Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Natural Resources and Agricultural Sciences

Landscape Factors Related to Performance of *In-Situ* Sensors for Prediction of Total Phosphorus

- A statistical analysis of data from 194 streams in Sweden

Cham Hoang

Department of Aquatic sciences and assessment

Master´s thesis • 30 HEC
Soil and water management – Master’s program
Uppsala 2017

Landscape Factors Related to Performance of In-Situ Sensors for Prediction of Total Phosphorus

Cham Hoang

Supervisor: Jens Fölster, Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment

Examiner: Martyn Futter, Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment

Credits: 30 HEC
Level: Second cycle, A2E
Course title: Independent Project in Environmental Science
Course code: EX0431
Program/education: Soil and water management – Master's program

Place of publication: Uppsala
Year of publication: 2017
Online publication: http://stud.epsilon.slu.se

Keywords: Surface water monitoring, Surrogate parameters, Total Phosphorous, Turbidity, Catchment, Land-Use, Soil Type, Linear Regression, Multivariate Analysis

Swedish University of Agricultural Sciences

Faculty of Natural Resources and Agricultural Sciences
Department of Aquatic sciences and assessment
Geochemistry and Hydrology

Abstract

Accurate monitoring data of phosphorus concentrations in the water will benefit the evaluation of measures taken to reduce the problems of eutrophication. Today, the quantification of phosphorus loads from diffuse sources involves large uncertainties. Traditional grab sampling could underestimate the loads exported from diffuse sources due to the flow dependence of particulate phosphorus (PP) concentrations. Technical innovations allow continuous monitoring of certain parameters in the water and by that generate data of high temporal resolution. These parameters have the potential to act as surrogate measurements for total phosphorus (TP). One parameter that is often argued to have strong correlations with total phosphorus is turbidity, which might reflect the PP fraction in the water. However, poorer correlations have also been reported. Possibly due to land-use or soil type within the catchments. By finding landscape factors related to good/poor performance of in-situ sensors, in the prediction of TP, will benefit the strategical planning of monitoring programs. In this study, data from available sensor parameters were used to create regression models for TP at 194 monitoring stations in Sweden. Turbidity, water temperature and total organic carbon were most frequently included in the significant regression models. The correlation coefficient from the most significant TP regression model at each station was compiled into a response dataset for multivariate analysis. The correlation coefficients were then predicted with landscape data from the catchments of the monitoring stations with multivariate analysis. A separate dataset of landscape data within 100 m buffer zone was also used for comparison. The results indicated weak associations between the landscape factors and the performance of the TP prediction models. The landscape factor of most significance relation to high correlation coefficients was forest on mire, whereas water and open wetlands along the buffer zone were related to lower correlation coefficient values. No preferences were found between significant catchment factors and sensor parameters included in the regression model at the stations. Data of the whole catchment explained larger variations in the TP regression model’s strength than the buffer zone data, possibly due to low data resolution in the buffer zone.

Populärvetenskaplig sammanfattning

Miljöövervakningen av sjöar och vattendrag är ett sätt för oss att ta tempen på rådande trender, bra liksom dåliga. Ett rådande problem är övergödningen av sjöar och vattendrag som orsakas av utsläpp av näringsämnen från jordbruk, skogsbruk, reningsverk, avlopp och tätorter.

Vattenmätningar som ofta tas vid utvalda tillfällen riskerar att missa ökade koncentrationer av partikulär fosfor vid plötsligt ökad avrinning. Idag finns det möjlighet att ta kontinuerliga prover med sensorer av vissa parametrar i vatten. Provresultaten från dessa parametrar kan användas för beräkning av missade höga halter av t.ex. fosfor på grund av ökad avrinning. Det har t.ex. påvisats starka samband mellan turbiditet (ett mått vattnets grumlighet) och partikulär fosfor i vattendrag.

I den här studien testades olika kombinationer av tillgängliga sensorparameterar, turbiditet, elektrisk konduktivitet, organisk kol, vattentemperatur och pH, för att hitta den optimala beräkningsmodellen av total fosfor för 194 mätstationer i Sverige. De parametrar som oftast förekom bland de optimala beräkningsmodellerna var turbiditet, organisk kol och vattentemperatur. Starka samband mellan organisk kol och total fosfor kan innebära en stor mängd organisk fosfor i den totala halten. Vattentemperaturen kan reflektera en säsongsbaserad fosforkoncentration i vattnet. Lägre flöden i vattendrag under sommaren kan bidra till högre koncentrationer av löst fosfor från reningsverk eller enskilda avlopp. En mindre andel stationer fick beräkningsmodeller med låg förklaringsgrad, vilket kan bero på olika markanvändningstyper inom stationernas avrinningsområden.

Markanvändning och jordtyper inom stationernas avrinningsområden beräknades tillsammans med fosformodellernas förklaringsgrad. Resultatet visade att skog på myr och kalhyggen ofta förekom tillsammans i avrinningsområdena och en stor andel av dessa ofta sammanföll med starka fosformodeller. Fosfor kan exporteras till vattendrag från blöta dikade kalhyggen, och då främst i organisk form, vilket kan förklara varför organisk kol ofta var inkluderad i de starka fosformodellerna. Ett negativt samband hittades också mellan starka fosformodeller och större andel vatten och öppen våtmark. En förklaring är att sjöar och öppna våtmarker har en förmåga att bromsa ner effekten av höga flöden som bär med sig partikulär fosfor.

Table of Contents

List of tables

List of figures

Figure 14. One component PLS model loading plot of the land-use/cover and JG2 data. Buffer zone 100m dataset, all variables and observations included.

Figure 15. VIP and coefficient diagrams of the whole catchment PLS model. All landscape variables and observation included.

Figure 16. Coefficient plot of the model component of the most significant PLS models, only VIP>0.5. Buffer zone data.

Abbreviations

A	Agriculture
AG	Artificial green areas
CF	Final felling
CLC	Coordination of information of information on the environment
Cl-rich	Finer grained soils
CV	Cross validation
DA	Developed areas (urban and settlements)
DOP	Dissolved organic phosphorus
DP	Dissolved phosphorus
EC	Electric conductvity
EEA	Europen Environemntal Agency
Erod.	Soil erodibility data
EU	European Union
F	Forest on mineral soils (all types)
FM	Forest on mire
FNU	Formazin Nephelometric Unit
GIS	Geografic Information System
H	Potentially high erodibility
H_0	Null hypothesis
H_1	Alternative hypothesis
JG2	Jord Grundlager 2
JY1	Jord Ytlager 1
L	Low erodibility
Lat.	Latitude
LU	Land-Use/Cover
M	Moderate erodibility
M	Moderate erodibility
M	Moderate erodibility
MA	Miscellaneous agriculture (i.e. orchards)
MLR	Multiple Linear Regression

MVM	Mark-Vatten-Miljöcentrum
n	observations
No	No/very low erodibility
NTU	Nephelometric Turbidity Unit
OV	Open vegetation,
P	Pastures
P.Em.	Mean phosphorus emissions 2013/2014 (kg/ha)
PC	Principal Component
PCA	Principal Component Analysis
PIP	Particulate Inorganic Phosphorus
PLS	Partial Least Squares
POP	Particulate Organic Phosphorus
PP	PP
PRESS	Predictive Residual Sum of Square
PSU	Penn State University
Q^2	Prediction capacity
R^2	SLR correlation coefficient
R^2 adjusted	MLR adjusted correlation coefficient
R^2X	PCA Correlation coefficient
R^2Y	PLS Correlation coefficient
RMSEP	Root Mean Square Error of Prediction
RO	Rock Outcrops
Sa-rich	Coarser grained soils
SCB	Statistics Sweden
SEPA	Swedish Environment Protection Agency
SGU	Geological Survey of Sweden
Simp.soil	Simplified Soil Data
SLR	Simple Linear Regression
SLU	the Swedish University of Agricultural Sciences
SLU-GET	SLU Geodata Extraction Tool
SMD	Svensk Marktäcke Data
SMED	Swedish Environment Emission Data

SMHI	Swedish Meteorological and Hydrological Institute
SRP	Soluble Reactive Phosphorus
SSE	Sum of Squares error
SWAM	Swedish Agency of Marine and Water Management
Te	Water Temperature
TO	TOC
TOC	Total Organic Carbon
TP	TP
TSS	Total suspended solids
Tu	Turbidity
USGS	United States Geological Survey
W	Water and wetlands
WFD	European Water Framework Directive
VIF	Variance Inflation Factor
VIP	Variances Influence on Projection
ViVaN	Virtual Watercourse Network
WQ	Water Quality
X	Independent Variable
Y	Response Variable
β_0	Regression Intercept
β_1	Regression Coefficient

1 Introduction

Environmental monitoring is important to detect water pollution or evaluate measures taken to reduce the problems of water quality deterioration. One of the largest problem related to poor water quality is eutrophication, caused by excessive export of phosphorus from different land-use systems. Sweden is obligated to take measure to abate the problem of eutrophication according to the national environmental objectives as well as the EU Water Framework Directive. In the 1970's nutrient stripping was installed at all larger sewage treatment plant in Sweden, which efficiently reduced the amount of dissolved phosphorus (DP) emitted. Despite, the problem of eutrophication remains due to the presence of other sources within the catchment(Fölster et al., 2014). Today, the main challenge is to account for the diffuse sources, such as agriculture and urban areas.

Export of phosphorus from diffuse sources is often associated with increased runoff events, causing soil erosion and in extension export of particulate phosphorus (PP). Traditional sampling methods, such as grab sampling or flow-proportional sampling, could give inaccurate estimations of the total loads of phosphorus. The concentration of phosphorus in surface water is not only influenced by the emissions from different sources but also complex cycling processes within the streams. All processes combined will cause unexpected fluctuations of total phosphorus (TP) concentrations in the water that are difficult to detect (Cassidy et al., 2017, Withers and Jarvie, 2008).

Continuous research investigates the possibility to develop sound monitoring methods at low costs to enable quantitative and qualitative data collection (Rekolainen et al., 1991, Jones et al., 2011, Fölster and Rönnback, 2015). Advancements in technology have provided efficient tools for water monitoring e.g. sensor technology. One sensor could be submersed in the water to continuously log a set of variables. The results will be data of high temporal resolution, capturing diurnal variations. To date, only several water quality parameters are possible to continuous monitor with sensors but it is possible to use these as proxy for TP. Strong correlations have been found between TP and turbidity (Fölster and Rönnback, 2015, Jones et al., 2011). However, there are differences in the correlation pattern, possibly due to catchment properties. In theory, soil type and land-use are influencing the amount and mobility of TP in the landscape, especially the particulate fraction (Ulén et al., 2012, Line et al., 2002, McDowell, 2006).

Numerous studies have argued for a strong correlation between landscape factors and surface water quality (Amiri and Nakane, 2009, Haidary et al., 2013, Buck et al., 2004, Villa Solís, 2014, Fölster and Rönnback, 2015). For instance, Amiri and Nakane (2009) did an attempt to find statistical correlation between landscape factors and several water quality variables. Their landscape attributes models explained up to 92% of the water quality variations. However, application of extensive landscape datasets in e.g. multiple linear regression could be problematic due the often noisy, skewed and multicollinear data properties as pointed out by Nash and Chaloud (2011). In that given situation, Partial Least Square (PLS) could be a method to consider. PLS is a robust statistical method not affected by the mentioned data inadequacies due to the extraction of latent data structures (Cassel et al., 1999). In a study by Nash and Chaloud (2011), a large landscape GIS dataset were used to predict several surface water biota variables with PLS. From their results, the authors concluded PLS as a suitable methodology when having fragmented and inadequate data at hand.

This study aims to find landscape factors that are related to the performance of sensor parameters in predicting TP concentrations. The result will provide some foundations for strategical planning of monitoring programs, as it could indicate the types of catchment suitable for TP surrogate measurements.

1.1 Objectives

The main objective of the study is to identify important landscape factors related to the strength in correlation between TP and turbidity (alone or in combination with other available sensor parameters). Water quality data from 194 monitoring stations in Sweden will be used to create TP regression models at each site. The correlation coefficients from the regression models will then be used as the response variable in multivariate analysis with landscape factors as prediction variables.

In addition, the following research questions will be addressed:

- How strong are the correlations between sensor parameters and TP in Swedish streams?
- Could the correlation with TP be further improved by combining multiple sensor parameters?
- What landscape factors best explain strong/weak correlation patterns between measured sensor variables and TP?
- Is the difference in correlation strength best explained by factors of the whole catchment or factors within the 100m buffer zone?
- Is it possible to deduce suitable sensor parameters based on the characteristics of the catchment?

2 Background

The following section is a literature review of the topic and will provide some theoretical background for the study.

2.1 Swedish Environment Monitoring Program of Surface Water

Monitoring of the environment was initiated by the Swedish Environment Protection Agency (SEPA) in 1965. Today, the task is divided between different authorities. Data on the state of the environment is collected in ten different monitoring programs, including a program for surface waters, which is implemented on national, regional, and local scale. The National Monitoring Program is today performed by the Swedish University of Agricultural Sciences (SLU), with finance from The Swedish Agency of Marine and Water Management (SWAM). The purpose is to render a general picture of the current state of surface waters. The monitoring of water bodies included in the national program are distributed to assure an extensive spatial coverage. Some of the fresh water bodies monitored in the national programs are considered somewhat pristine, making the results suitable to use as reference for the regional and local monitoring. The regional and local monitoring programs for surface water are supervised by the regional authorities and adapted to the regional environment or more specific purposes, such as possible impact from land management measures. (Sahlsten et al., 2014)

Regarding national and EU regulations the monitoring programs should generate cohesive and high-quality data to allow the utility of data from different monitoring programs to assess the water quality status. Sweden has adopted 16 environmental objectives, where several are connected to the state of the freshwaters, such as “Zero eutrophication”, “Natural acidification only”, “Flourishing lakes and streams”, “A Balanced Marine Environment” and “Flourishing Coastal Areas and Archipelagos”. To achieve the objectives, the monitoring programs work as a mean to assess the state of the waters to determine suitable measures, or evaluate the effect of measures taken. In line with the European Water Framework Directive (WFD) Sweden also has the responsibility to report status of the water bodies within each water management district, which is under the supervision of the water authorities. As the water authorities are cooperation among several regional authorities, the data used in the assessment for reporting is a miscellany of different monitoring programs. For this reason, cohesivity in the data is important. There is a large coverage of water quality data in Sweden due to extensive environmental regulations (encouraging the monitoring of natural resources) and state appropriations (funding the national and regional monitoring programs). However, with increased cooperation frequency on both national and international level, and stricter environmental regulations, improvement of methods are encouraged to obtain larger spatial coverage as well as higher degree of temporal detail. (Sahlsten et al., 2014)

2.1.1 Monitoring Methods

Occasional grab sampling, which is a common method today, could generate data of too coarse temporal resolution. Neal et al. in 2012 did an experiment in which samples were collected every 7 hours for 18 months from different streams within a catchment in Wales. The samples were analysed for a vast number of chemical parameters. The results showed that the concentrations responded to the fluctuations of the flow, indicating the potential of high frequency sampling to capture shorter periodic transport. The costly practice of surface water monitoring and sample analysis often compromise the temporal coverage on behalf of the spatial, or vice versa. For this reason, SWAM has reported a desire to develop sampling methods to increase the overall efficiency (Sahlsten et al., 2014). Another commonly applied method is flow-proportional sampling, which regulates the sampling frequency depending on the flow rate in the stream. The samples are often taken based on flow intervals, i.e. every 6mm and composited into a larger tank, from which the total load of the monitored period is estimated. In catchments where the transport of pollutants from non-point sources via overland flow is dominating, this method has been argued to generate relatively accurate load estimations without increasing the number of samples (Rekolainen et al., 1991). However, a study by Harmel and King (2005) indicated some possible errors coupled to the method, as they found that correlation between suspended sediment and flow rate could vary.

Many studies have tried to evaluate the suitability of surrogate measurements to obtain cost efficient and greater time resolution water quality data (Christensen et al., 2000, Horsburgh et al., 2010). Surrogate measurements might be automatized through sensors located permanently in the stream. The data obtain can be used as proxy for other chemical variables in the water. Christensen et al. (2000) used surrogate measurements of specific conductivity, temperature, dissolved oxygen, and turbidity to construct regression models for other parameters in interest (sulphate, bacterial activity, biota exposure and total suspended solids). Data for model calibration was collected the first two years to generate a regression model to predict new data, which was compared with sampled data from the following year. Their results showed that the regression models successfully predicted the new data and a high potential of the method in terms of gathering detailed temporal water quality data (Christensen et al., 2000).

2.2 Phosphorus

Anthropogenic activity has made a pronounced alteration to the natural phosphorus cycle, polluting aquatic environments by accelerating and sustaining eutrophication. The mining of inorganic phosphorus from the mineral apatite, for agricultural purposes to feed a large population and consumption, increases the load transported to aquatic systems (Bennett et al., 2001).

2.2.1 Forms of Phosphorus

There are two main forms of phosphorus, particulate (PP) and dissolved (DP). DP can either be inorganic orthophosphates (PO_4^{3-}, HPO_4^{2-} , $H_2PO_4^{-}$) (DIP) or organic molecules of esters or phosphonate (DOP) (Ruttenberg, 2003) and is often defined as the fraction of total phosphorus (TP) readily filtered through a 0,45µm membrane filters (Spivakov et al., 1999). As for PP, an organic (POP) and inorganic (PIP) division can also be made. PIP might occur as primary minerals, such as apatite, or precipitated from chemical reaction with Ca^{2+} to form the less soluble secondary mineral hydroxyapatite in neutral to alkaline conditions (Eriksson, 2011). In acidic soils, phosphorus might precipitate to form the mineral Strengite ($FePO_4 \cdot 2H_2O$) with Fe or/and Wavellite ($Al_3(PO_4)_2(OH)_3 \cdot 5H_2O$ with Al. However, the different forms of orthophosphates are most commonly associated with surface adsorption by ferric Fe and Al hydroxides in low pH soils and sediments, creating strong inner sphere complexes with low solubility (Essington, 2015). POP often consist detrital fragments of organic matter (Yohismura, 2007).

In soil and water, microorganisms and primary producers easily assimilate the high soluble DP, leaving the less soluble PP to accumulate. The fixation of phosphorus, through precipitation of secondary minerals and mineral surface sorption in certain redox conditions and soil composition, further limits the phosphorus availability as a nutrient for primary producers (Ruttenberg, 2003). These limiting processes encourage increased addition of mineral or organic phosphorus to the soil to increase productivity.

2.2.2 Soil Transport

In general, DP is transported in the soil through matrix flow, which is a vertical flow from the topsoil to the subsoil, also referred to as leaching, whereas PP is transported laterally on the soil surface with runoff water or vertically through macrospores in the soil profile (Haygarth and Sharpley, 2000). A gross estimation of 90% of the total amount of phosphorus transported to the oceans is in particulate form (Ruttenberg, 2003), which highlights the importance of the interactional processes in the soil. In fact, although soil accumulation through mineral sorption of phosphorus will reduce the availability to plants, it helps to limit the losses to surface waters through leaching processes. Djodjic et al. (2004) conducted a plot scale investigation of the relationship between phosphorus losses, soil type and soil phosphorus content. In terms of phosphorus concentration in the topsoil, which might represent the scheme of fertilizer amendment, no clear correlation was observed with the leaching amount. The water conducting properties of the soil seem to be of larger importance. The authors found that low pace of moving water from topsoil to subsoil will increase the capability of the subsoil to trap the phosphorus through particle surface absorption (Djodjic et al., 2004). However, losses of PP have been observed to occur at large extent from clayey soils (Villa Solís, 2014), as sandy soils in Sweden often characterize as acidic, which might enhance the surface sorption process in the presence of Fe- and Al hydroxides (Kyllmar et al., 2006). Moreover, the presence of macro pores structures in clay soils will enhance a fast movement of water, causing internal erosion. Fine clay particles also has the ability to disperse into the water making it easy to transport from terrestrial system (Ulén et al., 2001). Detachment of PP might also occur on bare soils with low infiltration capacity, especially in water saturated arable soils, through sheet or rill erosion (Villa Solís, 2014).

2.2.3 Transport and Cycling in the Water Column

Rivers and streams are important carriers of phosphorus to lakes and oceans. Depending on the chemical composition and hydrological regime of the river, the forms of phosphorus transported could vary (Ruttenberg, 2003). Although difficult to generalize, the main form of TP might be represented by the least reactive and bioavailable form (Villa Solís, 2014). Withers and Jarvie (2008) compiled the different processes affecting the phosphorus transport and retention in river systems, which include both biotic and abiotic factors (fig.1). Terrestrial derived phosphorus might

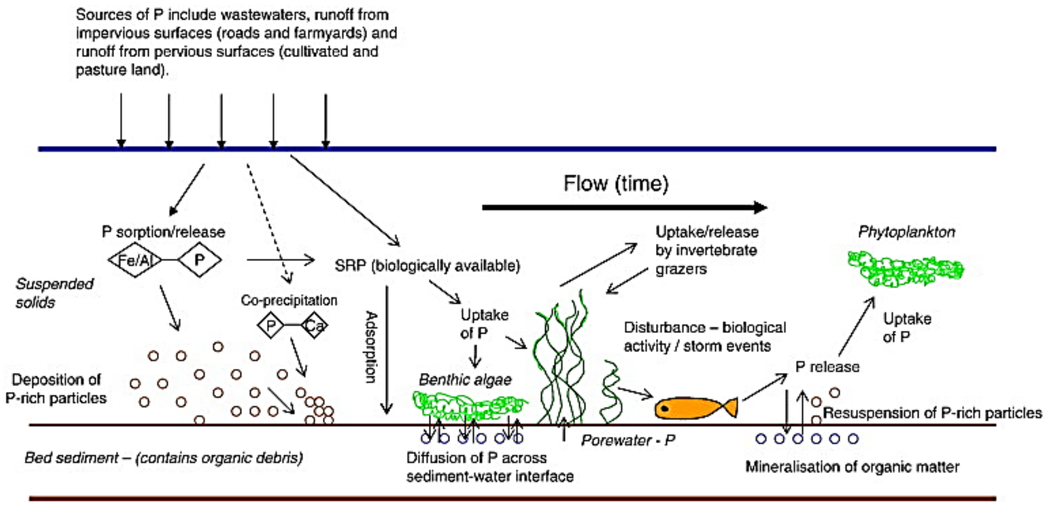


Figure 1. In stream processes of Phosphorus (Withers and Jarvie, 2008).

undergo transformation in the water column as depicted in figure 1, where the flow time is the key factor regulating the rate of uptake, sorption, desorption, and sedimentation. Concluded in several studies (Ulén et al., 2012, Svendsen and Kronvang, 1993, Turner and Haygarth, 2000) is the importance of high flows in the generation of phosphorus transport from soils, which extends to the stream network. In slow moving water, retention is promoted by increased biological uptake, chemical reactions and sedimentation (Reddy et al., 1999). Wetlands have also been argued to promote the retention of phosphorus in the catchment and by that limit the transportation. The richness in vegetation slows down the water velocity and enables sedimentation of PP. In fact, the construction of wetlands as abatement of nutrient export from agricultural land-use systems is a much supported method (Reddy et al., 1999, Kynkäänniemi, 2014), however, there is a notion of a finite efficiency from these systems due to saturation (Hickey and Doran, 2004).

The removal of DIP and DOP in the water column is efficiently performed by primary producers, either through direct assimilation or by enhancing the mineral sorption process due to their production of oxygen in the respiration process. Moreover, the presence of macrophytes in the water could limit the flow velocity and hence enables sedimentation (Withers and Jarvie, 2008).

In similarity with the soil processes, the chemical reactions involved in phosphorus retention includes sorption and desorption of SRP by Fe and Al hydroxide particles in the water column or mineral precipitation with Ca^{2+}. Apart from requiring the presence of these chemical components, often derived from the watershed, redox conditions and pH also plays an important role (Withers and Jarvie, 2008). The coprecipitation of phosphorus and Ca^{2+} is most efficient in alkaline stream conditions (House, 1999) whereas surface sorption by Fe/Al oxide minerals are most efficient in oxidative conditions. For this reason, the fate of buried PP might become reverse if the redox conditions changes, i.e. due to microbial respiration in the sediment (Sondergaard et al., 2001). The depletion of oxygen is also associated with eutrophication, which promotes a positive feed-back loop by the reduction of photosynthesis in the water and hence creating anoxic conditions (Spears et al., 2008).

Another way in which sediment deposited PP might emit DIP to the water column is through diffusion. A steep concentration gradient between the sediment pore water and bulk water body will trigger the diffusion process. Concentrations of DIP in the sediment is influenced by i.e. microbial activity of mineralization/assimilation, respiration (altering the redox condition) (Withers and Jarvie, 2008). Regarding the concentration of PP in the water column, re-suspension induced by abiotic or biotic processes plays an important role (Søndergaard et al., 2003). A study by Tammeorg et al. (2015) estimated that 62-68% of the phosphorus sedimentation was accounted for by the PP re-suspension in the lake Peipsi, Estonia. Although strongly followed by a sedimentation trend, the re-suspended PP might become a source of SRP if pH values are high in the water column (Koski-Vähälä and Hartikainen, 2001).

2.2.4 Emissions From Different Land-use Systems

Anthropogenic sources of phosphorus can be distinguished into “point sources” (sewage treatment plants, domestic septic systems and pipe drained agricultural field or diary drainage) (Edwards and Withers, 2007), and “diffuse sources” (agricultural fields, rural and urban impervious areas and forests) (Withers and Jarvie, 2008). As the sources might emit different forms of phosphorus to surface waters, an assessment of these within the catchment might be of importance in this

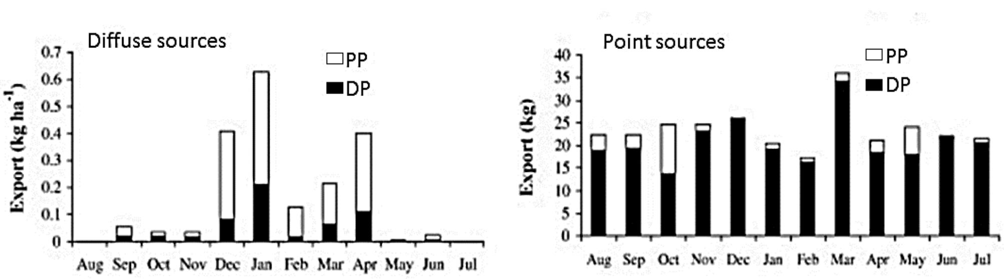


Figure 2. Estimated loads of PP and DP from point (kg) and diffuse sources (kg/ha), adapted from Withers and Jarvie (2008).

study. In general point sources are often coupled with the emissions of DP whereas diffuse sources represent the delivery of PP (fig. 2).

Impact from point sources has decreased in recent decades, as they are easily defined and controlled i.e. through effluent management at sewage treatment plants (Carpenter et al., 1998). Some point sources, such as the vast amount of domestic septic tanks, are difficult to control and quantify, making them a substantial contributor of phosphorus to water bodies. Furthermore, artificial hydrological network, such as sewer systems below urban areas, might overflow as a consequence from heavy precipitation and by that becoming an uncontrolled source of DP to surface water bodies (Withers and Jarvie, 2008).

Today, the highest concern is the quantification of the diffuse sources. Indicated by figure 2, delivery of phosphorus has a high seasonal variation. This is mainly due to the hydrological dependency of transport and seasonal agricultural inputs. In a study by Dorioz et al. (1998), the transport of TP from both point and diffuse sources was assessed from four different hydrological regimens within the catchment of river Foron. Their study showed that surface transportation of TP increased with increasing precipitation and runoff, especially from the diffuse sources of the agricultural and urban areas. Similar trends have been observed in Sweden, where elevated levels of discharge are followed by an increase in PP loads from agricultural systems (Ulén et al., 2012, Ulén and Persson, 1999).

Agricultural fields

The form of phosphorus emitted from agricultural fields depends on soil type and management practices. In Sweden, the most common field drainage practice is through tile drains at 1 m depth with open inlets colleting surface runoff (Ulén and Jakobsson, 2005). This type of drains will increase the risk of PP losses to surface waters, as it prevents the sieving effect by the soil matrix (Ulén et al., 2007). In addition, poorly drained soils also contribute to detachment and transport of PP due to surface erosion. Ulén and Jakobsson (2005) identified silty and clayey agricultural soils as prone to erosion and by that contribute to high losses of PP, especially during seasons with sparse/no vegetation cover and high rates of runoff. Moreover, leaching of DP might also occur from agricultural fields with low sorption capacity i.e. sandy soils. However, the behaviour of DP concentration differs from PP regarding its relationship to flow conditions. Concentration of PP in streams tend to increase as a result from high flows, whereas concentrations of DP might be diluted due to the increased amount of water (Novak et al., 2003).

Pastures

Pastures have also been identified as contributors of phosphorus to surface waters. A study was conducted by Mcdowell et al. (2006) on phosphorus losses from grazed pastoral systems in New Zealand. Their results indicated that 30-40% of PP was generated from dung, 10% from fertilizers and 50% from the soil itself. Mcdowell et al. (2007) found in another study that a relatively larger fraction (38%) of TP was PP when dung and treading were prominent elements in the pastoral grazing system. Their results imply the importance of soil structure in terms of surface runoff as high degree of treading will prevent infiltration.

Urban areas

Regarding urban areas, the rate and amount of phosphorus transport might increase due to conversion of pervious into impervious surfaces. The low infiltration capacity properties of these surfaces will limit the sub-surface sorption processes. Here, phosphorus might derive locally from fertilized lawns, constructions sites or sewer overflow, or transported from remote agricultural fields through hydraulic connectivity. Line et al. (2002) estimated that the phosphorus export from a multi-use urban catchment could be 302% larger than pre-developed conditions. A study by Cowen and Lee (1976) found that the fraction of PP in the TP transported from urban areas varies, with the largest fractions of PP deriving from erosive construction sites.

Forest

The extensive coverage of forests in Sweden makes the quantification of the phosphorus losses from these systems interesting in the scope of this study. Most documented losses of phosphorus from Swedish forests are at background values (Uggla and Westling, 2003). At some locations, an increase in organic phosphorus losses due to soil erosion after final felling has been documented (Löfgren, 2007). Forests in the south of Sweden are more prone to higher export TP of concentrations, possibly due to higher concentrations of organic matter in the streams (Uggla and Westling, 2003). In water quality management practices, forested buffer along streams are commonly applied, which helps to trap eroded soil in surface runoff. These systems might be comparable to natural forest systems (Borin et al., 2010).

2.2.5 Importance of Buffer Zones

Arguments in literature regarding the importance of near stream buffer zones on the water quality vary (Borin et al., 2010, Nava-López et al., 2016, Johnson et al., 1997, Sliva and Williams, 2001). Johnson et al. (1997) compared the influence on water quality of landscape factors between whole catchments and 100 m ecotones. Their results indicated that the total variation in TP found in the streams was better explained by the landscape data of the 100m ecotone. In a different study, Sliva and Williams (2001) showed that 100 m buffer strip accounted for less variation in water quality than data from the whole catchment, especially for total solids regardless of the season.

2.3 Sensor Parameters for Surrogate Measurements

2.3.1 Turbidity

In water quality assessment, nephelometry is often used to measure the attenuation of light in water, which translates into units of turbidity (Davies- Colley and Smith, 2001). The theoretic framework behind is based on the light scattering capacity of particles in water when exposed to an incident light beam captured by a detector. A high degree scattering and attenuation will result in a high reading of turbidity. The most commonly used units for turbidity is FNU (Formazin Nephelometric Unit with an infrared light source) or NTU (Nephelometric Turbidity Unit with a white light source), both of which detect the scattering light at a 90° with a single detector (USGS, 2017). The 90° angle to the detector is preferred due to its efficiency in capturing scattered light from a wide range of different particle sizes. Other methods are also available, however not as frequently applied as the FNU and NTU methods (Sadar, n.d.). The measured turbidity is relative in a sense as the different methods and technology will generate different values and hence not comparable in data analysis (Ankcorn, 2003).

Apart from instrumental variations of turbidity values. Discussed above, variations due to catchment, in-stream and particle properties might also influence the readings on the turbidimeter. The land-use and soil properties within a catchment will have an impact on the material transported to the water and the flow rate within the stream will determine the materials in suspension. In streams, variables such as organic matter, sediment, organisms and dissolved material will generate readings of turbidity but as these particles varies in terms size, shape and colour, the optical properties will vary (Perkins, 2013). Waters high in organic material might result in unexpectedly low turbidity readings due to their capacity to absorb the incident light beam (Ankcorn, 2003). However, this error might be accounted for by using an infrared light source, such as in the FNU method, which operates within high and wide range of wavelengths (USGS, 2016).

Turbidity as a proxy for total phosphorus

Several previous studies have implied the suitability of using turbidity as surrogate measurement for suspended particles (Gippel, 1995, Christensen et al., 2000, Fölster and Rönnback, 2015). Recent instrumental development of turbidity sensors allow efficient continuous *in situ* monitoring of short sampling frequencies (Lewis, 1996). Apart from generating data of high temporal resolution, logistical circumstances of samples handling will also reduce. Successful findings of correlation between total suspended material (TSS), TP and turbidity have been done. Jones et al. (2011) set up monitoring stations at two locations (upstream and downstream) in the Little Bear River in Utah, with different hydrological and water quality conditions. Data was collected on TSS, TP and turbidity to develop correlation models to estimate TP and TSS as function of turbidity. The results were promising in terms of TSS estimates from turbidity at both locations. However, the relationship between TP and turbidity was stronger at the upstream location, which might be due to the higher export of PP from the catchment. Furthermore, they suggest improvement of the method, and relationship between TP and turbidity, by accounting for the DP fraction with surrogate measurements of conductivity (EC). A more recent study by Fölster and Rönnbäck (2015) also showed different correlation patterns at different locations. Turbidity, EC and TP data from up to 255 stations of the Swedish National Monitoring Program between 2010 to 2012 was used in the study. The authors found that the combination of conductivity and turbidity generated better correlation coefficients than when only using turbidity. In conclusion, the author suggested that the difference in correlation was due to soil type and land-use within the different catchment.

2.3.2 Other Sensor Variables

Other water quality parameters possible to monitor continuously with sensors are electric EC, water temperature, total organic carbon (TOC) and pH. EC is associated with dissolved solids in the water, most likely different types of salts, as it measures the ability of water to conduct electric currents in units of mS/m (Chapman, 1996). The charged properties of DIP and DOP might justify the inclusion of EC, to account for the fraction SRP in TP (Schilling et al., 2017). Temperature is a strong regulator in terms of biological activity and chemical reactions in the water. There is i.e. a strong relationship between EC and water temperature, as warmer waters will show higher values of the EC due to increased ionic activity and consequently dissociation of larger molecules (Barron and Ashton, 2005). Warmer waters often enhance the biological activity and increases primary production and thereby the uptake of i.e. DP. Strong complexation between phosphorus and hydroxides prevail at low pH, which makes pH a contender in the explanation of the fraction DP in TP.

TOC has also shown strong correlations with TP, making it a suitable surrogate for TP, especially if the phosphorus is in organic form (Miguntanna et al., 2010). However, most of the variables mentioned are correlated in natural waters. Correlation between prediction parameters could cause a risk of overestimation of each parameters prediction capacity. EC is for instance strongly correlated to water temperature and TOC and TOC in turn could show strong correlation to turbidity (Chapman, 1996).

3 Materials and Methods

Available secondary datasets were downloaded and analysed with different statistical methods.

- To construct the total phosphorus (TP) prediction models, linear regression methods were applied on the water quality data. The method is efficient in constructing prediction equations (Christensen et al., 2000). Moreover, the method requires certain data properties, such as normality, independent variance, and linear relationship, which was fulfilled by the water quality data used here.
- Principal component analysis (PCA) was applied as an initial step to find latent structures within the landscape dataset. The method is often preferred for data of many variables that might be correlated. These correlations are easily detected with the method (Eriksson et al., 2006).
- Partial least squares (PLS) was used to find explanatory landscape factors that influence the strength of the TP regression models at the different sites. The method is similar to PCA and allows the use of datasets of many variables, which might be correlated and of lower quality (skewed, noisy, few observations). The main difference between PLS and PCA is that the PLS model explains both the variation in a prediction and a response dataset. This enables the detection of defined variable relationships between the two datasets (Nash and Chaloud, 2011).

3.1 Total Phosphorus Models

3.1.1 Water Quality Data

Water quality data, containing parallel measurements of total phosphorus (TP) and turbidity, between 2010 to 2016 was extracted from MVM environmental data at SLU. The database stores data from different surface water monitoring programs nationwide. Due to the different purposes and actors behind the monitoring programs, a great variation in data content and quality was retrieved, which required some refinement. To reduce modelling error due to missing values, some criterions were set. Data relevant for the analysis, turbidity (FNU), TP (µg/l), conductivity (EC) at 25° (mS/m), pH, TOC (mg/l) and water temperature (ºC) were required to be present for each year for the stations to be included in the study. The selection resulted in 194 stations from different monitoring programs (table 1). All water quality parameters included in the dataset are possible to measure with sensors, except from TP. The data was used to create regression models, suitable for prediction of TP, at the different stations.

Table 1. Different monitoring programs in which the stations of the water quality data in this study are included

Monitoring program	No. of stations included
National Monitoring Program	142
Regional Monitoring Program	35
Coordinated Recipient Control	10
Program of Värnens Affluent	9
Program of Mälarens Affluent	8
Other Water Quality Programs	65

3.1.2 Statistical Methods

Simple Linear Regression (SLR)

SLR was first applied to get an overview of correlations between the different sensor variables and TP. The analysis was performed in the statistical software JMP 12.1.

In SLR one independent variable is plotted against one response variable to find a linear relationship. To obtain a viable linear model some data criterion must be fulfilled, such a normal distribution of the data, constant residual distribution and independent observations, where the two former criteria might be managed through normalization of the data. The general model of SLR consists of Y, the response variable, X, the independent variable, the regression coefficient, β_1, the intercept β_0 and ε, the residuals. The basic theory of SLR is that Y should change with X in β_1 units for a linear relationship.

$$Y = \beta_0 + \beta_1 \times X_1 + \varepsilon$$

The regression is then evaluated, in term of fit to the data, by the R^2 value. A R^2 close to 1 means a good fit and the model can explain a large variation within the data. Furthermore, the regression should be tested for significance in a F-test, where the alternative hypothesis (H_1) is favoured over the null hypothesis (H_0) if the obtained p-value is < 0.05 (see further details in MLR section below).

Multiple Linear Regression (MLR)

In the present study, turbidity, EC, TOC, water temperature and pH were added in in stepwise approach to find the best correlation model for TP at each of the 194 stations. Turbidity was included in all model combinations due to the assumption of its strong relationship with particulate phosphorus (PP). EC was considered the second most important parameter, as it might account for the dissolved phosphorus (DP) fraction of TP. TOC, pH and water temperature were added in the attempt to increase the R^2 adjusted values further. To obtain a normal distribution of the data, and to enable the comparison between the variables, logarithmic values of turbidity, EC and TOC were used. The highest correlation coefficient at each station was then compiled into a dataset to use as the response variable in Partial Least Squares analysis, where the landscape factors were used as prediction variables.

Regression models for TP (Y) and the different possible sensor variables (X) were computed with Multiple Linear Regression (MLR) in the statistical software JMP 12.1, for each station. In MLR, multiple independent variables (X) can be included in a stepwise approach to find explanatory model parameters for the response (Y)

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \dots + \beta_i \times X_i + \varepsilon$$

Where β_0 is the intercept and β_i is the regression coefficient for the i^{th} variable. Additions of explanatory variables will automatically increase the R^2 value, which is a result from an increase in the number of observations (n). However, only looking at the R^2 could result in overlooking the risk of adding unexplanatory variables to the model (Grandin, 2012). To account for this, a R^2 adjusted value can be estimated by

$$R^2_{adjusted} = 1 - \left(\frac{n-1}{n-p}\right) \times (1 - R^2)$$

where n-1 is the degree of freedom and p is the reported result from a two- sided t-test for significance for the variable regression.

The models are then further evaluated in terms of significance using F-statistic, which utilize the calculated sum of squares error (SSE) from two hypothetical models

$$H_A = Y = \alpha + (\beta_1 \times X_1) + \varepsilon$$

$$H_0 = Y = \alpha + \varepsilon$$

In theory, the alternative hypothesis (H_A) will generate a smaller SSE than the null hypothesis (H_0) due to addition of model parameter (variable coefficient). The question is how small, which might be answered by the F-ratio

$$F = \left(\frac{SSE_{H_0} - SSE_{H_A}}{df_{H_0} - df_{H_A}}\right) \div \left(\frac{SSE_{H_A}}{df_{H_A}}\right)$$

A large F- ratio will indicate a small SSE in model of the H_A, justifying the rejection of the H_0 (PSU, 2017). The definition of large is however determined by the degrees of freedom, which will influence the critical value of the F-ratio. Here, the model significance was evaluated with the Prob>F value. Values of Prob> F <0.05 indicate a possibility of significant influence from at least one variable in the regression.

When performing MLR, it is important to pay caution to multicollinearity, strong relationships between the independent variables, which creates error during the estimation of the contribution of the model parameters (Graham, 2003). One way to detect multicollinearity in the data is to examine the Variance Inflation Factor (VIF), which is calculated from

$$VIF_k = \frac{1}{1 - R_k^2}$$

The R^2_k is the R^2 value of the regression line of the investigated variables plotted against the remaining independent variables (PSU, 2017). High VIF values, >5, are indicating multicollinearity and exclusion of the variable should be considered (Stine, 1995).

For all multi-parameter models, the VIF was investigated along with the significance of the different parameters in the model. If a variable did not fulfil the confident interval of significance (p>0.05), the parameter was removed from the regression model and another parameter combination was tested. The same was applied for variables with VIF>5.

3.2 Explanatory Landscape Data Models

3.2.1 Landscape Data Collection

Land-use/Land-cover

For each monitoring station, the coordinate system obtained in SWEREF99 from MVM environmental data was converted into RT90 2.5g V, which could be combined into a catchment id with a delineated catchment stored in the internal AROS database at SLU. The catchments have been processed within the ViVaN project (Virtual Watercourse Network). A digital elevation model (GSD+50) was combined with a road map (1:100 000), to visualize the water occurrences, and manipulated by exaggeration of high and low points, to create a defined drainage network with flow directions. Exaggeration of the high point created well-defined water divides and enhanced the extraction of catchments (Nisell et al., 2007). Some catchments were extracted from Swedish Meteorological and Hydrological Institute (SMHI) GIS database. The Svaro_2012_2 catchment dataset was downloaded, which contains a polygon layer of sub-catchments aro_y_2012_2 as well as the delineations of the larger main catchments to which they belong. Each sub-catchment has been generated from a 50m grid elevation model by manual identification of water divides. Each catchment also has an inlet and outlet point (SMHI, 2004). In ArcGIS 10.4.1, relevant sub-catchments from the aro_y_2012_2 polygon layer were extracted by selecting the catchment polygons containing the stations in interest. Luckily, most stations were positioned closed to the outlet of the sub-catchments. These were then matched, by the specific AU_CD id, with the main catchment in the Svaro_2012_2 dataset.

Land-use data was extracted from the Svensk Marktäcke Data (SMD), which was downloaded as an 8-bit raster dataset, containing an attribute table with different identified land-uses in Sweden. The data was developed as result from the Europen Environemntal Agency (EEA) program Corine (Coordination of information of information on the environment) initiated in 1985. Part of the program is the Corine Land-use Cover (CLC), for which the SEPA and Lantmäteriet have made contributions by compilation of the SMD data layer. The identified land-use classes and spatial coverage in the SMD layer has the reference year of 2000. The mapping of the SMD was through remote analysis using LANDSAT TM images with a grid of 25x25m (i.e. Landsat 7 ETM, Landsat 5 TM) and map material from

Lantmäteriet, SLU, SMHI and Statistics Sweden (SCB). Within the SMD raster 58 land-use and cover classes can be found. Quality evaluation of the data was done between 2004 and 2005 and the results indicated a high standard. (SEPA, 2014)

SMD data for each catchment was tabulated in ArcGIS 10.4.1 by masking the raster area with the obtained catchment polygons. The data was then calculated from m^2 to % of land-use of the catchment area. A downscaling of the 58 land-use classes into 11 more relevant classes was also done, including point source emissions. The downscaling was done by combining land-use classes with similar properties, i.e. all forest classes as forest etc. (table 2). The catchments areas in km^2 (Total Area) and latitudinal positions (Lat.) are also included in the dataset, where the latitude could explain mean air temperature, due to the temperature gradient from north to south in Sweden.

Data on the distribution and emissions of point sources within each catchment was obtained from Swedish Environment Emission Data (SMED) for 2013 and 2014 and represented the largest point sources. A quantification of point sources within each catchment and the specific emissions based on the catchment area was made (kg/ha). Mean emission values of the two years was used in the analysis.

Table 2. All landscape factors (excl. soils) used as prediction dataset for strong total phosphorus models. Land-use variables in % of catchment area

Land-cover	**Id**
Mean P emissions 2013/2014 (kg/ha)	P.Em.
Latitudinal position of station	Lat.
Catchment areas (km^2)	Total Area
Water and wetlands	W
Miscellaneous agriculture (i.e. orchards)	MA
Pastures	P
Final felling	CF
Agriculture	A
Forest on mineral soils (all types)	F
Forest on mire	FM
Artificial green areas	AG
Developed areas (urban, residential, and exploited areas)	DA
Open vegetation (dense grasslands, heathlands, or general sparse vegetation cover)	OV

Soil Data

To account for the impact from different fractions of soil type within a catchment, soil data was extracted from the Geological Survey of Sweden (SGU) via SLU

Geodata Extraction Tool (GET). The downloaded soil data contained ESRI shapefiles of different soil classes in the scale of 1:25 000-100 000 of the soil base layer (JG2). These layers have been adapted from material of varying degree of detail due to the long and ongoing process of soil mapping in Sweden. Data of highest resolution has been generated from on-ground mapping, which covers more densely populated areas, whereas data of low resolution has been processed from orthophotos. For some features of the soil data, errors up 70m can be expected, such as areas adapted from old material or undefined boundaries between different soil types. The JG2 soil data gives a comprehensive coverage of the different soil types in Sweden found at 0.5m depth, which is the mapping depth in field surveys. In some cases, the soil data in JG2 might coincide with the dataset of surface soil (JY1), occurring at depths less than 0.5m. (SGU, 2016)

The JG2 dataset contained two different attribute tables, JG2_Legend and JG2_TX, with minor differences in soil classification. The attributes in JG2_TX was chosen due to the higher content of clay represented. The extraction of the soil data was executed in ArcGIS 10.4.1 by tabulation of the soil layer by the delineated catchment areas. All values were recalculated from m^2 to % of the whole catchment area. Three soil datasets were compiled to use in the statistical analysis. One with the original soil classes from JG2_TX and a simplified version with clay rich soil (Cl-rich), coarser grain soils (Sa-rich) and rock out crops (RO) [1]. The third dataset represented the potential erodability of the soil and was a reclassification based on the SGU shoreline soil data (table 3). SGU's shoreline soils dataset does not consider potential water flow velocity, impact from waves and morphology but only the physical properties of the soil, such as cohesion (SGU, 2016). Thus, relatively

Table 3. Soil erodibility variables adapted from SGU JG2_TX soil classification layer and SGU classification for shoreline soils dataset.

JG2	JG2_TX	J_ENKEL	J_ENKEL_TX	EROD	Erodibility	Id
50	Glacifluvial	87	Sand or gravel	4	Potentially high erodibility	H
75	Peat	75	Peat or gyttja	3	Moderate erodibility	M
84	Postglacial sand--gravel	87	Sand or gravel	4	Potentially high erodibility	H
86	Clay--Silt	86	Clay--Silt	4	Potentially high erodibility	H
100	Till	100	Till	2	Low erodibility	L
888	Rock	888	Rock	1	No/very low erodibility	No
9792	Clayey till/till	100	Till	3	Moderate erodibility	M
9800	Till and/weathered soil	87	Sand or gravel	3	Moderate erodibility	M

[1] Water and peat were only included in the erodibility dataset. For the JG2 and the simplified soil datasets, the coverage of water and wetlands from the SMD was used instead due to higher resolution.

acceptable to apply, despite minor differences in the classification between the two datasets.

Buffer Zones 100m

For the 100m buffer zone, a polyline layer Vd_l_2012_2 (flowlines in all streams, rivers and lakes occurrences in Sweden compiled) was downloaded from SMHI GIS database. In addition, a polygon layer of all water surface occurrences, Vy_y_2012_2, was downloaded. This layer contains all the water surfaces, islands and mires in Sweden with significant size. These layers were then buffered with 100m on both sides in ArcGIS 10.4.1 to generate a polygon layer of the streams buffer zones. The following step was to clip the SMD and soil datasets with the buffer polygon. Finally, the SMD and soil data were tabulated using the defined catchment polygons, from AROS and SMHI Svaro_2012_2, to match the data with belonging station. The data was then transformed from m^2 to % of the buffer zone area within each catchment. To enable the comparison with the whole catchment data, similar land-use and soil classification were compiled for the 100m buffer zone. A total of 175 catchments are included in the buffer zone dataset.

The landscape data of land-use/land-cover, soil type, spatial location and catchment areas were used to explain the variation in the correlation coefficient values from the TP regression models.

3.2.2 Statistical Methods

Principal Component Analysis (PCA)

The landscape datasets were overviewed in a PCA model together with catchment area and latitudinal position of the stations. From the model, characteristics of the catchment could be detected as well as the spatial distribution. Distinct correlation patterns between the landscape variables included and catchment of more extreme land-use/soil types values could also be detected. The PCA was performed with MKS multivariate analysis tool SIMCA-P 14.0 by an autofitting function. When autofitting the data, SIMCA-P calculates the principal components (PC) with cross validation (CV), see details on CV in the PLS section below. Moreover, PCA requires data normality, and the variables with severe skeweness ($>\pm 1$) were log transformed to obtain relatively normal distribution. The loading and score plots of the PCA models were examined to find correlation between the landscape variables as well as the observation distribution pattern.

PCA is an efficient method to look at the structures within the dataset by finding latent structures in the data, which is expressed as axis (PC). The PCs are translated from the actual variables and their relationships within the dataset. As multicollinearity is a common problem when dealing with multiple variables, the extraction of latent structures allows the representation of all data despite the inner correlations (Jolliffe, 2002). The first computed PC explains most of the variation within the data. Several PC might be computed, as many as the actual number of variables in the dataset, but with decreasing explanatory capacity. PCA is performed on a dataset X with a matrix of n observations and m variables. The principal model of PCA is

$$X = TP + E$$

Where T are the scores, which explains the positions of the observations along the PC. Loadings, P, are weights of the variables and describe how there are contributing to the component. Variables with strong relationship are clustered together along the PC. The residuals are represented by E, which are not explained by the PC (Yu et al., 2010). For further details on the method see Joliffe, 2002.

Partial Least Squares (PLS)

In the scope of this study, PLS was used to investigate the correlation between the obtained R^2 adjusted values of the TP regression models and the landscape variables (land-use, soil type, latitudinal position, catchment area and point sources). The dataset of the R^2 adjusted values was used as the response dataset and the landscape data as prediction dataset. The modelling was performed in MKS multivariate analysis tool SIMCA-P 14.0, which calculates components (latent structures) with cross validation (CV) when the data is autofitted. In CV, the dataset is divided into groups. Each group is then excluded from the data. The remaining data is used to predict the excluded data. The predictions are then compared with the observed values in the excluded group, by looking at the Predictive Residual Sum of Square (PRESS). A small value indicates a model with high prediction ability as the predicted values are close to the observed data. This is done until all groups are tested (Eriksson et al., 2006). In SIMCA-P, the PRESS value is transformed into the Q^2 parameter by division with the original sum of squares, then subtracted from 1. Outliers in the data were detected in the score plot together with the Hotelling's T2 plot, which displays each observation's location from the model origin in the score plot against a 95 and 99% confidence limits. Observations beyond the 99% limit were considered severe outliers. The most important variables for the model was found in the Variance Influence on Projection (VIP) plot.

The heart of the PLS method is to find latent structure within the predictor dataset that also explains variations in the response dataset (Y). Many variables might be included in the analysis as the problem of multicollinearity is neglectable due to the extraction of components, reflecting hidden correlation pattern between the X/Y variables. Each component extracted brings forth information of the most important variables as well their correlation with the remaining variables. The principal model of PLS is

$$X = T * P' + E$$
$$Y = U * C' + F$$

Where T and U are score vectors P and C are loading vectors and E and F are residuals, not explained by the latent variables. The X matrix contains columns with different variables (K, denoted x_{ik}) and rows of observations (N, i). In a similar manner is the Y matrix structured, with the variables (M denoted y_{im}) in columns and observations (N, i) in rows. Prior to PLS modelling, the data is auto-scaled, meaning a subtraction of the average values. The pre-treatment will prevent the different scales to become a problem, i.e. larger scales, or variations, will seemingly have larger importance, which can be controlled by the auto-scaling (Wold et al., 2001). In finding good correlations between the Y and X data, PLS construct latent variables, also referred to as components (a) of X-scores, from the original X data with coefficients (w^*_{ka})

$$t_a = \sum_k W^*_{ka} X_{ik}$$

The X-score could then be multiplied with the loading value (p_{ak}), which should be good summaries of the X data, to limit the residual E (e_{ik})

$$X_{ik} = \sum_a t_{ia} p_{ak} + e_{ik}$$

By using the X-scores, Y can be predicted as

$$y_{im} = \sum\nolimits_a c_{ma} t_{ia} + f_{im}$$

Where c_{ma} are loading values for the different variables in the Y data and f_{im} the residuals (Eriksson et al., 2006). When analysing the data, the score-plot and loading plot might be combined to relate the properties of the observations to the variables.

It is also possible to take advantage of the calculated loading plot to determine which predictor variables that contribute most to the model. Variables located far from the model origin contributes the most, whilst variables with no/little contributions are positioned towards the plot origin. The score plot alone can be used to find patterns in the data as observations located close to each other in the plot have similar data properties. It is also possible to find outliers by examining the score plot by looking at observations that are positioned far from the rest. Evaluation of the PLS model was done by the cumulated R^2Y and Q^2 values, which express the combined model fit by all added components. An R^2Y close to 1 means a well-fitted model. Although the addition of components will increase the R^2Y value, the Q^2 might stay the same or even decrease due to model overfitting. Large Q^2 values (>0,5) will indicate a good prediction outcome of the created model. However, due to the risk of model overfitting, the difference between R^2Y and Q^2 should not exceed 0.2-0.3 (Miljöstatistik.se, 2017). The VIP plot can be used to evaluate each variable, where VIP<0.5 indicate irrelevant contributions and values >1 indicate strong contributors.

An initial model was created from different landscape variable combinations, where all the variables are kept despite VIP value, for both the whole catchment and buffer zone data. From this model, all variables could be evaluated in terms of significance and possible relationship to the response variable (R^2 adjusted values). A second model was created from the same landscape variables combination but excluding the observation outliers present in the first model. This step was performed to see if the outliers had a large impact on the model. A third model was created using only variables with mean VIP>0.5 as this might create a better model (Nash and Chaloud, 2011). The landscape variable combination with the most significant model was used as the prediction dataset in the modelling of separate groups of observations. These groups were defined from the different combinations of the water quality parameters in the TP models, which could help to indicate the landscape factors influence on the selection of surrogates.

3.3 Limitations

- The landscape data was a combination of data from two different databases, with different degree of resolution. This could have an influence on the interpretation of the results.
- The water quality data was required to have parallel measurements of turbidity and TP, which reduced the amount of stations available to use. Moreover, the stations were reduced further due to the lack of delineated catchment areas.
- The soil data used was mapped at 0.5m, which might not fully represent the soil layer that is affected by surface runoff.
- The land-use data had a relatively high resolution, which makes it suitable to use. However, important factors such as management practices in the agriculture category and forest were not available.
- A time discrepancy existed between the prediction and response datasets. The land-use data had a reference year of 2000, whereas the water quality data was collected between 2010 and 2016.
- Some of the statistical methods used required data normality (PCA), which the landscape data lacked in many of the variables. Normalization of the data improved the skewness only slightly.
- In statistical analysis, the possibility to combine variables and datasets are infinite. Due to time constraints, only a few data combinations and subsets were tested.

4 Characteristics of the Catchments

4.1 General Site Description

The selected monitoring stations are distributed from north to south (between latitudes 68°28' N and 55°39' N), as well as west to east, all over Sweden (fig. 3). Clusters of stations can be found along near coastal areas, with higher density in the south-western and north-eastern regions. The belonging catchments of the stations

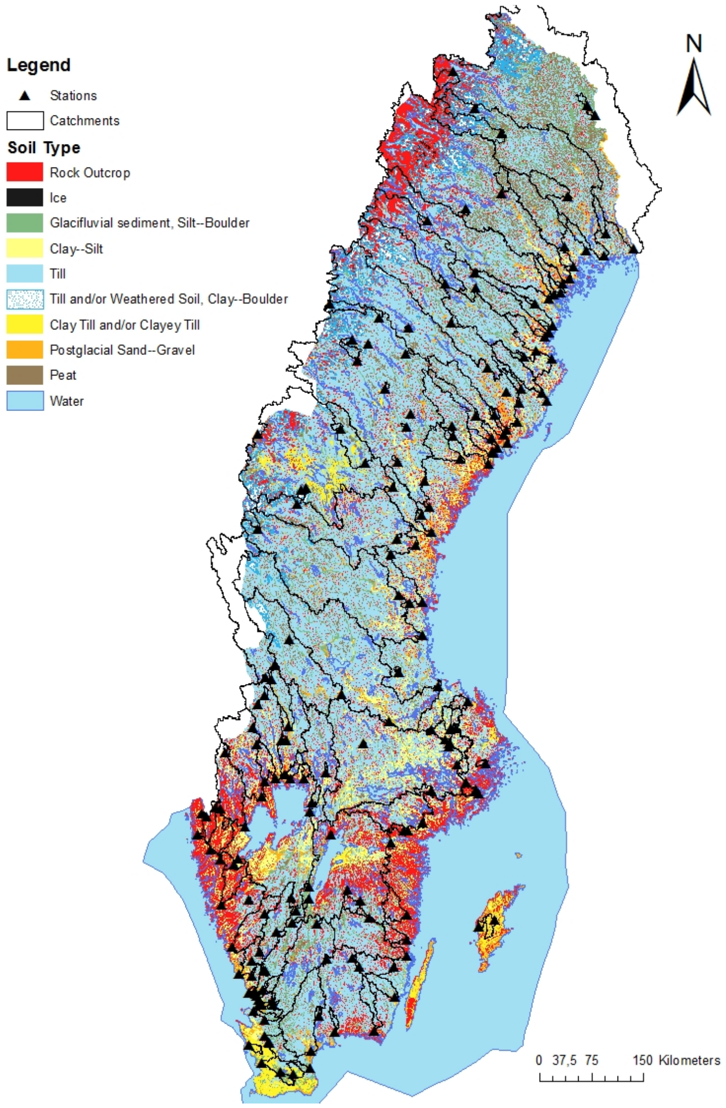


Figure 3. Distribution of soil types in Sweden and locations of the monitoring stations and their belonging catchments. Soil data adapted from SGU JG2_TX layer.

range from 0.5 km^2 to 48,152 km^2 and cover larger areas of the north-western parts and less coverage towards the coastal zones along the east and south coast (fig. 3). A gradient of mean temperature from -3°C to 8°C range from north to south. The mean annual precipitation also shows a distinct gradient, from west to east, ranging from 1,300 mm to 500 mm (SMHI, 2017). Till of varying properties dominates the surficial geology in Sweden (fig. 3). Weathered till with grain size from clay to boulder is found along the mountain ranges in the north-western regions. Soils high in clay/silt content are found at defined places towards the eastern, southern, and western parts of the country, as well as on the two islands Öland and Gotland. The highest concentrations of wetlands can be found far in the north. Veins of sandy deposits are also common characteristics throughout. Forests covers the largest parts of Sweden at approximately 69% followed by water and wetlands at around 9%, respectively. Agricultural lands have been estimated to approximately 8% with the distribution mainly in the southern parts (appendix 1). Developed and urban areas together, account for approximately 3%, mainly concentrated further south and around the larger cities (SCB, 2013).

4.2 Catchments and Buffer Zones Data Properties

All variables of the landscape data were skewed except from forest (F), which was relatively normal distributed (appendix 2). Figure 4 shows the distribution of the land-use and soil type fractions in the catchments of the 194 stations. More than 75% of the catchments had less than 10% of agriculture (A). Some stations, however, had catchments of larger coverage of agriculture, approaching 90%. Final felling (CF) was found at 8% or more in almost half of the catchments, where the largest overage approached 40% in a few catchments. Almost all catchments had zero or minor percentages of developed areas (DA), with only two catchments with significant coverage of developed areas at 14 and 33%. Forest (F) was present in all catchments, where 75% of the catchments had >50% forest coverage. More than 75% of the catchments had <7% forest on mire (FM). A few catchments had larger fractions, exceeding 20% coverage. Open vegetation (OV) was absent in most catchments, with a few exceptions of catchments with very high coverage approaching 100%. Only 25% of the catchments had >5% pasture (P). Water and open wetland (W) was the variable, which had the second most even coverage among the catchments, after F. Almost half of the catchments had >10% coverage and 25% of the catchments had about 20% coverage or more. Soils of low erodibility showed a large range among the catchments, where 75% of the catchments had 40% or more of this type of soils in the catchments. Highly and medium erodible soils

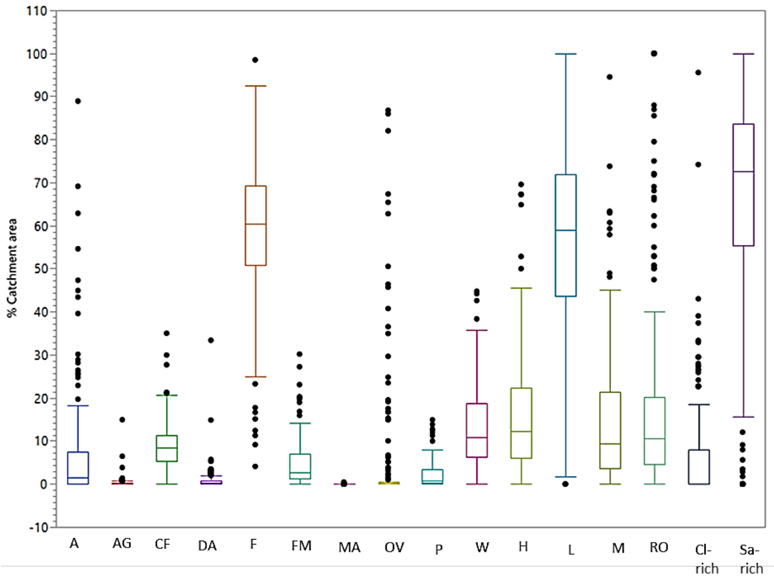


Figure 4. Box plot of percentage land-use data distribution of the whole catchments. Boxes contain 50% of the catchments, with the median represented by the bar across the box. The whiskers represent 25% of the catchments each, were the end-points are the minimum and maximum values. Black dots are outliers and represent catchments of deviating values of the variable. **A**=agriculture, **AG**=artificial green areas, **CF**=final felling, **DA**=developed areas, **F**=forest, **FM**=forest on mire, **MA**=miscellaneous agriculture, **OV**=open vegetation, **P**=pasture, **W**=water and wetlands **H**=high erodible soil**,** **L**=low erodible soil, **M**=medium erodible soil, **RO**=rock outcrops, **Cl-rich**= clay and silt, **Sa-rich**= coarser grained soils.

were equally common in the catchments, where 25% of the catchments had > 20% of these soil types. There were catchments of more extreme coverage of highly and medium erodible soil, approaching 80% of the catchment area. Finer grained soils were present in half of the catchments, 25% had fractions between 10 to 20%. There were, however, several catchments with extreme values between 25 to almost 100%. Coarser grained soils were more common and had larger coverage in a high number of catchments. Almost all catchments had fractions >10% with a few exceptions of catchments that had less. 50% of the catchments had > 75% coverage of coarser grained soils.

The 100m buffer zone area of the streams of the stations showed similar distribution pattern as the whole catchments (fig. 5). The largest difference between the two datasets was higher frequency of extreme values within the buffer zone dataset, especially for the W and soil variables.

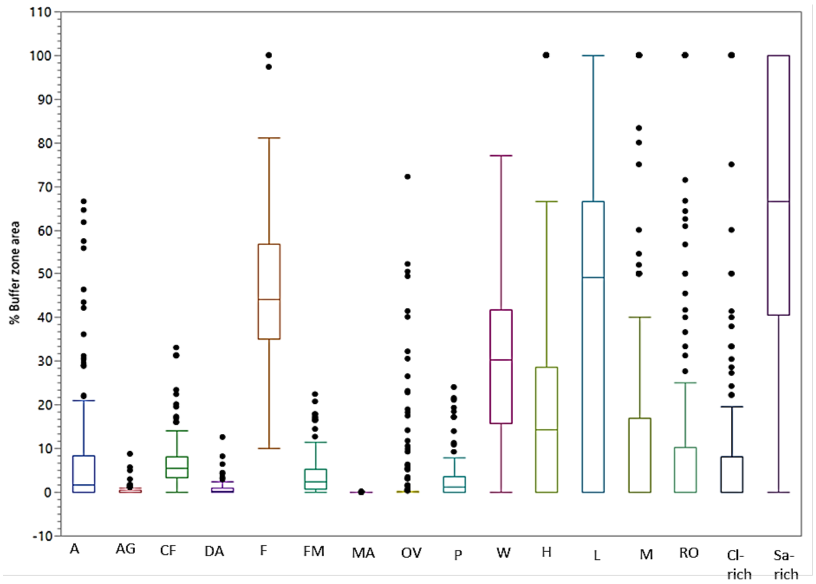


Figure 5. Box plot of percentage land-use data distribution of the buffer zones 100m, Boxes contain 50% of the stations, with the median represented by the bar across the box. The whiskers represent 25% of the stations each, were the end-points are the minimum and maximum values. Black dots are outliers and represent stations of deviating values of the variable within their buffer zone area. **A**=agriculture, **AG**=artificial green areas, **CF**=final felling, **DA**=developed areas, **F**=forest, **FM**=forest on mire, **MA**=miscellaneous agriculture, **OV**=open vegetation, **P**=pasture, **W**=water and wetlands **H**=high erodible soil**, L**=low erodible soil, **M**=medium erodible soil, **RO**=rock outcrops, **Cl-rich**= clay and silt, **Sa-rich**= coarser grained soils.

4.3 Correlated Landscape Factors & Observation Patterns

4.3.1 PCA Model

A four-principal component PCA model was autofitted to the whole catchment data, with all landscape factors included. From the PCA model correlations between the landscape factors were detected, as well as the distribution of the observations. The 1st PC of the model explained the largest variation in the data (table 4). Overfitting was not a problem due to the acceptable difference between the R^2X and Q^2 (<0.3).

Table 4. The different components in the PCA model of the whole catchment data (all variables included) and their associated R^2X and Q^2 values. R^2X (cum) and Q^2(cum) enables the evaluation of the model.

Component	R2X	R2X(cum)	Q2	Q2(cum)
1	0.296	0.296	0.233	0.233
2	0.154	0.45	0.0723	0.288
3	0.145	0.595	0.166	0.406
4	0.0872	0.683	0.0947	0.462

A cluster of anthropogenic variables along the negative end of the 1st PC, indicated correlation among them (fig. 6, top). All variables, which reflected the finer soil types from JG2_TX data were also correlated (Cl-rich, Cl-Si, H). In fact, all variables located in the cluster could be correlated to each-other. On the opposite

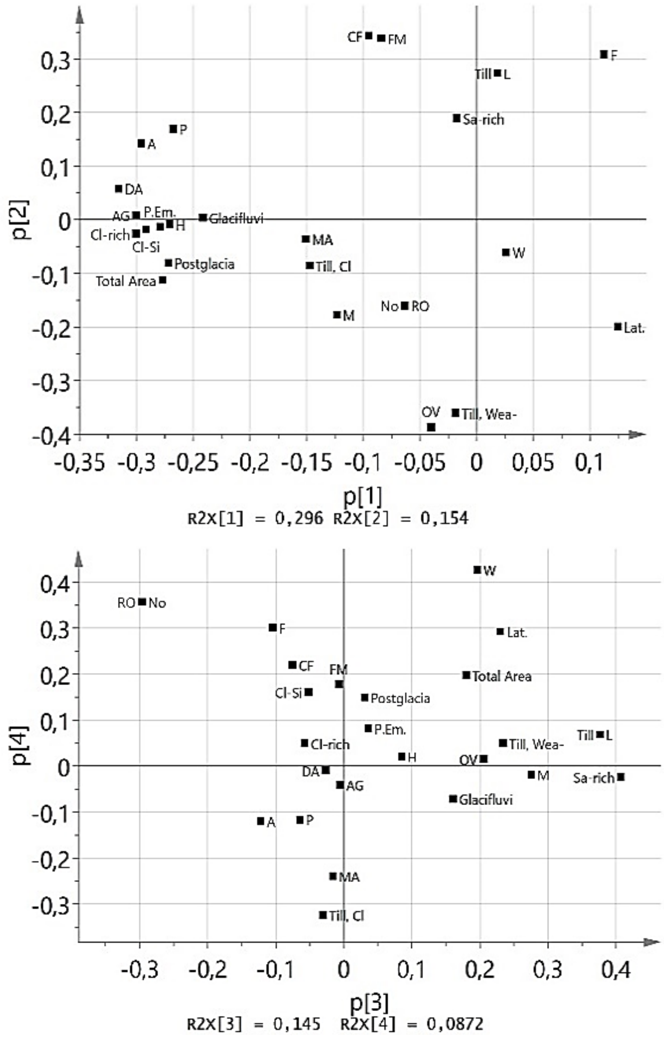


Figure 6. Loading plots of the four-principal components PCA model of the whole catchment dataset, all variables included. **Left)** 1st and 2nd PC. **Right)** 3rd and 4th PC. **Sa-rich**= coarser grained soils, **Cl-rich**= clay and silt, **RO**=rock outcrops, **Glacifluvi**=glacifluvial sediment, **Postglacia**=postglacial sediment, **Till, Cl**= clayey till, **Cl-Si**= clay and silt, **No**= non-erodible soil, **L**=low erodible soil, **M**=medium erodible soil, **H**=high erodible soil, **Till,We**= till or weathered soils**, FM**=forest on mire, **CF**=final fellings, **MA**=miscellaneous agriculture, **P**=pasture, **OV**=open vegetation, **F**=forest, **A**=agriculture, **P.Em**.=point source emissions, **Lat**.=latitudinal position, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands

side of the plot origin of the 1st PC, forest (F) and latitude (Lat.) were prominent variables, indicating a negative correlation with the cluster of variables opposite. The 2nd PC was dominated by final felling (CF) and forest on mire (FM), which might be strongly correlated. Open vegetation (OV) and till/weathered soil were located opposite, far to the PC negative end, which implied negative correlation to CF and FM. The relationship of the discussed variables was further confirmed in the loading plot of the 3rd and 4th PC (fig. 6, bottom), i.e. the cluster of variables along the 1st PC all co-moved towards the origin of the plot. Variables of importance in the 3rd PC were till and coarser grained soils (Sa-rich) and in the 4th PC rock outcrops (RO), water and wetlands (W) contributed to large weights on the positive side of the component. RO and W seemed to have a negative correlation to clayey till (Till, Cl) and miscellaneous agriculture (MA).

The PCA model of the buffer zone data explained only 58% of the data variation and showed slightly different correlations patterns of the landscape factors (appendix 3). The model consisted of four components, where the first component explained the most variation. Here, FM and CF also seemed to be correlated as well as the anthropogenic factors. The 3rd and 4th PC differed the most from the PCA model of the whole catchment data, i.e. highly erodible soils (H) was more important here (appendix 3).

An overview of the observations (stations) in the dataset indicated some preferences in the distribution pattern. A larger number of observations was located on the positive side of the 2nd PC (fig. 7, top, fig. 6, top). There was also a dense group of observations cluster on the negative side of the 1st component, close to the origin. From the score plot outliers were detected. The presence of outliers was expected, as the PCA model only explained 68% of the data variation. Outliers in 1st and 2nd PC might all be located further north in the country with high fractions of open vegetation (OV) due to their position along the plot axis, coinciding with the OV variable in the loading plot (fig. 7, top). The outliers in the 3rd and 4th PC might have higher fractions of rock outcrops (RO) in the belonging catchments due to the coinciding position in the loading and score plots (fig. 7, bottom, fig. 6, bottom). The outlier, not at all explained by the model PC is Råån Helsingborg. The catchment of this station had the highest fractions of agriculture (88%) and clayey till (95%) in the whole dataset. Regarding the water quality correlation properties (gradient of R^2 adjusted), a more heterogeneous group of observations are found in the lower left corner of the score plot (fig. 7, top), and a higher frequency of high R^2 adjusted on the positive side of the 2^{nd} PC.

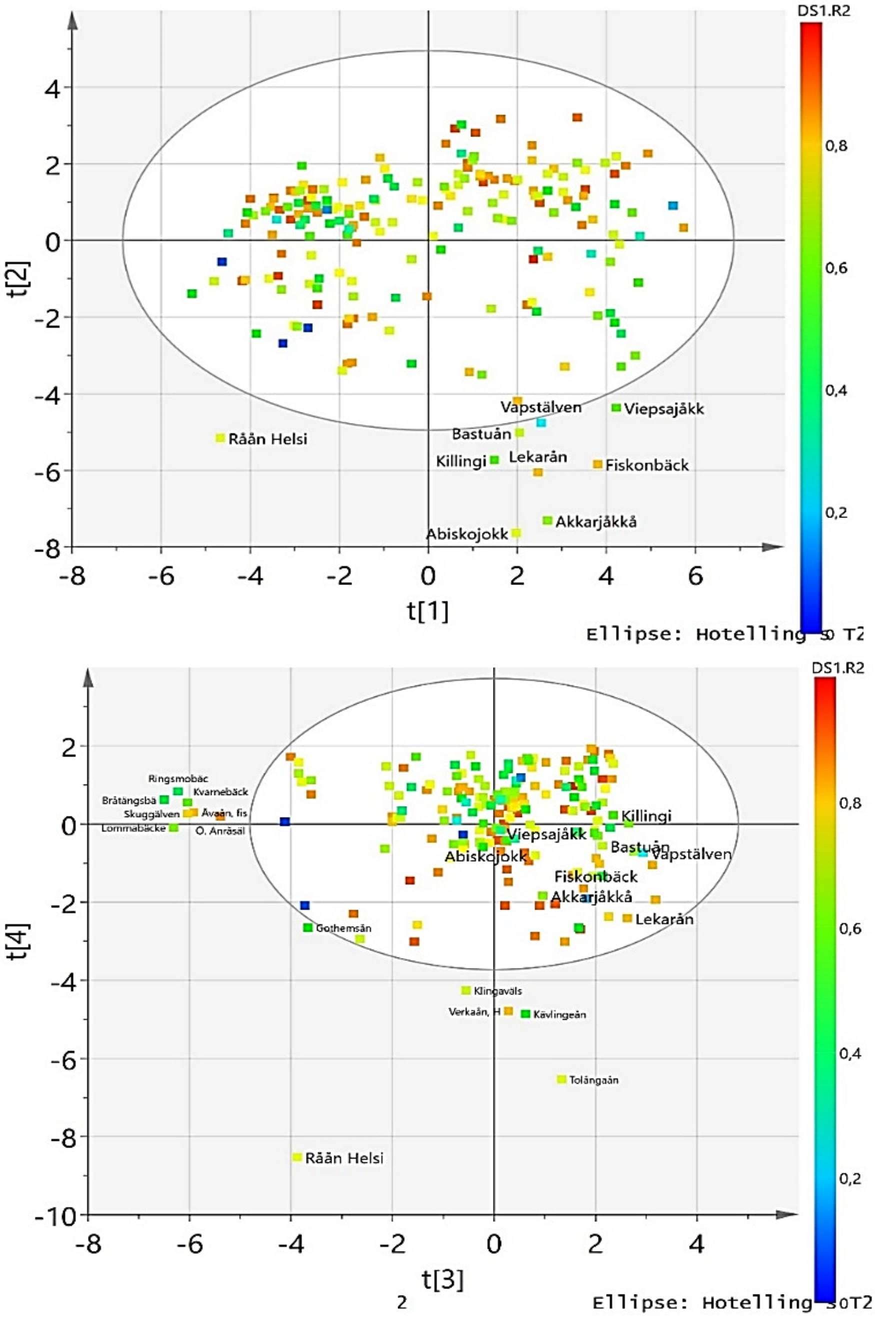


Figure 7. Score plots of the observations (194) in the whole catchment dataset. The observations are colored according to their R^2 adjusted values from the MLR analysis. The outliers from the 1^{st} and 2^{nd} PC are name-labelled for identification. The white ellipse represents the 95% significant range in the Hotelling's T2 significant test, observations located outside the ellipse are considered outliers. **Top**) 1^{st} and 2^{nd} PC **Bottom)** 3^{rd} and 4^{th} PC

4.3.2 Interpretation of Landscape Data Structures

The most significant structure of the model reflected a north to south gradient. Variables associated with anthropogenic activity were all located far opposite from the latitude variable. In Sweden, the largest population is found in the south, which is where most of the developed areas, agriculture and point sources can be found. The model also revealed possible correlation between the anthropogenic factors and the clayey soil types, which indicates that they often co-exist at equal fractions within the catchments. As hypothesized in this study, larger fractions of agriculture, point sources, pastures, urban areas, and clay soils might result in higher export of TP to surface water. Together with the frequently found strong correlations between turbidity and TP in literature, strong correlation models were expected for catchments high in these variables. However, phosphorus mobilization and export pathways in the catchment are complex and the co-existence of the mentioned variables could add to the complexity. One possible outcome from the correlation pattern is the suppression of the expected impact from each individual source. I.e. during high runoff rates, concentrations of DP from point sources are expected to be diluted, whereas concentrations of PP from diffuse sources could increase (Bowes et al., 2008). This could in extension create large variability in the correlation pattern between TP and the selected surrogates.

The second component could also be a north to south gradient. Open vegetation was a significant contributing factor in the second component and the north-west of Sweden has extensive covers of open vegetation (appendix 1).

A larger number of observations were found opposite to the latitude an open vegetation variable in the score plots. This indicates that most of the stations included in this study are located towards the south. These regions also seemed to hold a larger number of sites with strong TP models. Outliers were detected outside of the significant range towards the latitude factor, indicating their northerly position in the country.

5 Results

5.1 Total Phosphorus Regression Models

Simple linear regression analysis (SLR), of all observations, showed that total phosphorus (TP) had the strongest correlation with turbidity (fig. 8). However, the SLR analysis of the separate stations in the data showed a large range of R^2 values, 0.01 to 0.9, between TP and turbidity (table 5). About 38% of the sites had $R^2 \geq 0.65$ (table 5). Turbidity was at all stations positive correlated to TP.

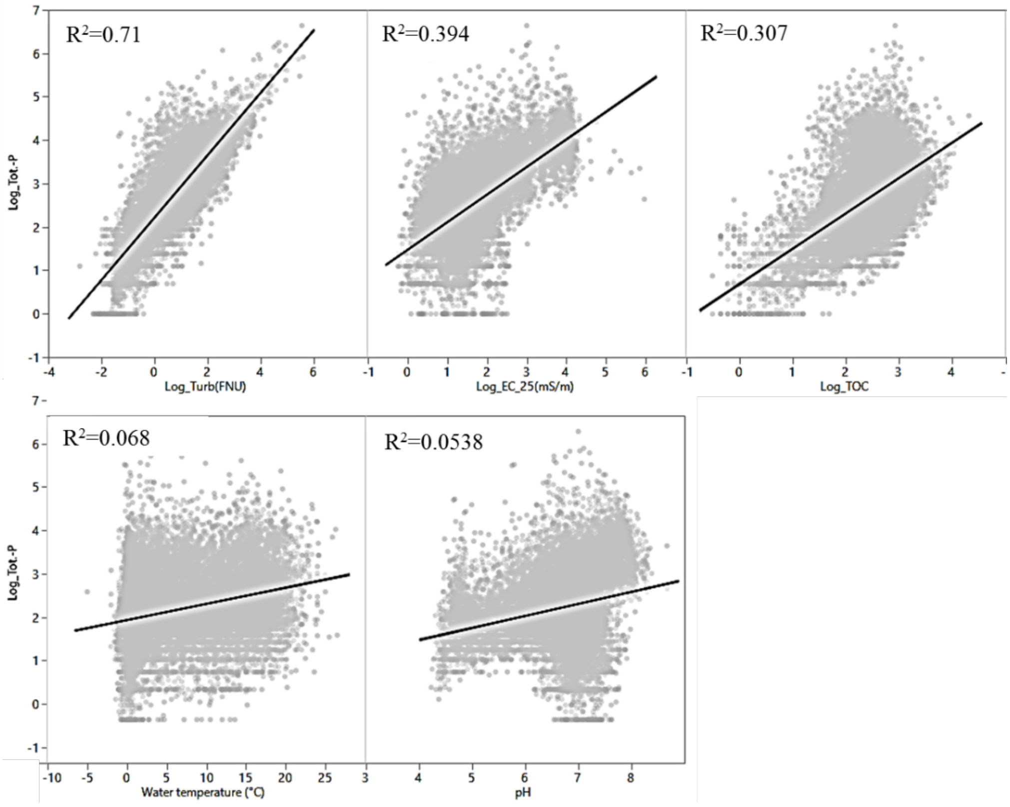


Figure 8. Bivariate SLR plots of all observations in the water quality dataset of the five different sensor parameters and total phosphorus.

When adding electric conductivity (EC) as a second parameter (TuEC), a higher percentage (46%) of stations obtained R^2 adjusted values ≥ 0.65 (table 5). Moreover, the range of the R^2 adjusted values increased in min and max values, 0.097-0.92. However, the addition of EC resulted in less significant models. EC was negatively correlated to TP at 71% of the stations with a significant model where EC was included as a parameter.

Table 5 Different configurations of the independent variables in the TP models and the obtained number of total significant models from each configuration. The number of stations with R^2 adjusted ≥ 0.65 as well as the range are also presented. The models are sorted with decreasing number of variables included. **Tu**=turbidity, **EC**=conductivity, **TE**=water temperature, **TO**=TOC. The model configurations discussed above are in bold

Model	No. of Variables	Tot. Significant Models	% of Stations with R2 adj. ≥ 0.65	Range R2 adj.
Tu	**1**	**194**	**38**	**0.01–0.9**
Tu EC	**2**	**105**	**46**	**0.097–0.92**
Tu TO	2	130	54	0.10–0.92
Tu pH	2	95	40	0.08–0.89
Tu Te	2	109	58	0.13–0.92
Tu EC Te	**3**	**60**	**62**	**0.34–0.91**
Tu EC TO	3	57	65	0.4–0.95
Tu EC pH	3	35	60	0.27–0.88
Tu TO pH	3	54	69	0.3–0.95
Tu TO, Te	3	64	70	0.36–0.92
Tu pH Te	3	91	69	0.39–0.91
Tu EC Te TO	**4**	**21**	**86**	**0.45–0.90**
Tu EC TO pH	4	15	73	0.3–0.89
Tu TO pH Te	4	28	82	0.39–0.92
Tu EC Te pH	4	29	83	0.29–0.92
Tu EC Te TO pH	**5**	**7**	**86**	**0.4–0.88**

When water temperature (Te) was added to the independent variables, with turbidity and EC (TuECTe), further increase in percentage (62%) of R^2 adjusted values ≥ 0.65 was obtained (table 5). Although, a reduction in total significant models, the range interval of R^2 adjusted improved 0.34 – 0.91 (table 5). Water temperature was positive correlated to turbidity at 92% of stations.

The addition of TOC to the TuECTeTO model generated only 21 stations with significant regressions but with a smaller range of R^2 adjusted values (0.4 – 0.9) and much higher percentage (86%) of R^2 adjusted values ≥ 0.65 (table 5). Positive correlations between TOC and turbidity were found at 90% of the stations.

In a TP model including all five sensor variables (TuECTeTOpH) only 7 significant models were generated. There was no improvement in terms of range or percentage of higher values compared to the previous MLR configuration of TuECTeTO (table 5). Negative correlations between TP and pH prevailed at 71% of the station with pH included in the most significant model.

Main reasons behind the decrease in the number of significant models when variables were added to the regression models were insignificant parameters as well as multicollinearity. The model combination of TupH had the least significant regressions (table 5), implying frequent insignificant contributions of pH to the regression. Among the water quality variables used in this study, TOC, EC and pH often showed multicollinearity with each other. Table 5 indicates this trend, models with a combination of these had lower numbers of significant models, even when compared to models with equal number of variables. The problem of multicollinearity also existed between other variables to a smaller extent i.e. between turbidity and TOC or turbidity and EC.

Subsets of stations from the whole dataset are presented in table 6. Each subset contained stations which had their relative highest R^2 adjusted value from the associated model in the table. Despite, large ranges of the R^2 adjusted values remained. However, 66% of all included stations (194) have R^2 adjusted values ≥ 0.65.

Table 6. Number of stations with significant MLR models by each variables combination. The selected stations have their relative highest R^2 adjusted values from the selected model. The numbers of stations sum up 194 stations, which is the whole dataset. **Tu**=turbidity, **EC**=conductivity, **TE**=water temperature, **TO**=TOC

Model	No. Significant Models	% of Stations with R2 adj. ≥ 0.65	Range R2 adj.
Tu TO	23	48	0.1 – 0.91
Tu TO pH Te	20	85	0.53 – 0.92
Tu TO Te	19	74	0.37 – 0.93
Tu pH Te	19	58	0.28 – 0.91
Tu	18	44	0.08 – 0.9
Tu EC Te	17	41	0.15 – 0.87
Tu EC Te TO	14	86	0.4 – 0.9
Tu TO pH	12	58	0.41 – 0.87
Tu EC	11	64	0.33 – 0.88
Tu EC TO	10	80	0.4 – 0.95
Tu Te	10	70	0.42 – 0.92
Tu EC Te pH	9	100	0.69 – 0.92
Tu EC Te TO pH (all)	5	80	0.41 – 0.88
Tu EC To pH	3	33	0.47 – 0.72
Tu EC pH	2	50	0.42 – 0.66
Tu pH	2	50	0.48 – 0.76

From the 194 significant models in table 6, a combination of three parameters represented most stations (fig. 9), 53 of these had R^2 adjusted ≥ 0.65. As indicated by table 4, the risk of obtaining multicaollinearity in the data was high whenever more than 3 variables were used as independent variables. Out of all significant models, TuTO represented most stations followed by the more complex TuTOpHTe (table 6). The latter had the highest number of stations with R^2 adjusted values ≥ 0.65. Most frequently occurring variables in the significant models were, except from turbidity, water temperature and TOC.

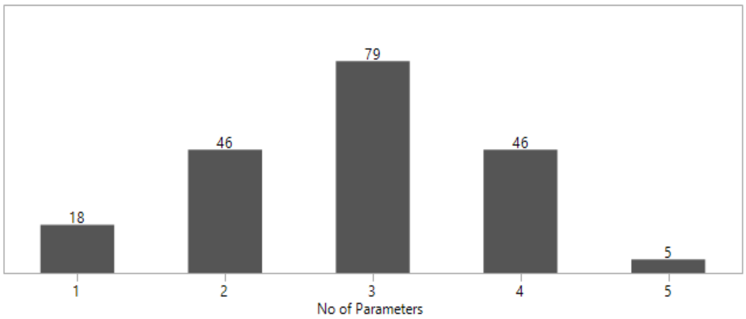


Figure 9. Distribution of stations, from the selection of significant models in table 4, and the number of parameters in the variables combination representing each station. Models with three parameters represent the highest number of stations.

A trend of higher R^2 values at stations of higher turbidity and TP concentartions existed in the group of stations with the Tu model (fig. 10). No such defined trend was found among the other groups.

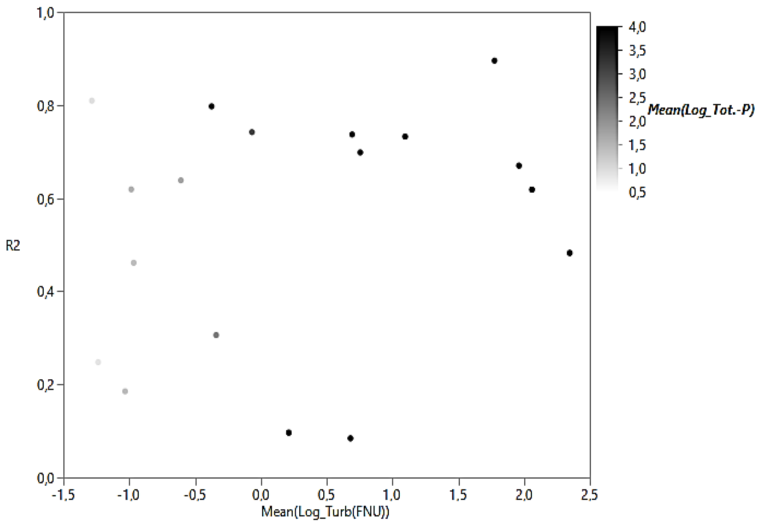


Figure 10. Mean log concentration of turbidity (x-axis) and TP (colour gradient) against R^2 adjusted values (y-axis) of 18 stations with only turbidity as the independent variable.

There were many stations that gained a significant increase in correlation whenever a variable was added. Examples are given in figure 11, where the inclusion of EC, water temperature and TOC increased the correlation coefficient substantially from the previous configuration.

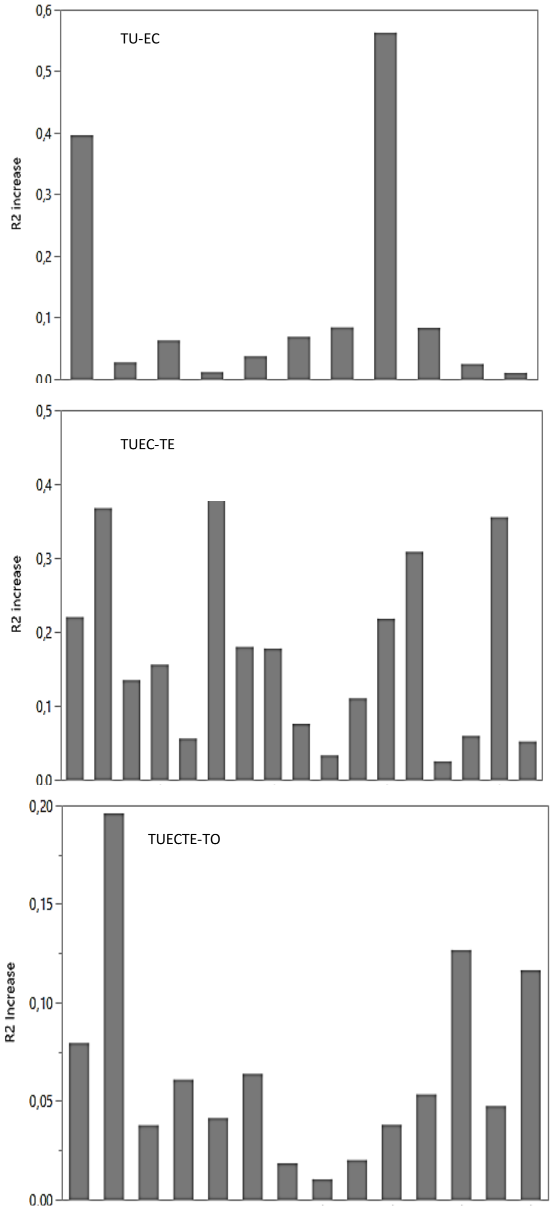


Figure 11. The increase in R^2 adjusted values at each station when EC (top), Te (middle) or TOC (bottom) is included as an additional independent variable. **TU**=turbidity, **EC**=electric conductivity, **Te**=water temperature, **TOC**=total organic carbon. Each bar represents a station with the given total phosphorus model. Y-axis is show units of R^2 adjusted values.

5.2 Explanatory Landscape Data Models

5.2.1 Whole Catchments

Whole dataset

All combinations of the whole catchment data in PLS, all 194 stations included, showed poor correlations with the R2 dataset (R^2 adjusted values from the significant TP models in table 5) (table 7). The removal of outliers did not improve the model fit nor the prediction capacity (table 7). Land-use/cover (LU) in combination with the simple soil data (Cl-rich, Sa-rich, RO) had the best model fit outcome, $R^2Y = 0.174$ and $Q^2 = 0.0648$.

Table 7. PLS correlation models of the whole catchment dataset. Different combinations of the landscape factors (X data) and their correlation to the MLR R^2 adjusted values (Y-data) are presented. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. **LU**=land-use, **JG2**=original soil classes, **simp.soil**= reduced soil classes, **Erod**.=soil erodibility classes

Data/Model	Components	Observations	R2X(cum)	R2Y(cum)	Q2(cum)
LU JG2	1	194	0.112	0.174	0.025
LU JG2 No Outliers	1	192	0.126	0.119	-0.00236
LU simp.soil	**1**	**194**	**0.137**	**0.174**	**0.0648**
LU simp.soil No Outliers	1	192	0.14	0.122	0.0409
LU Erod.	1	194	0.121	0.17	0.0396
LU Erod. No Outliers	1	192	0.167	0.1	0.0127
LU	1	194	0.137	0.174	0.0547
LU No Outliers	2	192	0.377	0.153	0.0704
JG2	1	194	0.149	0.0625	-0.0421
JG2 No Outliers	1	190	0.149	0.0699	0.0376
Erod.	1	194	0.412	0.0163	-0.0176
Erod. No Outliers	1	186	0.401	0.0208	-0.0382

The most influential variables in the component, extracted from the LU and simple soil dataset, were forest on mire (FM) and developed areas, which includes urban, residential, and exploited areas (DA) (fig. 12). Most of the variables are uncertain due to the large range of error bars in the VIP plot (fig. 13, top). In fact, only FM and coarser grained soils (Sa-rich) are significant. The coefficient plot shows high levels of uncertainty in the variable's relationship to the R^2 adjusted values (fig.13, bottom). Sa-rich and FM were the only certain variables, which seemed to have positive correlations to the R2 variable.

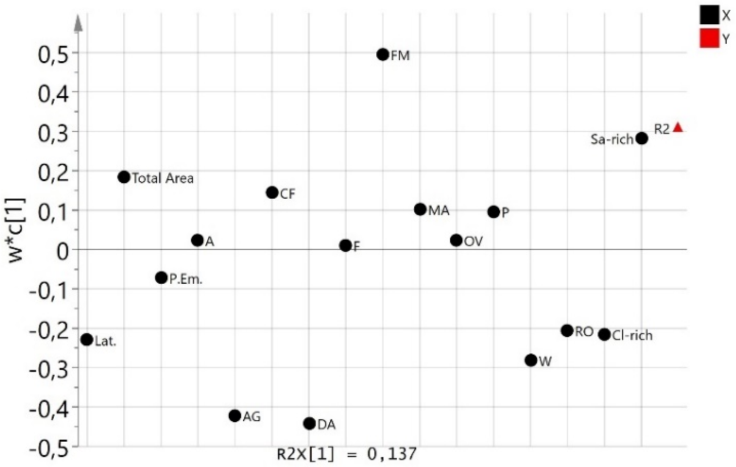


Figure 12. One component PLS model loading plot of the land-use/cover and simplified soil data, only three soil classes (**Sa-rich**= coarser grained soils, **Cl-rich**= clay and silt, **RO**=rock outcrops) Land-use/cover variables; **FM**=forest on mire, **CF**=final fellings, **MA**=miscellaneous agriculture, **P**=pasture, **OV**=open vegetation, **F**=forest, **A**=agriculture, **P.Em**.=point source emissions, **Lat**.=latitudinal position, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands, **R2**= R^2 adjusted values from the MLR.

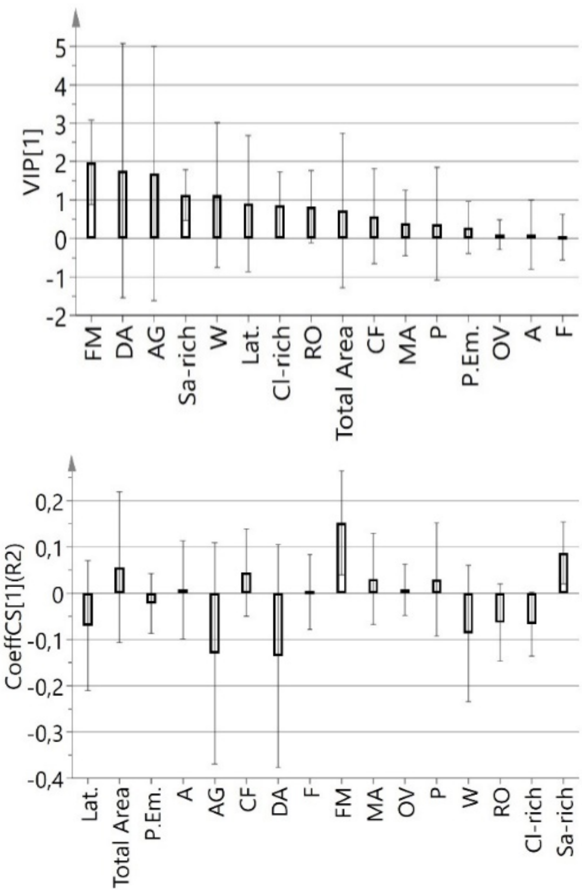


Figure 13. Top) VIP plot of all the landscape variables included in the one component PLS model. Significant variables have VIP>0.5 and error bars not crossing 0.5 limit. **Bottom)** Coefficient plot of the model component. Variables with negative correlation to the R^2 adjusted values have coefficient values <0. Error bars crossing 0 indicate non-significant coefficients. **Sa-rich**= coarser grained soils, **Cl-rich**= clay and silt, **RO**=rock outcrops. Land-use/cover variables; **FM**=forest on mire, **CF**=final fellings, **MA**=miscellaneous agriculture, **P**=pasture, **OV**=open vegetation, **F**=forest, **A**=agriculture, **P.Em**.=point source emissions, **Lat**.=latitudinal position, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands.

The models in table 8 only include variables with mean VIP>0.5. A slight increase in Q^2 was obtained in all models, although, the R^2Y decreased with the removal of non-contributing variables. Forest on mire (FM) was a contributing factor in most models and significant at two, with positive coefficients in the PLS model (table 8).

Table 8. PLS correlation models of the whole catchment dataset. Different combinations of the landscape factors (X data) and their correlation to the MLR R^2 adjusted values (Y-data) are presented. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. Only variables with mean VIP ≥1, with error bar not crossing zero. +/- correlation coefficient, *** non - significant coefficient, crossing zero.**LU**=land-use, **JG2**=original soil classes, **simp.soil**= reduced soil classes, **Erod**.=soil erodibility classes

Data/Model	Components	Observations	R2Y(cum)	Q2(cum)	VIP>1 **
LU JG2	1	194	0.167	0.0772	FM (+)
LU simp.soil	**1**	**194**	**0.17**	**0.0957**	**FM (***)**
LU Erod.	1	194	0.166	0.0875	FM (***)
LU	1	194	0.152	0.0888	FM (+)
JG2	1	194	0.0628	0.0393	W (***)
Erod.	1	194	0.0163	-0.0176	NO (***)

Data subsets

When PLS was performed on separate groups of observations, based on the TP model's parameter configuration, higher R^2Y and Q^2 were generated than for the dataset of all 194 observations (table 9). Here, the data combination of the simplified soil and land-use data was used due to its relative highest R^2Y and Q^2 from the modelling of the whole dataset. The groups with very few observations were not autofitted with CV, instead a component was force fitted to the data (*). Seven of the models were valid and had significant contributing landscape variables. All PLS models presented in table 9 are free from outliers and variables with mean VIP<0.5. Compared to the PLS models of the whole dataset, a larger diversity of prominent component variables was found here. Among the significant PLS models, <0.3 difference between Q^2 and R^2Y, the landscape factors that explained most observations were forest (F), rock outcrops (RO), with negative correlations to the R2. Latitude (Lat.) also showed significant negative correlation to R2 in two models. Coarser grained soils (Sa-rich) was the only landscape factor with positive correlation, also significant at in two models. The forest on mire variable (FM) was less prominent and only significant, with negative coefficient, in one of the models (table 9). No obvious preference pattern was detected between the TP model's parameter combination and the landscape factors.

Table 9. PLS correlation models of data subsets, based on the TP model variable combinations, of the whole catchment data. The X-data is a combination of land-use/cover and simplified soil data. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. No variables with mean VIP<0.5 are included in the model. **= relative significant VIP value, error not crossing zero *=force fitted component. The most significant landscape variables are also presented for each model and their coefficient, (+)=positive, (-)=negative, (***)=not significant. **Tu**=turbidity, **EC**=conductivity, **TE**=water temperature, **TO**=TOC. **FM**=forest on mire, **W**=water and wetlands, **RO**=rock outcrops, **CF**=final fellings, **F**=forest, **Sa-rich**=coarser grained soil, **DA**=developed areas**,** **P**=pasture, **OV**=open vegetation, **Lat**.=latitude. Significant models in italic font

Model	Comp.	Obs.	R2Y(cum)	Q2(cum)	VIP>1 **
Tu	3	18	0.85	0.428	**FM** (+)
Tu EC*	1	11	0.344	0.019	-
Tu EC Te	*1*	*17*	*0.51*	*0.282*	***Lat.** (-), **W** (-)*
Tu EC TO*	1	10	0.385	-0.1	-
Tu EC pH*	1	2	1	1	-
Tu EC TO Te	*1*	*14*	*0.363*	*0.207*	***RO** (-), **CF** (-), **F** (-)*
Tu EC TO pH*	1	3	0.923	0.564	**Sa-rich** (***)
Tu TO	*1*	*23*	*0.382*	*0.235*	***F** (-), **Area** (+)*
Tu TO pH	*2*	*12*	*0.703*	*0.412*	***Sa-rich** (+), **RO** (***), **F** (***)*
Tu TO Te	1	19	0.566	0.222	-
Tu TO pH Te	1	20	0.495	0.171	**Lat.** (-)
Tu pH*	1	2	1	1	-
Tu pH Te	*1*	*19*	*0.412*	*0.152*	***DA** (***), **AG** (***)*
Tu Te	*1*	*10*	*0.659*	*0.487*	***CF** (***), **Sa-rich** (+), **RO** (-)*
Tu EC Te pH	2	9	0.685	0.306	**CF** (-), **P** (+)
Tu EC TO Te pH	*1*	*5*	*0.843*	*0.769*	***OV** (-), **Lat.** (-)*

5.2.2 Buffer Zone 100 m

Whole dataset

Results from the PLS analysis of the buffer zone data, all 175 observations, showed weaker correlation than data of the whole catchment (table 10). The autofitting did not generate any components in many of the landscape variable combinations. In that case, one component was force fitted to the data to give some implications on the structure. Data of the LU and JG2 combination had the best model among the tested datasets.

Table 10. PLS correlation models of the buffer zone 100m dataset. Different combinations of the landscape factors (X data) and their correlation to the MLR R^2 adjusted values (Y-data) are presented. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. **LU**=land-use, **JG2**=original soil classes, **simp.soil**= reduced soil classes, **Erod**.=soil erodibility classes. *=force fitted component

Data/Model	Components	Observations	R2X(cum)	R2Y(cum)	Q2(cum)
LU JG2	**1**	**175**	**0.113**	**0.123**	**0.0107**
LU JG2 No Outliers*	1	172	0.142	0.0982	-0.00317
LU simp.soil	1	175	0.156	0.112	0.0201
LU simp.soil No Outliers	1	172	0.184	0.0901	0.0151
LU & Erod.*	1	175	0.129	0.114	-0.00448
LU & Erod. No Outliers	1	172	0.169	0.0876	0.00172
LU	1	175	0.178	0.112	0.0294
LU No Outliers	1	172	0.217	0.0886	0.0248
JG2 *	1	175	0.185	0.0125	-0.0344
JG2 No Outliers*	1	165	0.168	0.01	-0.0642
Erod.*	1	175	0.316	0.00197	-0.0467
Erod. No Outliers*	1	166	0.309	0.00188	-0.0437

One PLS component model was extracted from the LU and JG2_TX data combination and described 2 variables, FM and W, strongest. Furthest out on the positive end of the component is FM (fig. 14), indicating positive correlation to the R^2 adjusted values. Far off on the negative end is W with negative correlation (fig. 14).

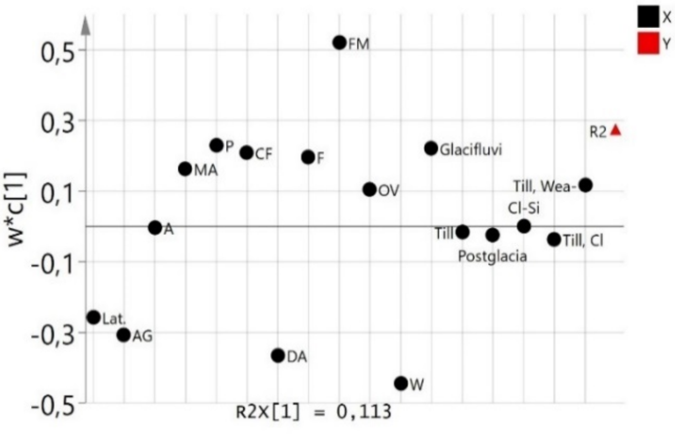
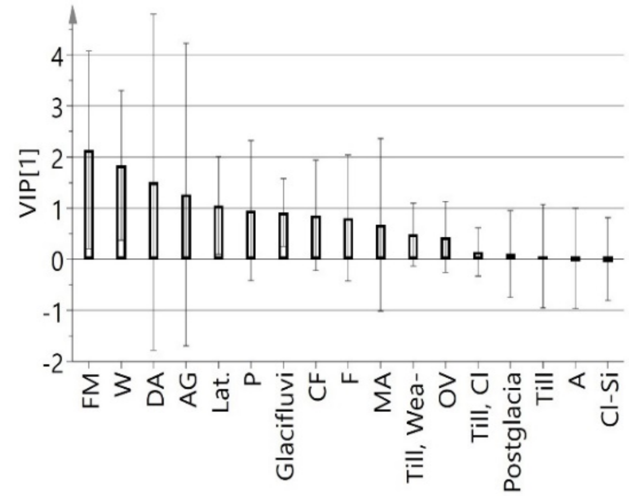


Figure 14. One component PLS model loading plot of the land-use/cover and JG2 soil buffer zone data. **Glacifluvi**=glacifluvial sediment, **Postglacia**=postglacial sediment, **Till, Cl**= clayey till, **Cl-Si**= clay and silt, **Till,We**= till or weathered soils**,** **FM**=forest on mire, **CF**=final fellings, **MA**=miscellaneous agriculture, **P**=pasture, **OV**=open vegetation, **F**=forest, **A**=agriculture, **Lat**.=latitudinal position, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands, **R2**= R^2 adjusted values from the total phosphorus models.

The VIP plot further established the importance of FM and W to the model structure (fig. 15, top), as they are the only significant variables together with glacifluvial sediment. However, the coefficient plot also indicated large uncertainties in the variables (fig. 15, bottom).



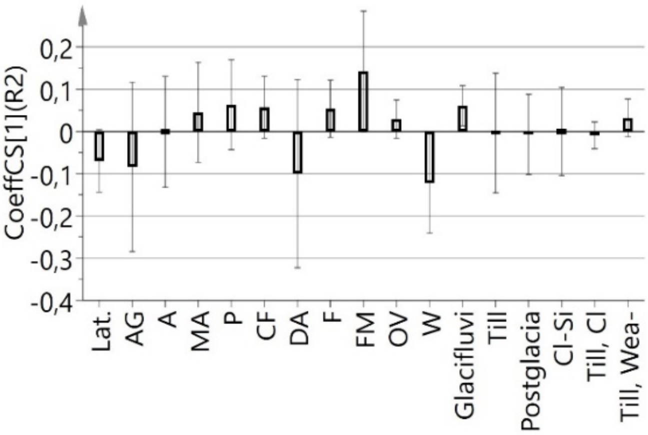
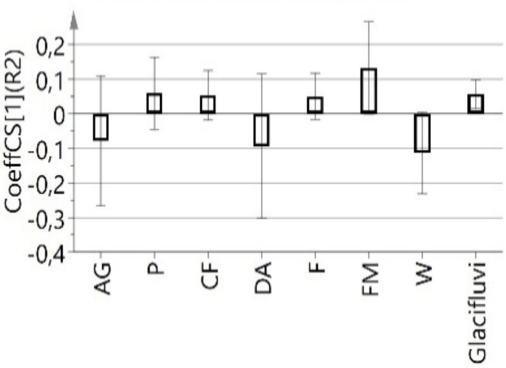


Figure 15. Top) VIP plot of all the landscape variables included in the one component PLS model of the land-use/cover and JG2 soil data. Significant variables have VIP>0.5 and error bars not crossing 0,5 limit. **Bottom)** Coefficient plot of the model component. Variables with negative correlation to the R^2 adjusted values have coefficient values < 0. Error bars crossing 0 indicate non-significant coefficients. **Glacifluvi**=glacifluvial sediment, **Postglacia**=postglacial sediment, **Till, Cl**= clayey till, **Cl-Si**= clay and silt, **Till,We**= till or weathered soils**,** **FM**=forest on mire, **CF**=final fellings, **MA**=miscellaneous agriculture, **P**=pasture, **OV**=open vegetation, **F**=forest, **A**=agriculture, **Lat**.=latitudinal position, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands, **R2**= R^2 adjusted values from total phosphorus models.

The models in table 11 were generated from the same landscape variables combinations as in table 10 but without variables with VIP<0.5. Exclusion of non-contributing variables increased the Q^2 values slightly, whereas the R^2Y decreased. Here, the buffer zone data also generated weaker PLS models than data from the whole catchment. In the LU and simple soil data model, no soil variables showed significant contributions and are all excluded from the model. The data combination with soil erodibility (Erod.) was not fitted with any model due to no significant contributing variables. Variables that often contributed significantly to the models were FM and W. In the two strongest models (LU JG2 and LU) mean coeffcient values indicate positive correlation between R2 and FM and negative correlation between R2 ad W (fig. 16).

Table 11. PLS correlation models of the buffer zone 100m dataset. Different combinations of the landscape factors (X data) and their correlation to the MLR R^2 adjusted values (Y-data) are presented. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. Only variables with mean VIP ≥1, with error bar not crossing zero. +/- correlation coefficient, *** non - significant coefficient, crossing zero, *forced component.**LU**=land-use, **JG2**=original soil classes, **simp.soil**= reduced soil classes, **Erod**.=soil erodibility. No soil variables are significant in the LU Erod. model.

Data/Model	**Components**	**Observations**	**R2Y(cum)**	**Q2(cum)**	**VIP>1 ****
LU JG2	**1**	**175**	**0.101**	**0.0405**	**FM (+) W (***)**
LU simp.soil	1	175	0.0953	0.0346	FM (***) W (***)
LU & Erod.*	1	175	-	-	-
LU	1	175	0.0953	0.0346	FM (***) W (***)
JG2	1	175	0.0128	0.007	Glaci (***)
Erod.*	1	175	0.002	-0.0319	-



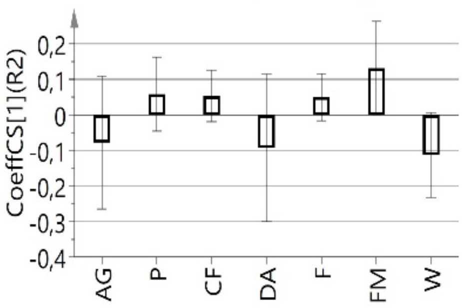


Figure 16. Coefficient plot of the model components of the LU JG2 and LU model in table 11. Variables with negative correlation to the R^2 adjusted values have coefficient values < 0. Error bars crossing 0 indicate non-significant coefficients. **Glacifluvi**=glacifluvial sediment, **P**=pasture, **F**=forest, **AG**=artificial green areas, **DA**=developed areas, **W**=water and wetlands, **CF**=final fellings, **FM**=forest on mire. **Left**) LU & JG2 **Right**) LU and simp.soil. **LU**= land-use, **JG2**=original soil classes**, simp.soil**=simplified soil data

Data subsets

Grouping of the observations, based on the TP regression model parameter combinations, using the buffer zone land-use (LU) and JG2 data generated four significant models (table 12). Both the R^2Y and Q^2 were higher than the buffer zone dataset with all observations included. However, less significant contributing variables were determined due to the large range of VIP errors, compared to the models of the whole catchment data. Among the significant models, developed areas (DA) contributed significantly in two models with negative correlation to R2. The variables with pronounced contributions in the models with all 175 observations (FM and W) were not as prominent here.

Table 12. PLS correlation models of groups, based on the TP model variable combinations, of the buffer zone 100m data. The X-data is a combination of land-use/cover, JG2 soil layer and soil erodibility classes. Model evaluation is done by the R^2Y(cum) and Q^2(cum) values. No variables with mean VIP<0.5 are included in the model. **= relative significant VIP value, error not crossing zero, *=force fitted component. The most significant landscape variables are also presented for each model and their coefficient, (+)=positive, (-)=negative, (***)=not significant. **Tu**=turbidity, **EC**=conductivity, **TE**=water temperature, **TO**=TOC. **FM**=forest on mire, **W**=water and wetlands, **Glaci**=glacifluvial sediment, **DA**=developed areas, **A**=agriculture, **Cl-Si**=clay and silt, **P**=pasture, **Lat**.=latitude. Significant models are in italic font.

Model	Components	Observations	R2Y(cum)	Q2(cum)	VIP>1**
Tu*	1	17	0.545	0.0191	**FM** (***)
Tu EC	*1*	*11*	*0.399*	*0.146*	***Glaci.*** *(+)*
Tu EC Te	1	9	0.655	0.352	**W** (-)
Tu EC TO	*1*	*10*	*0.569*	*0.343*	***DA*** *(-)*
Tu EC pH*	1	2	1	1	-
Tu EC TO Te	1	14	0.28	0.0735	-
Tu EC TO pH*	1	3	0.922	0.783	**Glaci**. (-)
Tu TO*	1	22	0.598	-0.027	-
Tu TO pH*	1	9	0.491	-0.1	-
Tu TO Te*	1	15	0.508	-0.1	-
Tu TO pH Te	1	15	0.538	0.15	-
Tu pH*	1	2	1	1	-
Tu pH Te	*1*	*16*	*0.48*	*0.33*	***DA*** *(***)*
Tu Te	1	9	0.655	0.352	**W** (***), **P** (***)
Tu EC Te pH	2	9	0.855	0.479	**P** (***), **A** (***)
Tu EC TO Te pH	1	5	0.874	0.641	**Cl-Si** (***)

6 Discussion

6.1 Total Phosphorus Regression Models

Turbidity had the strongest correlation to total phosphorus (TP), which confirmed previous studies. However, the strong correlation found between the two variables might be biased due to the selection of stations. Turbidity is included as a parameter for all stations in the Swedish National Monitoring Program of Surface Waters since 2010. For stations in other programs in this study, turbidity could have been added for specific purposes i.e. complementing TP measurements. The 18 stations of the Tu model, had stronger correlations at stations with higher concentrations of TP and turbidity. Only half of these stations is part of the National Monitoring Program. This finding could further indicate a biased selection of the water quality data. However, the specific locations and monitoring program for each station was not investigated due to the limitations of the study.

As mentioned, only 18 stations were selected to have turbidity as the sole explanatory variable as models with more parameters generated stronger correlations at the remaining stations. Among the TP models with a second independent variable, the combination of turbidity and TOC resulted in the highest number of significant models as well as R^2 adjusted values ≥ 0.65. Moreover, the absolute strongest correlations were generated by TP models with three independent variables in which TOC was included. Kronvang (1992) found significant relationship between particulate organic matter and phosphorus in two agricultural catchments with sandy loam in Denmark. One of their objectives was to estimate the fraction of phosphorus in different forms of particulate matter (inorganic and organic) at different flow conditions. Their results show that the phosphorus content in inorganic matter fluctuates with flow conditions whereas the phosphorus content in organic matter was relatively constant. Miguntanna et al. (2010) also found strong correlation between TOC and TP (r^2=0.757) in their search of surrogate parameters for urban storm water monitoring. The high frequency of strong correlation models where TOC was included as a parameter could be explained by the findings of Kronvang (1992) and Miguntanna et al. (2010). A more stable content of phosphorus in the particulate organic matter, regardless flow conditions, will result in more constant correlation patterns between TOC and TP. Further indication on this notion was the positive correlation between TOC and TP at almost all sites.

The model with the second highest number of significant regressions was TuTe, where 63 % of the stations had R^2 adjusted values ≥ 0.65. This makes turbidity and water temperature potential surrogates for TP. The frequent positive correlation between the two variables could reflect a shared positive relationship with stream flow. For example, due to distinct seasonal runoff patterns in Sweden, higher runoff rate with particle transport could be expected during the spring and autumn. In autumn, increased rainfall could increase the delivery of plant residues to the streams, containing organic phosphorus. Water temperatures positive correlation with TP could also reflect stream water flow rate. In summer, lower stream flow could be expected and hence higher concentrations of dissolved phosphorus (DP), especially with point sources in connection to the stream (Jarvie et al., 2006). In addition, higher temperature in reduced light conditions has been argued to increase the fluxes of phosphorus from the sediment active layer (Jiang et al., 2008) and the recent brownification of Swedish surface waters (Mormul et al., 2012) could be a second explanation of the positive correlation between TP and water temperature.

The two remaining variables, electric conductivity (EC) and pH, were included in least significant models. However, EC were included in slightly more models with R^2 adjusted ≥ 0.65 compared to pH. EC and pH were negative correlated to TP at approximately 70% of the sites. As conductivity could serve as an estimator for dissolved compounds in the water, it was expected to explain the DP fraction by being positive correlated to TP. Possible reasons for the negative correlations found between TP and EC is dilution of ions during intense runoff, which could increase TP export to streams. The TupH model had the least significant regressions as well as R^2 adjusted values ≥ 0.65 compared to the other models. This indicates relatively poor surrogate potential of pH in general.

Evaluations of the different MLR model configurations showed that additions of parameter to the model decreased the number of significant regressions. However, models with more parameters had smaller ranges of R^2 adjusted values as well as higher values. The decrease in significant models was both due to insignificant model parameters and multicollinearity. In the PLS analysis, the stations were represented by the model from which the strongest significant regression was obtained. The most frequently chosen models had three parameters and more than half of these had R^2 adjusted values ≥ 0.65. This result indicates the ample use of three parameters to avoid model parameter errors and at the same time obtain strong correlation models. Water temperature and TOC were included as a parameter at the highest number of sites, compared to pH and EC. Moreover, the addition of water temperature and TOC to the TP models also generated the most frequent increase in

R^2 adjusted values. From this, it might be possible argue important impact from landscape factors related to TOC and water temperature on the regression models.

6.2 Explanatory Landscape Data Models

Important Factors - Whole Catchment Models

The most significant PLS model of the whole catchment factors was created from the data combination of land-use/cover and simplified soil data. Only three soil classes were included in the simplified soil data (coarse grained, fine grained and rock outcrops). These were created to generalize the different soil types. Although significant, the model only explained 17% of the variation in the TP models correlation coefficients. To sustain an honest model, all outliers were kept as they were not considered error values but rather a representation of deviating catchments. Furthermore, the exclusion of the outliers did not improve the model fit. The most significant landscape factors in the model were forest on mire and coarser grained soils, both with positive correlation to the R^2 adjusted values. The positive correlation between strong TP models and forest on mires could be explained by the more stable phosphorus content in organic compared to inorganic matter. Moreover, data from monitoring programs of Swedish forest streams indicates highest concentrations of phosphorus in the south, often associated with organic carbon in the streams (Uggla and Westling, 2003).

Dillon and Molot (1997) presented export data of Dissolved Organic Carbon (DOC), TP and iron (Fe) from 20 forested catchments in Ontario, Canada between 1980 and 1992. The data revealed strong positive correlation patterns between percentage peatlands and Fe export, Fe export and TP export, DOC and peatlands and between DOC and Fe. Presence of Fe in natural waters enhances the complexation of DP to DOC, which the authors argued to confirm their findings. A lab scale experiment by Johnson (2010) further established this relationship. The experiment showed that DOC and Fe complexes, which are formed under oxygenated conditions, dissociated under ultra violet radiation due to photolysis of the DOC. The dissociation generated positively charged iron particles, which efficiently adsorbed the DP in the water. These findings could support the possible positive influence of forest on mire on strong TP models.

In literature, the relationship between TP and clay/silt soils is frequently argued for, which contradicts the findings here. The variables coupled to clay/silt soils were insignificant. Instead, the PLS model showed that coarser grained soils are significant related to stronger TP prediction models. The coarser grained soil category of the PLS model contain weathered soil, till, glacifluvial and postglacial sediments, which vary from silt to boulder in texture and possibly also structure. These variations might be both temporal i.e. changes in agricultural practices, which improve soil structure or spatial due to different processes of deposition. According to previous discussion, the TP measurements included in this study might be high in organic phosphorus rather than inorganic. This makes the use of the JG2 soil classes irrelevant if the soil structure is more important than the general soil type.

The remaining hypothesized important landscape factors for strong TP regression models (agriculture, pastures, final felling, urban areas, and point sources), were not significant in the PLS model. The TP exported from these areas might be very different in composition and concentrations. From the PCA model, it became evident that these often coexist at equally high fractions in the same catchments. Temporal variations in the different forms of phosphorus in TP could lead to a large variation in the correlation patterns with selected surrogates. However, there was a strong correlation between forest on mire and final felling in PCA model, indicating that these two factors were often found at equal fractions in the catchments. Increase in organic phosphorus transport due to final felling has been documented in Sweden (Löfgren, 2007). Temporary ditches after final felling is a common practice in Sweden, which drains surplus water to benefit productivity. This activity will further increase the export of organic phosphorus (Swedish Forest Agency, 2017). The close relationship between the forest on mire and final felling variables could imply a larger influence from the final felling variable than shown by the PLS model.

Important Factors - Buffer Zone Models

The most significant buffer zone PLS model explained only 10.1% of the R^2 adjusted values from the TP models. Data combination of land-use/cover and JG2 was used in the model. Important component factors were forest on mire, water, and open wetlands. Forest on mire had positive correlations to high correlation coefficients of the TP regression models, whereas water and wetlands showed negative correlation.

Presence of open wetlands along the stream buffer zone will limit the transport of PP to the main stream, as it will help to increase retention by reducing the flow velocity. In a modelling study by Yang et al. (2016), different scenarios of riparian wetlands restoration were tested (0 to 100% restoration) to find the effect on water quality. The authors found decreased TP and sediment export from the watershed with an increase in restoration fraction, which provides some validity to the findings in the present study.

Forest on mire was indicated to have positive correlation with strong TP models. The explanations given for the strong influence in the whole catchment PLS model could be applied even here. In addition, forest on mire and final felling were strongly correlated in terms coverage and spatial distribution, indicate by the buffer zone PCA model. Along the 100 m buffer zones, temporary ditching practices of final felling on mire could have a large impact on the export of organic phosphorus. The differences between forest on mire and open forest could be explained by the properties of wetland and the surrounding landcovers. The open wetland category in the SMD might consist of highly saturated mires or wetlands compared to the forest on mire, which could be forest with patches of wetlands (SEPA, 2014).

The suggested important landscape factors, agriculture, clayey/silty soils, pasture and developed areas did not make any significant contributions to the buffer zone PLS model. Strong correlations were found between these variables in the buffer zone PCA model. The presence of many different sources within the same catchment will cause irregular correlation patterns between TP and the chosen surrogate parameters.

Important Factors - Subset Models

Subsets of whole catchment data generated the most significant models with good fit. The grouping of stations in subsets was done according to the TP model assigned to each station from the MLR analysis. Variations in the PP fractions in TP could be accounted for by adding variables to complement turbidity in the TP models. Relationships between the sensor parameter included in the TP model and landscape factors, was expected to be found, such as EC and point sources.

Seven significant PLS models were computed by CV. Compared to the whole dataset models, a larger variation of significant factors was found among the subset models. The better model fits indicate that the data subsets are more homogenous in land-cover properties, whereas the variation in significant contributing factor could imply catchment specific properties of the phosphorus exported. Latitude was a significant contributing factor in two models, with positive relationship to the correlation coefficients from the TP models. Water temperature and EC were both included in the TP model as parameters. The higher temperatures in southern Sweden will result in higher water temperatures as well as higher rate of. This could create periods of low flow, which will magnify the emissions from the larger point sources. Moreover, the results from the PCA model indicated that most of the larger point sources in the southern regions. From this, it is possible to suggest that water temperature and EC are suitable as surrogates for TP in the southern regions of

Sweden. The coarser grained soil factor was a significant contributing factor with positive correlation to the TP model's strength in two of the PLS models. The TP models had relatively different parameter configurations, which prevents the conclusion of any significant relationship with the significant landscape factors. The forest factor was also more pronounced here, than in the whole dataset model, and show consistent negative correlation to the strength of the TP models. Export of TP from Swedish forests is relatively constant at background values if no drastic management practices, such as ditching is applied (Uggla and Westling, 2003). Moreover, phosphorus is often stored in the humus layer of the soil profile and is easily taken up by the vegetation in a thriving forest and by that limit the export to surface water (Giesler et al., 2002).

Less significant PLS models were created from the buffer zone data subsets. Among the four significant PLS models, developed areas was significant contributing in two models, correlated to weaker TP models. The reason for the poor TP models from the developed areas, included in this study, could be due to irregular emission rate from these catchments. Moreover, developed areas often co-exist with other types of sources, which will further add to the irregularities of phosphorus transported to the streams.

One interesting finding was the PLS model of the data subsets of the TuEC model, with 11 observations. Here glacifluvial sediment was a significant factor with positive correlation to strong TP models. This could suggest a suitability of using EC as a surrogate for TP in catchments with high fractions of glacifluvial soils. The pronounced positive correlation between coarser grained soil classes, in the simplified soil data, might also be explained. However, the PLS model only explained 40% of the data variation. This finding is in line with previous research, indicating higher leaching rates of DP from coarser grained soils.

6.3 Data Uncertainties and Future Studies

Data from the whole catchment explained more of the variation in the R^2 adjusted values than the buffer zone data, both when using the whole dataset and the TP model subsets. The whole catchment data was slightly less skewed, which might have influenced the outcome, despite the argued robustness of PLS. The high frequency of extreme values in the landscape datasets made the extraction of significant components difficult. The importance of buffer zone vs whole catchment properties differs in literature. Some report greater importance (Johnson et al., 1997, Sliva and Williams, 2001). One factor that has been reported to play a crucial role in the correlation between water quality and landscape factors is slope (Nava-López et al., 2016, Nash and Chaloud, 2011), which was not considered in this study. Soil erodibility data adapted from the JG2 at low resolution did not show significant contributions in any of the PLS models. Moreover, the data of the buffer zone 100m had very low resolution, especially for the soil classes, making it non-representative of reality. Suggestions for further studies are to include soil data of higher resolution, especially if impact from buffer zones is to be determined. A soil map, which covers the different agricultural soils in Sweden at high resolution is available. This data could be suitable to use for the buffer zones if the possibility to use surrogate measurements is assessed for agriculture land-use. Here, the buffer zone was delineated to 100 m along the streams. In future studies different width of the buffer zone can be included as it could give different results in terms of explaining the water quality (Nava-López et al., 2016).

The low explanatory capacity of the PLS models created from the whole dataset made the interpretation of important landscape factors uncertain. However, some landscape factors re-occurred as important model components. This indicates some general spatial preference of strong correlations between TP and turbidity and other sensor parameters. The low model fit could be due to the use of secondary spatial data although better models from the use of similar data have been reported. Nash and Chaloud (2011) obtained stronger correlation in their analysis of GIS landscape data and water biota ($R^2Y>0.34$), and they suggested suitable use of this kind of data and method in future studies. One of the reasons why better results were obtained in the study by Nash and Chaloud (2011) might be the inclusion slope relief, which they argued to be the most important factor. Moreover, the use of data subsets according to eco regions generated better model fit than data of the whole basin in their study, which was in line with the findings here.

The subsets in this study were defined from the TP models to deduce relationships between water quality parameters and landscape factors but no such trend was detected. In future studies, different standard of data division could be applied. I.e. subsets based on the land-use/cover could generate more homogenous data, which could enhance the extraction of significant components and models of better fit. In this study, the main attempt was to find general landscape characteristics explaining the correlation between sensor variables and TP. By including different types of catchments, interesting factors could be revealed, i.e. the positive influence of forest on mire and the possible importance of organic phosphorus for strong TP models.

Among the different landscape dataset combinations, the simple soil and land-use data combination generated the strongest PLS model (whole catchment data). The buffer zone data was best modelled using the JG2 and land-use data combination. In the buffer zone dataset, less JG2 soil classes were included. Discussed in previous sections, the low resolution of the soil data used might not capture the true distribution of the soil in the buffer zone as only a few classes from JG2 were present. Regardless, soil types had little influence in general compared to the land-use data in most PLS models.

The soil dataset, JG2, represent the ground-layer and mapped at 0.5m depth. Surface runoff processes, which is essential in PP export, could be better represented by more surficial soil distribution. However, the sparse coverage of the surface soil layer (JY1) motivated the use of the JG2 data. Although not reflecting surface runoff processes, the soil data used could have captured possible subsoil properties. Another source of data uncertainty was the time discrepancy between SMD land-use data and the water quality data. The reference year of the land-use classes defined in the SMD layer was 2000 and the water quality data extracted was for 2010 to 2016. According to SCB (2017), large changes occur during these years in terms of land-use in Sweden, i.e. agricultural fields were often converted to developed areas. Another, obvious change in land-use classes is final felling to forest coverage. Depending on the forestry practices, the time from final felling to young forest is approximately 10 years (Egnell, 2011). Unfortunately, this temporal discrepancy between the water quality data and the land-use data creates large uncertainties, as the extracted important component variables might be reflecting other type of current land-uses.

6.4 Applicability of Methods

Applying analysis on large secondary datasets compiled from different sources is an efficient and cheap method. In Sweden, both water quality and GIS data are abundant and easily accessed. Results from this study indicate that secondary data might be suitable to find general trends i.e. to synthesize hypotheses. Based on the discussion of the low data resolution and possible changes in surface soil and land-use properties, application of secondary data could result in large uncertainties. If the purpose is to address defined questions, data generated for the sake of the objectives would be more suitable.

Multiple linear regression as a method to develop models for water quality correlation is a useful method as it allows the inclusion of several independent variables. Furthermore, the regression equations created might be used as prediction models. A larger number of variables increases the explanatory capacity of the model, which could give precise prediction if conditions are stable. The main problem here was the multicollinearity and insignificant variable regression when multiple variables were included. In other studies (Eklöf et al., 2012, Dahlén et al., 2000), water quality correlation has been developed with PLS, which could generate robust models with less time invested. PLS does not assume data normality nor independent variance, which allow inclusion of many predictors. Moreover, the SIMCA-P software, in which PLS was performed, also provide simple prediction functions that allow instant validation of models.

The data overview with PCA revealed important structures in the landscape data. As PCA finds latent structures in the X-data (landscape factors in this study) correlation patterns are easily detected. In PLS, correlations between the explaining variables are less certain as the model also tries to explain the Y-data (TP regression model strength in this study). From the PCA it was possible to find underlying reasons for the contributions from some of the significant landscape factors in the PLS model. I.e. the importance of latitude in many of the PLS models could result from strong negative correlation with the anthropogenic factors as the largest distribution are found in the south. The PCA model also established strong correlations between the anthropogenic factors and enabled the discussion of their poor representation in the PLS models. The interesting correlation between forest on mire and final felling was also found with PCA, which revealed possible importance of final felling on mires.

The application of PLS on the landscape data to deduce the water quality correlation involved large uncertainties. The catchment data used was severely skewed, one possible reason for the poor model outcomes. Despite, several significant components were extracted, which helped to address some of the questions in this study. PLS is often argued to be a robust method with no assumptions on data normality (Cassel et al., 1999) and allows the use of few and noisy observations. Improvements in the use of the method could be made by better defined datasets i.e. inclusion of relevant variables or exclusion of irrelevant ones. However, PLS allowed the exclusion of unimportant variables on the go by the evaluation of the VIP values. More specific objectives cold also be defined to improve the methods applicability, i.e. by looking at more type-specific catchments with more defined land-use systems.

7 Conclusions

The main aim of this study has been to identify landscape factors that influence the applicability of surrogate parameters for TP. TP regression models were created for 194 monitoring stations, using available data of five different sensor parameters. The regression coefficients, of the TP models, were then analysed together with landscape data of the belonging catchments and buffer zone areas. The findings of the study suggest that

- There is a good potential of using available sensor parameters as surrogate for TP in most Swedish streams.

- Turbidity was the best surrogate parameter for TP in most cases. The addition of further sensor parameters helped to increase the correlation coefficients at many sites.

- A smaller fraction of the stations obtained weak TP correlation models, possibly due to differences in catchment and buffer zone properties.

- Forest on mire was the most important landscape factor for strong TP correlation models, which could explain the frequent inclusion of TOC in the significant TP models.

- Large fractions of water and wetlands in the catchment or along the buffer zone could have negative influence of the TP model's prediction capacity.

- The choice of sensor parameter must be evaluated prior application at the specific sites. If the catchment has large fractions of forest on mire, turbidity, and TOC could be considered among the sensor parameters. Except this, no preference pattern of landscape factor and TP model parameters was found.

- Data from the whole catchment explained more of the variation in the TP model's strength than the 100m buffer zone data.

- Complex landscape data (many extreme values) hindered the extraction of significant components with PLS. Weak models were generated and hence large model uncertainties.

- Future studies should include other landscape factors related to hydraulic connectivity, such as slope relief and distribution of ditches in the catchment.

- To evaluate the importance of buffer zone, soil- and land-use data of higher resolution should be considered.

References

AMIRI, B. J. & NAKANE, K. 2009. Modeling the linkage between river water quality and landscape metrics in the Chugoku district of Japan. *Water resources management,* 23, 931-956.

ANKCORN, P. D. Clarifying turbidity—The potential and limitations of turbidity as a surrogate for water-quality monitoring. 2003. Georgia Institute of Technology.

BARRON, J. J. & ASHTON, C. 2005. The effect of temperature on conductivity measurement. *TSP,* 7.

BENNETT, E. M., CARPENTER, S. R. & CARACO, N. F. 2001. Human Impact on Erodable Phosphorus and Eutrophication: A Global PerspectiveIncreasing accumulation of phosphorus in soil threatens rivers, lakes, and coastal oceans with eutrophication. *BioScience,* 51, 227-234.

BORIN, M., PASSONI, M., THIENE, M. & TEMPESTA, T. 2010. Multiple functions of buffer strips in farming areas. *European Journal of Agronomy,* 32, 103-111.

BUCK, O., NIYOGI, D. K. & TOWNSEND, C. R. 2004. Scale-dependence of land use effects on water quality of streams in agricultural catchments. *Environmental Pollution,* 130, 287-299.

CARPENTER, S. R., CARACO, N. F., CORRELL, D. L., HOWARTH, R. W., SHARPLEY, A. N. & SMITH, V. H. 1998. Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological applications,* 8, 559-568.

CASSEL, C., HACKL, P. & WESTLUND, A. H. 1999. Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics,* 26, 435-446.

CASSIDY, R., DOODY, D. G. & WATSON, C. J. 2017. Impact of legacy soil phosphorus on losses in drainage and overland flow from grazed grassland soils. *Science of The Total Environment,* 575, 474-484.

CHAPMAN, D. V. 1996. Water quality assessments: a guide to the use of biota, sediments, and water in environmental monitoring.

CHRISTENSEN, V. G., JIAN, X. & ZIEGLER, A. C. 2000. *Regression analysis and real-time water-quality monitoring to estimate constituent concentrations, loads, and yields in the Little Arkansas River, south-central Kansas, 1995-99*, US Department of the Interior, US Geological Survey.

COWEN, W. F. & LEE, G. F. 1976. Phosphorus availability in particulate materials transported by urban runoff. *Journal (Water Pollution Control Federation)*, 580-591.

DAHLÉN, J., KARLSSON, S., BÄCKSTRÖM, M., HAGBERG, J. & PETTERSSON, H. 2000. Determination of nitrate and other water quality parameters in groundwater from UV/Vis spectra employing partial least squares regression. *Chemosphere,* 40, 71-77.

DAVIES- COLLEY, R. & SMITH, D. 2001. Turbidity Suspeni) ed Sediment, and Water Clarity: a Review. *JAWRA Journal of the American Water Resources Association,* 37, 1085-1101.

DILLON, P. & MOLOT, L. 1997. Effect of landscape form on export of dissolved organic carbon, iron, and phosphorus from forested stream catchments. *Water Resources Research,* 33, 2591-2600.

DJODJIC, F., BÖRLING, K. & BERGSTRÖM, L. 2004. Phosphorus leaching in relation to soil type and soil phosphorus content. *Journal of Environmental Quality,* 33, 678-684.

EDWARDS, A. & WITHERS, P. 2007. Linking phosphorus sources to impacts in different types of water body. *Soil Use and Management,* 23, 133-143.

EGNELL, G. 2011. Is the productivity decline in Norway spruce following whole-tree harvesting in the final felling in boreal Sweden permanent or temporary? *Forest Ecology and Management,* 261, 148-153.

EKLÖF, K., FÖLSTER, J., SONESTEN, L. & BISHOP, K. 2012. Spatial and temporal variation of THg concentrations in run-off water from 19 boreal catchments, 2000–2010. *Environmental pollution,* 164, 102-109.

ERIKSSON, J. 2011. *Marklära,* Lund, Lund : Studentlitteratur.

ERIKSSON, L., ANDERSSON, P. L., JOHANSSON, E. & TYSKLIND, M. 2006. Megavariate analysis of environmental QSAR data. Part I–A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular diversity,* 10, 169-186.

ESSINGTON, M. E. 2015. *Soil and water chemistry: an integrative approach*, CRC press.
FÖLSTER, J., JOHNSON, R. K., FUTTER, M. N. & WILANDER, A. 2014. The Swedish monitoring of surface waters: 50 years of adaptive monitoring. *Ambio,* 43**,** 3-18.
FÖLSTER, J. & RÖNNBACK, P. 2015. Turbiditet som mått på suspenderat material
och totalfosfor. Swedish University of Agricultural Sciences: Department of Aquatic Sciences and Assessment.
GIESLER, R., PETERSSON, T. & HÖGBERG, P. 2002. Phosphorus limitation in boreal forests: effects of aluminum and iron accumulation in the humus layer. *Ecosystems,* 5**,** 300-314.
GIPPEL, C. J. 1995. Potential of turbidity monitoring for measuring the transport of suspended solids in streams. *Hydrological processes,* 9**,** 83-97.
GRAHAM, M. H. 2003. CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION. *Ecology,* 84**,** 2809-2815.
GRANDIN, U. 2012. Dataanalys och hypotesprövning för statistikanvändare. SEPA
HAIDARY, A., AMIRI, B. J., ADAMOWSKI, J., FOHRER, N. & NAKANE, K. 2013. Assessing the impacts of four land use types on the water quality of wetlands in Japan. *Water Resources Management,* 27**,** 2217-2229.
HARMEL, R. & KING, K. 2005. Uncertainty in measured sediment and nutrient flux in runoff from small agricultural watersheds. *Transactions of the ASAE,* 48**,** 1713-1721.
HAYGARTH, P. M. & SHARPLEY, A. 2000. Terminology for phosphorus transfer. *Journal of environmental quality,* 29**,** 10-15.
HICKEY, M. B. C. & DORAN, B. 2004. A review of the efficiency of buffer strips for the maintenance and enhancement of riparian ecosystems. *Water Quality Research Journal of Canada,* 39**,** 311-317.
HORSBURGH, J. S., SPACKMAN JONES, A., STEVENS, D. K., TARBOTON, D. G. & MESNER, N. O. 2010. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. *Environmental Modelling & Software,* 25**,** 1031-1044.
HOUSE, W. A. 1999. The Physico-Chemical Conditions for the Precipitation of Phosphate with Calcium. *Environmental Technology,* 20**,** 727-733.
JARVIE, H. P., NEAL, C. & WITHERS, P. J. A. 2006. Sewage-effluent phosphorus: A greater risk to river eutrophication than agricultural phosphorus? *Science of The Total Environment,* 360**,** 246-253.
JIANG, X., JIN, X., YAO, Y., LI, L. & WU, F. 2008. Effects of biological activity, light, temperature and oxygen on phosphorus release processes at the sediment and water interface of Taihu Lake, China. *Water Research,* 42**,** 2251-2259.
JOHNSON, D. J. 2010. *Influences of Dissolved Organic Carbon, Iron, and Aluminum Photochemistry on Phosphorus Cycling in Fresh Surface Waters.* The University of Maine.
JOHNSON, L., RICHARDS, C., HOST, G. & ARTHUR, J. 1997. Landscape influences on water chemistry in Midwestern stream ecosystems. *Freshwater Biology,* 37**,** 193-208.
JOLLIFFE, I. 2002. *Principal component analysis*, Wiley Online Library.
JONES, A. S., STEVENS, D. K., HORSBURGH, J. S. & MESNER, N. O. 2011. Surrogate measures for providing high frequency estimates of total suspended solids and total phosphorus concentrations1. Wiley Online Library.
JORDAN, P. & CASSIDY, R. 2011. Technical Note: Assessing a 24/7 solution for monitoring water quality loads in small river catchments. *Hydrol. Earth Syst. Sci.,* 15**,** 3093-3100.
KOSKI-VÄHÄLÄ, J. & HARTIKAINEN, H. 2001. Assessment of the Risk of Phosphorus Loading Due to Resuspended Sediment. *Journal of Environmental Quality,* 30**,** 960-966.
KYLLMAR, K., CARLSSON, C., GUSTAFSON, A., ULÉN, B. & JOHNSSON, H. 2006. Nutrient discharge from small agricultural catchments in Sweden: Characterisation and trends. *Agriculture, Ecosystems & Environment,* 115**,** 15-26.
KYNKÄÄNNIEMI, P. 2014. *Small wetlands designed for phosphorus retention in Swedish agricultural areas*.
LEWIS, J. 1996. Turbidity- controlled suspended sediment sampling for runoff- event load estimation. *Water resources research,* 32**,** 2299-2310.
LINE, D. E., WHITE, N. M., OSMOND, D. L., JENNINGS, G. D. & MOJONNIER, C. B. 2002. Pollutant Export from Various Land Uses in the Upper Neuse River Basin. *Water Environment Research,* 74**,** 100-108.

LÖFGREN, S. 2007. *How to estimate N and P losses from forestry in northern Sweden: report from a workshop at the Royal Swedish Academy of Agriculture and Forestry 14 March 2006*, Kungl. Skogs-och Lantbruksakademien.
MCDOWELL, R. 2006. Contaminant losses in overland flow from cattle, deer and sheep dung. *Water, Air, and Soil Pollution,* 174**,** 211-222.
MCDOWELL, R., NASH, D. & ROBERTSON, F. 2007. Sources of phosphorus lost from a grazed pasture receiving simulated rainfall. *Journal of Environmental Quality,* 36**,** 1281-1288.
MIGUNTANNA, N. S., EGODAWATTA, P., KOKOT, S. & GOONETILLEKE, A. 2010. Determination of a set of surrogate parameters to assess urban stormwater quality. *Science of The Total Environment,* 408**,** 6251-6259.
MILJÖSTATISTIK.SE, 2017 available at: http://www.miljostatistik.se/PLS.html
MORMUL, R. P., AHLGREN, J., EKVALL, M. K., HANSSON, L.-A. & BRÖNMARK, C. 2012. Water brownification may increase the invasibility of a submerged non-native macrophyte. *Biological Invasions,* 14**,** 2091-2099.
NASH, M. S. & CHALOUD, D. J. 2011. Partial least square analyses of landscape and surface water biota associations in the Savannah River Basin. *ISRN Ecology,* 2011.
NAVA-LÓPEZ, M. Z., DIEMONT, S. A., HALL, M. & ÁVILA-AKERBERG, V. 2016. Riparian Buffer Zone and Whole Watershed Influences on River Water Quality: Implications for Ecosystem Services near Megacities. *Environmental Processes,* 3**,** 277-305.
NOVAK, J., STONE, K., WATTS, D. & JOHNSON, M. 2003. Dissolved phosphorus transport during storm and base flow conditions from an agriculturally intensive southeastern coastal plain watershed. *Transactions of the ASAE,* 46**,** 1355-1363.
PENN STATE EBERLY COLLEGE OF SCIENCE, 2017. STAT 501. REGRESSION METHODS. available at: https://onlinecourses.science.psu.edu/stat501/node/316
PERKINS, R. L. 2013. *Turbidity Monitoring on Minnesota Construction Sites: Insight into the Factors Influencing the Turbidity and TSS Relationship.* UNIVERSITY OF MINNESOTA.
REDDY, K. R., KADLEC, R. H., FLAIG, E. & GALE, P. M. 1999. Phosphorus Retention in Streams and Wetlands: A Review. *Critical Reviews in Environmental Science and Technology,* 29**,** 83-146.
REKOLAINEN, S., POSCH, M., KÄMÄRI, J. & EKHOLM, P. 1991. Evaluation of the accuracy and precision of annual phosphorus load estimates from two agricultural basins in Finland. *Journal of Hydrology,* 128**,** 237-255.
RUTTENBERG, K. 2003. The global phosphorus cycle. *Treatise on geochemistry,* 8**,** 682.
SADAR, M. n.d. Turbidity measurement: A simple, effective indicator of water quality change. *HACH Application Notes About the Autbor (s)*.
SAHLSTEN, E., STENSDOTTER BLOMBERG, U., EJDUNG, G. & HÅKANSSON, B. 2014. Nationell akvatisk miljöövervakning 2015 – 2020. *In:* SAHLSTEN, E. (ed.). Gothenburg: The Swedish Agency of Marine and Water Management.
SCHILLING, K. E., KIM, S.-W. & JONES, C. S. 2017. Use of water quality surrogates to estimate total phosphorus concentrations in Iowa rivers. *Journal of Hydrology: Regional Studies,* 12**,** 111-121.
SLIVA, L. & WILLIAMS, D. D. 2001. Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water research,* 35**,** 3462-3472.
SONDERGAARD, M., JENSEN, P. J. & JEPPESEN, E. 2001. Retention and internal loading of phosphorus in shallow, eutrophic lakes. *The Scientific World Journal,* 1**,** 427-442.
SPEARS, B. M., CARVALHO, L., PERKINS, R. & PATERSON, D. M. 2008. Effects of light on sediment nutrient flux and water column nutrient stoichiometry in a shallow lake. *Water Research,* 42**,** 977-986.
SPIVAKOV, B. Y., MARYUTINA, T. & MUNTAU, H. 1999. Phosphorus speciation in water and sediments. *Pure and Applied Chemistry,* 71**,** 2161-2176.
STINE, R. A. 1995. Graphical interpretation of variance inflation factors. *The American Statistician,* 49**,** 53-56.
SVENDSEN, L. M. & KRONVANG, B. 1993. Retention of nitrogen and phosphorus in a Danish lowland river system: implications for the export from the watershed. *Hydrobiologia,* 251**,** 123-135.
SWEDSIH ENVIRONMENT PROTECTION AGENCY, 2014. Svenska Marktäckedata 1.2th ed. Stockholm Sweden.

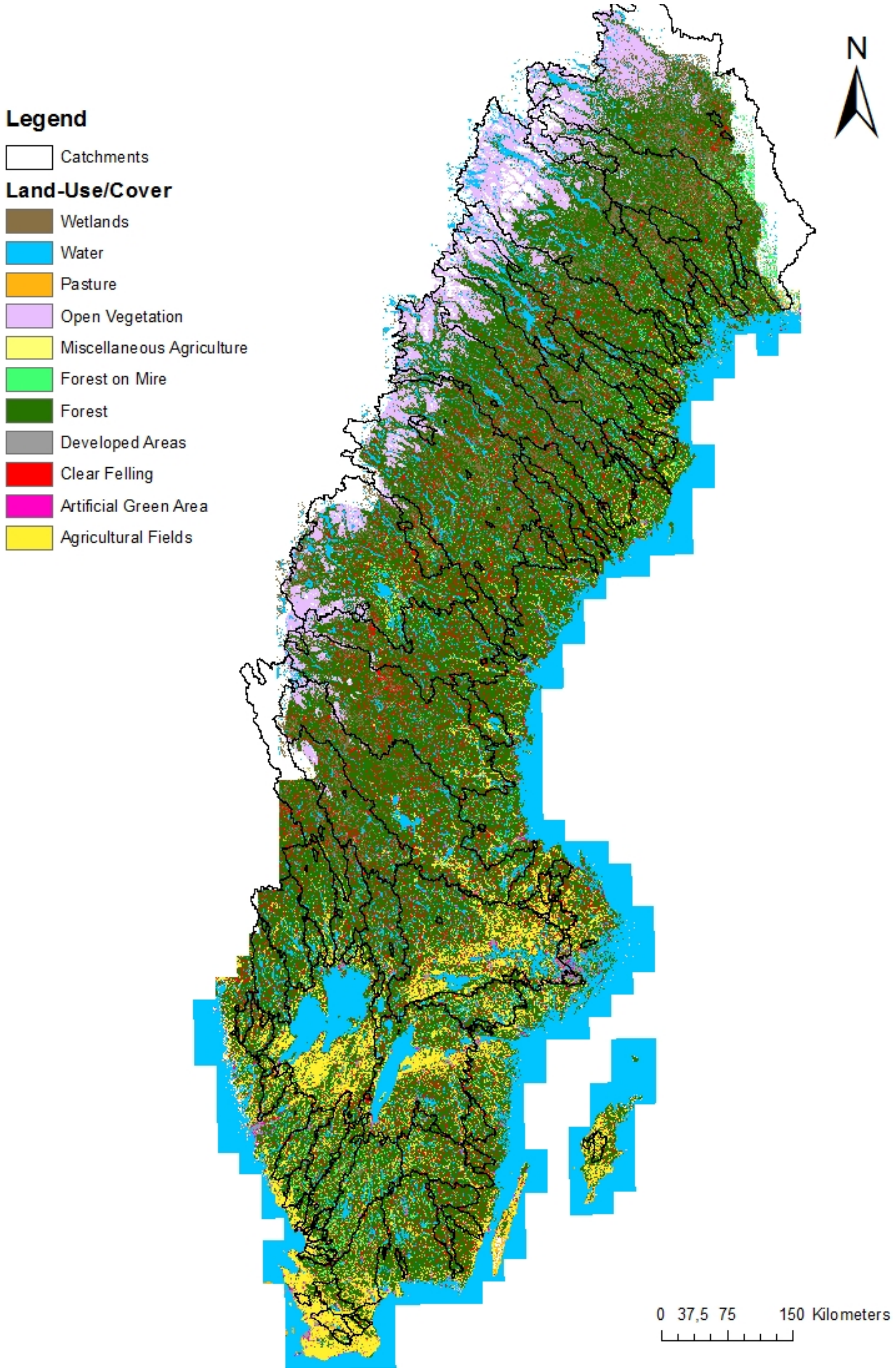
SWEDISH FOREST AGENCY, 2017. available at: https://www.skogsstyrelsen.se/bruka-skog/dikning/dikning/
SWEDISH GEOLOGICAL SURVEY, 2016. Production description JORDARTER 1:25 000-1:100 000. Version 1.1.
SWEDISH METEOROLOGICAL AND HYDROLOGICAL INSTITUTE, 2017. available at: https://www.smhi.se/klimatdata/meteorologi/kartor/monYrTable.php?par=nbdAvv
SWEDISH METEOROLOGICAL INSTITUTE, 2004. GIS och GIS-databaser 2004. Fact Sheet no. 19.
SWEDISH STATISTICS, 2013. Markanvändningen i Sverige, 6th ed. Örebro, Sweden
SØNDERGAARD, M., JENSEN, J. P. & JEPPESEN, E. 2003. Role of sediment and internal loading of phosphorus in shallow lakes. *Hydrobiologia,* 506**,** 135-145.
TAMMEORG, O., HORPPILA, J., LAUGASTE, R., HALDNA, M. & NIEMISTÖ, J. 2015. Importance of diffusion and resuspension for phosphorus cycling during the growing season in large, shallow Lake Peipsi. *Hydrobiologia,* 760**,** 133-144.
TURNER, B. L. & HAYGARTH, P. M. 2000. Phosphorus forms and concentrations in leachate under four grassland soil types. *Soil Science Society of America Journal,* 64**,** 1090-1099.
UGGLA, E. & WESTLING, O. 2003. Utlakning av fosfor från brukad skogsmark. *IVL Report B 1549.* Stockholm: IVL Swedish Environmental Research Institute Ltd.
ULÉN, B., BECHMANN, M., FÖLSTER, J., JARVIE, H. & TUNNEY, H. 2007. Agriculture as a phosphorus source for eutrophication in the north- west European countries, Norway, Sweden, United Kingdom and Ireland: a review. *Soil Use and Management,* 23**,** 5-15.
ULÉN, B. & JAKOBSSON, C. 2005. Critical evaluation of measures to mitigate phosphorus losses from agricultural land to surface waters in Sweden. *Science of The Total Environment,* 344**,** 37-50.
ULÉN, B., JOHANSSON, G. & KYLLMAR, K. 2001. Model predictions and long-term trends in phosphorus transport from arable lands in Sweden. *Agricultural Water Management,* 49**,** 197-210.
ULÉN, B. & PERSSON, K. 1999. Field- scale phosphorus losses from a drained clay soil in Sweden. *Hydrological Processes,* 13**,** 2801-2812.
ULÉN, B., VON BRÖMSSEN, C., KYLLMAR, K., DJODJIC, F., STJERNMAN FORSBERG, L. & ANDERSSON, S. 2012. Long-term temporal dynamics and trends of particle-bound phosphorus and nitrate in agricultural stream waters. *Acta Agriculturae Scandinavica, Section B–Soil & Plant Science,* 62**,** 217-228.
US Geological Survey, 2017. Avaialble at: https://or.water.usgs.gov/grapher/fnu.html
VILLA SOLÍS, A. 2014. *Risk assessment of erosion and losses of particulate phosphorus.*
WITHERS, P. J. A. & JARVIE, H. P. 2008. Delivery and cycling of phosphorus in rivers: A review. *Science of The Total Environment,* 400**,** 379-395.
WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems,* 58**,** 109-130.
YU, H., WU, Z., BAO, T. & ZHANG, L. 2010. Multivariate analysis in dam monitoring data with PCA. *Science China Technological Sciences,* 53**,** 1088-1097.

Acknowledgements

To start, I’d like to thank my supervisor Jens Fölster, which introduced me to this interesting topic. The supervising has been impeccable as several drafts of the report were sent in and returned with valuable comments, every time. Much help was also given in the process of hypothesis synthesising as well as finding aspects for discussion of the results. I’d also like to thank Emil Back, which was a huge support during the extraction of GIS data. Not only did Emil help me to get the data but he also gave valuable advice in working with ArcGIS. Further thanks are directed to Hampus Markensten, which provided me with a much- needed script for the GIS data, making my life easier. I’d also like to mention Maria Tranvik and her supervisor Karin Eklöf, which kindly allowed me to join their meeting where they discussed statistical methods. Also thanks to Emma Lannergård for her helpful notes on some of the statistical methods used in the study. At last, I’d like to send my love and thanks to my family and friends for your constant support. Not only during this master project but throughout all, sometimes very challenging, years at the university.

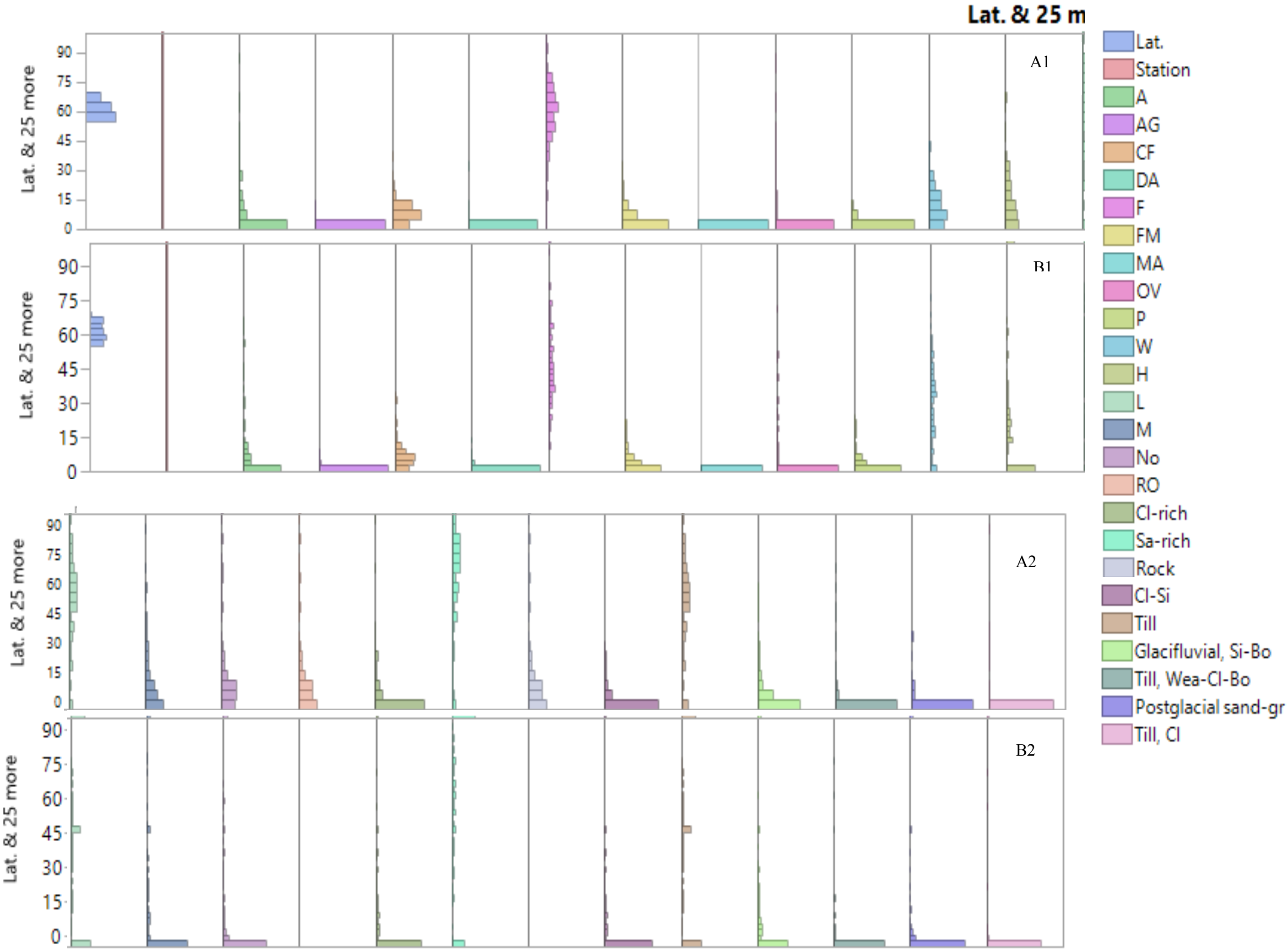
Appendices

Appendix 1.



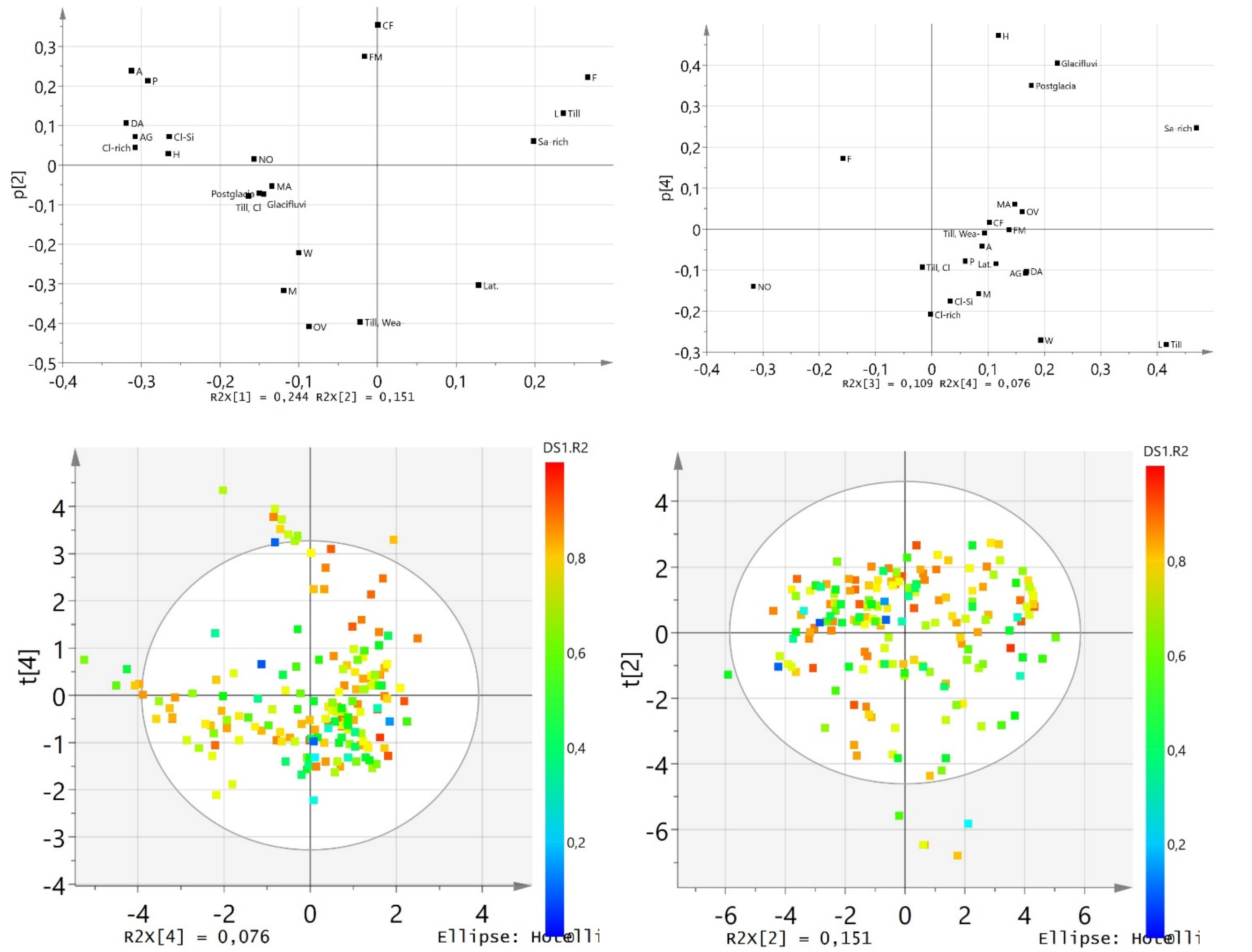
Land-use/cover in sweden. Adapted from Corine SMD (SEPA)

Appendix 2



Data distribution of the whole catchment (A1,2) and buffer zone 100 m (B1,2).

Appendix 3.



Top) PCA loading plots of the buffer zone data (left 1st and 2nd PC & right 3rd and 4th PC). Bottom) PCA score plots of the buffer zone data (left 1st and 2nd PC & right 3rd and 4th PC).