we consider NRMSE. In conclusion, our use of multiple criteria shows the advantage of providing more information of both correlation and errors between simulations and observations.

3.4.1 Uncertainties of retention modeling

Uncertainty may arise from a lack of data availability and data representativeness. For instance, when assessing model performance in different geographical zones, retention models perform worse in the Torrid Zone than at the global level, which might be due to a misrepresentation of the nutrient states throughout the Torrid Zone (it covers only 1.5% of all TN samples and 3.4% of TP samples). In the South Temperate Zone, despite

a sufficient amount of data, the data lack representativeness, as most of the samples were collected in the Murray Darling Basin in Australia. We included NO3/DIN and PO4 data and used nutrient ratios to deal with a lack of data availability. However, the imposed nutrient ratio may introduce uncertainty into observation data as well. For instance, Turner et al. (2003), Meybeck (1982), and Goolsby et al. (1999) estimated global NO_3/TN ratios to vary from 59% to 86% and PO_4/TP ratios from 46% to 70% by investigating world's rivers. However, other literature (e.g., Liu et al. 2018; Yan et al. 2001) provided specific ratios for different rivers. To lower the uncertainty raised by these ratios, we used specific ratios firstly, and if no specific ratios were found, we employed the recommended global ratio from Turner et al. (2003). As more data become available, these retention models can be further evaluated and improved.

The ability of the model to reproduce the hydrological conditions is also crucial for the performance of modeled retention. For instance, although the Torrid Zone and the North Frigid Zone had almost the same amount of data, the performance of these two regions was quite different. Better retention predictions in the North Frigid Zone are related to more accurate PCR-GLOBWB discharge simulations in Europe, North America, and monsoon-dominated regions due to more precise meteorological forcing. In contrast, the least accurate results in the Torrid Zone are probably linked to the unsatisfactory simulation of discharge in African rivers since PCR-GLOBWB likely overestimates the groundwater recession rates and underestimates African inland delta evaporation (Sutanudjaja et al., 2018). In addition, due to faster rates of hydrological change in humid tropics, the hydrological condition is harder to describe precisely by a yearly-step model (Wohl et al., 2012).

On the other hand, the processing of water storage in PCR-GLOBWB introduced more uncertainties into the estimation of the hydraulic load than of specific runoff that was only affected by the discharge. Assuming reservoirs serve hydropower generation, PCR-GLOBWB overestimates the real reservoir volume by maximizing storage capacity under full power generation due to a lack of data from power plants on a global scale (Haddeland et al., 2006; Adam and Lettenmaier, 2008). However, PCR-

GLOBWB sometimes underestimates the total water volume by ignoring small reservoirs when combining multiple water bodies located within the same cell (Beusen et al., 2015). These uncertainties may explain why retention estimates from hydraulic-load-driven retention equations deviate more from observations than when based on specific-runoff-driven equations.

3.4.2 The effect of driving forces on P and N retention

The reason that specific-runoff-driven models perform better than hydraulic-load-driven models lies mainly in accuracy of the predictions on their driving force. IMAGE-GNM can better predict specific runoff that is composed of discharge and area since discharge was validated with observation in PCR-GLOBWB and area was obtained from geo-information (Van Beek et al., 2011). In contrast, hydraulic load works worse due to the uncertainties of reproducing water volume and water body depth.

Temperature has been shown to be an important driving force of P retention (D'angelo et al., 1991) to compensate for the difference between predicted and observed concentration but works secondary to hydraulic drivers, as Figure 3.7 shows those retention models considering temperature factor (i.e., Wollheim et al. (2006) and De Klein (2008)) lower the difference between predictions and observations within the models with the same hydraulic drivers. The effect of temperature works via influencing PO_4 release from sediments in streams and lakes (Fillos and Swanson, 1975; Holdren and Armstrong, 1980; Jensen and Andersen, 1992; Kim et al., 2003) and the physical properties of the water (Jeppesen et al., 2009). In contrast, N retention may also be affected by temperature, given NH_4 release from sediments (Shinohara et al., 2021), but the temperature effect on N is less substantial than P, since the N content ratio between sediments and other mediums (e.g., water) was found to be much lower than P (Downing and McCauley, 1992).

Future scenarios point to a global temperature increase due to greenhouse gas emissions (IPCC, 2018). Under a warmer climate, higher water temperature increases the time windows of biological activities and intensifies the interaction of the physical

environment and the biogeochemical properties in the hydrosphere (Jeppesen et al., 2009; Withers and Jarvie, 2008). This would likely lead to more nutrient release from aerobic sediments and an increase in nutrient concentrations in freshwaters.

3.4.3 Limitations and future improvement

River damming causes a decrease in the specific runoff and the hydraulic load, which leads to sediment trapping and an increase in nutrient retention (Maavara et al., 2015). While empirical equations capture the effects of changing hydrological parameters, they do not include biogeochemical mechanisms. These limitations act on both N and P. With respect to biogeochemical mechanisms, limitations relate to the lack of accounting for interactions among nutrient species, interactions with other elements, and for instance remobilization of P into water bodies due to the long-term accumulation of anthropogenic P retention in sediments. The errors between modeled and observed riverine P are larger than for N in our study. The larger error of P may result from the complexity of P transformations between unneglectable particle forms and dissolvable species, and the complex exchange between the water column and the sediment, which statistical regression equations of TP cannot reproduce or predict.

As such, model developers should search for ways to incorporate mechanistic geochemical dynamics into modeling nutrient retention in aquatic systems, so that models can better estimate N/P fate by distinguishing the specific forms and by including the transformations among different nutrient species. For instance, Vilmin et al. (2020) proposed a framework to describe the interactive processes between nutrient species and examined the model performance of N fate by splitting TN into ammonium (NH_4^+), nitrate (NO_3^-), nitrite (NO_2^-), and organic nitrogen. Future research into process-based biogeochemical dynamics is needed to better assess P retention.

3.4.4 Implications for the global assessment of nutrient retention

The global assessment of retention equations that was conducted in our study can improve the accuracy of global nutrient models: compared to the currently used

retention equation, applying the best-fit retention equation can reduce the NRMSE of riverine N, lake P, and riverine P in IMAGE-GNM by 41%, 66%, and 107%, respectively. By comparing the performance of empirical equations in different geographical regions, our study provided a possible way for model developers to further consider integrating regional retention modeling into global nutrient simulations. Further, the analyses of errors in performance, having distinguished the role of driving forces, function form, and equation coefficients, can constitute a step forward to the future development of empirical retention equations.

3.5 Conclusion

In this study, we used NRMSE to evaluate the error of model outcomes and Pearson's r of log-transformed data. We employed ANOVA and post hoc analyses to evaluate the under- or overestimates of different retention models.

Our results showed that global retention derived from different retention equations generates different patterns: the hydraulic-load-driven equations differ considerably from specific-runoff driven models and predicted relatively lower retention. The hydraulic driver is thus the most important factor that affects predicted TN/TP concentrations. Globally, empirical equations perform better for N than P. The retention models of Behrendt and Opitz (1999) (q) generate the lowest NRMSE for both N and P, being the best option for estimating riverine retention of TN/TP, while De Klein (2008)'s model is recommended for simulating P retention in lakes and reservoirs.

This global assessment allows model developers to choose empirical retention equations that best fit their region, thus improving the accuracy of modeling global nutrient fate and the N or P exports to coastal waters. Such improvements provide a better insight into the eutrophication in aquatic systems and support decision-makers to formulate environmental policies. The analysis on the driving force of retention constitutes a basis for the development of retention models for future nutrient fate and waterborne eutrophication-related studies.

Supporting Information

Table S3.1 Python packages, versions, and functions used for ANOVA and post hoc test. The analysis was accomplished by Python 3.7.

	Package	Version	Subpackage/Module	Function
ANOVA	scipy	1.6.2	stats	f_oneway
Welch's ANOVA	pingouin	0.5.1	/	welch_anova
Tukey's HSD	statsmodels	0.12.2	stats.multicomp	MultiComparison, tukeyhsd
Games-Howell post hoc test	pingouin	0.5.1	/	pairwise_gameshowell

Table S3.2 Best-performing retention models on a global and regional scale

Region and Observation Type	N	P
Lakes* Global	/	De Klein (2008)
Rivers* Global	Behrendt and Opitz (1999) (*q*)	Behrendt and Opitz (1999) (*q*)
North Frigid Zone	Behrendt and Opitz (1999) (*q*)	Wollheim et al. (2006) combined with the lake retention models of Kirchner and Dillon (1975) or Chapra (1975)*
North Temperate Zone	Behrendt and Opitz (1999) (*q*)	Behrendt and Opitz (1999) (*q*)
Torrid Zone	Venohr et al. (2005)	Behrendt and Opitz (1999) (*q*)
South Temperate Zone	Behrendt and Opitz (1999) (*q*)	Wollheim et al. (2006)

* The column names “Lakes” and “Rivers” indicate the classification of observed data, not the retention model types. The combination of riverine and lake retention model for P is used for validation with observed data of river stations.

Table S3.3 Multiple-comparison of the residual averages of N retention models by Tukey's HSD

Model 1	Model 2	Mean difference	p-value	Lower limit	Upper limit	Reject null hypothesis
Behrendt & Opitz (W_L)	Kelly et al.	0.585	0.001	0.170	1.000	TRUE
De Klein	Kelly et al.	0.736	0.001	0.321	1.151	TRUE
De Klein	Wollheim et al.	0.606	0.001	0.191	1.021	TRUE
Kelly et al.	Venohr et al.	-0.528	0.003	-0.943	-0.113	TRUE
Behrendt & Opitz (q)	Kelly et al.	0.489	0.009	0.074	0.904	TRUE
Behrendt & Opitz (W_L)	Wollheim et al.	0.454	0.021	0.039	0.869	TRUE
Venohr et al.	Wollheim et al.	0.398	0.070	-0.017	0.813	FALSE
Kelly et al.	Seitzinger et al.	-0.388	0.084	-0.803	0.027	FALSE
Behrendt & Opitz (q)	Wollheim et al.	0.359	0.141	-0.056	0.774	FALSE
De Klein	Seitzinger et al.	0.348	0.171	-0.067	0.763	FALSE
Seitzinger et al.	Wollheim et al.	0.258	0.520	-0.157	0.673	FALSE
Behrendt & Opitz (q)	De Klein	-0.247	0.568	-0.662	0.168	FALSE
De Klein	Venohr et al.	0.208	0.730	-0.207	0.623	FALSE
Behrendt & Opitz (W_L)	Seitzinger et al.	0.196	0.780	-0.219	0.611	FALSE
Behrendt & Opitz (W_L)	Behrendt & Opitz (q)	0.095	0.900	-0.320	0.510	FALSE
Behrendt & Opitz (W_L)	De Klein	-0.151	0.900	-0.566	0.264	FALSE
Behrendt &	Venohr et al.	0.057	0.900	-0.358	0.472	FALSE

Model 1	Model 2	Mean differ-ence	p-value	Lower limit	Upper limit	Reject null hypothesis
Opitz (W_L) Behrendt & Opitz (q)	Seitzinger et al.	0.101	0.900	-0.314	0.516	FALSE
Behrendt & Opitz (q)	Venohr et al.	-0.039	0.900	-0.454	0.376	FALSE
Kelly et al.	Wollheim et al.	-0.130	0.900	-0.545	0.285	FALSE
Seitzinger et al.	Venohr et al.	-0.140	0.900	-0.555	0.275	FALSE

*The family-wise error rate (FWER) is set to be 0.05, which means a 5% of probability rejecting the null hypothesis when it is true. If p-value $\geqslant$ 0.05, it fails to reject the null hypothesis after adjustment for the multiple comparisons. Since Tukey's HSD generates constant SE and $q_{\alpha,k,\nu}$, lower and upper limit equal to Mean difference $\pm$ SE $\cdot\, q_{\alpha,k,\nu}$.

Table S3.4 Multiple-comparison of the averages of P retention models by Games-Howell post hoc test

Model 1	Model 2	Mean differen-ce	SE	t_w	Degree of freedom	p-value	Reject null hypothesis
Category 1: retention models applied to rivers and lakes							
Behrendt & Opitz (W_L)	De Klein	0.139	0.033	4.2	1381.2	0.001	TRUE
Behrendt & Opitz (q)	De Klein	0.103	0.021	4.9	1551.9	0.001	TRUE
De Klein	Kelly et al.	-0.243	0.039	-6.2	1215.6	0.001	TRUE
De Klein	Wollheim et al.	-0.136	0.031	-4.3	1436.4	0.001	TRUE
Behrendt & Opitz (q)	Kelly et al.	-0.140	0.037	-3.8	1051.1	0.004	TRUE

Model 1	Model 2	Mean differen-ce	SE	t_w	Degree of freedo m	p-value	Reject null hypoth esis
Kelly et al.	Wollhei m et al.	0.107	0.044	2.4	1574.6	0.232	FALSE
Behrendt & Opitz (W_L)	Kelly et al.	-0.104	0.045	-2.3	1620.4	0.294	FALSE
Behrendt & Opitz (W_L)	Behrendt & Opitz (q)	0.036	0.031	1.2	1156.7	0.900	FALSE
Behrendt & Opitz (W_L)	Wollhei m et al.	0.003	0.039	0.1	1688.7	0.900	FALSE
Behrendt & Opitz (q)	Wollhei m et al.	-0.034	0.029	-1.2	1196.5	0.900	FALSE
Category 2: retention models applied to lakes only							
Brett &Benjamin	Chapra	-0.013	0.039	-0.3	1694.9	0.900	FALSE
Brett &Benjamin	Kirchner & Dillon	-0.004	0.038	-0.1	1696.0	0.900	FALSE
Brett &Benjamin	Wollhei m et al.	0.033	0.038	0.9	1695.1	0.900	FALSE
Chapra	Kirchner & Dillon	0.009	0.039	0.2	1694.6	0.900	FALSE
Chapra	Wollhei m et al.	0.046	0.038	1.2	1692.1	0.900	FALSE
Kirchner & Dillon	Wollhei m et al.	0.037	0.038	1.0	1695.3	0.900	FALSE

* If p-value $\geqslant$ 0.05, it fails to reject the null hypothesis after adjustment for the multiple comparisons. Degree of freedom that varies between pairwise groups was derived from Welch's Anova. $q_{\alpha,k,\nu}$ is a constant due to the large degree of freedom. Since Games-Howell post hoc test generates variable $SE_{i,j}$, which t_w varies according to, we listed them here.