Received: 7 April 2020 | Revised: 13 July 2020 | Accepted: 31 October 2020

DOI: 10.1002/gdj3.111

DATA PAPER

Geoscience Data Journal RMetS WILEY

Global database of diffuse riverine nitrogen and phosphorus loads and yields

Rich W. McDowell[1,2] | **Alasdair Noble**[1] | **Peter Pletnyakov**[1] | **Luke M. Mosley**[3]

[1]AgResearch, Lincoln Science Centre, Christchurch, New Zealand

[2]Faculty of Agriculture and Life Sciences, Lincoln University, Lincoln, New Zealand

[3]School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia

Correspondence
Rich McDowell, AgResearch, Lincoln Science Centre, Private Bag 4749, Christchurch 8140, New Zealand.
Email: richard.mcdowell@agresearch.co.nz

Abstract

Human activities have increased the input of nitrogen and phosphorus into riverine systems. These inputs can increase algal growth that degrades aquatic ecosystems. We constructed a global database of diffuse loads (kg) and yields (kg ha^{-1} yr^{-1}) of dissolved and total nitrogen and phosphorus forms for 7 years (centred around 2008) in 1,421 catchments. Yields were calculated from 640,950 measurements that were checked, filtered and harmonized from readily available sources. We used the yield data to create a georeferenced model to calculate yields of nitrogen and phosphorus forms across 6,020 catchments, globally. The database can be used to assess and inform policy to reduce nitrogen and phosphorus losses from land to freshwater, improve nutrient use efficiency on farms, and help calibrate global models being used to explore scenarios such as nutrient management efficiency in a changing climate. The source data and R code are provided at https://doi.org/10.25400/lincolnuninz.11894697.

KEYWORDS

dissolved reactive phosphorus, eutrophication, modelling, nitrate

1 | INTRODUCTION

The loads (kg) and annual yields (kg ha^{-1} yr^{-1}) of nitrogen (N) and phosphorus (P) are key factors controlling the eutrophication of streams and rivers (Dodds and Smith, 2016). Nutrients can originate from point and non-point sources. Reducing loads lost from point sources can be efficiently managed with engineering solutions (Macintosh *et al*., 2018). However, loads lost from

Data set

Identifier: https://doi.org/10.25400/lincolnuninz.11894697

Creator: Richard McDowell

Dataset correspondence: Richard.mcdowell@agresearch.co.nz

Title: Global database of riverine nitrogen and phosphorus loads and yields

Publisher: Figshare

Publication year: 2020

(Resource type): Database

(Version): 1.0

non-point sources, otherwise known as diffuse sources, reflect how a range of climate and catchment characteristics interact with land use decisions (Carpenter *et al*., 1998; Álvarez *et al*., 2017). To manage diffuse losses, we require knowledge of how these interactions vary in space, beginning at the coarsest (global) level and becoming increasingly targeted and tailored to finer spatial scales as more information is available.

To calculate loads, discharge (e.g. L s^{-1}) is multiplied by the concentration (e.g. mg L^{-1}) of nutrients. Discharge is usually measured continuously, while owing to cost and logistics, nutrient concentrations are often measured on a fortnightly, monthly or quarterly basis. There are several methods available to combine continuous discharge with less frequent concentration measurements to interpolate daily loads over a year or years. In examining the Ratio method and five- and seven-parameter rating methods, Snelder *et al*. (2017) found that the seven-parameter rating method produced estimates with the highest precision and representativeness from continuous discharge data and monthly or quarterly concentration data. The seven-parameter method can account for seasonal differences in concentration data and has been widely used in countries like the US and New Zealand (Alexander *et al*., 2002; Cohn, 2005) for the last 20 years; other methods have been developed but are not as widely used (Lee *et al*., 2016).

Few studies have calculated N and P loads for a range of streams and rivers across either single or multiple countries (Meybeck, 1982; Smith *et al*., 2003; Smith *et al*., 2003; Alvarez-Cobelas *et al*., 2009; Worrall *et al*., 2016). These studies were done several decades ago, focused on large catchments, usually in developed countries, and used a variety of methods to calculate loads. Although they have formed the backbone of hundreds of subsequent studies, more recent data over a wider geographic area are required to address contemporary questions such as the consequences of significant land use changes especially in developing countries on nutrient losses (Lambin and Meyfroidt, 2011).

Models can address data gaps, which is especially important to provide information for many developing countries. There are many good models that estimate the load and yield of N and P at a catchment scale (Jeppesen *et al*., 2011; Elliott *et al*., 2016); however, few exist at a regional or global scale (Seitzinger *et al*., 2010; Bouwman *et al*., 2013). Models, including those that estimate load and yield, can be categorized as either 'lumped' or 'distributed'. Lumped models often describe loads with statistical relationships to catchment characteristics such as soil type, and land use (Stevenson *et al*., 2012; Julian *et al*., 2017) but are restricted by the spatial and temporal scale of measurement. Distributed models rely on operational functions to describe process dynamics at a fine scale (Kroeze *et al*., 2012). However, unless a lot of data are available at a fine spatial and temporal scale, distributed models are seldom well resolved. This issue has been noted for the comprehensive and well-used Global NEWS models of nutrient flows where calibration data were taken from one source covering 100-200 rivers (Meybeck and Ragu, 1997; Kroeze *et al*., 2012).

Here, we provide and describe a database of measured nitrite-nitrate-N (NO_x-N), total N, dissolved reactive P (DRP) and total P loads and yields for 1,421 streams and rivers distributed worldwide. We combined these data with data on catchment characteristics and land use in a model to estimate the load and yield of these nutrient forms across the globe. We focus our analysis on catchments that are likely to have major diffuse sources of nutrient losses although we cannot discount some input from point sources, which recent evidence has shown to be highly variable in magnitude and contributing factors (van Puijenbroek *et al*., 2019). Our global data set is expected to be useful for multiple purposes, for example: (a) in the catchment-specific comparison of loads or yields to address the impact of land use on surface water quality and informing policy to target measures to prevent eutrophication or toxicity (McDowell *et al*., 2016); (b) on questions of resource efficiency such as what farming systems make the best use of nutrient inputs especially where resources like P are finite? (Herrero *et al*., 2017); or (c) in the calibration or development of earth system models that can be used to estimate and explore nutrient losses under future scenarios like a changed climate or across different spatio-temporal scales (Greaver *et al*., 2016; Fetzel *et al*., 2017). Our georeferenced global data set is an order of magnitude larger than those of Meybeck and Ragu (1997) or Alvarez-Cobelas *et al*. (2009) and provides much more data for Africa, Asia and South America. A range of rivers from Australasia is used to test the representativeness and performance of the modelled outputs.

2 | DATA DESCRIPTION AND DEVELOPMENT

The methods are described in six steps depicted in Figure 1 and outlined in detail below.

2.1 | Database acquisition and checking (step 1)

Nutrient concentration and discharge data were obtained from publicly available databases or by email requests to researchers from relevant peer-reviewed literature (GEMStat, GLORICH, New Zealand Water Quality database [NZWQ] and the Murray-Darling database; Table 1). Load data were already available for the NZWQ database (Snelder *et al*., 2018). Since these loads used the same checks, georeferencing and filters as outlined below, the corresponding yields were input directly into the global yield models (step 5).

The databases were chosen for their geographic representativeness and use of systems to ensure that the quality of the data was robust. Each database had its own reporting conventions, measurement units and laboratory detection limits. To produce a globally consistent nutrient data set, we checked the data using a multi-stage process that involved:

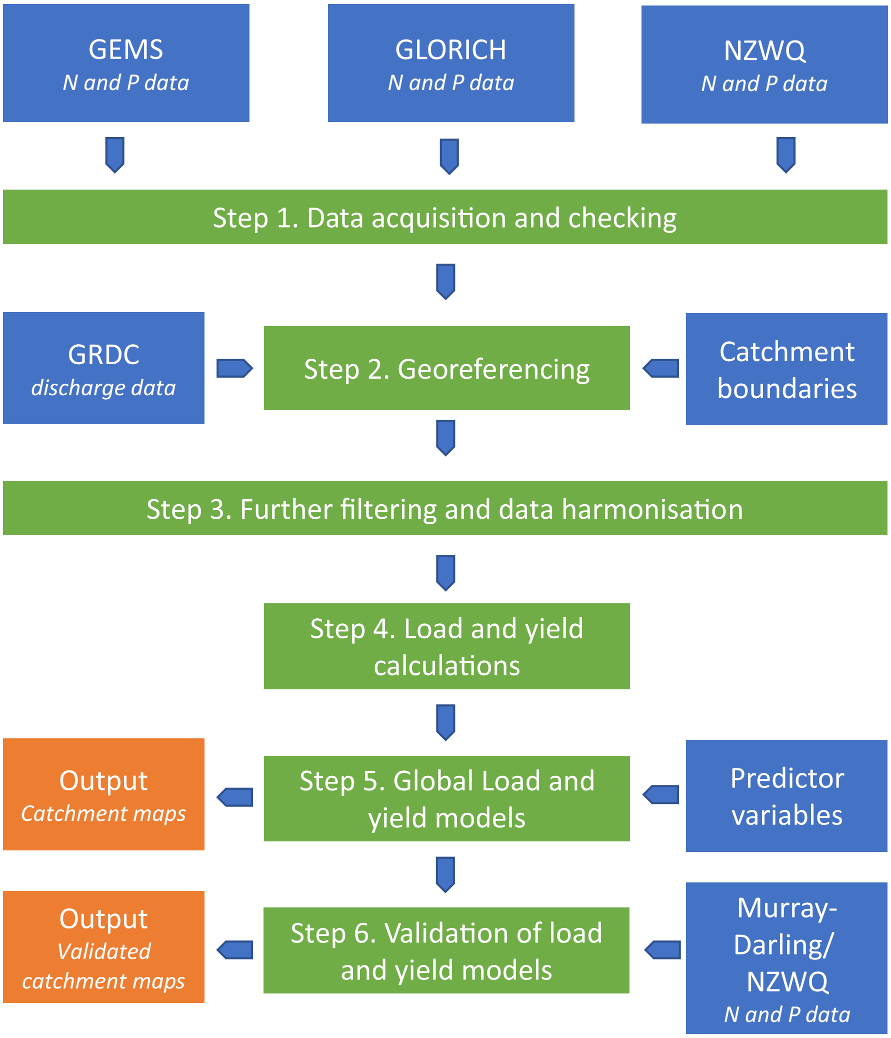

FIGURE 1 Flowchart of the steps involved in checking, filtering, georeferencing, harmonizing and analysing the data to produce a database of loads and yields for up to 1,421 different catchments and models to estimate yields globally. Note that blue boxes are input data

(a) checking data at each site for outliers; (b) Imputing replacement values for data below the detection limit; and (c) Checking that analytical methods were acceptable. Details of these checks are found in an aligned publication that explored concentration data only (McDowell *et al*., 2020).

2.2 | Georeferencing (step 2)

We checked the location and name of each site within the 6,020 catchments of the fourth level of the HydroSHEDS drainage network (Lehner and Grill, 2013) (Figure 2). For the GLObal RIver Chemistry Database, (GLORICH) (Hartmann *et al*., 2014) and Global Environmental Monitoring System (GEMS) (United Nations Environment Programme, 2018) databases we located sites within each catchment where discharge had been recorded that were within 50 km downstream of sampling sites but had no major urban centre (population > 100,000) between the sampling site and the discharge site. We obtained discharge from these sites from the Global Runoff Data Centre (GRDC) (Federal Institute of Hydrology, 2018) and assigned a GRDC identifier to each sampling site.

We did not apply a GRDC identifier to these data as discharge was not needed to calculate loads or was available from other sources. However, we did reference them to a corresponding catchment at the fourth level of the HydroSHEDS drainage network shapefiles (Lehner and Grill, 2013).

2.3 | Further filtering and data harmonization (step 3)

For each variable at each site, we examined the availability of N and P data from 1990 to 2016. Sites were further filtered from our database if they:

Did not have ≥36 sampling dates spread over at least a recent 3-year period, and where the latest sampling date was before 1995. The 36-sample size minimum and date filter is a trade-off between the good spatial coverage and representativeness of sites and having too few recent data to produce precise load and yield calculations (Snelder *et al*., 2017).

TABLE 1 Data sources and richness at several steps used to calculate the load and yields of phosphorus and nitrogen forms

Database	N or P fraction	Step 1: Number of sites (data records) after checking	Step 3: Number of sites (data records) after filtering and harmonization	Steps 5 and 6: Number of catchments with predictor variables contributing to global yield models	Source of data
GEMStat	DRP	1,392 (59,065)	111 (12,310)	107	United Nations Environment Programme (2018)
	TP	2,268 (80,120)	640 (40,897)	254	
	NO_x-N	605 (35,870)	136 (12,863)	107	
	TN	2,476 (119,526)	744 (55,181)	76	
GLORICH	DRP	818 (137,461)	579 (104,590)	486	Hartmann *et al.* (2014)
	TP	770 (110,260)	481 (74,020)	481	
	NO_x-N	717 (152,886)	517 (99,944)	378	
	TN	517 (68,167)	316 (39,892)	158	
Murray-Darling	DRP	25 (34,642)	25 (18,325)	10[a]	Biswas and Mosley (2018)
	TP	25 (35,889)	25 (18,611)	10	
	NO_x-N	25 (34,454)	24 (17,733)	9	
	TN	30 (45,852)	29 (22,100)	10	
NZWQ	DRP	723 (74,573)	449 (31,121)	438	McDowell *et al.* (2013); Larned *et al.* (2016)
	TP	723 (74,571)	449 (31,121)	311	
	NO_x-N	723 (74,573)	449 (31,121)	86	
	TN	723 (74,573)	449 (31,121)	26	

[a] Only used as part of the technical validation.

Were associated with blanks, replicates or control samples (where identified in the source databases).

Were recorded in the GLORICH database and were already present in the GEMS, NZWQ or Murray-Darling databases. Likewise, we removed sites from the GEMS database that were recorded in the NZWQ or Murray-Darling databases. We did this because the NZWQ and Murray-Darling databases contained more data for the same site listed in either the GLORICH or GEMS databases.

Once checked, georeferenced and filtered, the resulting data were harmonized for the most recent 7 years of data. This resulted in a data set whose mean age was 2008. The 7-year period was chosen to obtain annual means that were less likely to be skewed by 1 year, while avoiding substantial variation associated with continental or global climatic trends over longer periods. For example, in the Southern Hemisphere, the Southern Oscillation Index results in distinct trends in loads associated with wet and dry years (Scarsbrook *et al.*, 2003). The 7-year period also minimizes, but does not exclude, any likely influence from anthropogenic trends (Larned *et al.*, 2016). A similar approach is used on a 9-year rolling average by the US Geological Survey (Lee *et al.*, 2017).

After harmonization, a total of 1,004,233 nutrient concentration measurements were available at 1,492, 1,915, 1,412 and 2,052 sites for DRP, TP, NO_x and TN, respectively (Table 1). Discharge data were available for, on average, 98% of the 7-year period for which concentration data were available. Loads were truncated where discharge data were not available.

2.4 | Load and yield calculations (step 4)

For each sampling site, we calculated mean annual loads for DRP, TP, NO_x-N and TN using the commonly used seven-parameter rating method rating method (Alexander *et al.*, 2002). We chose the seven-parameter method as it provides estimates with the highest precision and representativeness from continuous discharge data and monthly or quarterly concentration data (Snelder *et al.*, 2017). Rating methods derive a relationship between the sampled nutrient concentrations (C_i, mg L^{-1}) and discharge (q_i, m^3 s^{-1}), which are then used to estimate concentration for each day of the entire sampling period (Cohn *et al.*, 1992; Hirsch, 2014). The model is based on fitting seven parameters:

$$\begin{aligned}\ln(\widehat{C_i}) = \beta_1 + \beta_2\left[\ln\left(q_i\right) - \overline{(\ln(q))}\right] + \beta_3\left[\ln\left(q_i\right) - \overline{(\ln(q))}\right]^2 \\ + \beta_4\left(t_i - \overline{T}\right) + \beta_5\left(t_i - \overline{T}\right)^2 + \beta_6 \sin\left(2\pi t_i\right) + \beta_7 \cos\left(2\pi t_i\right)\end{aligned} \quad (1)$$

where, $\beta_{1,2,\ldots,7}$ are regression coefficients, t_i is the time in decimal years, $\overline{T}$ is the mean value of time in decimal years, $\overline{(\ln(q))}$ is the mean of the natural log of catchment discharge on the days sampled, and $\widehat{C}_i$ is the estimated concentration at time *i*.

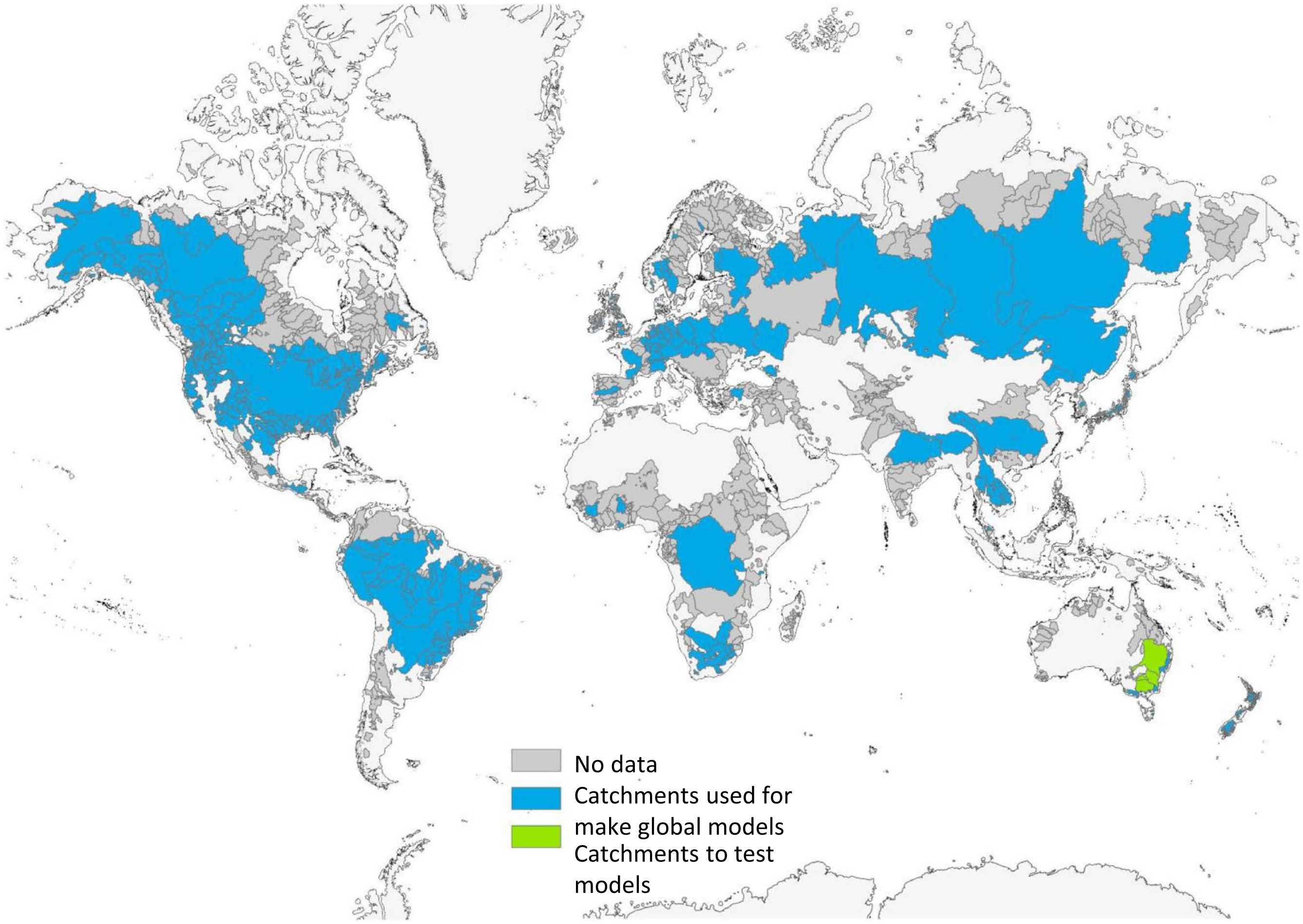

FIGURE 2 Catchment boundaries at the fourth level of HydroSHEDs ($n = 6{,}020$; grey) and for those used to generate ($n = 1{,}421$; blue) and test ($n = 30$; green) global models of nitrogen and phosphorus yields

Coefficients were estimated by multiple linear regression of the sample data and used to calculate the concentration on each day of the 7-year period of record. Estimates of $\widehat{C}_i$ are back-transformed and corrected for retransformation bias by adding a smearing estimate (Duan, 1983):

$$S=\frac{1}{n}\sum_{i=1}^{n}e^{\widehat{\varepsilon}_i} \tag{2}$$

where n is the number of data points and $\widehat{\varepsilon}_i$ are the residuals of the regression models. The correction factor (S) is applied over the whole range of estimations as it is assumed that the residuals are homoscedastic.

The total load (L), expressed as a yield (kg ha^{-1} yr^{-1}), was calculated by multiplying and summing discharge and estimated concentrations over the 7-year time series as:

$$L=\frac{K}{A_c}\left(\sum_{j=1}^{N}S\frac{\widehat{C}_j q_j}{N}\right) \tag{3}$$

where A_c is catchment the area in ha obtained from (United States Department Of The Interior - United States Geological Survey, 2008), K is a unit conversion factor equal to 31.6 kg s mg^{-1} yr^{-1}, C_j is the concentration for each day in period of record estimated from the above model in mg m^{-3}, q_j is the daily mean discharge for each day in period of record in m^3 s^{-1}, and N is the number of days in the period of record.

Equation 1 was fitted using both a linear model using log-transformed data and a generalized linear model with a Gaussian distribution and log-link. In general, the two models gave similar results, but in a few cases, one or other generated results that were not plausible. In these cases, and to avoid poorly estimated yields from biasing the global yield models, we inspected the results to determine if they were within expected ranges. The generalized linear model was chosen for further analysis except where yields fell outside the range for large catchment yields of 0-30 kg ha^{-1} y^{-1} for P and 0-100 kg ha^{-1} y^{-1} for N (Hale *et al*., 2015). Where this occurred the yield from the simple linear model was chosen ($n = 341$). Where neither method produced yields within these ranges ($n = 32$) we substituted them with average yields for large agricultural catchments across Asia, Australasia, Europe, and North and South America (McDowell and Wilcock, 2008; Billen *et al*.,

TABLE 2 List, units and sources of variables used to estimate the yields of nitrogen and phosphorus forms

Variable	Units	Source
Catchment area	km^2	United States Department Of The Interior - United States Geological Survey (2008)
Population	People in catchment	
Mean rainfall per month	mm month^{-1}	Fick and Hijmans (2017)
Mean altitude	m above sea level	United States Department Of The Interior - United States Geological Survey (2008)
Mean slope	%	United States Department Of The Interior - United States Geological Survey (2008)
Soil order	FAO soil order code	IUSS Working Group WRB (2015)
Land cover classification	–	Sayre *et al*. (2014)
Mean runoff per month	mm	Fekete *et al*. (2018)
Soil Olsen P stock	kg ha^{-1}	McDowell *et al*. (2020)
Crop cover 2009	% of catchment	European Space Agency (2010)
Forest cover 2009	% of catchment	European Space Agency (2010)
Lentic cover 2009	% of catchment	European Space Agency (2010)
Pasture cover 2009	% of catchment	European Space Agency (2010)
Rangeland cover 2009	% of catchment	European Space Agency (2010)
Urban cover 2009	% of catchment	European Space Agency (2010)
Mean runoff each month	mm month^{-1}	Fekete *et al*. (2018)
Terrestrial biomes	Categorical	Dinerstein *et al*. (2017)
Ecoregion within biomes	Categorical	Dinerstein *et al*. (2017)
Global ecological land units - bioclimate	Categorical	Sayre *et al*. (2014)
Global ecological land units – landform	Categorical	Sayre *et al*. (2014)
Global ecological land units - lithology	Categorical	Sayre *et al*. (2014)
Evapotranspiration	mm yr^{-1}	Willmott and Matsuura (2001)
Potential evapotranspiration	mm yr^{-1}	Willmott and Matsuura (2001)
Soil wetness	mm over profile	Willmott and Matsuura (2001)
Population density - 1990	Persons km^{-2}	Center for International Earth Science Information Network - CIESIN - Columbia University and Centro Internacional de Agricultura Tropical - CIAT (2005)
Population density - 1995	Persons km^{-2}	Center for International Earth Science Information Network - CIESIN - Columbia University and Centro Internacional de Agricultura Tropical - CIAT (2005)
Population density - 2000	Persons km^{-2}	Center for International Earth Science Information Network - CIESIN - Columbia University and Centro Internacional de Agricultura Tropical - CIAT, (2005)
Zone class (temperate, tropical, polar)	–	–
Hemisphere (north, south)	–	–

2011; Strokal *et al*., 2015; Hale *et al*., 2015; Bustamante *et al*., 2015) of 0.25, 0.5, 2, and 2 kg ha^{-1} y^{-1} for DRP, TP, NO$_x$-N and TN, respectively.

2.5 | Global yield models (step 5)

We modelled yields, instead of loads, globally. While modelling loads results in greater coefficients of determination than modelling yields, this is because loads are principally a function of catchment size and specific runoff (mm yr^{-1}) not because they better account for processes. We chose to model yields so that relative nutrient losses could be easily compared between catchments (Smith *et al*., 2005). Yields were log$_{10}$-transformed to remove skewness and derive an approximately normally distributed dataset. Predictor variables were obtained from a wide variety of sources, listed in Table 2. These variables reflected known categorical and continuous factors likely to affect N and P losses

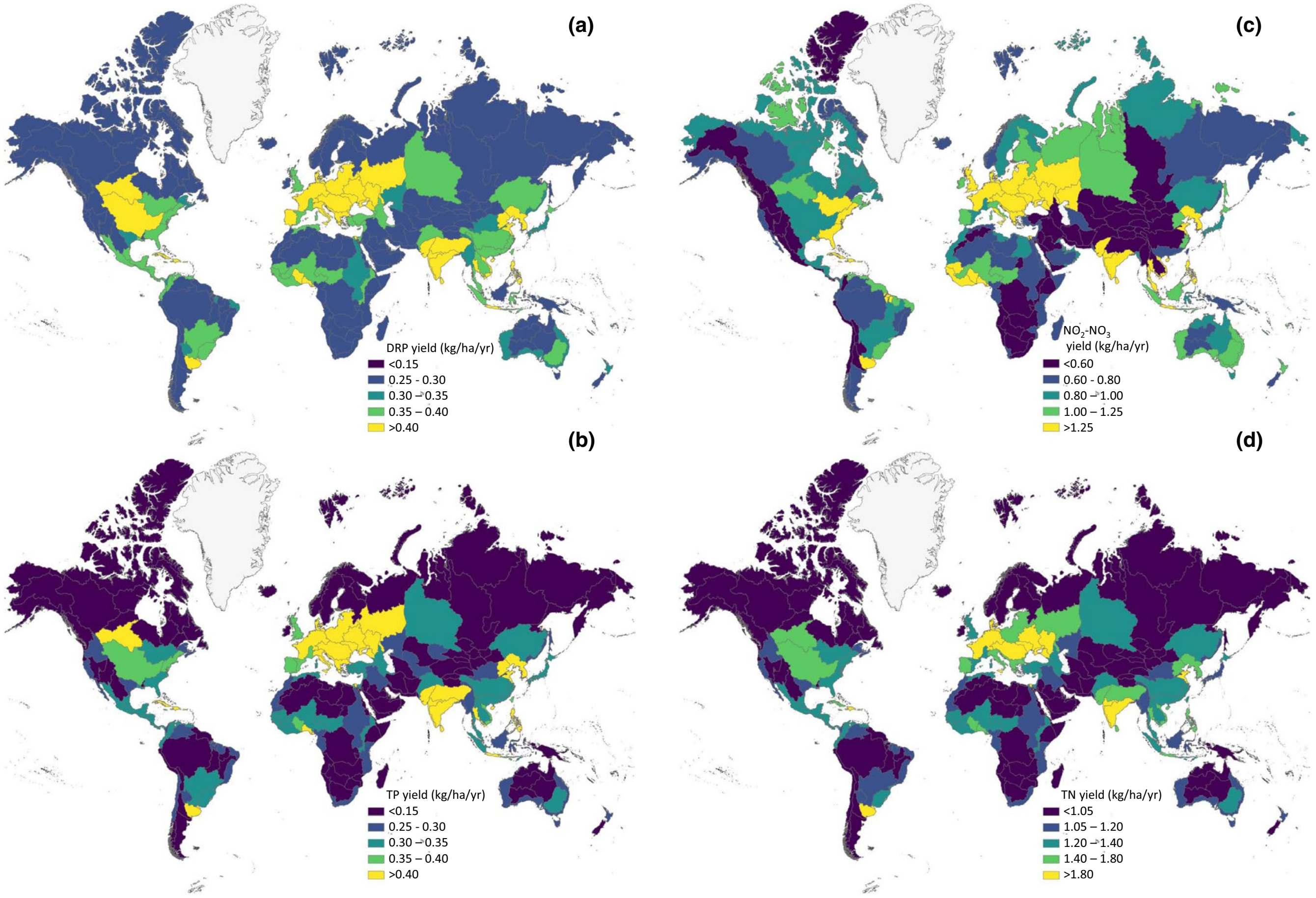

FIGURE 3 Relative range in estimated (a) dissolved reactive P (DRP), (b) total P, (c) nitrite-nitrate-N, and (d) total N yields (kg ha^{-1} yr^{-1}) across the globe at the third level of HydroBASINS (*n* = 288). Yield classes were determined using the Natural Breaks option within ARC GIS

to water such as land use and land use intensity (Julian *et al*., 2017; Álvarez *et al*., 2017), catchment (e.g. soil wetness) and climate characteristics (van Meter and Basu, 2017). Continuous variables (e.g. percentage cropland) were expressed as a mean proportion across each catchment. Where more than one categorical variable (e.g. hierarchical USEPA ecoregion classifications Dinerstein *et al*. (2017)) was present in a catchment, the catchment was assigned to the dominant category for that catchment.

After yields were combined with predictor variables, we had data for 1,421 catchments. This is fewer than those output after harmonization because not all predictor variables were available at all sites or there was more than one site in a catchment (Figure 2). The number of sites is much greater than have been used in the analysis of nutrient loadings to the ocean (165 rivers) (Smith *et al*., 2003), and the 100-200 rivers used to calibrate the Global NEWS model or the Integrated Model to Assess the Global Environment – Global Nutrient Model (Mayorga *et al*., 2010; Beusen *et al*., 2015).

We used a best subsets routine in the R-programming language (www.r-project.org) to generate an over-fitted model to estimate catchment yields which was then reduced by sequentially removing the least significant variable until we were left with a parsimonious model where all variables were significant. An optimal regression output (i.e. avoiding over-parameterization) was generated with the aid of the Mallows' Cp statistic. Significant ($p < .05$) variables and their coefficients for the equations predicting concentration and load of N and P forms are available in Supplementary Tables S1-S4.

Models estimating yields of DRP, NO$_x$-N, TP and TN were extended globally in ArcGIS using the datasets in Table 2. Raster grids were created with a spatial resolution of 0.025 degrees, which corresponded to the coarsest grid cell associated with input data listed in Table 2. To provide consistent global coverage for presentation, grid values were averaged within the boundaries of 288 catchments in the HydroBASINS Level 3 (Lehner, 2014) shapefiles. However, note that the model is produced and can be used to estimate yields at the fourth level of the HydroSHEDS network. Catchments were assigned to continents based on majority area covered within a continental boundary. Global nutrient yield data are presented in Figure 3, estimated yields were back-transformed and corrected for retransformation bias by adding the smearing estimate (Duan, 1983).

TABLE 3 Variables, coefficient of determination, and the significance of models used to estimate log-transformed global dissolved reactive P, total P, NO_x-N and total N yields

Model	Variables output from the model	Coefficient of determination (p value)	Bias correction
Dissolved reactive P	Crop land (%), Potential evapotranspiration (mm yr^{-1}), Soil Wetness (mm over profile), Ecoregion	0.36 (<.001)	5.99
Total P	Crop land (%), Urban land (%), Ecoregion	0.36 (<.001)	2.96
NO_x-N	Mean altitude (msl), Crop land (%), Soil Wetness (mm over profile), Ecoregion	0.57 (<.001)	1.97
Total N	Crop land (%), Soil Wetness (mm over profile), Ecoregion	0.59 (<.001)	0.86

Note: Coefficients for each variable, including US EPA Ecoregions (Dinerstein *et al*., 2017), are found in the file 'Supplementary information yield models.xlsx' hosted at https://doi.org/10.25400/lincolnuninz.11894697. Outputs must be multiplied by the bias correction factor after back-transformation.

2.6 | Model performance (step 6)

The models accounted for between 35% and 59% of the variation in the data (Table 3; Supplementary Tables S1-4), with apparent higher reliability for TN and NOx compared to TP and DRP. In hindsight, an alternative approach such as machine learning may have produced greater coefficients of determination. However, the selected approach was like those used previously thereby enabling the quality of the data to be compared and not the method. The coefficient of determination is similar to or better than the best performing empirical models (r^2 for TP = 0.38) of global nutrient yields by others but contains many more catchments (Meybeck, 1982; Meybeck and Ragu, 1997; Smith *et al*., 2005). The exception is for the study by Alvarez-Cobelas *et al*. (2009) who produced coefficients of 0.60 DRP and 0.73 for TP for 104 and 132 rivers, respectively. However, their models included runoff (viz, discharge) which is not an independent variable given it is a critical component of calculating loads and yields. Coefficients of determination were also like those of data-hungry process-based models like Global NEWS 2 (Mayorga *et al*., 2010) (r^2 = 0.54 for dissolved inorganic N, which is mostly comprised of NO_x) or IMAGE-GNM (Beusen *et al*., 2015) (r^2 = 0.58 for TN) who used far fewer sites to calibrate their estimates.

As expected, our models contained variables thought by a range of studies to control nutrient losses such as the percentage of cropland in a catchment, hydrological parameters such as soil wetness or potential evapotranspiration, and ecoregions which capture climate by soil by vegetation interactions (Julian *et al*., 2017; Álvarez *et al*., 2017; van Meter and Basu, 2017). Their importance differed between nutrient forms (see 'Supplementary information yield model outputs.xlsx') but a detailed investigation is beyond the scope of this paper.

Unless continuous discharge can be combined with continuous concentration measurements, yields will always be influenced by the intervals at which concentrations are sampled. Hence, the validation of estimated yields against 'true' yields is not strictly possible (Jordan *et al*., 2005). We chose the seven-parameter method to try and account for the paucity or frequency of data at some sites. The performance of this method is discussed elsewhere (Cohn, 2005; Hirsch, 2014; Lee *et al*., 2016; Snelder *et al*., 2017). We did, however, examine the ability of our models to estimate yields calculated from an independent dataset of nine sites in the Murray-Darling River Basin (Biswas and Mosley, 2018) and a 20 rivers from the NZWQ database (Snelder *et al*., 2018) that were not included in database used for the global yield models. We chose these catchments for their size (stream order 6 or 7) to match the mean stream order in our global database of 6.7, and for their wide coverage of biomes: Deserts and Xeric Shrublands and range of Temperate and Tropical biomes, but no Tundra.

The output (Figure 4) showed similar or better coefficients of determination amongst N and P forms to that produced in the global yield models, although estimates generated using the global model overestimated DRP and underestimated total P, NO_x-N and total N.

3 | DATASET ACCESS

The database is provided within the zip folder 'Global_Nutrient_Yields.zip', hosted at https://doi.org/10.25400/lincolnuninz.11894697. A list and description of the data and files contained within the zip file are given in the file—'Description.txt'. Briefly, the file contains Excel files of the checked, filtered and harmonized DRP, TP, NO_x-N and TN concentrations at each site and data set used in the calculation of loads and yields: 'Global River Data with flow sites.xls' for the GEMS data, 'GLORICH_hydrochemistry_GRDC_Order_3.xls' for the GLORICH data and the Murray-Darling data were housed in the '/MD_Dat' folder. The GLORICH

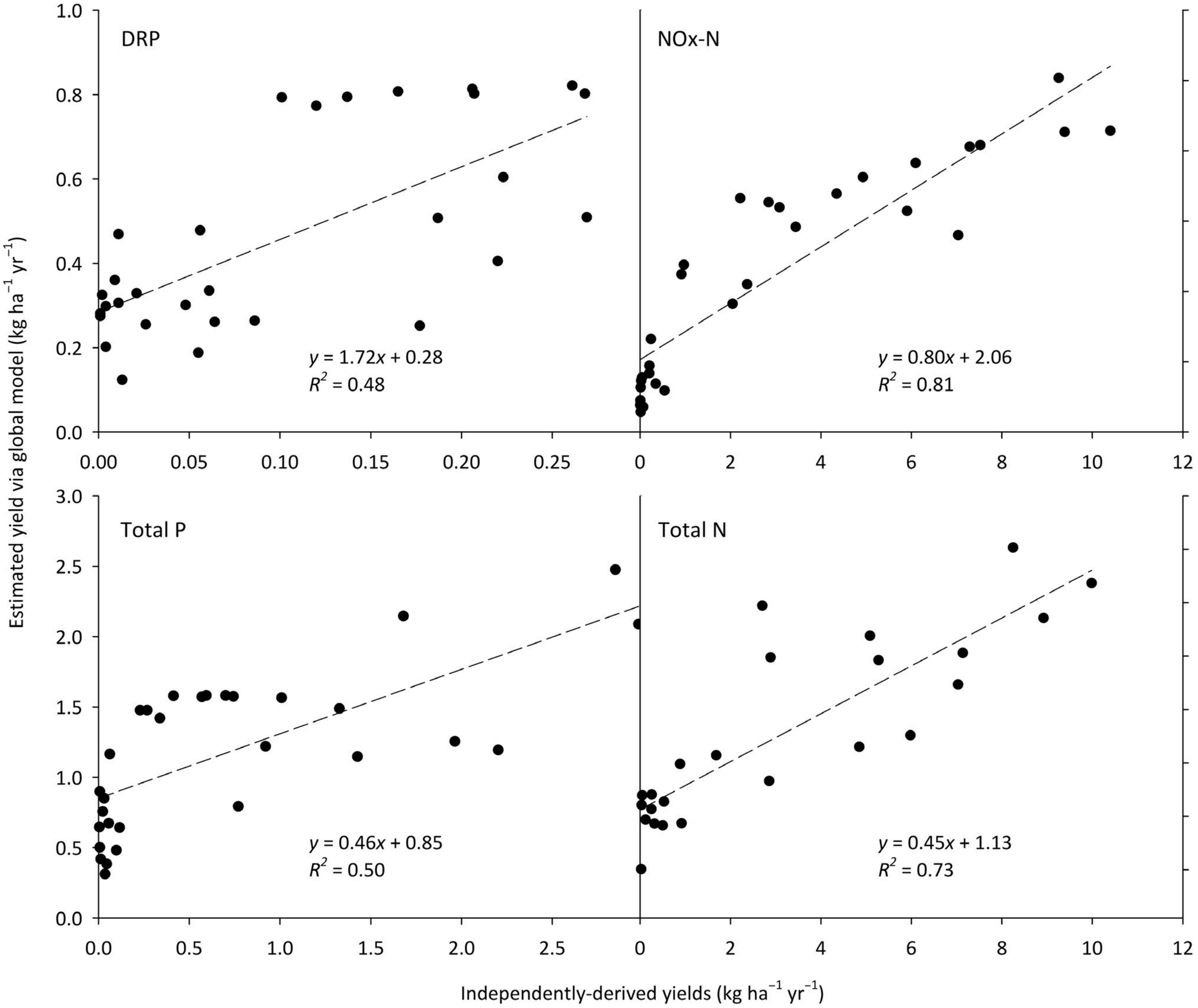

FIGURE 4 Relationship between estimated yields using the global model and yields generated via an independent data set of order 6 and 7 rivers from Australia and New Zealand

and GEMS sites are linked to GRDC site identifiers, while the NZWQ sites have a 6-8 digit code and the Murray-Darling sites are identified by name.

The R code is available to calculate nutrient loads and yields as 'Site yield estimates_GEMS.R', 'Site yield estimates_GLORICH.R' and Site yield estimates_Murray_Darling.R'. The code to model yields globally is available in the file 'Global yield models.R' Data and scripts are also given to calculate Total Kjeldhal N (TKN) and Total Suspended Solids yields, but these are not used in the present global model. Additional Excel files give the load and yield of N and P forms for all filtered and harmonized catchments separately (Global load dataAN.xls) and merged with predictor variables (Global yield and predictor merge.xls; step 6). Each Excel file contains metadata associated with their listed abbreviations, a full description of the parameter and their units of measurement.

The processed nutrient load and yield database generated in this paper can be used for any purpose provided the source is acknowledged. Use of the concentration and discharge data should acknowledge the original source. Furthermore, the original data sources should be contacted before using the concentration and discharge data for commercial purposes. Updated discharge data are available, free of charge, from the GRDC River Discharge Data service at https://www.bafg.de/GRDC/EN/02_srvcs/21_tmsrs/riverdischarge_node.html, and for the NZWQ database from https://hydrowebportal.niwa.co.nz/. Updated nutrient concentrations for the NZWQ database are available from: https://www.lawa.org.nz/download-data/#river-water-quality, and for the GEMS database from: https://gemstat.org/data/data-portal/.

The units of concentration are μg m^{-3} in the NZWQ database, μmol L^{-1} in the GLORICH database, and mg L^{-1} in the Murray-Darling and GEMS databases. We converted µmol L^{-1} into mg L^{-1} by multiplying the GLORICH data for N and P by 0.014007 and 0.030974, respectively. If obtaining updated nutrient data from their data sources note that we treated dissolved inorganic P as analogous to DRP, and nitrate or NO_2-NO_3-N as NO_x-N.

It is intended that this database is updated every 5 years.

4 | POTENTIAL DATA SET USE AND REUSE

Possible applications of our nutrient loads and yields database include, but are not limited to, the isolation of nutrient loss hotspots, inputs to marine systems and the calculation of nutrient use efficiencies (Mueller *et al*., 2012; Roy *et al*., 2016; Lu and Tian, 2017; Powers *et al*., 2019). However, our estimates are for large rivers. The mean Strahler stream order across the database is 6.7. These data are therefore most appropriately used at large scales such as biomes, countries or continents. If making comparisons at these scales note that many of the catchments could cross more than one biome or jurisdiction.

We also caution against the use of the data beyond the years we could calculate robust yields and loads. On average, this was 2005-2012. Although we constrained the number of years to avoid trends in climate influenced by continental and oceanic weather systems (Scarsbrook *et al*., 2003), we cannot guarantee that localized or regional trends or events (e.g. storms or droughts) may have impacted on calculated loads and yields (Ockenden *et al*., 2016). These data can, however, be used as a baseline from which past or future yields can be compared.

We did not specifically account for atmospheric deposition or point source discharges in our calculations of yield and load. Deposition and point source discharges can affect yields and influence biotic growth (Elser *et al*., 2009; Tipping *et al*., 2014; Bowes *et al*., 2014). However, deposition data were only available for a few catchments or lakes. We therefore considered their effect to be included within riverine exports. It is possible to calculate the contribution of point source discharges (Bowes *et al*., 2008), although we expect its influence to be low in this instance given we filtered out most sites up to 50-km downstream of an urban centre and likely wastewater discharge.

ACKNOWLEDGEMENTS

We are grateful to The Global Runoff Data Centre, 56068 Koblenz, Germany, for discharge data, the International Centre for Global Water Resources and Global Change for the GEMStat water quality data, Prof Dr J Hartmann for the GLORICH water quality data, the Murray-Darling Basin Authority for the Murry-Darling catchment data, and the New Zealand Government for the New Zealand Water Quality database. Funding to write this manuscript was provided by the Our Land and Water National Science Challenge (contract C10X1507 from the Ministry of Business, Innovation and Employment).

OPEN PRACTICES

This article has earned an Open Data badge for makingpublicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.25400/lincolnuninz.11894697 Learn more about the Open Practices badges from the Center for OpenScience: https://osf.io/tvyxz/wiki.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Rich W. McDowell https://orcid.org/0000-0003-3911-4825

REFERENCES

Alexander, R.B., Johnes, P.J., Boyer, E.W. & Smith, R.A. (2002) A comparison of models for estimating the riverine export of nitrogen from large watersheds. *Biogeochemistry*, 57, 295–339.

Álvarez, X., Valero, E., Santos, R.M.B., Varandas, S.G.P., Sanches Fernandes, L.F. & Pacheco, F.A.L. (2017) Anthropogenic nutrients and eutrophication in multiple land use watersheds: Best management practices and policies for the protection of water resources. *Land Use Policy*, 69, 1–11.

Alvarez-Cobelas, M., Sánchez-Carrillo, S., Angeler, D.G. & Sánchez-Andrés, R. (2009) Phosphorus export from catchments: a global view. *Journal of the North American Benthological Society*, 28, 805–820.

Beusen, A.H.W., van Beek, L.P.H., Bouwman, A.F., Mogollón, J.M. & Middelburg, J.J. (2015) Coupling global models for hydrology and nutrient loading to simulate nitrogen and phosphorus retention in surface water – description of IMAGE–GNM and analysis of performance. *Geoscientific Model Development*, 8, 4045–4067.

Billen, G., Silvestre, M., Grizzetti, B., Leip, A., Garnier, J., Voss, M. *et al*. (2011). *Nitrogen flows from European watersheds to coastal marine waters The European nitrogen assessment: Sources, effects and policy perspectives*. Cambridge University Press.

Biswas, T.K. & Mosley, L.M. (2018) *From Mountain Ranges to Sweeping Plains, in Droughts and Flooding Rains; River Murray Water Quality over the Last Four Decades*. Water Resources Management.

Bouwman, A.F., Beusen, A.H.W., Griffioen, J., Van Groenigen, J.W., Hefting, M.M., Oenema, O. *et al*. (2013). Global trends and uncertainties in terrestrial denitrification and N2O emissions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368, 20130112

Bowes, M.J., Jarvie, H.P., Naden, P.S., Old, G.H., Scarlett, P.M., Roberts, C. *et al*. (2014) Identifying priorities for nutrient mitigation using river concentration–flow relationships: The Thames basin, UK. *Journal of Hydrology*, 517, 1–12.

Bowes, M.J., Smith, J.T., Jarvie, H.P. & Neal, C. (2008) Modelling of phosphorus inputs to rivers from diffuse and point sources. *Science of the Total Environment*, 395, 125–138.

Bustamante, M.M.C., Martinelli, L.A., Pérez, T., Rasse, R., Ometto, J.P.H.B., Siqueira Pacheco, F. *et al*.(2015) Nitrogen management challenges in major watersheds of South America. *Environmental Research Letters*, 10, 065007.

Carpenter, S.R., Caraco, N.F., Correll, D.L., Howarth, R.W., Sharpley, A.N. & Smith, V.H. (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 8, 559–568.

Center for International Earth Science Information Network - CIESIN - Columbia University & Centro Internacional de Agricultura Tropical - CIAT (2005). *Gridded Population of the World, Version 3 (GPWv3): Population Density Grid*. NASA Socioeconomic Data and Applications Center (SEDAC).

Cohn, T.A. (2005) Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resources Research*, 41. https://doi.org/10.1029/2004WR003833.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D. & Summers, R.M. (1992) The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research*, 28, 2353–2363.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E. *et al*. (2017) An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67, 534–545.

Dodds, W.K. & Smith, V.H. (2016) Nitrogen, phosphorus, and eutrophication in streams. *Inland Waters*, 6, 155–164.

Duan, N. (1983) Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78, 605–610.

Elliott, A.H., Semadeni-Davies, A.F., Shankar, U., Zeldis, J.R., Wheeler, D.M., Plew, D.R. *et al*. (2016) A national-scale GIS-based system for modelling impacts of land use on water quality. *Environmental Modelling & Software*, 86, 131–144.

Elser, J.J., Andersen, T., Baron, J.S., Bergström, A.-K., Jansson, M., Kyle, M. *et al*. (2009) Shifts in Lake N: P stoichiometry and nutrient limitation driven by atmospheric nitrogen deposition. *Science*, 326, 835–837.

European Space Agency (2010). *European Space Agency GlobCover Portal - GlobCover 2009*. ESA and Université Catholique de Louvain. Retrieved from http://due.esrin.esa.int/page_globcover.php [Accessed DEcember 2018]

Federal Institute of Hydrology (2018). *Global river data centre*. Federal Institute of Hydrology. Retrieved from https://www.bafg.de/GRDC/EN/01_GRDC/grdc_node.html [Accessed 22 December 2018]

Fekete, B.M., Vörösmarty, C.J. & Grabs, W. (2018). *UNH/GRDC composite runoff fields v1.0*. University of New Hampshire. Retrieved from http://www.compositerunoff.sr.unh.edu/ [Accessed December 2018]

Fetzel, T., Havlik, P., Herrero, M., Kaplan, J.O., Kastner, T., Kroisleitner, C. *et al*. (2017) Quantification of uncertainties in global grazing systems assessment. *Global Biogeochemical Cycles*, 31, 1089–1102.

Fick, S.E. & Hijmans, R.J. (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37, 4302–4315.

Greaver, T.L., Clark, C.M., Compton, J.E., Vallano, D., Talhelm, A.F., Weaver, C.P. *et al*. (2016) Key ecological responses to nitrogen are altered by climate change. *Nature Climate Change*, 6, 836.

Hale, R.L., Grimm, N.B., Vörösmarty, C.J. & Fekete, B. (2015) Nitrogen and phosphorus fluxes from watersheds of the northeast U.S. from 1930 to 2000: Role of anthropogenic nutrient inputs, infrastructure, and runoff. *Global Biogeochemical Cycles*, 29, 341–356.

Hartmann, J., Lauerwald, R. & Moosdorf, N. (2014) A brief overview of the GLObal RIver chemistry database, GLORICH. *Procedia Earth and Planetary Science*, 10, 23–27.

Herrero, M., Thornton, P.K., Power, B., Bogard, J.R., Remans, R., Fritz, S. *et al*. (2017) Farming and the geography of nutrient production for human use: a transdisciplinary analysis. *The Lancet Planetary Health*, 1, e33–e42.

Hirsch, R.M. (2014) Large biases in regression-based constituent flux estimates: causes and diagnostic tools. *Journal of the American Water Resources Association*, 50, 1401–1424.

IUSS Working Group WRB (2015). *World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming soils and creating legends for maps*. World Soil Resrouces Reports No. 106. Food and Agrculture Orgnisation.

Jeppesen, E., Kronvang, B., Olesen, J., Audet, J., Søndergaard, M., Hoffmann, C. *et al*. (2011) Climate change effects on nitrogen loading from cultivated catchments in Europe: implications for nitrogen retention, ecological state of lakes and adaptation. *Hydrobiologia*, 663, 1–21.

Jordan, P., Arnscheidt, J., McGrogan, H. & McCormick, S. (2005) High-resolution phosphorus transfers at the catchment scale: the hidden importance of non-storm transfers. *Hydrology and Earth System Sciences*, 9, 685–691.

Julian, J.P., de Beurs, K.M., Owsley, B., Davies-Colley, R.J. & Ausseil, A.G.E. (2017) River water quality changes in New Zealand over 26 years: response to land use intensity. *Hydrology and Earth System Sciences*, 21, 1149–1171.

Kroeze, C., Bouwman, L. & Seitzinger, S. (2012) Modeling global nutrient export from watersheds. *Current Opinion in Environmental Sustainability*, 4, 195–202.

Lambin, E.F. & Meyfroidt, P. (2011) Global land use change, economic globalization, and the looming land scarcity. *Proceedings of the National Academy of Sciences*, 108, 3465–3472.

Larned, S.T., Snelder, T., Unwin, M.J. & McBride, G.B. (2016) Water quality in New Zealand rivers: current state and trends. *New Zealand Journal of Marine and Freshwater Research*, 50, 1–29.

Lee, C.J., Hirsch, R.M., Schwarz, G.E., Holtschlag, D.J., Preston, S.D., Crawford, C.G. *et al*. (2016) An evaluation of methods for estimating decadal stream loads. *Journal of Hydrology*, 542, 185–203.

Lee, C.J., Murphy, J.C., Crawford, C.G. & Deacon, J.R. (2017). *Methods for computing water-quality loads at sites in the U.S. geological survey national water quality network*. U.S. Geological Survey.

Lehner, B. (2014) *HydroBASINS*. McGill University.

Lehner, B. & Grill, G. (2013) Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27, 2171–2186.

Lu, C. & Tian, H. (2017) 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.

Macintosh, K.A., Mayer, B.K., McDowell, R.W., Powers, S.M., Baker, L.A., Boyer, T.H. *et al*. (2018) Managing diffuse phosphorus at the source versus at the sink. *Environmental Science & Technology*, 52, 11995–12009.

Mayorga, E., Seitzinger, S.P., Harrison, J.A., Dumont, E., Beusen, A.H.W., Bouwman, A.F. *et al*. (2010) Global Nutrient Export from WaterSheds 2 (NEWS 2): Model development and implementation. *Environmental Modelling & Software*, 25, 837–853.

McDowell, R.W., Dils, R.M., Collins, A.L., Flahive, K.A., Sharpley, A.N. & Quinn, J. (2016) A review of the policies and implementation of practices to decrease water quality impairment by phosphorus in New Zealand, the UK, and the US. *Nutrient Cycling in Agroecosystems*, 104, 289–305.

McDowell, R.W., Noble, A., Pletnyakov, P., Haggard, B.E. & Mosley, L.M. (2020) Global mapping of freshwater nutrient enrichment and periphyton growth potential. *Scientific Reports*, 10, 3568.

McDowell, R.W., Snelder, T.H., Cox, N., Booker, D.J. & Wilcock, R.J. (2013) Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. *Marine & Freshwater Research*, 64, 387–400.

McDowell, R.W. & Wilcock, R.J. (2008) Water quality and the effects of different pastoral animals. *New Zealand Veterinary Journal*, 56, 289–296.

Meybeck, M. (1982) Carbon, nitrogen, and phosphorus transport by world rivers. *American Journal of Science*, 282, 401–450.

Meybeck, M. & Ragu, A. (1997). *Presenting the GEMS-GLORI, a compendium of world river discharge to the oceans* (pp. 3–14). Freshwater Contamination 1997 Rabat, Morocco. IAHS Publications.

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N. & Foley, J.A. (2012) Closing yield gaps through nutrient and water management. *Nature*, 490, 254.

Ockenden, M.C., Deasy, C.E., Benskin, C.M.H., Beven, K.J., Burke, S., Collins, A.L. *et al*. (2016) Changing climate and nutrient transfers: Evidence from high temporal resolution concentration-flow dynamics in headwater catchments. *Science of the Total Environment*, 548–549, 325–339.

Powers, S.M., Chowdhury, R.B., Macdonald, G.K., Metson, G.S., Beusen, A.H.W., Bouwman, A.F. *et al.* (2019) Global opportunities to increase agricultural independence through phosphorus recycling. *Earth's Future*, 7, 370–383.

Roy, E.D., Richards, P.D., Martinelli, L.A., Coletta, L.D., Lins, S.R.M., Vazquez, F.F. *et al*. (2016) The phosphorus cost of agricultural intensification in the tropics. *Nature Plants*, 2, 16043.

Sayre, R., Dangermond, J., Frye, C., Vaughan, R., Aniello, P., Breyer, S. *et al*. (2014) *A new map of global ecological land units - An ecophysiographic stratification approach*. Association of American Geographers.

Scarsbrook, M.R., McBride, C.G., McBride, G.B. & Bryers, G.G. (2003) Effects of climate variability on rivers: Consequences for long term water quality analysis. *JAWRA Journal of the American Water Resources Association*, 39, 1435–1447.

Seitzinger, S.P., Mayorga, E., Bouwman, A.F., Kroeze, C., Beusen, A.H.W., Billen, G. *et al.* (2010) Global river nutrient export: A scenario analysis of past and future trends. *Global Biogeochemical Cycles*, 24, https://doi.org/10.1029/2009GB003587.

Smith, R.A., Alexander, R.B. & Schwarz, G.E. (2003) Natural background concentrations of nutrients in streams and rivers of the conterminous United States. *Environmental Science & Technology*, 37, 3039–3047.

Smith, S.V., Swaney, D.P., Buddemeier, R.W., Scarsbrook, M.R., Weatherhead, M.A., Humborg, C. *et al*. (2005) River nutrient loads and catchment size. *Biogeochemistry*, 75, 83–107.

Smith, S.V., Swaney, D.P., Talaue-Mcmanus, L., Bartley, J.D., Sandhei, P.T., McLaughlin, C.J. *et al*. (2003) Humans, hydrology, and the distribution of inorganic nutrient loading to the ocean. *BioScience*, 53, 235–245.

Snelder, T.H., Larned, S.T. & McDowell, R.W. (2018) Anthropogenic increases of catchment nitrogen and phosphorus loads in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 52, 336–361.

Snelder, T.H., McDowell, R.W. & Fraser, C.E. (2017) Estimation of catchment nutrient loads in New Zealand using monthly water quality monitoring data. *JAWRA Journal of the American Water Resources Association*, 53, 158–178.

Stevenson, R.J., Bennett, B.J., Jordan, D.N. & French, R.D. (2012) Phosphorus regulates stream injury by filamentous green algae, DO, and pH with thresholds in responses. *Hydrobiologia*, 695, 25–42.

Strokal, M., Kroeze, C., Li, L., Luan, S., Wang, H., Yang, S. *et al*. (2015) Increasing dissolved nitrogen and phosphorus export by the Pearl River (Zhujiang): a modeling approach at the sub-basin scale to assess effective nutrient management. *Biogeochemistry*, 125, 221–242.

Tipping, E., Benham, S., Boyle, J.F., Crow, P., Davies, J., Fischer, U. *et al*. (2014) Atmospheric deposition of phosphorus to land and freshwater. *Environmental Science: Processes & Impacts*, 16, 1608–1617.

United Nations Environment Programme (2018). *GEMStat database of the Global Environment Monitoring System for Freshwater (GEMS/Water) Programme*. International Centre for Water Resources and Global Change. Retrieved from https://gemstat.org [Accessed 22 December 2018].

United States Department of the Interior - United States Geological Survey (2008). *HydroSHEDS*. U.S. Dept. of the Interior, U. S. Geological Survey.

van Meter, K.J. & Basu, N.B. (2017) Time lags in watershed-scale nutrient transport: an exploration of dominant controls. *Environmental Research Letters*, 12, 084017.

van Puijenbroek, P.J.T.M., Beusen, A.H.W. & Bouwman, A.F. (2019) Global nitrogen and phosphorus in urban waste water based on the Shared Socio-economic pathways. *Journal of Environmental Management*, 231, 446–456.

Willmott, C.J. & Matsuura, K. (2001). *Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950 - 1999)* [Online]. University of Delaware: University of Delaware. Retrieved from http://climate.geog.udel.edu/~climate/html_pages/download.html [Accessed December 2018]

Worrall, F., Jarvie, H.P., Howden, N.J.K. & Burt, T.P. (2016) The fluvial flux of total reactive and total phosphorus from the UK in the context of a national phosphorus budget: Comparing UK river fluxes with phosphorus trade imports and exports. *Biogeochemistry*, 130, 31–51.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: McDowell RW, Noble A, Pletnyakov P, Mosley LM. Global database of diffuse riverine nitrogen and phosphorus loads and yields. *Geosci Data J*. 2020;00:1–12. https://doi.org/10.1002/gdj3.111