Drivers of global microplastic concentrations in rivers

A comparison of the multiple linear regression and random forest regression approach

Master Thesis Simon B. M. Smits, BSc. 6260438

First supervisor: Dr. Michelle van Vliet Second supervisor: Dr. Marcel van der Perk Master: Earth Surface and Water Track: Coastal Dynamics and Fluvial Systems Number of ECTS: 30 February 7th, 2022 – August 21st, 2022



Abstract

Microplastic particles in rivers negatively influence the overall water quality and have a harming effect on biotic species living in aquatic ecosystems. Global microplastic concentrations in rivers are often determined by various area-specific drivers, from different driver categories. Examples are population density, gross domestic product and the amount of mismanaged plastic waste (socio-economic driver category). Other examples are precipitation, surface runoff and river streamflow (hydrologic driver category). Improved understanding of these drivers is therefore urgently needed in determining and counteracting the increasing amount of riverine plastic pollution. Previous literature has shown that both human- and natural processes tend to have a strong influence on microplastic concentrations in rivers and oceans. In this study, an extensive literature study on microplastic hotspots worldwide is carried out. Microplastic hotspots in the world can be found in East and Southeast Asia, India, Central Africa, Europe and North America. Estimates of global riverine microplastic outputs are between 1.15 and 2.41 million tonnes per year with the largest contribution from rivers in Asia.

In addition, a data analysis of microplastic concentrations in rivers worldwide is carried out in relation to four socio-economic drivers, three hydrologic drivers and six land use drivers. For these analyses, the multiple linear regression, and the random forest regression approach are applied in order to determine the most important and significant drivers. Both regression approaches showed that drivers from the socio-economic category had the highest contribution in explaining concentrations of microplastic in rivers, with gross domestic product and mismanaged plastic waste production being the most important drivers. In this study, conclusions are drawn on the importance of various drivers on microplastic concentrations, which can contribute to understand the origin and fate of microplastics in the environment and set up mitigation strategies in the future.

Abbreviations

MLR		Multiple linear regression
RF		Random forest
PD		Population density
GDP		Gros domestic product
MPW		Mismanaged plastic waste production
WWT	I	Wastewater treatment
R		Surface runoff
HDI		Human development index
PP		Polypropylene
PE		Polyethylene
PET		Polyethylene terephthalate
PS		Polystyrene
PA		Polyamide
MSE		Mean squared error
nRMSE	I	Normalised root mean squared error

1.	INTR	ODUCTION	4			
2.	LITE	RATURE STUDY	7			
	2.1	FATE AND CHARACTERISTICS OF (MICRO)PLASTIC IN THE ENVIRONMENT	8			
	2.2	APPROACHES TO ESTIMATE MICROPLASTICS IN RIVERS WORLDWIDE	11			
	2.3	MICROPLASTIC HOTSPOTS IN RIVERS WORLDWIDE	13			
3.	MET	HODS FOR REGRESSION ANALYSES	17			
	3.1	Monitoring data	17			
	3.2	DRIVER DATA	18			
	3.2.1	Socio-economic driver data	19			
	3.2.2	Hydrologic driver data	21			
	3.2.3	Land use data	21			
	3.3	DATA CONVERSIONS	22			
	3.4	MULTIPLE LINEAR REGRESSION	24			
	3.5	RANDOM FOREST REGRESSION	25			
	3.6	MICROPLASTIC PREDICTIONS	28			
4.	REG	RESSION RESULTS	30			
	4.1	GLOBAL DRIVER DATA CORRELATION	30			
	4.1.1	Population density	32			
	4.1.2	Mismanaged Plastic Waste production	32			
	4.1.3	Wastewater treatment	32			
	4.1.4	Gross domestic product	33			
	4.1.5	Hydrologic driver data correlations	33			
	4.2	MULTIPLE LINEAR REGRESSION	36			
	4.3	RANDOM FOREST REGRESSION	37			
	4.4		39			
_	4.5		40 45			
5.	DISC		45			
	5.1	MAIN SOURCES OF UNCERTAINTIES	45			
	5.2	IMPLICATIONS AND RECOMMENDATIONS OF FUTURE MICROPLASTIC RESEARCH	4/			
6.	CONCLUSIONS4					
7.	APPE	NDICES	52			
	7.1	OVERVIEW AND DESCRIPTION OF THE APPENDICES	52			
	Арре	ndix A1 Global gridded population density data in 2020 in number of persons per km ²	53			
	Арре	ndix A2 Global gridded Mismanaged Plastic Waste production in 2019 in tonnes per year	54			
	Appe	ndix A3 Global gridded Wastewater Treatment discharge data in m³/day	55			
	Арре	ndix A4 Global gridded GDP per capita in 2015 in US dollars	56			
	Арре	ndix A5 Global gridded precipitation data form 2014 in mm	57			
	Арре	ndix A6 Global gridded surface runoff data from 2014 in mm/day	58			
	Арре	naix A7 Global gridded streamflow data from 2014 in m²/s	59			
	Арре	naix A8 Global gridded non-forest land use fraction in fraction per pixels	60			
	Арре	muix A9 Global gridded crop land use fraction in fraction per pixel	61			
	Appenaix A1U Global gridaea crop land use fraction in fraction per pixel					
	Appendix A11 Global gridded rangeland land use fraction in fraction per pixel					
	Appendix A12 Global gridded nastureland land use fraction in fraction per pixel					
	Appendix A15 Global griaded postal elana iana ase fraction in fraction per pixer					
	Ασος	ndix B2 Residuals versus model fit as determined by the MLR approach	00 67			
8	RFFF	RENCES	68			
· · ·						

1. Introduction

To meet the global human plastic demand, wide-scale use of plastic material has increased strongly since the 1950's. Where the share of plastic in municipal solid waste in 1960 was only 1%, this percentage has increased to almost 10% in 2005 (Geyer et al., 2017). Even though plastic consumption in developed countries shows a consistent pattern nowadays, the abundance of plastic in developing countries is still rising fast due to low production costs and the lack of alternative materials (Andrady & Neal, 2009). The abundance of plastic and thus the amount of mismanaged plastic waste (MPW) depends on multiple socio-economic drivers such as population density, gross domestic product (GDP), waste generation and the percentage of plastic present in general waste (Hoornweg & Bhada-Tata, 2012; Schmidt et al., 2017). Non-natural properties of plastic make MPW susceptible to accumulation in natural systems, such as rivers and oceans.

To improve the functionality of plastic, it is produced in different sizes. A distinction is made between macro-, meso-, micro- and nanoplastics. Macroplastics are plastic particles with a size larger than 5 cm. Mesoplastics have a size of 5 mm to 5 cm. In addition to this, microplastics (0.1 μ m to 5 mm) and nanoplastics (< 0.1 μ m) are distinguished (van Emmerik & Schwarz, 2020). Degradation of the two largest groups results in increased amounts of micro- and nanoplastics (van Wijnen et al., 2019). Therefore, micro- and nanoplastics are often categorized as plastic waste as they are originating from the larger plastic categories. The smaller size of the particles in micro- and nano-category makes collection and recycling less productive.

Accumulation of mismanaged plastic waste into river systems depends on multiple hydrologic drivers. Examples of these drivers are precipitation, surface runoff and river discharge. Precipitation mobilises plastic particles located on land surfaces (Meijer et al., 2021). Surface runoff after rainfall events can transport plastic particles towards rivers and streams (Lebreton et al., 2017). River discharge is important in transporting plastic particles from rivers towards the oceans (Schmidt et al., 2017).

Furthermore, land use characteristics and especially canopy cover determined by the type of land use, tend to influence soil erosion processes (Hartanto et al., 2003). Erosion of the soil by surface runoff and rainfall enhances microplastic particles to enter aquatic ecosystems. For other pollutants such as nitrogen, phosphorus and lead, the correlation between the land use type and the surface water quality has been outlined in multiple other studies (Adeola Fashae et al., 2019; Liu et al., 2009; Tong & Chen, 2002). A positive correlation has been found multiple times for 'open' land use types where human impact strongly changed the landcover.

Wastewater treatment (WWT) on domestic and industrial water can be applied to decrease the amount of plastic waste entering the environment. Depending on the method used, the efficiency of microplastic removal can reach a value between 80 and 90 percent (Ngo et al., 2019).

The urge to better understand the transport of (micro)plastics from land to rivers has been acknowledged, resulting in more research on this topic in recent years (Lebreton et al., 2017; Meijer et al., 2021; Schmidt et al., 2017; van Wijnen et al., 2019). These studies have resulted in different types of models to quantify the amount of micro- and macroplastics and specify the spreading of plastics that accumulate into aquatic systems. Examples are modelling studies which calculate pollutant loadings of plastics to streams and rivers using socioeconomic drivers such as population density, mismanaged plastic waste production and connection to wastewater treatment (Schmidt et al., 2017; van Wijnen et al., 2019). On the other hand, a hydrological modelling approach on this topic has also been used to calculate transport of plastics in aquatic systems (Meijer et al., 2021). Here, the land surface part of the hydrological cycle is represented and hydrological output (e.g., surface runoff, discharge) are used to quantify plastic mobilisation and transport both over land and through rivers to eventually calculate plastic concentrations in rivers. Both types of models are applied on a global scale (Lebreton et al., 2017; Meijer et al., 2021; Schmidt et al., 2017; van Wijnen et al., 2019). However, a combined approach to estimate the contribution of both socio-economic drivers, hydrologic drivers, and land use characteristics on microplastic pollution has not been developed yet. This thesis will investigate the difference in importance and contribution of these various drivers to fill the knowledge gap that is present to date.

One method to estimate the amount of microplastics in rivers is by using a statistical data analysis of microplastic monitoring data and driver data. Such analyses can be done by applying multiple approaches. Multiple linear regression (MLR) is a commonly used method to analyse the relation between a dependent (predicted) variable and multiple independent (predictor) variables, assuming a linear relation between the different predictor variables (Boy-Roura et al., 2013). Via this method and the eventual regression model performance analysis, importance and significance of the predictor variables can be estimated. Random forest regression (RF regression) is a more contemporary regression method based on a machine learning algorithm analysing data using large numbers of random decision trees (Breiman, 2001; Rodriguez-Galiano et al., 2014).

To fill the presented knowledge gap on microplastic concentrations in rivers and its drivers, this study will estimate the influence of socio-economic-, hydrologic- and catchment land use

drivers on global microplastic concentrations in rivers comparing the multiple linear regression and random forest regression approach. Both approaches are tested on monitoring data of riverine microplastic amounts of 59 locations and spatially explicit data of socio-economic-, hydrologic- and land use drivers. Such a combined approach using various driver categories has not yet been assessed in scientific research. To address this aim the following research questions will be answered:

- 1. What are hotspot regions of microplastic concentrations in rivers worldwide?
- 2. What is the contribution of socio-economic drivers, hydrologic drivers, and land use characteristics to microplastic loads in rivers, according to the multiple linear regression and random forest regression approach?
- 3. How do the results of estimated importance of the various drivers of both methods compare?
- 4. What is the performance of both methods for predicting microplastic loads in rivers in other regions of the world?

Answering these questions will eventually result in the ability to evaluate the most important drivers that are responsible for microplastic particles in rivers worldwide. These results will than make it possible to mitigate to microplastic pollution more efficiently, only considering the drivers that are most responsible. Research question 1 will be answered using a literature study presented in chapter 2. Here the spatial distribution of hotspots together with the origin, fate and the transport mechanisms of microplastics in the environment will be discussed. The methodology to gather the results for research question two to four is outlined in chapter 3. Research question two to four will be answered in chapter 4 by analysing the results of the multiple linear regression and the random forest regression. In chapter 5, the results of both regression analyses will be discussed, and implications of the used methods and approaches will be analysed.

2. Literature Study

A literature study has been conducted to answer the first research question concerning hotspots of microplastic concentrations worldwide. Search terms used to find literature in the Web of Science database and Google Scholar were 'microplastic* rivers' combined with 'Oceans', 'Transport', 'Modelling' and 'Hotspots'. Merging 'microplastic* rivers' with 'modelling' resulted in 38 articles of which seven applied a modelling approach to determine microplastic transport from rivers to the oceans worldwide. An overview of the articles is given in Table 1. Four of these studies used driver and monitoring data to build regression models. Three of the four studies using regression especially investigated the influence of socio-economic driver and ignored the possible influence of hydrologic drivers or land use characteristics (Jambeck Jenna R. et al., 2015; Mai et al., 2020; Schmidt et al., 2017). Only one study considered the influence of surface runoff on the mobility of microplastics and was therefore able to investigate seasonality in microplastic loads influenced by monsoonal precipitation (Lebreton et al., 2017). Three of the seven modelling studies developed a process-based model to determine microplastic loads in rivers. These studies only considered socio-economic drivers as model inputs, again ignoring the possible influence of hydrologic drivers (Meijer et al., 2021; Siggfried et al., 2017; van Wijnen et al., 2019). In addition to the seven regression or processbased studies, two monitoring studies were considered where visual counting of microplastic particles in rivers at different locations was implemented (Eriksen et al., 2014; van Calcar & van Emmerik, 2019). Eriksen et al. (2014) used neuston nets with a standard mesh size of 0.33 mm towed at the sea surface outside of a vessel. Van Calcar and Van Emmerik (2019) applied visual counting by observing all plastic particles passing through a predefined section of the river. These observations were done on bridges. Different measuring methods may show different results which makes it important to discuss the various results on microplastic counting in both rivers and oceans. In addition to the modelling studies described in Table 1, other studies investigating the fate, transport capacity and harming characteristics of microplastics will be used and referred to with in-text citations.

Table 1 Overview presenting the title, method, considered drivers and references of the applied literature. Abbreviations represent mismanaged plastic waste (MPW), population density (PD), surface runoff (R) and human development index (HDI).

Article title	Method	Drivers	Reference
River plastic emissions to the world's oceans	Regression	MPW, PD,	(Lebreton et al., 2017)
Export of Plastic Debris by Rivers into the Sea	Regression model	MPW	(Schmidt et al., 2017)
Plastic waste inputs from land into the ocean	Regression model	MPW, PD	(Jambeck Jenna R. et al., 2015)
Global Riverine Plastic Outflows	Regression model	PD, MPW, HDI	(Mai et al., 2020)
Modelling global river export of microplastics to the marine environment: Sources and future trends	Process-based model		(van Wijnen et al., 2019)
More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean	Process-based model		(Meijer et al., 2021)
Export of microplastics from land to sea. A modelling approach	Process-based model		(Siegfried et al., 2017)
Abundance of plastic debris across European and Asian rivers	Monitoring based on visual counting		(van Calcar & van Emmerik, 2019)
Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces weighing over 250.000 Tons Afloat at Sea	Monitoring based on visual counting		(Eriksen et al., 2014)

2.1 Fate and characteristics of (micro)plastic in the environment

Plastic is human-made material consisting of large polymers, produced using non-natural chemical reactions (van Emmerik & Schwarz, 2020). This production process enables plastic to be a flexible, strong, lightweight and thermoresistant material that can be used for a range of different purposes. Development of the modern-day plastics takes place since the 1950's. In this period, different polymer types were introduced. The combination of the advantageous properties and the relatively small production costs resulted in a quick emergence of plastic material in today's society. In 2008, the global plastic demand reached a value of 245 million tonnes (Andrady & Neal, 2009). Recent numbers on plastic production rates show an enormous increase in plastic production over the last decade. Where plastic demand in 2008 was 245 million tonnes, plastic production in 2018 reached a value of 348 million tonnes (van Emmerik & Schwarz, 2020).

Plastic material can be subdivided in different groups depending on chemical composition and size. The main types of plastic produced nowadays are polypropylene (PP), polyethylene (PE), polyethylene terephthalate (PET), polystyrene (PS) and polyamide (PA) (van Emmerik

& Schwarz, 2020). PP, PE and PET composites are mostly used in the packaging industry since these materials are easily formable, can be hermetically sealed and are hydrophobic. It is therefore often used to package food. The application of these plastic types in the packaging industry has resulted in reducing transport costs and expanding expiring dates. PS and PA are mostly applied in textiles (van Emmerik & Schwarz, 2020).

The non-natural chemical reactions used to produce plastic material, ensure that natural degradation cannot take place, making accurate collection of plastic waste extremely important. Where the share of plastic in municipal solid waste in 1960 was only 1%, this percentage has increased to almost 10% in 2005 (Geyer et al., 2017). These increasing numbers can be coupled to the rapid increase in plastic production since the 1950's (Andrady & Neal, 2009). When collection or recycling of plastic waste does not take place and the disposal of plastic waste is not well managed, solid plastic waste becomes mismanaged plastic waste (MPW). The abundance of plastic and thus the amount of MPW depends on multiple socio-economic factors. Examples are population density, gross domestic product (GDP), waste generation and the percentage of plastic present in general waste (Hoornweg & Bhada-Tata, 2012; Schmidt et al., 2017).

Regions that have a large population density can be found in East and Southeast Asia, Northern India, Northwest Europe and the east coast of the United States (CIESIN - Columbia University, 2018). A correlation can be found in the MPW values for the same regions. MPW



Figure 1 Mismanaged plastic waste generation in 2015 in tonnes per year as presented by Lebreton & Andrady (2019).

generation shows largest amounts in East and Southeast Asia and Northern India, where MPW generation easily exceeds 200.000 tonnes per year (Figure 1) (Lebreton & Andrady, 2019). However, for Northwest Europe and the east coast of the US, Lebreton and Andrady

(2019) visualise smaller MPW generation numbers (Figure 1). In such regions more investments are done to filter municipal solid waste in order to recycle plastic materials, preventing them from becoming MPW (Hoornweg & Bhada-Tata, 2012).

Accumulation of MPW into river systems depends on multiple hydrologic drivers. Examples of these drivers are precipitation, surface runoff and river discharge. Precipitation mobilises plastic particles located on land surfaces (Meijer et al., 2021). To explain this process, mobilisation and transport of plastic is coupled to the transport behaviour of natural sediments (Waldschläger et al., 2022). Waldschläger et al. (2022) compare the behaviour of natural sediments to tackle microplastic challenges. The impact of falling raindrops on the soil can start a process known as hydraulic soil erosion (Nearing et al., 2005). Natural sediments on land surfaces react to this precipitation impact in two different ways. It reacts to the total amount of rain that falls in the region. Therefore, long-term precipitation periods result in larger erosive potentials which enhances soil erosion capacities of a certain region (Nearing et al., 2005).

Surface runoff after rainfall events transports plastic particles towards rivers and streams (Lebreton et al., 2017). Areas where precipitation is able to (re)mobilise soil particles, surface runoff acts as transport mechanism for soil particles, resulting in higher soil erosion rates (Arnaez et al., 2007). When plastic particles are included in the soil or lay on the soil, the same transport mechanism is responsible for the movement of these particles.

When eventually plastic particles enter rivers and streams via these land surface processes, river discharge plays a role in transporting it through the rivers towards the oceans and seas (Schmidt et al., 2017). This driver is therefore important for investigating plastic amounts in oceans but also for analysing the movement of plastic from the upstream area of a river to the downstream area. During peak discharges, transport of microplastic follows the discharge trend, resulting in a higher transport capacity of plastic in the downstream direction (Hurley et al., 2018). In addition, peak discharges of river system often go hand in hand with increased surface runoff, which again enhances microplastic concentrations (Arnaez et al., 2007).

Land use characteristics tend to influence the transport and mobilisation of polluting substances by surface water processes. For other pollutants than microplastics, such as nitrogen, phosphorus and lead, the correlation between the land use type and the surface water quality has been the focus of multiple studies (Adeola Fashae et al., 2019; Liu et al., 2009; Tong & Chen, 2002). A positive correlation has been found multiple times for 'open' land use types where human impact strongly changed the landcover. Examples of these land use types are pastureland, cropland, urban areas and waste ground. Here, trees and shrubs are often missing which enhances surface runoff flow velocities, resulting in an increased transport

capacity and soil erosion. In addition to this, earlier mentioned land use types tend to have a decreased canopy cover, which normally reduces the impact of raindrops on the surface. A lack of leaves therefore increases mobilisation of polluting particles, resulting in a larger chance of these particles ending up in the water environment. This also applies for microplastics (Townsend et al., 2019).

Eventually all these drivers result in an increased abundance of plastic waste particles in aquatic environments worldwide. The harming effect of microplastic on biotic and abiotic species in these aquatic environments disturbs natural processes, resulting in reduced functionality of the internal systems of varying species (van Emmerik & Schwarz, 2020). Especially the smaller plastic particles (< 1cm) can easily infiltrate into the food chain of the small aquatic species which will also lead to a disturbed ecosystem for the larger species living in rivers and streams.

To reduce plastic pollution in the aquatic system, treatment of domestic and industrial wastewater can be applied (Talvitie et al., 2017). Especially for the larger plastic particles from the macro- and mesoplastic groups, purification of wastewater is more efficient. These larger groups are filtered out more easily, resulting in a reduced amount of macro- and mesoplastic particles in aquatic systems after the treatment plant (Rasmussen et al., 2021). Since known that a vast amount of microplastic originates from the degradation of the larger plastic particle groups, treatment of largest category will also result in a reduced amount of microplastics. However, treatment of wastewater is not applied in the same way around the globe and seems to have a strong correlation with GDP (Jones et al., 2021). Therefore, treatment of wastewater is done on larger scales in regions such as North America and Northwest Europe. In Asia and South America, a smaller part of wastewater is treated, resulting in more polluted water reentering the environmental systems (Jones et al., 2021). Numbers on wastewater treatment are 50-100% for North America and Northwest Europe and 25-50% for Asia and South American regions. Africa has the lowest wastewater treatment percentages varying between 0 and 25%. However, in Egypt, Tunis and Morocco 25% to 50% of the total amount of wastewater is treated (Jones et al., 2021).

2.2 Approaches to estimate microplastics in rivers worldwide

To date, the most robust estimates of plastic concentrations are obtained by research where plastic concentrations are visually monitored. With this method the number of plastic particles in riverine or ocean ecosystems is visually counted or investigated using nets to get the plastic out of the water (Eriksen et al., 2014; Mani et al., 2015). Microplastic samples can directly be elaborated and analysed. Therefore, the quality of this type of data often reaches the desired accuracy. However, the quantity of monitoring data is often scarce and forces researchers to

reduce the spatial scale of their study. For this reason, it is still not possible to investigate microplastic concentrations in rivers on a global scale using only monitoring data.

To work towards a solution for this problem, the development of process-based- and regression models to describe microplastic concentrations in rivers has become an important topic for researchers in the field of microplastics worldwide. Especially the recently increasing numbers of articles focussing on the development of these models demonstrate the urgency of large-scale microplastic research (Chen et al., 2021; Lebreton et al., 2017; Meijer et al., 2021; Schmidt et al., 2017). To add to this, marine microplastic research is far more extensive than freshwater microplastic research (Chen et al., 2021). Validated models give the opportunity to determine microplastic concentrations in rivers where no monitoring data is available. Therefore, the scale of research can be increased making it possible to not only focus on one river, but to broaden the scope to multiple rivers. It is also possible to development multiple models, with each model focussing on different drivers. For example, Lebreton et al. (2017) developed a regression model with mismanaged plastic waste, population density and surface runoff as drivers. On the other hand, Meijer et al. (2021) applies a process-based model to drivers such as plastic waste, land use, wind, precipitation and rivers. A disadvantage of these type of model is that most models do not consider all drivers that are influencing the amount of plastic in aquatic systems. This often leads to results with larger uncertainties.

Due to the improving quality of process-based models, a shift can be seen in modern-day microplastic research. As a consequence, monitoring research now often focuses on macroand mesoplastics, since those categories are better visible with the naked eye and do not require a microscope. Researching macro- and mesoplastic hotspots can provide understanding of the spatial distribution of microplastic (van Wijnen et al., 2019). However, for microplastics, research methodologies nowadays often use validated models.

During most recent years, the development of conceptual modelling and the increase in the amount of data that drive global plastic numbers, have created opportunities to analyse (micro)plastic, reducing the amount of monitoring equipment and expensive expeditions to gather data. New monitoring data will always remain necessary to validate the conceptual models build nowadays. However, feeding models with data of drivers that is already available, enables the possibility to calculate plastic amounts in regions where availability of monitoring data is scarce. Process-based models also allow the production of future projections of microplastic concentrations.

Regression models, as used in this study, have also shown increasing importance in recent microplastic research (Jambeck Jenna R. et al., 2015; Lebreton et al., 2017; Mai et al., 2020; Schmidt et al., 2017). Due to the easy applicability of varying regression equations, the statistical relationship between varying drivers and a predicted variable can be assessed. Therefore, it does not require complicated equations describing the process as is the case with process-based models. When a statistical relationship is found, new driver data can be used to determine plastic concentrations at locations where monitoring data is scarce. Also, when future driver data is available, future projections of microplastic concentrations can be produced. However, a disadvantage is that regression analyses need large amounts of initial driver data and predicted variable data to find the statistical relationship. When these data are not available, the application of regression models will not lead to the desired result.



Figure 2 Global plastic inputs from rivers to sea in tonnes per year according to Lebreton et al. (2017).

2.3 <u>Microplastic hotspots in rivers worldwide</u>

Global (micro)plastic loads exported from rivers to the oceans as determined by the regression studies are between 1.15 and 2.41 million tonnes per year (Lebreton et al., 2017; Schmidt et al., 2017). The 20 most polluting rivers are located in Asia and contribute for more than 2/3 of the global annual plastic input to oceans. Both studies considered micro- as well as macroplastics. Global annual input of only microplastics was estimated to be 0.16 million tonnes (Schmidt et al., 2017). However, regression analyses done by Schmidt et al. (2017) only considered the amount of mismanaged plastic waste as input variable for the regression model. Regression studies executed by Lebreton et al. (2017) also took into account monthly averaged runoff in addition to the amount of plastic waste in the catchment. (Micro)plastic

hotspots are located in Asia, Southeast Asia, India, Central Africa and the east coast of South America (Figure 2).

For studies that used conceptual modelling in their methodology, results on microplastic loads in rivers show different values. One of the most notable conclusions from one of these studies is that more than 1000 rivers account for 80% of the global riverine plastic emissions into the oceans (Meijer et al., 2021). These findings are completely different compared to the top 20 rivers accounting for 2/3 of the global annual plastic input as concluded by Lebreton et al. (2017). Conceptual modelling showed that annual emission of plastic was 0.8 to 2.7 million metric tonnes in 2015, which is the same order of magnitude as the findings in the regression studies (Lebreton et al., 2017; Schmidt et al., 2017). Again, hotspots are located in Asia, Southeast Asia, India and the east coast of South America where The Philippines and India are responsible for the largest pollution of plastic into the ocean with 356.000 and 126.000 metric tonnes per year, respectively. Results from Meijer et al. (2021) are shown in Figure 3.



Figure 3 National riverine plastic emissions in metric tonnes per year according to Meijer et al. (2021).

Determining plastic pollution hotspots from studies that only use visual counting requires evaluation of large amounts of studies to achieve an overview of where more plastic particles are counted. For this research, studies that used visual counting of (micro)plastic particles have been reviewed. The fact that most modelling and regression studies required monitoring data to validate their models, papers describing visual counting of plastics were easily findable in the already applied articles. For the analyses, the Rhine River, Danube River and Yangtze River are compared. The number of plastic particles per m³ for these rivers are 4.92·10⁰, 8.23·10⁻¹ and 4.14·10³, respectively (Schmidt et al., 2017). As shown in these results the Yangtze River, located in East Asia, contains the highest amount of microplastic particles, supporting the hypothesis of Asia being a (micro)plastic hotspot. However, only considering

these three rivers is not convincing enough to point out microplastic hotspots with visual counting techniques.

Therefore, the AdventureScientists.org database was used to support the studies that applied visual counting for determination of (micro)plastic concentrations. This database consists of a large amount of plastic particle measurements both in rivers and in oceans. Counting is done by volunteer researchers and it expresses the number of pieces microplastic per litre (Christiansen, 2018).

The data can be accessed via AdventureScientists.org/microplastics.html and shows a world map containing all measurements done for this project (Figure 4). It shows the large amount of datapoints that is analysed. Focus of the data is pointed to oceanic monitoring locations, especially around the USA and in the Atlantic Ocean between Central America and North Africa. Near the east coast of South America and in the Southeast Asian region, oceanic microplastic measurements were also executed. Riverine monitoring is lacking in most parts of the world, except for the USA and some locations in Europe, South America and India/Bangladesh. From the riverine datapoints, it is impossible to distinguish microplastic



Figure 4 Marine and freshwater microplastic concentrations as determined by the AdventureScientists.org monitoring campaign (Christiansen, 2018). Green dots represent marine microplastic measurements, blue dots represent freshwater microplastic measurements.

hotspots as the spatial distribution of the monitoring locations is too poor. High concentrations of microplastics in rivers can be found in the early stages of the Amazone River and in the Ganges/Brahmaputra Delta. However, from Figure 4 it is unclear whether these regions can be assigned as hotspots, considering the low amount of measuring points.

For the oceanic datapoints, the spatial distribution and the quantity of measurements allows a better interpretation of the data. Hotspots can be found around the east and west coast of the USA, in northern Canada, and in the Southeast Asian area. Especially in the Beaufort Sea, north of Canada and Alaska, multiple samples were collected with a microplastic concentration exceeding a value of 100 pieces per litre. For Southeast Asia, only one measurement contained more than 100 pieces per litre.

A disadvantage of these oceanic datapoints is the uncertainty of the origin of the plastic material. Due to the ocean circulation currents that are present around the globe, plastic that is found at a certain location can have a different origin location. For rivers, it can be said that the plastic particle has its origin within the upstream catchment area, ignoring the possible influence of wind on the spreading of plastic. For the oceanic plastic particles, the location where monitored plastic will eventually be measured is strongly depending on the ocean currents that are present in the part of the world where the datapoint is achieved. In addition, climatic circulation patterns vary over time and space, making the route of a plastic particle in the ocean unpredictable (Welden & Lusher, 2017).

3. Methods for regression analyses

3.1 Monitoring data

To estimate microplastic concentrations in rivers based on socio-economic-, hydrologic- and land use driver data, monitoring data of microplastic concentrations from rivers was gathered at 59 different locations. For the composition of this monitoring database, global spatial distribution of monitoring data was important in the process of collecting a heterogeneous dataset. Especially in the eventual process of combining monitoring data with driver data, a heterogeneous set of numbers on driver data is desired in order to reduce correlation between various drivers and produce a stable regression model (Disatnik & Sivan, 2016).

In this study, monitoring data from 59 locations presented in two different studies was implemented with a spatial distribution over North- and South America, Europe and Asia (Jiang et al., 2019; Schmidt et al., 2017). 55 Monitoring locations were obtained from Schmidt et al. (2017) spread across North America, South America, Europe, and Asia. Approximately 60% of these monitoring data points were conducted in rivers around the Great Lakes in the United States and 16 other samples were taken in the Rhine and Seine River. The other 4 monitoring locations, located on the Tibetan Plateau in Central Asia, were obtained from Jiang et al. (2019). An overview of the monitoring data locations is given in Figure 5.

It was considered important that monitoring data was obtained from rivers with varying flow regimes, hydroclimatic zones and socio-economic conditions. Therefore, rivers with large discharges were included such as the Danube River and the Yangtze River but also smaller scale rivers, such as the Biobio River in Chile were used in the monitoring dataset. Microplastic measurements as presented in the monitoring studies were transformed into concentration with the unit number of microplastic particles per m³ (p/m³). For the data from Schmidt et al. (2017) a unit conversion from n/1000m³ to p/m³ was applied. In other words, concentration magnitudes were multiplied by 0.001. Data from Jiang et al. (2019) was already presented in the unit p/m³, so no conversion was needed. A shapefile of all monitoring locations was added to ArcGIS Pro, containing information on the latitude, longitude and the microplastic concentrations at each location.

All microplastic measurements from Schmidt et al. (2017) were carried out between July 2013 and December 2014, except for the two measurements in Los Angeles, United States. These measurements were done in 2004. Monitoring locations that contained data from multiple samples were averaged, so that for each location one microplastic concentration value was obtained. Data sampling from Jiang et al. (2019) was obtained in July 2018 and consisted of four days of sampling. How many samples eventually were taken is not mentioned in the article. Therefore, it is assumed that these monitoring locations contained more than 17

samples, so more samples than the location from Schmidt et al. (2017) with the highest number of samples.



Figure 5 Overview of the monitoring data locations implemented in the regression analyses. Dots represent monitoring data from Schmidt et al. (2017), triangles represent monitoring data from Jiang et al. (2019). Colour of the dots represents the number of samples taken at each monitoring location.

3.2 Driver data

Monitoring data from the 59 locations was connected to spatially distributed gridded raster data of various drivers to estimate the influence of these drivers on global microplastic concentrations in rivers. Driver data was subdivided into socio-economic driver data (population density, gross domestic product, mismanaged plastic waste production and wastewater treatment) and hydrologic driver data (precipitation, surface runoff and river discharge). In addition, land use data was also considered to be a driver for microplastic input to rivers. Landcover classes implemented to the dataset were non-forest, forest, cropland,

urban land, rangeland and pastureland. The eventual regression models were fitted considering four socio-economic drivers, three hydro-geological drivers and six different land use types (Table 2). All driver data was achieved from publications in scientific research journals. Raster data was uploaded into ArcGIS Pro as NetCDF or TIFF file, so that raster calculations were executable. Raster calculations that were applied to the data will be explained in the next part of this thesis.

Table 2 Overview of the driver data implemented in the regression analyses. The table shows the resolution, unit and reference of the socio-economic, hydrologic- and land use driver data. Bottom row shows the applied catchment area data that is explained in section 3.3. The colours for each driver correspond to the flowchart in Figure 6.

Socio- economic data	Driver dataset	Resolution	Unit	Reference
	GPW population density	30 arcsec	People/km ²	(CIESIN - Columbia University, 2018)
	GDP per capita	5 arcmin	USD	(Kummu et al., 2018)
	MPW production	30 arcsec	Tonnes/year	(Lebreton & Andrady, 2019)
	WWT discharge	15 arcsec	m³/day	(Ehalt Macedo et al., 2022)
Hydrological data	Precipitation	30 arcmin	mm	(Schneider et al., 2015)
	Surface Runoff	30 arcmin	mm/day	(Ghiggi et al., 2019)
	Streamflow	30 arcsec	m³/s	(Barbarossa et al., 2018)
Land use data	Non-Forest Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
	Forest Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
	Crop Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
	Urban Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
	Rangeland Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
	Pasture Fraction	30 arcmin	Fraction per pixel	(Hurtt et al., 2020)
Upstream Catchment Area	HydroBASINS Catchment outlines	500 m spatial resolution	Km ²	(Lehner & Grill, 2013)

3.2.1 Socio-economic driver data

Socio-economic driver data consisted of four global datasets:

- 1) populations density data (CIESIN Columbia University, 2018);
- 2) Gross Domestic Product (GDP) per capita data (Kummu et al., 2018);
- 3) Mismanaged Plastic Waste (MPW) production data (Lebreton & Andrady, 2019); and,

4) wastewater treatment (WWT) data (Ehalt Macedo et al., 2022). An overview of the implemented data is given in Table 2, which also shows the original resolution, units and references of the applied driver datasets. Appendix A1-13 visualises all microplastic concentration drivers on a global grid.

Most monitoring data samples were measured at different moments in time. To deal with this, driver data was matched to the same period as the monitoring data. Therefore, monitoring data from 2004 was matched with driver data from 2004 and monitoring data from 2013/2014 was matched with driver data from 2013/2014. For some driver data sets, the corresponding year to match monitoring data was missing. In this case, the period closest to the monitoring data time period was used, so that temporal scales of monitoring and driver data were as closely linked as possible.

Population density data was obtained from the Center for International Earth Science Information Network (CIESIN) which presented a global gridded dataset for population density and total population for the year 2000, 2005, 2010, 2015 and 2020. Population density in 2015 was used for the monitoring data from 2013/2014. For monitoring data from 2004, population density data from 2005 was implemented. For the monitoring data from Jiang et al. (2019) that was measured in 2018, also population density data from 2015 was used. Raster data was available in TIFF-format, which enabled importation into the ArcGIS database.

Global gridded GDP per capita data was obtained from Kummu et al. (2018), which was a multiyear dataset from 1990 to 2015. The temporal spreading of the data enabled the possibility to couple monitoring data from 2004 and 2013/2014 with GDP per capita data from the same year, with the aim to improve the quality of the eventual regression model. For the monitoring data from 2018, GDP data from 2015 was used.

MPW production data was obtained from Lebreton and Andrady (2019) and consisted of TIFF raster data with a resolution of 30 arcseconds. A temporal dimension is missing in this data, making it impossible to match the MPW production data with the corresponding sampling year. WWT discharge data was obtained from Ehalt Macedo et al. (2022), also known as the HydroWASTE dataset for wastewater treatment data, which is a spatially explicit database consisting of the characteristics of 58,502 wastewater treatment plants. For this study only the discharge of treated wastewater in m³/day was used as influencing driver on the amount of riverine microplastics. HydroWASTE data was downloaded as CSV-file and transformed into raster data using ArcGIS Pro. Table 2 and Appendix A1-4 show the properties and a global plot of all socio-economic datasets used in the analysis of driver contribution.

3.2.2 Hydrologic driver data

Influence of hydrologic driver data was assessed by implementing three datasets: 1) precipitation data, 2) surface runoff data and 3) streamflow data. Precipitation data was obtained from the GPCC Global precipitation dataset including long-term monthly mean precipitation from 1891 until 2016. To correlate driver data with the right monitoring data, RStudio was used to calculate the yearly average precipitation for the mutual year. By doing this, it was possible to create a raster dataset containing the average precipitation for the year 2004 and 2013/2014, which were the years where most samples were taken. Again, precipitation data for the year 2018 was not included in the data, making it impossible to match precipitation data with the monitoring data from 2018. Therefore, precipitation data for the latest year 2016 was used for the monitoring data from 2018.

Surface runoff data was obtained from GRUN global gridded runoff dataset (Ghiggi et al., 2019). GRUN is an observation based monthly reconstruction of surface runoff covering the period from 1902 to 2014. The same method applied the precipitation data was used to calculate the yearly average runoff for the years 2004 and 2013/2014.

Global gridded streamflow data was used from FLO1K global annual streamflow dataset (Barbarossa et al., 2018). Here, an annual streamflow dataset at a resolution of 30 arcseconds (1 km) is presented covering a period from 1960 to 2015. Data was made time specific and yearly averaged for 2004 and 2014. Table 2 gives an overview of the properties of the used hydrologic datasets. Appendix A5-7 shows again global plots of the hydrological data.

3.2.3 Land use data

For the implementation of land use data, the Land Use Harmonization 2 (LUH2) historical dataset was used as third driver category (Hurtt et al., 2020). The data consists of historical land use states covering a period from 850-2015. It is based on the History of the Global Environment database (HYDE), presented by Klein Goldewijk et al. (2017) providing long-term historical, spatially explicit timeseries of population estimates and land use reconstructions on a 30 arcmin resolution. Present land use is reconstructed by analysing satellite images using remote sensing techniques. With the use of historical data of climate, soil, slope and neighbourhood of rivers and lakes, land use data analysis for the past are determined (Hurtt et al., 2020). The raw data consists of six different land use types describing non-forest, forest, cropland, rangeland, urban land and pastureland (Appendix A8-13). Non-forest and forest land use type were subdivided in primary and secondary forests. Primary forests represent woodlands that have not been modified by humans since the start of the data in 850. Secondary forest is now described as forest, but with the characteristic that it has been modified in the past. The same subdivision was applied to the non-forest land use types. Cropland was divided into C4 annuals, C3 annuals, C4 perennials and C3 perennials.

For this study, the LUH2 data for the year 2004 and 2013/2014 was taken corresponding to the microplastic monitoring data. Again, land use data from 2015 was connected to the monitoring data from 2018, since the land use data stopped in the year 2015. Primary non-forest and secondary non-forest data was summed so that the land use data considered all areas that were covered by non-forest or forest in the years 2004 and 2013/2014. This was also done for primary forest and secondary forest data. For the cropland data, only the C4 and C3 annual crops were summed so that only the yearly growing crops were considered, ignoring shrubs and small plants that have lifetime longer than 2 years. These were namely categorised as the perennials crop types. Rangeland, urban land and pastureland were not subdivided in different categories and could be included immediately into the dataset. LUH2 data was again downloaded as NetCDF file, containing for each land use type the fraction per pixel (number between 0 and 1) assigned to that certain land use type.

3.3 Data conversions

After extraction of the driver data for the specific year with available monitoring data, the cell size of all gridded driver datasets was set to a consistent spatial resolution of 0.05 degrees (180 arcseconds). Since all driver datasets consisted of continuous raster data instead of discrete class data, resampling in order to increase the resolution was carried out using the bilinear resampling technique in ArcGIS Pro. This technique was chosen because it uses the four surrounding grid cells to assign a new cell in the finer resolution raster grid, which makes the bilinear interpolation method better suitable for continuous raster data (Fewtrell et al., 2008). Converting the gridded driver data to a finer resolution was important to eventually determine driver data quantities on a catchment wide scale. These steps will be outlined in the next part of this study.

To determine the influence of the driver data on the monitored microplastic concentrations, the HydroBASINS global basin outline dataset was used to match monitoring point data with the upstream catchment area in which the monitoring data point was located (Lehner & Grill, 2013). By doing this, not only driver data at the monitoring location itself was taken into account, but driver data from the entire upstream catchment area was considered.

The Hydrobasins dataset was used to determine the upstream basin area for each monitoring point. Polygon features were drawn in ArcGIS Pro to visualise and quantify the upstream catchment areas that were connected to the monitoring locations. Via this method driver data of the entire upstream area was considered that will eventually flow through the monitoring location. All driver datasets were clipped on the HydroBasins feature polygons using the Zonal Statistics spatial analyst tool in ArcGIS Pro. For the socio-economic- and the hydrologic driver data, this tool summed all grid cells within the upstream catchment polygon. For land use

types, data was averaged over the upstream catchment area. This because land use data was given in fraction per pixel (between 0 and 1). Summing over the entire upstream area would therefore not correspond with the unit of the data. The result is an eventual dataset consisting of microplastic concentrations for all 59 monitoring locations and catchment summed and averaged driver data for the seven different drivers and six land use types with a consistent resolution and time period. Lastly, driver and land use data were normalized by the total upstream catchment area (km²) so that the order of magnitude of the complete dataset became smaller. For the driver data sets, all values were divided by the total upstream catchment area (km²) so that upstream catchment area. As already explained, the difference in unit between driver and land use data required a different normalization method. A flowchart of the data process is shown in Figure 6.



Figure 6 Flowchart of the data conversion process executed for the regression analyses. Colours of the various steps correspond to Table 2.

3.4 <u>Multiple linear regression</u>

Multiple linear regression (MLR) is a frequently used method for data analyses focussing on determining relationships between one dependent variable and multiple independent variables. The general form of an MLR equation that is commonly used is (Chenini & Khemiri, 2009):

(1)
$$V_d = \beta_0 + \beta_1 V_{i1} + \beta_2 V_{i2} + \dots + \beta_n V_{in}$$

For the determination of microplastic concentrations using socio-economic-, hydrologic- and land use drivers, this MLR formula can be described by the following equation:

(2)
$$C_{mp} = \beta_0 + \beta_1 P d + \beta_2 G D P + \beta_3 M P W + \beta_4 W W T + \beta_5 P + \beta_6 R + \beta_7 Q + \beta_8 L U_1 + \beta_9 L U_2 + \beta_{10} L U_3 + \beta_{11} L U_4 + \beta_{12} L U_5 + \beta_{13} L U_6$$

In this equation C_{mp} represents the microplastic concentrations in rivers. The independent variables in this equation consist of population density (PD), GDP, MPW, WWT, precipitation (P), runoff (R), streamflow (Q) and the six land use classes (LU₁₋₆). β_0 describes the intercept of the linear model. β_{1-13} describe the MLR coefficients, also known as the slopes of the linear model.

Before applying the MLR function on the data, correlations between the independent and depending variables were analysed. Pearson's correlation (r) was determined for all driver variables included in the data (Lee Rodgers & Nicewander, 1988). A correlation matrix was produced visualising the correlation between two variables as a number between r = +1 and r= -1. The meaning of r = +1 is that the variables have a strong positive correlation, 0 displays a weak correlation and r = -1 suggests that variables have a strong negative correlation. A strong correlation between independent variables means that these variables not only relate to the depending variable but also influence each other. For an optimal MLR analysis, correlation between the independent variables is low, so that changes in one of the variables do not influence the other variables and eventually the regression model. In the case of high correlation between multiple independent variables, multicollinearity of the data can be assessed. Multicollinearity explains the problem of high correlation between two or more independent variables, such that these variables do not provide unique or independent information in the regression model. The variance inflation factor (VIF) is a commonly used method to describe multicollinearity between independent variables. VIF was therefore calculated and plotted in RStudio. Multicollinearity becomes a problem when the VIF threshold of 5 is exceeded (Fotheringham & Oshan, 2016).

Multiple linear regression analyses in this study were conducted using the MLR function incorporated into RStudio. The dependent variable was taken as the monitored microplastic concentration in particles per m³. Independent variables were set to be all driver datasets used in this study. However, to better examine the functionality of the MLR analyses and the influence of driver data determined by this regression method, stepwise regression was carried out (Wang et al., 2016). This method allows to add the independent variables step by step to the regression function. It can therefore show statistical characteristics of each independent variable, but also the influence and importance of the added variables. Eventually, all variables are added in the stepwise regression resulting in the complete regression analysis with all independent variables included.

Output of the MLR function in RStudio gives a matrix displaying statistical information on the relationship between the dependent and the independent variables. The matrix contains the model residuals, which describe the difference between the actual observed response values and the response values that the linear regression model predicted. For the linear model to be accurate, it is desired to have the residuals equally distributed around the monitoring data points. In addition to this, the output matrix of the linear model describes the intercept and slope of the analysed variables within the statistical relationship, also known as the model coefficients. They can be used to make predictions with the multiple linear regression model that is produced. For each variable, the standard error of the coefficients is displayed in the matrix. P-values are used to determine the significance of the statistical relation. The smaller the p-value, the greater the statistical significance of the observed difference. Rule of thumb is p<0.05 means a significant difference. When p>0.05, the observed difference is not considered significant (Uyanık & Güler, 2013). R² of the eventual regression model was also analysed to check the quality of the model.

3.5 Random forest regression

Random forests (RF) and random forest regression were first introduced by Breiman, (2001). This machine learning algorithm was developed to produce classification and regression trees using independent and dependent variables in a structured dataset consisting of training and testing data (Figure 7). RF regression was executed using the 59 locations containing microplastic monitoring data and the socio-economic-, hydrologic- and land use drivers.

RF regression uses a randomly selected subset of the data to determine the relationship and importance of the independent variables, for which the drivers as discussed above were selected. The algorithm is trained by using a subset of the data as training data, also known as the InBag data. Another subset of the data is used to test the data, this subset is also known

as the Out of Bag (OOB) data. Application of RF regression in earlier studies have shown its improved functionality on the following characteristics (Rodriguez-Galiano et al., 2014):

- It is able to learn complex patterns in datasets, also taking into account any nonlinear complex relationships between explanatory and dependent variables. This characteristic is important in physical geographic research because statistical relations are often nonlinear.
- 2. It requires less computing time for large datasets, making it a perfect method for global statistical modelling.
- 3. It can handle enormous numbers of variables without variable deletion.
- 4. It gives estimated of what variables are important.



Figure 7 Flowchart of the random forest regression process (Rodriguez-Galiano et al., 2014).

To construct the machine learning decision trees, 90% of the data was categorised as as InBag training data and 10% of the data as OOB testing data. Monitoring and driver data of the first 54 rivers of the dataset were used as training data and separated from the last five rivers of the dataset. This latter group of rivers was assigned as test data in order to validate the model. The model constructed in this thesis used more training data compared to other studies applying RF to physical geographic problems (Rodriguez-Galiano et al., 2014; Thorslund et al., 2021). These studies used 80% and 66% of their data as InBag training data, respectively. Due to the smaller amount of input data that was available, it was not possible to fit our model with such a division of training and test data. For this reason, a larger amount

of data was used to train the model. As a consequence, a reduced amount of data remained to test the model. RF regression models were than built using the training data and validated with the testing data. Two required parameters in building an accurate RF regression model are the number of decision trees (ntree) used to build the RF model and the random subset of independent variables (mtry) used for each decision tree (Breiman, 2001). This latter process is called 'bagging' and is used to support the randomness of the RF regression which increases the quality of the training process (Rodriguez-Galiano et al., 2014). The number of trees needed to build a stable regression model depends on when the mean squared error (MSE) of the model stabilizes. Therefore, multiple model runs were done using 50, 100, 200, and 500 decision trees. MSE's were than plotted and the number of trees were chosen where MSE was stable. Next, mtry was chosen after modification of mtry after every run until MSE reached the lowest value. Via this method, the TuneRF function included in RStudio which is produced to find the optimal mtry for the analysed data, was ignored. Mtry, as suggested by the TuneRF function, did not produce the required results and therefore did not decrease the MSE of the model. Eventually, ntree was set to 100 and mtry was set to 10. After implementation of these settings, the model was run ten times, which was needed to tackle the randomness of the RF function. Due to this randomness, the output of the model is slightly different each time it runs. At ntree = 100 and mtry = 10, the variation between all 10 runs was small enough to be useful, without increasing the MSE or decreasing the R² value.

The run where MSE was minimal also resulted in the optimal R² value, suggesting that with this run, microplastic concentrations could be best explained by the driver data. From the MSE results of each run, the normalised root mean squared error (nRMSE) was calculated to evaluate the model performances. To do so, the square root of the MSE was used to calculate the RMSE. After this, RMSE was divided by the mean of all microplastic concentration observations in the dataset. A normalised RMSE represents the differences between the average of the summation of the squared error of the actual output value and the predicted output value (Singh et al., 2017). It can therefore determine whether the model reaches a high enough accuracy to predict output when only independent variables are used as input. After evaluating the RMSE, the output of the model can be used for multiple other purposes. In this study, the variable importance function of the model output will be used to analyse the importance of the drivers on the microplastic particles concentrations. A comparison of the multiple linear regression approach and the random forest approach will be used to answer research question 2 and 3.

3.6 <u>Microplastic predictions</u>

To answer the fourth research question concerning microplastic concentration predictions, both regression methods were applied to new driver data from other rivers around the globe. Prediction calculations of both regression methods were applied to eight rivers, on condition that these prediction rivers were not included in the initial monitoring data used to build the MLR and RF regression models. These eight rivers were the Mekong River, Indus River, Nile River, Congo River, Niger River, Amazon River, Paraña River and the Orinoco River (Figure 8).



Figure 8 Locations of the rivers applied in the regression predictions. Eight rivers were chosen with a spatial distribution across Asia, Africa and South America.

Predicted microplastic concentrations were calculated with driver data at the corresponding locations. Looking at the temporal scale, predictions were calculated with the most recent data from each driver dataset. Population density data from the year 2020 was applied, GDP from 2015, WWT discharge from 2022, MPW data from 2019, precipitation data from 2016, runoff data from 2014, streamflow data from 2015 and land use data from 2015. For the eight rivers, data on these thirteen drivers was implemented into an ArcGIS Pro raster map and an Excel file. The raster map was required for the MLR predictions. Random forest predictions were gathered using the Excel file and RStudio. Predicted microplastic concentrations determined by both approaches were than analysed and compared to existing literature.

Microplastic concentration predictions from the MLR approach were executed with the variable coefficients output of the model. These variable coefficients describe the statistical relationship of that specific variable with the concentration of microplastic (Boy-Roura et al., 2013). In other words, when driver coefficients are determined, it is possible to apply them on new data,

resulting in new microplastic concentration data that follows the new driver data. This method was applied for the MLR predictions and was executed by applying the MLR equation to the new driver data. Since MLR predictions were applied on an ArcGIS raster layer, it was possible to produce a global plot of the new predictions that came out of the MLR model. As a result, a 0.25 x 0.25-degree raster layer was produced containing the new microplastic concentration predictions (in particles/m³).

Microplastic concentration predictions for the RF approach required a different method. Random Forest in RStudio has an incorporated prediction function, which is a function that uses the produced RF model for predictions with new data. It is for this reason that new data required to be an Excel File instead of an ArcGIS Pro raster layer. After running the code, new microplastic concentration predictions (in particles/m³) are presented that allow comparison with the predictions from the MLR approach. Due to the raster structure of the new data and the incompatibility of the RF model with this raster data, RF predictions were only executed on the eight rivers with new driver data. A global plot of the RF predictions is therefore missing.

4. Regression Results

4.1 Global driver data correlation

For MLR to be accurate, a correlation matrix for all independent variables was produced (Figure 9). Rows and columns show the correlation between the different drivers. In the right column, correlation between all drivers and the microplastic concentrations is also presented. A complete correlation matrix that consists of at least two rows and columns is always a symmetrical square. Therefore, only showing half the correlation matrix is enough to analyse it. This upper half of the correlation matrix is shown in Figure 9. As can be seen, the correlation matrix consists of positive and negative correlations with all correlation coefficients between r = +1 and r = -1. The further away the correlation coefficient is from 0, the stronger the relationship is between the two drivers. As rule of thumb a correlation magnitude of r = 0.4 or -0.4 is considered weak, magnitudes outside of this range represent a strong correlation



Figure 9 Correlation matrix displaying Pearson's correlation (r) between all drivers and microplastic concentrations. Values range from +1 to -1, where +1 represents a strong positive correlation and -1 represents a strong negative correlation. The right column shows the correlation between microplastic concentration and the considered drivers.

(Taylor, 1990). However, the diagonal shows a line of maximum positive correlation which makes sense since the diagonal displays the correlation between two of the same variables. To discuss the spatial distribution and the correlation of each driver with respect to microplastic concentrations, Figure 10 represents global grids of the applied driver data.



Figure 10 A-G Global gridded plots of the socio-economic- and hydrologic driver data. Magnifications of these figures are added to Appendix A. Figure 10H visualises global microplastic hotspots with the amount of plastic input from rivers as determined by Lebreton et al. (2017).

Figure 10A-G show the socio-economic- and hydrologic drivers. Figure 10H displays the microplastic concentration hotspots according to Lebreton et al. (2017). Larger visualisations of these plots can be found in Appendix A1-13, where land use drivers are also presented on a global grid.

4.1.1 Population density

Figure 10A and Appendix A1 show a corresponding pattern between population density and microplastic concentration hotspots for the regions Southeast Asia, India and Bangladesh and Central Africa (Figure 10H). Within these geographical areas, multiple regions are present where population density exceeds 900 inhabitants per km². The most densely populated areas can be found in Northeast India and Bangladesh. However, correlation between population density and microplastic concentration in the correlation matrix (Figure 9) has a magnitude of 0.14. Since values within the range -0.4 and +0.4 represent small correlation, it can be said that population density and microplastic concentrations have a weak correlation according to the correlation matrix (Taylor, 1990). It can thus be said that visual correlation from Figure 10A and Figure 10H and the correlation given by the correlation matrix do not correspond. Probably since only three regions (Southeast Asia, India and Bangladesh and Central Africa) show visual correlation between population density and microplastic and microplastic concentration matrix and concentration as shown in Figure 10A and 10H.

4.1.2 Mismanaged Plastic Waste production

Mismanaged plastic waste production (Figure 10B, Appendix A2) shows strong correlation with the spatial distribution of microplastic hotspots. Again, Southeast Asia shows the largest amounts of mismanaged plastic waste (18.4 Mt/y), followed by Central Africa (5.89 Mt/y) and Europe (2.66 Mt/y). High correlations between MPW and microplastic concentration are supported by a Pearson's r value of 0.55 (Figure 9).

4.1.3 Wastewater treatment

Global distribution of wastewater treatment discharge (Figure 10C, Appendix A3) shows a different pattern when comparing this driver to the spatial distribution of microplastic hotspots. Figure 10C visualises that large areas can be distinguished where no wastewater treatment is taking place. These areas are present in Africa, South America and Asia. Areas where largescale treatment is applied are located in North America, Europe and East Asia and seem to correlate with the areas where the GDP is high. This argument is not supported by the correlation matrix, since WWT and GDP correlate with a magnitude of +0.23. According to Laerd Statistics (2020), this magnitude is not large enough to allocate it as a strong correlation. In addition, correlation between WWT and microplastic concentrations represents a

magnitude of 0.24, which again is not considered as strong correlation. In fact, a positive correlation between WWT and microplastic concentrations is remarkable since the application of WWT would result in a decrease in microplastic concentration and therefore in a negative correlation (Rasmussen et al., 2021).

4.1.4 Gross domestic product

GDP per capita (Figure 10D, Appendix A4) is differently distributed than the microplastic hotspots. Large GDP values can be found in North America, Europe, small parts of Russia, the Middle East and Australia. A low GDP is found in large parts of Africa, South America and Asia. Therefore, it seems that a low GDP has a stronger correlation with microplastic hotspots than regions with a high GDP. The same result is presented in the correlation matrix. GDP has strong negative correlation with microplastic concentration (r = -0.43). Therefore, it can be said that an increased GDP results in reduced amounts of microplastic in rivers.

4.1.5 Hydrologic driver data correlations

Hydrologic driver data is plotted in Figure 10E-G and Appendix A5-7, where global distribution of precipitation, surface runoff and streamflow are visualised. Correlation between the hydrologic drivers and microplastic concentration is rather small. R for precipitation is -0.2, for runoff -0.16 and for streamflow +0.1. However, from the plotted figures (Figure 10E-G) correlation between hydrologic drivers and microplastic concentration seems stronger. Especially for regions where precipitation and runoff reach large magnitudes, microplastic concentration rates seem to increase.

Mutual correlation between drivers themselves is worth investigating since strong correlations between drivers may results in a regression model with larger uncertainties, especially for the machine learning random forest regression (Nicodemus & Malley, 2009). As can be seen, the socio-economic drivers show a strong positive correlation to each other (r > 0.4). Especially population density, MPW and WWTdischarge represent high coefficients (population density-MPW, r = 0.69; population density-WWTdischarge, r = 0.71; MPW-WWTdischarge, r = 0.46). However, GDP is the only socio-economic driver that does not have a strong correlation with the other socio-economic drivers (r < 0.4). Strong correlations can also be found between the different land use drivers as can be seen in the bottom right corner of the correlation matric (Figure 9). Non-forest land cover fraction and crop land cover fraction have a positive correlation of 0.89. In addition, pastureland cover fraction shows strong correlations with non-forest, forest and cropland (r = 0.8, r = 0.89 and r = 0.96, respectively). GDP is the only driver showing stronger negative coefficients. Pearson's correlation for GDP was smaller than -0.4 with the non-forest land cover, urban land cover, rangeland cover and pastureland cover (r = 0.40).

-0.49, r = -0.43, r = -0.41, r = -0.41, respectively). Correlation between de independent drivers and the dependent driver (MP concentration) is shown in the right column of the correlation matrix. For microplastic concentration, the highest positive correlation is with MPW driver (r = 0.55), the most negative correlation is with GDP (r = -0.43). All other correlation coefficients are within the +0.4 or -0.4 range and are therefore weaker correlated with microplatic concentrations (Figure 9).

Multicollinearity was also plotted in RStudio. Results from this analysis are shown in Figure 11. This figure shows for each independent variable the multicollinearity this variable has with the other independent variables. Values larger than five suggest high multicollinearity that may influence regression analysis in a negative way. A vertical line is drawn which shows a multicollinearity of five. All bars that reach further than this line have a multicollinearity higher than ideal for MLR and RF analyses. However, considered that in this study multiple variables influence each other, for instance the different land use drivers, a multicollinearity smaller than ten is still considered acceptable (Salmerón et al., 2018). Multicollinearity between the various land use drivers is immediately visible. Especially pastureland and cropland have a severe



VIF Values

Figure 11 Multicollinearity plotted using the variance inflation factor (VIF). Dotted line represents VIF value of 5.

multicollinearity with VIF magnitudes exceeding 50. Since the unit of these drivers is in fraction per pixel, a change in value for one of the land use drivers also influences the value of another land use driver. In addition to this, population density has a multicollinearity larger than five (VIF = 8), suggesting that this variable can negatively influence the quality of the MLR model. However, since multicollinearity for this variable is still between five and ten, it is considered that the consequences of this higher multicollinearity are neglectable.
4.2 Multiple linear regression

As discussed, MLR was implemented using a stepwise method which executes the linear regression step by step adding one independent variable with each step. Therefore, thirteen separate regression steps were constructed, starting with the driver that had the largest Pearson's correlation coefficient with microplastic concentration (Figure 9). This method allowed researching the improvement of the model after each added driver. Results of the stepwise analyses are shown in Table 3. For each regression step, the added driver, R², the number of significant drivers and what drivers were significant is presented. Detailed results of the stepwise linear regression analyses can be found in Appendix B.

Step	Added Driver	<i>R</i> ²	N significant	Significant drivers	
			variables		
1	MPW	0.30	1/1	MPW	
2	GDP	0.43	2/2	MPW; GDP	
3	Forest fraction	0.44	2/3	MPW; GDP	
4	Pasture fraction	0.47	2/4	MPW; GDP	
5	WWT	0.47	2/5	MPW; GDP	
6	Precipitation	0.52	3/6	MPW; GDP; Precipitation	
7	Rangeland fraction	0.54	3/7	MPW; GDP; Precipitation	
8	Runoff	0.54	3/8	MPW; GDP; Precipitation	
9	Population density	0.58	4/9	MPW; GDP; Precipitation;	
				Population density	
10	Crop fraction	0.69	4/10	GDP; Pasture fraction; WWT;	
				Crop fraction	
11	Streamflow	0.69	4/11	GDP; Pasture fraction; WWT;	
				Crop fraction	
12	Urban fraction	0.73	5/12	GDP; Pasture fraction; WWT;	
				Crop fraction; Urban fraction	
13	Non-forest	0.73	5/13	GDP; Pasture fraction; WWT;	
	fraction			Crop fraction; Urban fraction	

Table 3 Detailed stepwise multiple linear regression results presenting R² after each step, the number of significant drivers and the type of significant driver.

Linear regression after thirteen steps shows an R² of 0.73 with five out of thirteen drivers being significant (p < 0.05) (Uyanık & Güler, 2013). These significant drivers are GDP (p = 0.005), Pasture fraction (p = 0.0008), WWT (p = 0.009), Crop fraction (p = 0.0003) and Urban fraction (p = 0.02). Each driver that was added to the stepwise linear regression model, resulted in varying improvements of the models R² value, which is often referred to as the coefficient of determination (Hopfe & Hensen, 2011; Nimon & Oswald, 2013). The first run with only MPW as driver resulted in an R² of 0.30. From the second run onwards R² gradually increased to 0.73. Table 4 shows the coefficient of determination (R² increase) for each driver. The

strongest improvement of the coefficient of determination was after step two, where GDP was added to the model. It went from 0.30 to 0.43 and therefore had an increase of 0.13. The second largest improvement of the model was after step 10 (crop fraction), where R² increased with 0.11. Step 5 (WWT), 8 (runoff), 11 (streamflow) and 13 (non-forest fraction) did not improve the model. Drivers added during these steps also did not increase the number of significant drivers.

Table 4 Increase in coefficient of determination after each step of the multiple linear regression approach. Colours follow the magnitude of increase.

Step	1	2	3	4	5	6	7	8	9	10	11	12	13
R ²	0.30	0.13	0.01	0.03	0	0.05	0.02	0	0.04	0.11	0	0.04	0
increase													

Remarkable is the change in significant drivers after step 10. Here, the crop fraction driver was added to the regression model, resulting in a switch in driver significance. Prior to step 10, MPW, GDP, precipitation and population density were the significant drivers. However, the consideration of crop fraction as driver of microplastic concentrations changed the statistical relationship making GDP, pasture fraction, WWT and crop fraction the significant drivers. Eight drivers did not show a significant relationship with microplastic concentrations after complete stepwise regression. These drivers were MPW, forest fraction, precipitation, rangeland fraction, runoff, population density, streamflow and non-forest fraction.

4.3 Random forest regression

Random forest analyses were done with the exact same dataset that was used for the MLR model. The RF function was run ten times and the run with the smallest normalised root mean squared error (nRMSE) and largest R-squared (R^2) was chosen to be the best fit for the data analysis. The reason that the code was run ten times is that because of the randomness of RF. Due to the random subset of variables that is used to build each tree, results of the model show minor differences each time it runs. Running the model more than ten times did not improve the model accuracy (R^2 and nRMSE) and did not reduce the difference in results.

Results showing the nRMSE, R^2 and variable importance are shown in Figure 12. As can be seen, 100 decision trees were used to build the model. From this point, the nRMSE and R^2 stabilise (Peters et al., 2007). Increasing the number of trees did not decrease the error rate and thus improve the results. NRMSE for the optimal run stabilized around 24.2. R^2 for this model run became stable with a value of 0.55. Both the nRMSE and R^2 figure also show the irregularity of the model during the first twenty decision trees. nRMSE of the first two decision trees had the smallest magnitude (nRMSE = 17 and nRMSE = 12). After producing three

decision trees, the error rate strongly increases to a magnitude of 38.1, reaching the highest error of the entire run. From this point onwards, the extra decision trees resulted in a gradual reduction of the nRMSE. R^2 results displayed in Figure 12B show an opposite trend. After formation of two decision trees, R^2 reaches a maximum value of 0.88. From this point it strongly decreases to a value of -0.12 from which it gradually increases towards the stabilized value of 0.55.

Figure 12C shows the importance in percentage increase in mean squared error (%IncMSE) of each driver variable that is considered in the RF model runs. It therefore displays the influence that each driver has on the nRMSE of the model and thus on the microplastic concentrations. GDP comes out as the most important independent variable with an %IncMSE of 6.6. Variations in GDP will lead to the largest increase in MSE, making the model less accurate. A reduced accuracy represents a model that is less capable of describing microplastic concentrations considering the driver data that is imported into the model. Second most important is the production of mismanaged plastic waste (%IncMSE = 4.2), followed by



Figure 12 A) Normalised root mean squared error. B) R^2 and C) relative variable importance in percentage mean squared error increase of the random forest regression approach.

the population density (%IncMSE = 3.5). It is worth mentioning that all hydrologic drivers (precipitation, runoff and streamflow) show a very small importance. From this driver category, precipitation has the largest importance with an %IncMSE of 0.54 followed by streamflow (%IncMSE = -0.23) and runoff (%IncMSE = -0.41). For the land use driver category, importance of the different variables varies widely. Crop fraction, non-forest fraction, rangeland fraction and pasture fraction have a rather strong importance with a %IncMSE value between 2.3 and 1.0. However, urban fraction, which is part of the land use driver category represents the smallest importance with a %IncMSE magnitude of -0.58.

4.4 Variable importance comparison

As can be seen, both regression approaches have led to dissimilar results in both the performance and the variable importance analysis. Both models did not result in a similar magnitude for R², pointing out that one model better described the relationship between the drivers and microplastic concentrations. MLR showed that for this dataset, it was able to predict 73% of the monitoring microplastic data (namely, $R^2 = 0.73$). For RF, only 55% of the microplastic monitoring data can be explained by the driver data. It can thus be said that for this amount of data, the more standard MLR approach resulted in the most accurate model to describe microplastic concentrations in rivers. In addition, both models differently analysed the importance of the drivers. For MLR, relative importance of drivers was assessed via the influence that each driver has on the change in R², also known as the coefficient of determination (Hopfe & Hensen, 2011; Nimon & Oswald, 2013). As is shown in Table 4, the strongest variable weight can be assigned to MPW, GDP and cropland fraction. After addition of these drivers to the stepwise regression, the coefficient of determination reached the largest value and R² increased strongly. However, MPW was the first variable added to the stepwise MLR, which probably exaggerates the weight of this driver and therefore the relative importance on the regression (Hopfe & Hensen, 2011). Implementation of the WWT, runoff, streamflow and the non-forest land use fraction step did not lead to an increased R², suggesting that these drivers had a smaller relative importance according to the MLR approach.

As discussed, relative importance for the RF approach is determined using the percent increase in MSE that each driver has (%IncMSE). Figure 12C determines GDP, MPW, population density and cropland fraction as drivers with the largest %IncMSE. WWT, streamflow, runoff and urban land fraction showed the least influence on the MSE.

Both regression approaches do show varying results on relative driver importance. Where GDP, MPW and cropland fraction are important drivers for both models, population density is more important for the RF model than for the MLR model (%IncMSE = 3.5 against a coefficient of determination of 0.04). Also, the least important drivers of both regression models seem to

correspond. However, one strong difference is the influence of the urban land use fraction. For MLR, this driver resulted in a coefficient of determination of 0.04. It is a small increase in R^2 , but it did influence the regression model. However, for RF regression, urban land use is assigned as the driver with the smallest %IncMSE.

4.5 Microplastic predictions

Global plotted MLR predictions are shown in Figure 14 where also the location of the eight prediction rivers is pointed out. Figure 15 presents a zoom on Asia since this is the geographical region where microplastic accumulate most strongly. As can be seen in Figure 14, large regions represent a microplastic concentration of zero. This is true for North America, large areas in Europe, Northern Africa and Australia. However, in or near rivers with a high flow regime, MLR predicts accumulation of microplastics. Examples are the Mississippi River, Danube and Volga River. In Central Africa, India and Bangladesh, Southeast Asia and East Asia microplastic concentrations increase gradually. Maximum microplastic concentrations as determined by the MLR model occur in the Amazon River region and the Ganges Brahmaputra Delta. For both regions, microplastic concentrations exceed 2500 particles per m³. Also, in these regions microplastic concentrations trends follow the river paths of multiple rivers. Examples are the Amazon River, the Congo River and the Paraña River. An explanation for this trend can be that the streamflow driver data has strongly influenced the prediction outcomes. Table 5 and Figure 13 present the results of both regression approaches (in particles/m³) for each prediction river. The order of magnitude difference between both predictions is also outlined.

MLR predicts the largest amount of microplastic in the Amazon River (179.1 particles/m³). The second most polluted river is the Nile with 116 particles/m³ microplastic. With a concentration of 0.79 particles/m³, the Indus Rivers is determined as the least polluted river.

Predictions as calculated by the RF approach have a much smaller magnitude compared to the MLR predictions. RF determines the Nile as most polluted river. Here, microplastic concentrations are 0.54 particles/m³. Furthermore, the Mekong and Congo River have the second highest microplastic concentration with 0.41 particles/m³. The Amazon, Paraña and Orinoco River have the least amount of microplastic with 0.033, 0.032 and 0.032 particles/m³, respectively.

For none of the prediction locations the difference in predicted microplastic concentration between both approaches was zero orders of magnitude. For the Mekong and Indus River, the prediction difference was one order of magnitude. For the Congo and Niger River, the difference was two orders of magnitude and for the Nile, Paraña and Orinoco River prediction outcomes varied over three orders of magnitude. For the Amazon River, microplastic prediction from both approaches resulted in the largest dissimilarity. Here, the difference between MLR and RF predictions contained four orders of magnitude.

It can thus be stated that both regression approaches predict microplastic pollution on a wide range of magnitudes. Comparing both results with existing literature, predictions given by the MLR approach seem to be the best estimation of micoplastic concentrations for these eight rivers. Lebreton et al. (2017) has shown that plastic inputs from rivers, in tonnes per year, within the mentioned pollution hotspots reach a 10³ and 10⁴ order of magnitude. Even though this study has determined plastic pollution in the unit particles/m³, it becomes clear that the magnitudes determined by the MLR approach are needed to reach the number of tonnes per year river input as calculated by Lebreton et al. (2017). Riverine plastic emissions (in metric tonnes per year) for the mentioned pollution hotspots, as determined by Meijer et al. (2021), lay in the range between 50.000 and 200.000 metric tonnes per year. These results show even larger orders of magnitudes than the outcomes of Lebreton et al. (2017), which again is in favour of the plastic predictions determined by the MLR approach. The higher model performances that are shown in section 4.2 support these findings. Comparison of our prediction results with the AdventureScientists.org database (section 2.3) was not possible, since none of the predicted rivers considered in this study is included in the Adventure Scientist monitoring campaign.

Location	Multiple Linear Regression	Random Forest Regression	
	MP concentration $[p/m^3]$	MP concentration $[p/m^3]$	ΔMP [order of
			magnitude]
Mekong River	7.7	0.41	1
Indus River	0.79	0.027	1
Nile River	116	0.54	3
Congo River	21.1	0.41	2
Niger River	4.2	0.032	2
Amazon River	179.1	0.033	4
Paraña River	17.03	0.032	3
Orinoco River	17.41	0.038	3

Table 5 Microplastic concentration predictions of both regression approaches for the eight considered prediction rivers [particles/m3]. The right column presents the difference, in orders of magnitude, between the MLR and the RF approach.



Figure 13 Diagram visualising the difference in microplastic concentration for all prediction rivers in particles per m³. Blue diagrams visualise prediction results from the MLR approach. Orange diagrams represent the RF prediction results.



Figure 14 Global plot of the MLR microplastic concentration predictions in particles per m³. Red dots show the eight prediction locations.



Figure 15 MLR prediction plot focussing on Asia. Again, the predicted microplastic concentration is plotted in particles per m³. Red dots show the prediction locations in Asia.

5. Discussion

From the literature study and the global multiple linear regression predictions, it can be stated that microplastic concentration hotspots are present in multiple regions of the world. Considering the literature review, microplastic concentration amounts have the largest magnitudes in Southeast Asia, South Asia, Central Africa, Europe and North America (Figure 2). Multiple linear regression in this study assigns microplastic hotspots to regions in Southeast Asia, South Asia, Central Africa and South America (Figure 14). It can be concluded that MLR proves to be the best approach to determine a statistical relationship between the considered drivers and global microplastic concentrations with this amount of data. R² results and the significance of various drivers support this outcome. Random forest does not reach the same performance as the MLR approach and therefore shows a smaller R², resulting in an increased normalised root mean squared error. Both approaches show that GDP, MPW and cropland fraction are the three most important drivers for microplastic concentrations. For MLR this was visible in the increase of the models' R² after stepwise addition of drivers. RF visualised this in the variable importance plot that was produced. Urban land use, runoff and streamflow are the least important drivers as was shown by both models.

Predictions that were determined with both regression models resulted in rather strong dissimilarities. MLR predictions were able to point out microplastic hotspots in regions corresponding to Lebreton et al. (2017) and Meijer et al. (2021). In addition, the orders of magnitude that MLR predictions showed seem to be more precise compared to the RF predictions (Lebreton et al., 2017; van Calcar & van Emmerik, 2019; van Wijnen et al., 2019).

5.1 Main sources of uncertainties

As always, data studies and the application of regression models show results with varying uncertainties (Reis & Saraiva, 2005). Such uncertainties can occur in the data that is implemented into the model (both monitoring data and driver data), the data conversions that were used to modify the data and in the models itself. For this study, the first two reasons for uncertainty will likely have reduced the accuracy of both regression models. As is explained in the method section (section 3.2) the driver datasets that were used for both regression analyses were implemented from scientific literature. Therefore, it is assumed that these driver datasets met the required quality standards for regression analysis. This applies the same for the monitoring data. However, since multiple measurements were achieved at each monitoring location, it was required to average the microplastic concentrations to obtain one measurement for each location. For some locations (Yangtze River and Hanjiang River in China) this average number was received from one measurement (Schmidt et al., 2017).

Other monitoring locations consisted of more than 17 samples from which the average microplastic concentration was determined. Examples are the Rhine River Delta near Rotterdam from Schmidt et al. (2017) and the monitoring data from the Tibetan Plateau (Jiang et al., 2019) (Figure 5). Averaging over a varying number of samples increases the uncertainty of the eventual microplastic concentration magnitude. In addition, a potential temporal scale included in data with multiple samples is often eliminated when averaging has been taking place.

Inconsistency in the data also occurs within the temporal differences between monitoring data and driver data. For instance, monitoring data from Jiang et al. (2019), that was measured in 2018, was mostly matched with driver data from the year 2015. This because driver data from the year 2018 was not available for most driver datasets. It is very likely that this approach has led to increased inaccuracy of both models. This type of inconsistency can be reduced when monitoring data is matched with driver data from the same year.

Lastly, the conversions that were applied to the driver datasets may have caused uncertainty in both models. All data conversions executed in this study are outlined in Figure 6. Even though these data conversions were supported by scientific literature, some of these actions can have influenced the homogeneity and resolution of driver data. Especially the implementation of the HydroBASINS dataset to determine upstream catchment areas (Lehner & Grill, 2013). Per monitoring location, the upstream basin area was determined with hand-drawn polygons in ArcGIS Pro. Within the data conversion process, this step has probably resulted in the strongest uncertainty since it was sometimes not completely clear where the border of the catchment area was located, forcing us to use other literature to determine where a catchment border was located. These uncertainties have influenced the area over which the driver data was calculated.

Regression results of both approaches have also shown a difference in performance. This difference in performance, and especially the lower accuracy of the RF model (R² = 0.55), can be a result of the amount of monitoring and driver data that was used in this study. To date, the volume of monitored microplastic data is small and mostly focussing on marine monitoring campaigns (Wagner et al., 2014). And even though the topic of microplastic pollution is becoming more and more important in scientific research, large-scale monitoring campaigns are still lacking. Therefore, finding monitoring data for the 59 locations used in this study already was a challenging task. It resulted in monitoring data that was focussed on a study which mainly described monitoring data measured in the United Stated (Schmidt et al., 2017). Comparing our data amount with other physical geographic research that applied RF regression, suggests that an increase in monitoring- and driver data would increase the RF

performances. For instance, Rodriguez-Galiano et al. (2014) applied RF to 175 wells of monitoring data and 24 driver variables to analyse nitrate pollution in groundwater. The same is true for Thorslund et al. (2021). Here, water salinity from 401 sub-basins and more than 400.000 measurements from 1980 to 2010 are analysed using 25 different driver variables. For the MLR, a smaller amount of data does often not reduce the performances of the model (Bonett & Wright, 2011). It is for this reason that the MLR model performed better than the RF model.

Results on microplastic predictions have shown strong differences between both regression methods. Especially the small magnitudes of the RF predictions are remarkable, knowing that the eight considered rivers are among the largest on the globe (Lehner & Grill, 2013). Schmidt et al. (2017) concluded that the largest rivers in the world do have the highest microplastic pollution rates. Comparing our results with monitoring studies from the Indus and Amazon River, it becomes clear that the MLR method shows the most corresponding results. Microplastic concentrations as determined by the RF approach are too small and not supported by any literature (Gerolin et al., 2020; Tsering et al., 2021). Gerolin et al. (2020) and Tsering et al. (2021) examined the amount of microplastic in river sediments and determined that the number of particles per kg sediment lay in the same order of magnitude as the MLR outcomes in this study. Comparison of our results with modelling studies focussing on the microplastic output from rivers to oceans is complicated due to the difference in unit that those studies use. Lebreton et al. (2017) and Meijer et al. (2021) both show that for the eight predicted rivers, microplastic output to sea exceeds a magnitude of 2000 tonnes per year. Such an amount of plastic will never be reached with the concentrations predicted by the RF approach, taking into account the small mass of microplastic particles and the discharges that these eight predicted rivers have. Even with the predicted MLR concentrations it is probably not possible to reach those amounts of microplastic output.

5.2 Implications and recommendations of future microplastic research

Results that are presented in this study can be used for multiple purposes. First, the dissimilarity of model performance of both regression approaches has shown the better quality of the MLR approach for this amount of monitoring and driver data. It is therefore recommended to analyse same-sized datasets with the MLR methods instead of the RF regression approach. When a larger scale dataset is investigated, RF regression may be a more suitable approach. Especially when the number of independent variables increases (Luan et al., 2020).

Second, the extensive results on variable importance open the opportunity to apply these outcomes to water quality policy and microplastic reduction in freshwater systems. Information on the correlation between dependent and independent variable can tell what the effect of changing driver data is on the other drivers and on the microplastic concentrations itself. Small changes in one driver can result in large changes in other drivers if the value of correlation is high enough. When the modified drivers have a significant influence on the amount of microplastics in rivers, it can eventually lead to a reduction of the microplastic concentrations. It needs to be said that due to the lower accuracy of the RF regression, its results on variable importance need to be analysed with caution, since the findings may not represent the right values. However, due to the stable model that is produced (Figure 12) and the similarities with MLR for the most important drivers, the variable importance outcomes are considered useful (Peters et al., 2007). From these results, changes in GDP, MPW, population density and cropland fraction will most strongly influence the amount of microplastic in rivers. Multiple studies focussing on microplastic drivers have shown that the socio-economic drivers are responsible for microplastic input to rivers (Lebreton et al., 2017; Meijer et al., 2021; Schmidt et al., 2017; van Wijnen et al., 2019). However, the importance of certain land use drivers on the input of microplastics has never been assessed and is therefore a promising result for future plastic pollution policy.

6. Conclusions

This study estimates the drivers and hotspot regions of microplastic concentrations worldwide based on literature study and regression analyses including both multiple linear regression and random forest regression. Spatial explicit driver data of four socio-economic drivers, three hydrologic drivers and six land use drivers was applied in combination with microplastic concentration data from 59 monitoring locations spread across the globe. The following conclusions for each research question can be stated.

What are hotspot regions of microplastic concentrations in rivers worldwide?

Global plastic loads from rivers to sea was estimated to be between 1.15 and 2.41 million tonnes per year, considering the modelling studies from Lebreton et al. (2017) and Schmidt et al. (2017). The 20 most polluting rivers were located in Asia and contributed for more than 2/3 to the global annual plastic input towards oceans. Degradation of macro- and mesoplastics is a common process which implies high microplastic concentrations in these regions (van Wijnen et al., 2019). Microplastic concentration hotspots can thus be found in East and Southeast Asia, India and Bangladesh, Central Africa, Europe and North America. Drivers that were clearly correlated with microplastic concentrations were MPW and GDP. This latter one showed the largest negative correlation. Microplastic hotspots with a high GDP (Europe and North America) clearly have lower microplastic concentrations than the other hotspots with a lower GDP (India and Bangladesh).

What is the contribution of socio-economic drivers, hydrologic drivers, and land use characteristics to microplastic loads in rivers, according to both statistical methods?

Stepwise multiple linear regression of the 13 driver variables with the 59 monitoring locations resulted in a regression model with a varying R² and significance for each driver. Drivers were added according to the correlation the driver had with microplastic concentration. After addition of all drivers into the stepwise multiple linear regression, R² reached a value of 0.73 and the significant drivers were GDP, pasture fraction, WWT discharge, GDP, crop fraction and urban fraction. Steps where MPW, GDP and crop fraction were added resulted in the largest increase in R². Random forest regression resulted in a model with an R² of 0.55. Normalised root mean squared error stabilised around 24.2. From the model, a variable importance analysis was done making it possible to conclude that GDP, MPW, population density and crop fraction were the most important drivers for microplastic concentrations in rivers.

This study therefore showed the difference in results of both regression approaches but has also shown that three of the four socio-economic drivers (GDP, MPW, Population Density) contributed most to pollution of microplastics in rivers. In addition, the crop fraction land use

drivers also proved to be a large contributor according to both approaches. Crop fraction is followed by the other land use drivers, which are located in the middle of the contribution spectrum. Hydrologic drivers were the least important contributors of microplastic pollution as was determined by both regression models.

How do the results of estimated importance of the various drivers of both methods compare?

Multiple linear regression as well as random forest regression determined that GDP and MPW were the most important drivers of microplastic pollution in rivers. For MLR, these two drivers are followed by the relative importance of crop fraction, where for RF regression population density comes after the two most important drivers. Comparison of both approaches also outlined the difference in non-forest fraction importance. This driver did not result in an R² increase of the MLR model, suggesting a relatively small influence on microplastic pollution. However, RF regression determined a much stronger influence of this driver, making it the fifth strongest driver of microplastics in rivers. Both approaches showed similar results for the relatively small importance of the hydrologic drivers, making it the least important driver category.

What is the performance of multiple linear regression and random forest regression techniques for predicting microplastic loads in rivers in other regions of the world?

As presented, most recent driver data was used to produce microplastic concentration predictions of eight global rivers, calculated from both the MLR and RF regression approach. Results of the two regression methods leaded to strong variations in predicted microplastic concentrations. MLR predictions showed the most promising outcomes when compared with microplastic concentration research (Lebreton et al., 2017; van Calcar & van Emmerik, 2019; van Wijnen et al., 2019). RF predictions presented much smaller magnitudes than the MLR predictions. For the Amazon River, the difference in microplastic predictions was four orders of magnitude, making it the most dissimilar of the considered rivers. The higher accuracy and the better correspondence with existing literature, shows that the MLR approach was the best predicting model for global microplastic concentrations in rivers with the amount of data used in this study.

Acknowledgements

First, I want to thank Dr. Michelle van Vliet for supervising and reviewing this study and maintaining close contact with me throughout this research process. Thank you, Dr. Marcel van der Perk, for being the second supervisor on this project. This study was made possible thanks to the availability of multiple open-source databases including GPCC, GRUN, HydroWaste, Hydroshed and FLO1K. I would also like to thank the Utrecht University for providing me with the right software (ArcGIS Pro and Rstudio) to do this thesis.

Finishing this thesis would never have succeeded without the help of my family. My parents always supported me throughout this journey. It was never a problem to use their study room when I did not want to go the university library or study at home in my student room. I really liked the coffee they made me when I was working there. Also, a big thank you to my sisters Sophie and Isabel, for reading the final version of this thesis and calling me when I needed some extra support.

Lastly, I thank my girlfriend, Marie-Christine, for standing next to me during the entire period. She helped me when I lost one of my best friends last May and made sure that after a small period of rest, I continued writing this thesis. Without you it would have taken me much longer to finish this research.

Paul, my friend, you tragically passed away on the 29th of May. It was hard not thinking about you during the last phase of this research, often making it hard to keep on going. However, it also gave me the strength to finish it. I will never forget you.

Data availability

- Complete database as used in both regression analyses can be accessed via following link:

Complete Database Excel File Download

In this file monitoring data, driver data and catchment area and location of all locations is presented.

- Codes for the multiple linear regression and random forest regression are given in the following R scripts:
 - Multiple linear regression: <u>Link to stepwise multiple linear regression code</u> (<u>Rscript</u>)
 - Random forest regression: Link to random forest regression code (Rscript)

7. Appendices

7.1 Overview and description of the appendices

This study contains 2 appendices referred to with in-text references to maintain readability when large figures or multiple consecutive figures are used. An overview and description of the appendices is given on this page.

Appendix A: Global gridded raster data of the microplastic concentration drivers that are analysed in the two regression methods.

Appendix A1: Global gridded population density data in 2020 in number of persons per km² (CIESIN - Columbia University, 2018).

Appendix A2: Global gridded Mismanaged Plastic Waste production in 2019 in tonnes per year (Lebreton & Andrady, 2019).

Appendix A3: Global gridded Wastewater Treatment discharge data in m³/day (Ehalt Macedo et al., 2022).

Appendix A4: Global gridded GDP per capita in 2015 in US dollars (Kummu et al., 2018).

Appendix A5: Global gridded precipitation data form 2014 in mm (Schneider et al., 2015).

Appendix A6: Global gridded surface runoff data from 2014 in mm/day (Ghiggi et al., 2019).

Appendix A7: Global gridded streamflow data from 2014 in m³/s (Barbarossa et al., 2018).

Appendix A8: Global gridded non-forest land use fraction in fraction per pixels (Hurtt et al., 2020).

Appendix A9: Global gridded forest land use fraction in fraction per pixel (Hurtt et al., 2020). *Appendix A10:* Global gridded crop land use fraction in fraction per pixel (Hurtt et al., 2020).

Appendix A11: Global gridded urban land use fraction in fraction per pixel (Hurtt et al., 2020). *Appendix A12:* Global gridded rangeland land use fraction in fraction per pixel (Hurtt et al., 2020).

Appendix A13: Global gridded pastureland land use fraction in fraction per pixel (Hurtt et al., 2020).

Appendix B1: Extended results of the stepwise multiple linear regression analyses. Table shows for each driver the standard error and p-value.

Appendix B2: Residuals versus model fit as determined by the MLR approach



Appendix A1 Global gridded population density data in 2020 in number of persons per km²







Appendix A3 Global gridded Wastewater Treatment discharge data in m³/day



Appendix A4 Global gridded GDP per capita in 2015 in US dollars



Appendix A5 Global gridded precipitation data form 2014 in mm



Appendix A6 Global gridded surface runoff data from 2014 in mm/day



Appendix A7 Global gridded streamflow data from 2014 in m³/s

Appendix A8 Global gridded non-forest land use fraction in fraction per pixels



Appendix A9 Global gridded forest land use fraction in fraction per pixel











Appendix A12 Global gridded rangeland land use fraction in fraction per pixel





Appendix A13 Global gridded pastureland land use fraction in fraction per pixel

Appendix B1 Extended results of the stepwise multiple linear regression analyses.

Residuals: 1Q Median Min 3Q Мах -529.31 -68.15 16.85 67.56 455.52 Coefficients: Estimate Std. Error t value Pr(>|t|) 1.712 0.093718 . 2.570e+02 1.501e+02 (Intercept) MPWmod 2.428e+00 2.352e+00 1.032 0.307529 -1.709e-01 5.903e-02 -2.896 0.005814 ** GDPmod ForestFracMod -3.978e-03 2.592e-03 -1.534 0.131961 5.833e-02 1.354e-02 PastureFracMod 4.308 8.83e-05 *** 6.015e-02 2.188e-02 2.750 0.008557 ** WWTmod Precipmod 1.546e+01 1.940e+01 0.797 0.429623 9.887e-04 1.723e-02 0.057 0.954494 RangeFracMod -6.887e+00 7.572e+00 -0.910 0.367880 Runoffmod -4.758e+00 5.396e+00 -0.882 0.382552 Popdensmod -2.490e-02 6.403e-03 -3.889 0.000328 *** CropFracMod 1.190e+00 1.308e+00 0.910 0.367754 Streamflowmod -1.835e-02 7.600e-03 -2.414 0.019904 * UrbanFracMod NonForestFracMod 1.759e-03 2.511e-03 0.700 0.487379 _ _ _ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207.3 on 45 degrees of freedom Multiple R-squared: 0.73, Adjusted R-squared: 0.652 F-statistic: 9.358 on 13 and 45 DF, p-value: 6.149e-09



Appendix B2 Residuals versus model fit as determined by the MLR approach

8. References

- Adeola Fashae, O., Abiola Ayorinde, H., Oludapo Olusola, A., & Oluseyi Obateru, R. (2019). Landuse and surface water quality in an emerging urban city. *Applied Water Science*, 9(2), 25. https://doi.org/10.1007/s13201-019-0903-2
- Andrady, A. L., & Neal, M. A. (2009). Applications and societal benefits of plastics.
 Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1526),
 1977–1984. https://doi.org/10.1098/rstb.2008.0304
- Arnaez, J., Lasanta, T., Ruiz-Flaño, P., & Ortigosa, L. (2007). Factors affecting runoff and erosion under simulated rainfall in Mediterranean vineyards. *Soil and Tillage Research*, *93*(2), 324–334. https://doi.org/10.1016/j.still.2006.05.013
- Barbarossa, V., Huijbregts, M. A. J., Beusen, A. H. W., Beck, H. E., King, H., & Schipper, A. M.
 (2018). FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015. *Scientific Data*, 5(1), 180052. https://doi.org/10.1038/sdata.2018.52
- Bonett, D. G., & Wright, T. A. (2011). Sample size requirements for multiple regression interval estimation. *Journal of Organizational Behavior*, *32*(6), 822–830. https://doi.org/10.1002/job.717
- Boy-Roura, M., Nolan, B. T., Menció, A., & Mas-Pla, J. (2013). Regression model for aquifer vulnerability assessment of nitrate pollution in the Osona region (NE Spain). *Journal of Hydrology*, *505*, 150–162. https://doi.org/10.1016/j.jhydrol.2013.09.048
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, H. L., Selvam, S. B., Ting, K. N., & Gibbins, C. N. (2021). Microplastic pollution in freshwater systems in Southeast Asia: Contamination levels, sources, and ecological impacts. *Environmental Science and Pollution Research*, 28(39), 54222–54237. https://doi.org/10.1007/s11356-021-15826-x

- Chenini, I., & Khemiri, S. (2009). Evaluation of ground water quality using multiple linear regression and structural equation modeling. *International Journal of Environmental Science & Technology*, *6*(3), 509–519. https://doi.org/10.1007/BF03326090
- Christiansen, K. (2018). Global and gallatin microplastics initiatives. *Adventure Scientists*, *531*, 532.
- CIESIN Columbia University. (2018). *Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11*. NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H49C6VHW
- Disatnik, D., & Sivan, L. (2016). The multicollinearity illusion in moderated regression analysis. *Marketing Letters, 27*(2), 403–408. https://doi.org/10.1007/s11002-014-9339-5
- Ehalt Macedo, H., Lehner, B., Nicell, J., Grill, G., Li, J., Limtong, A., & Shakya, R. (2022).
 Distribution and characteristics of wastewater treatment plants within the global river network. *Earth Syst. Sci. Data*, *14*(2), 559–577. https://doi.org/10.5194/essd-14-559-2022
- Eriksen, M., Lebreton, L. C. M., Carson, H. S., Thiel, M., Moore, C. J., Borerro, J. C., Galgani,
 F., Ryan, P. G., & Reisser, J. (2014). Plastic Pollution in the World's Oceans: More
 than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea. *PLOS ONE*,
 9(12), e111913. https://doi.org/10.1371/journal.pone.0111913

- Fewtrell, T. J., Bates, P. D., Horritt, M., & Hunter, N. M. (2008). Evaluating the effect of scale in flood inundation modelling in urban environments. *Hydrological Processes*, 22(26), 5107–5118. https://doi.org/10.1002/hyp.7148
- Fotheringham, A. S., & Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: Dispelling the myth. *Journal of Geographical Systems*, 18(4), 303– 329. https://doi.org/10.1007/s10109-016-0239-5
- Gerolin, C. R., Pupim, F. N., Sawakuchi, A. O., Grohmann, C. H., Labuto, G., & Semensatto, D. (2020). Microplastics in sediments from Amazon rivers, Brazil. *Science of The Total Environment*, *749*, 141604. https://doi.org/10.1016/j.scitotenv.2020.141604
- Geyer, R., Jambeck, J. R., & Law, K. L. (2017). Production, use, and fate of all plastics ever made. *Science Advances*, *3*(7), e1700782. https://doi.org/10.1126/sciadv.1700782
- Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2019). GRUN: an observation-based global gridded runoff dataset from 1902 to 2014. *Earth Syst. Sci. Data*, *11*(4), 1655–1674. https://doi.org/10.5194/essd-11-1655-2019
- Hartanto, H., Prabhu, R., Widayat, A. S. E., & Asdak, C. (2003). Factors affecting runoff and soil erosion: Plot-level soil loss monitoring for assessing sustainability of forest management. *Forest Ecology and Management*, *180*(1), 361–374. https://doi.org/10.1016/S0378-1127(02)00656-4
- Hoornweg, D., & Bhada-Tata, P. (2012). What a waste: A global review of solid waste management.
- Hopfe, C. J., & Hensen, J. L. M. (2011). Uncertainty analysis in building performance simulation for design support. *Energy and Buildings*, 43(10), 2798–2805. https://doi.org/10.1016/j.enbuild.2011.06.034

Hurley, R., Woodward, J., & Rothwell, J. J. (2018). Microplastic contamination of river beds significantly reduced by catchment-wide flooding. *Nature Geoscience*, *11*(4), 251–257. https://doi.org/10.1038/s41561-018-0080-1

Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk,
J., Fujimori, S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A.,
Humpenöder, F., Jungclaus, J., Kaplan, J. O., Kennedy, J., Krisztin, T., Lawrence, D., ...
Zhang, X. (2020). Harmonization of global land use change and management for the
period 850–2100 (LUH2) for CMIP6. *Geosci. Model Dev.*, *13*(11), 5425–5464.
https://doi.org/10.5194/gmd-13-5425-2020

- Jambeck Jenna R., Geyer Roland, Wilcox Chris, Siegler Theodore R., Perryman Miriam, Andrady Anthony, Narayan Ramani, & Law Kara Lavender. (2015). Plastic waste inputs from land into the ocean. *Science*, *347*(6223), 768–771. https://doi.org/10.1126/science.1260352
- Jiang, C., Yin, L., Li, Z., Wen, X., Luo, X., Hu, S., Yang, H., Long, Y., Deng, B., Huang, L., & Liu, Y. (2019). Microplastic pollution in the rivers of the Tibet Plateau. *Environmental Pollution*, *249*, 91–98. https://doi.org/10.1016/j.envpol.2019.03.022
- Jones, E. R., van Vliet, M. T. H., Qadir, M., & Bierkens, M. F. P. (2021). Country-level and gridded estimates of wastewater production, collection, treatment and reuse. *Earth System Science Data*, *13*(2), 237–254. https://doi.org/10.5194/essd-13-237-2021
- Kummu, M., Taka, M., & Guillaume, J. H. A. (2018). Gridded global datasets for Gross
 Domestic Product and Human Development Index over 1990–2015. *Scientific Data*, 5(1), 180004. https://doi.org/10.1038/sdata.2018.4
- Lebreton, L., & Andrady, A. (2019). Future scenarios of global plastic waste generation and disposal. *Palgrave Communications*, 5(1), 6. https://doi.org/10.1057/s41599-018-0212-7
- Lebreton, L., Zwet, J. van der, Damsteeg, J.-W., Slat, B., Andrady, A., & Reisser, J. (2017). River plastic emissions to the world's oceans. *Nature Communications*, 8(1), 15611. https://doi.org/10.1038/ncomms15611
- Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, *42*(1), 59–66.

https://doi.org/10.1080/00031305.1988.10475524

- 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*(15), 2171–2186. https://doi.org/10.1002/hyp.9740
- Liu, Z., Li, Y., & Li, Z. (2009). Surface water quality and land use in Wisconsin, USA a GIS approach. *Journal of Integrative Environmental Sciences*, *6*(1), 69–89. https://doi.org/10.1080/15693430802696442
- Luan, J., Zhang, C., Xu, B., Xue, Y., & Ren, Y. (2020). The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research*, 227, 105534. https://doi.org/10.1016/j.fishres.2020.105534
- Mai, L., Sun, X.-F., Xia, L.-L., Bao, L.-J., Liu, L.-Y., & Zeng, E. Y. (2020). Global Riverine Plastic Outflows. *Environmental Science & Technology*, *54*(16), 10049–10056. https://doi.org/10.1021/acs.est.0c02273
- Mani, T., Hauk, A., Walter, U., & Burkhardt-Holm, P. (2015). Microplastics profile along the Rhine River. *Scientific Reports*, *5*(1), 17988. https://doi.org/10.1038/srep17988

- Meijer, van Emmerik Tim, van der Ent Ruud, Schmidt Christian, & Lebreton Laurent. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. *Science Advances*, 7(18), eaaz5803. https://doi.org/10.1126/sciadv.aaz5803
- Nearing, M. A., Jetten, V., Baffaut, C., Cerdan, O., Couturier, A., Hernandez, M., Le
 Bissonnais, Y., Nichols, M. H., Nunes, J. P., Renschler, C. S., Souchère, V., & van Oost,
 K. (2005). Modeling response of soil erosion and runoff to changes in precipitation
 and cover. *Soil Erosion under Climate Change: Rates, Implications and Feedbacks, 61*(2), 131–154. https://doi.org/10.1016/j.catena.2005.03.007
- Ngo, P. L., Pramanik, B. K., Shah, K., & Roychand, R. (2019). Pathway, classification and removal efficiency of microplastics in wastewater treatment plants. *Environmental Pollution*, *255*, 113326. https://doi.org/10.1016/j.envpol.2019.113326
- Nicodemus, K. K., & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, *25*(15), 1884–1890. https://doi.org/10.1093/bioinformatics/btp331
- Nimon, K. F., & Oswald, F. L. (2013). Understanding the Results of Multiple Linear
 Regression: Beyond Standardized Regression Coefficients. *Organizational Research Methods*, *16*(4), 650–674. https://doi.org/10.1177/1094428113493929
- Peters, J., Baets, B. D., Verhoest, N. E. C., Samson, R., Degroeve, S., Becker, P. D., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, *207*(2), 304–318. https://doi.org/10.1016/j.ecolmodel.2007.05.011
- Rasmussen, L. A., Iordachescu, L., Tumlin, S., & Vollertsen, J. (2021). A complete mass balance for plastics in a wastewater treatment plant—Macroplastics contributes

more than microplastics. Water Research, 201, 117307.

https://doi.org/10.1016/j.watres.2021.117307

Reis, M. S., & Saraiva, P. M. (2005). Integration of data uncertainty in linear regression and process optimization. *AIChE Journal*, *51*(11), 3007–3019.

https://doi.org/10.1002/aic.10540

- Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L.
 (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of The Total Environment*, 476– 477, 189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001
- Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, *88*(12), 2365–2384. https://doi.org/10.1080/00949655.2018.1463376
- Schmidt, C., Krauth, T., & Wagner, S. (2017). Export of Plastic Debris by Rivers into the Sea. Environmental Science & Technology, 51(21), 12246–12253.

https://doi.org/10.1021/acs.est.7b02368

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & Ziese, M. (2015). GPCC Full Data Reanalysis Version 7.0 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data.

https://doi.org/10.5676/DWD_GPCC/FD_M_V7_050

Siegfried, M., Koelmans, A. A., Besseling, E., & Kroeze, C. (2017). Export of microplastics from land to sea. A modelling approach. *Water Research*, *127*, 249–257. https://doi.org/10.1016/j.watres.2017.10.011

- Singh, B., Sihag, P., & Singh, K. (2017). Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Modeling Earth Systems and Environment*, *3*(3), 999–1004. https://doi.org/10.1007/s40808-017-0347-3
- Talvitie, J., Mikola, A., Koistinen, A., & Setälä, O. (2017). Solutions to microplastic pollution –
 Removal of microplastics from wastewater effluent with advanced wastewater
 treatment technologies. *Water Research*, *123*, 401–407.
 https://doi.org/10.1016/j.watres.2017.07.005
- Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39.

https://doi.org/10.1177/875647939000600106

- Thorslund, J., Bierkens, M. F. P., Oude Essink, G. H. P., Sutanudjaja, E. H., & van Vliet, M. T.
 H. (2021). Common irrigation drivers of freshwater salinisation in river basins
 worldwide. *Nature Communications*, *12*(1), 4232. https://doi.org/10.1038/s41467-021-24281-8
- Tong, S. T. Y., & Chen, W. (2002). Modeling the relationship between land use and surface water quality. *Journal of Environmental Management*, *66*(4), 377–393. https://doi.org/10.1006/jema.2002.0593
- Townsend, K. R., Lu, H.-C., Sharley, D. J., & Pettigrove, V. (2019). Associations between microplastic pollution and land use in urban wetland sediments. *Environmental Science and Pollution Research*, *26*(22), 22551–22561.

https://doi.org/10.1007/s11356-019-04885-w

Tsering, T., Sillanpää, M., Sillanpää, M., Viitala, M., & Reinikainen, S.-P. (2021). Microplastics pollution in the Brahmaputra River and the Indus River of the Indian Himalaya.

Science of The Total Environment, 789, 147968.

https://doi.org/10.1016/j.scitotenv.2021.147968

- Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *4th International Conference on New Horizons in Education*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027
- van Calcar, C. van, & van Emmerik, T. van. (2019). Abundance of plastic debris across European and Asian rivers. *Environmental Research Letters*, *14*(12), 124051.
- van Emmerik, T., & Schwarz, A. (2020). Plastic debris in rivers. *WIREs Water*, 7(1), e1398. https://doi.org/10.1002/wat2.1398
- van Wijnen, J., Ragas, A. M. J., & Kroeze, C. (2019). Modelling global river export of microplastics to the marine environment: Sources and future trends. *Science of The Total Environment*, 673, 392–401. https://doi.org/10.1016/j.scitotenv.2019.04.078
- Wagner, M., Scherer, C., Alvarez-Muñoz, D., Brennholt, N., Bourrain, X., Buchinger, S., Fries,
 E., Grosbois, C., Klasmeier, J., Marti, T., Rodriguez-Mozaz, S., Urbatzka, R., Vethaak,
 A. D., Winther-Nielsen, M., & Reifferscheid, G. (2014). Microplastics in freshwater
 ecosystems: What we know and what we need to know. *Environmental Sciences Europe*, *26*(1), 12. https://doi.org/10.1186/s12302-014-0012-7
- Waldschläger, K., Brückner, M. Z. M., Carney Almroth, B., Hackney, C. R., Adyel, T. M., Alimi,
 O. S., Belontz, S. L., Cowger, W., Doyle, D., Gray, A., Kane, I., Kooi, M., Kramer, M.,
 Lechthaler, S., Michie, L., Nordam, T., Pohl, F., Russell, C., Thit, A., ... Wu, N. (2022).
 Learning from natural sediments to tackle microplastics challenges: A
 multidisciplinary perspective. *Earth-Science Reviews*, *228*, 104021.
 https://doi.org/10.1016/j.earscirev.2022.104021

Wang, M., Wright, J., Brownlee, A., & Buswell, R. (2016). A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis.
 Energy and Buildings, 127, 313–326. https://doi.org/10.1016/j.enbuild.2016.05.065

Welden, N. A., & Lusher, A. L. (2017). Impacts of changing ocean circulation on the distribution of marine microplastic litter. *Integrated Environmental Assessment and Management*, 13(3), 483–487. https://doi.org/10.1002/ieam.1911