DRAFT

Deltares

Prediction of the suspended particle matter in the Wadden Sea using Machine Learning

Prediction of the suspended particle matter in the Wadden Sea using Machinel Learning

Client	Internal Project
Contact	
Reference	
Keywords	Wadden Sea; suspended particles; Machine learning; predictions

Document control

Version	0.1
Date	20-12-2019
Project nr.	-
Document ID	-
Pages	25
Status	draft This is a draft report, intended for discussion purposes only. No part of this report may be relied upon by either principals or third parties.

Author(s)

Carmen Martinez Barbosa	Willem Stolte	Peter Herman
Bob Smiths	Julia Vroom	Fine Wilms

Summary

Suspended matter concentration in water systems is of enormous importance for water management and ecology, as it influences mud deposition, dynamics of saltmarshes, development of intertidal areas and the functioning of the food web (primary production, grazing by benthic filter feeders). Suspended matter concentration is influenced by various factors, including dynamics of the water system (e.g. discharge, boundary concentrations, current and wave fields); meteorological effects (e.g. rainfall, wind, temperature); ecological processes (e.g. growth of algae and other organisms); among others. Several studies have investigated suspended matter transport in water systems; however, mechanistic modelling of the process, although much improved compared to earlier attempts, remains fraught with difficulties.
The aim of this project is to create a data-driven predictive model of the suspended matter concentration in the Wadden Sea considering the meteorological and environmental effects in this region, using state-of-the-art machine learning techniques. The goal of this study is to have a quantitative understanding of some of the key drivers that lead to the change of suspended matter concentrations in the Wadden Sea.

Contents

1 Introduction

The Wadden Sea is an intertidal zone in the southeastern part of the North Sea which has coasts in The Netherlands, Germany and Denmark. The physical processes that occur in the Wadden Sea are very complex; particularly the dynamics of Suspended Particulate Matter (spm). Herman et al. (2018) made a statistical analysis of the spm content under the effect of mud characteristics (i.e. size and fraction of sediment) and biomass distribution of microphytobenthos. They found that not only these two factors are important in the spm dynamics; spm variations also occur at different time scales. Short term variations of spm (e.g 0.5- 14 days) are produced due to exchange of material with the North Sea. The processes happening at short time scales are well known and they can be modelled via numerical models. However, spm variations at intermediate time scales (e.g. months, seasons) or at long time scales (years; decades) are very difficult to model numerically, not only for constraints in computation timing, also because the processes governing the spm variation at those time scales are difficult to infer analytically. Therefore, there is a big uncertainty in the hydrodynamical models on the interplay between the spm content and physical, chemical and biological processes. In this picture machine learning models are of great advantage, since they can model complex relationships without the need of having an analytical formulation of the system being studied.

We make use of state-of-the-art machine learning techniques to model the spm content of the Wadden Sea. We account for different factors such as meteorological conditions; chemical composition of water; biological processes and spatial-temporal features. To understand the intermediate and long-term evolution of the spm content, we use in-situ measurements that cover a period of almost 20 years. The predictive model of spm will lead to understand the key drivers that govern the spm dynamics in the Wadden Sea. This model will also bring a quantitative estimation of the change of spm due to the variation of each of its drivers.

2 Data

To understand some of the key drivers that explain the spm content in the Wadden sea, we search for different data sources that provide information on four different factors:

1. **Chemistry and water quality:** We search for data that describes the concentration of different chemical elements in the Wadden Sea surface water. As examples, we gather data on the concentration of Oxygen, Phosphorus, Nitrogen, Carbon. We also gather information on the salinity content.
2. **Weather:** We obtain information on meteorological conditions in the Wadden Sea. Lettmann et al. (2009) found that wind is an important factor that determines the spm content in the Wadden Sea. We also use the historical measurements of water levels to infer the surge in that region.
3. **Biology:** Variables such as phytoplankton biomass and chlorophyll-a concentration in the surface water are biological factors that we include in the analysis of the spm content.
4. **Space and time:** The spm data that we collected covers different regions in the Wadden Sea and it spans a time range of almost 20 years. With this vast amount of information, we build predictors in time such as months or seasons. We also use the location of the measurements to infer the tidal inlets where the data belong to.

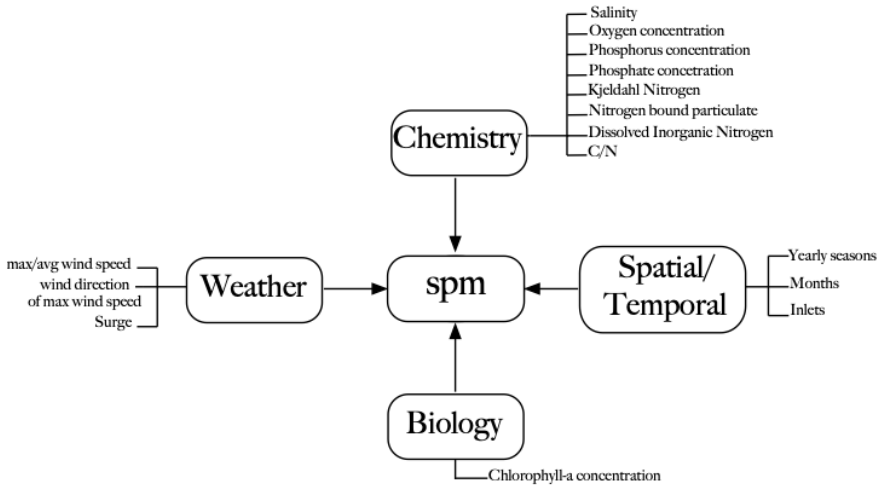


Figure 2.1: Features included in the study of the spm in the Wadden Sea

In Figure 2.1 we show some of the data that we collected to study the spm content in the Wadden Sea. Throughout this manuscript, we refer the spm to as *target*, since this is the feature that we want to predict. We refer the other features to as *predictors*, because they are the variables used to predict the target. In the following sections we explain in more detail the properties of the data gathered for this study.

2.1 spm

We obtain the measurements of spm content in the Wadden Sea by accessing to the Dutch archive of monitoring water data through the Python's code *ddlpy*: (https://github.com/openearth/ddlpy/tree/carmen).

The raw data cover measurements carried out in 88 stations during a variable period of time. The stations with best quality data have measurements every 2 weeks; other stations have scarce measurements, for one year only. During the data preprocessing, we disregard those last stations together with all the measurements taken in years before 1990. We also resample each timeseries to be daily and we apply non-linear interpolation in the missing data. We compare the preprocessed timeseries with those shown in Hermann et al. (2018). The timeseries are statistically the same, with updated measurements from 2016 onwards. After preprocessing, the number of stations with clean data is 24. In figure 2.2 we show their locations.

Figure 2.2: Stations where spm and the water level are measured.

2.2 Chemistry and water quality data

Through the *ddlpy* code we gather all the information on the Wadden Sea that is provided by the Dutch archive of monitoring water data. This archive contains data on sediment size; sediment diameter; concentration of different chemical elements in the surface water and other biological parameters. However, many of these features have either few measurements or they are not measured in the stations where we have spm data. Therefore, we made an analysis of completeness (see appendix A) and we selected the features that were measured in the spm stations which have a threshold of missing values smaller than 50%. The features that fulfill these requirements are the following:

- salinity_surface_water
- transparency_surface_water [dm]
- carbon_mass_concentration_surface_water (DOC, mg/l)
- carbon_bound_concentration_surface_water (POC, mg/l)
- total_Kjeldahl_nitrogen_concentration_surface_water [mg/l]
- total_nitrogen_concentration_surface_water [mg/l]
- ammonium_concentration_surface_water [mg/l]
- nitrite_concentration_surface_water [mg/l]
- nitrate_concentration_surface_water [mg/l]
- total_nitrogen_concentration_surface_water [mg/l]
- nitrate_and_nitrite_dissolved_fraction_surface_water [mg/l]
- total_nitrogen_bound_particulate_ surface_water [mg/l]

- carbon_concentration_surface_water [mg/l]
- oxygen_concentration_surface_water [mg/l]
- total_phosphorus_concentration_surface_water [mg/l]
- total_phosphorus_bound_concentration_surface_water [mg/l]
- phosphate_concentration_surface_water [mg/l]
- total_phosphorus_dissolved_fraction_surface_water [mg/l]
- chlorophyll-a_concentration_surface_water [μg/l]

These features have measurements carried out in different frequencies over time. During data preprocessing, we resample the time series daily, using linear interpolation to impute missing values. We also remove measurements taken before 1990 and remove outliers using domain expertise. We transform all the concentrations of chemical elements to [g/l] and the chlorophyll-a concentration is converted to mg/l.
From the features listed above, we created additional predictors such as:

- Dissolved Inorganic Nitrogen (DIN): NO_{2-} +NO_{3-} +NH_4
- POC fraction: POC/(POC+DOC)
- DOC fraction: DOC/(POC+DOC)
- C/N: (POC+DOC)/N

The POC fraction and DOC fraction are used for analyzing the data, but these features will not be used to build the data-driven model of spm. We also remove the concentrations of nitrite, nitrate and ammonium, since we account for them in the dissolved inorganic nitrogen.

2.3 Weather data

Figure 2.3: Stations where wind speed and wind direction are measured. For comparison, the spm stations are also plotted. The black lines correspond to the Voronoi of each wind station. The spm stations located within a Voronoi will get the value of wind speed and wind direction corresponding to the associated wind station.

2.3.1 Integrated Wind speed and wind direction

Meteorological data of the Wadden Sea can be found through the KNMI web portal[1]. This portal contains daily measurements of different meteorological parameters that span a time-range from 1990 to May-2019. We select the timeseries of wind speed and wind direction, as these parameters can influence the spm content in the Wadden Sea. In Figure 2.3. we show the locations where the wind speed and direction are measured. For each of those stations, we create a Voronoi which is used to join spatially the spm data with the wind parameters. During the merging of the data, we assume that all the spm stations within a Voronoi have the value of wind speed and wind direction of the associated meteorological station.

Once the data is merged, we have measurements of wind speed and wind direction per time and spm location. With this information, we create different aggregations of wind speed over time; e.g. maximum/average wind speed 1,2,…,9 days ago. We then compute which of these features were correlated the most with spm (with 99% confidence). We found that the *maximum wind speed 8 days ago* and *the average wind speed 8 days ago* were the features most correlated with the spm[2].

To obtain the wind direction, we select the wind direction corresponding to the most correlated maximum wind speed. Since this feature is cyclical, we decompose the angle to their x and y components, being x= Cos(θ) and y= Sin(θ).

2.3.2 Storm surge

To obtain the storm surge, we used the measurements of water level data provided by the Dutch archive. The raw measurements have non-uniform frequencies. Some timeseries contain a mixture of daily, hourly, and ten-minutely measurements. We Pre-processed the water level data by removing any duplicate time-stamp measurements. We keep only the first measurement. The data was then resampled to 10-minutely intervals. We use linear interpolation in cases where the measurement resolution was coarser than 10 minutes. In cases where an entire day of data was missing, we infilled those space with Not-a-Number (NaN) values. We keep the data of those stations that have less than 1000 missing values. In Figure 2.2 we show the location of the water level stations that satisfy this condition.

The timeseries of postprocessed water level data were decomposed to obtain the storm surge using the t_tide toolbox (Pawlowicz et al, 2002). We use krige spatial interpolation at every timestamp to obtain the storm surge in the location of the spm stations. Note that by using this approach, we assume that the variation of surge in the cross-shore direction is of minor importance.

2.4 Biological data

From the Dutch archive of monitoring water data, we obtained the chlorophyll-a concentration in surface water for some of the stations for which spm data is available. We preprocessed the data in the same fashion as explained in Section 2.2. We first remove the measurements prior 1990. We then resample each timeseries daily, using linear interpolation to impute missing values. The outliers were removed using expert knowledge.
We also obtained data of phytoplankton biomass from the Dutch archive. However, these data are scarce in space and time; therefore, we decided not to add this feature in the modelling.

[1] http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi.

[2] We used the Spearman correlation, which accounts for non-linearities in the data.

2.5 Variables describing time and space

We create temporal variables such as *month* and *season of the year*. We also used the description of tidal basins in the Wadden Sea (Elias et al. 2012) to add a spatial feature in the model. We assume that change in shape of the tidal basins in the Wadden Sea has not been significant during the period 1990-2019.

2.6 Data enrichment

Once all the predictors are preprocessed, we gather the clean datasets into a single table. During the merging, we remove both features and observations that have a percentage of missing values higher than 50%. The resulting table contains 79,547 observations that contain spm information corresponding to 14 stations in the Wadden Sea, and 24 features that describe the chemical composition of the water; weather conditions; biological indicators and spatial-temporal variations. In table 2.1 we show a list of these predictors, which will be used to build a data-driven model of the spm content of the Wadden Sea.

Chemistry and Water quality	Total Salinity surface water [dimensionless] DOC [g/l] TKN [g/l] Oxygen concentration surface[g/l] water [g/l] Phosphorus concentration surface water [g/l] Phosphate concentration surface water [g/l] DIN [g/l] TOC/N [dimensionless]	**Weather**	avg wind speed 8 days ago [km/h] max wind speed 8 days ago [km/h] x wind dir 8 days ago [rad] y wind dir 8 days ago [rad] surge [m]
Spatial/temporal	month autumn spring summer winter ameland_inlet eems-dollard_inlet frisian_inlet texel_inlet vlie_inlet	**Biology**	chlorophyll-a concentration surface water [mg/l]

Table 2.1. Predictors and target in the final dataset.

3 Analysis

An important step in building a data-driven model is to validate the post-processing made on the data. This can be carried out by comparing trends observed in the data with those published in the literature. In this report we focus the analysis on: 1. The distribution of spm across stations; the content of Dissolved Organic Carbon (DOC) and Particulate Organic Carbon (POC). 2. Correlations of spm content with some of the model predictors and 3. The observed organic matter partitioning in the Wadden Sea.
In the top-left panel of Figure 3.1 we observe the distribution of spm content in the stations of the Wadden Sea. As expected, regions with high turbidity have higher spm content on average. An example is the station BOCHTVWTM, located in the Eems-Dollar inlet (Middelburg & Herman (2007)).

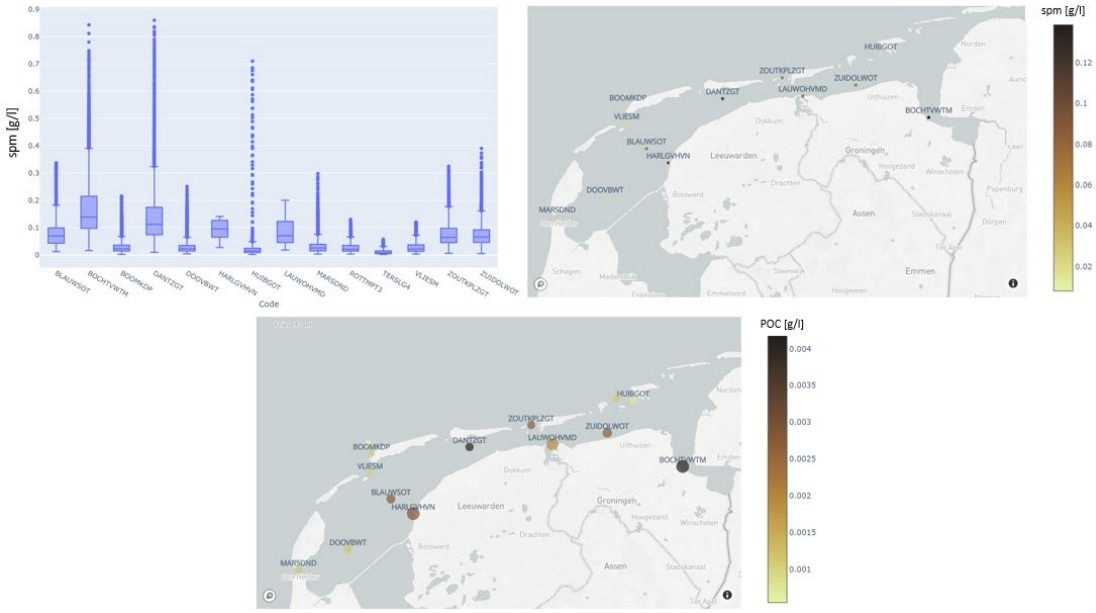


Figure 3.1. Top: Box plots of the distribution of spm in different locations of the Wadden Sea (left). Location of the spm stations in the post-processed dataset. The color represents the average value of spm. **Bottom:** Location of the spm stations. The color corresponds to the POC measured in that station. The size of the points corresponds to the measurements of DOC.

The POC content is also higher in stations where the spm is higher on average. On the other hand, we note that the stations that are close to the continent, have the largest DOC contents (e.g. BOCHTVWTM; HARLGGVHVN; LAUWOHVMD. See bottom panel of figure 3.1).
We also look at the correlations of spm with some of the model predictors. In the left panel of figure 3.2 we show the correlation plot of spm with the total salinity of surface water and with the phosphorus concentration. We can see that spm is weakly correlated with the water's salinity (Pearson's correlation= -0.54). This weak correlation is found across all the stations of the Wadden Sea. Middelburg & Herman (2007) arrive to the same conclusion by analyzing the data of other tidal estuaries along the western coast of Europe.
Contrary to the salinity, we found that the spm is highly correlated with the phosphorus concentration. This strong and positive correlation is found across all the stations in the Wadden Sea. This might be an indicate of a relation between synthesis of amino-acids in phytoplankton and spm content. We will discuss this topic in more detail in Section 4.1. Salinity

and phosphorus content are just two of all the features that we have gathered. In Appendix B we show a more complete plot of correlations of spm with other model predictors.

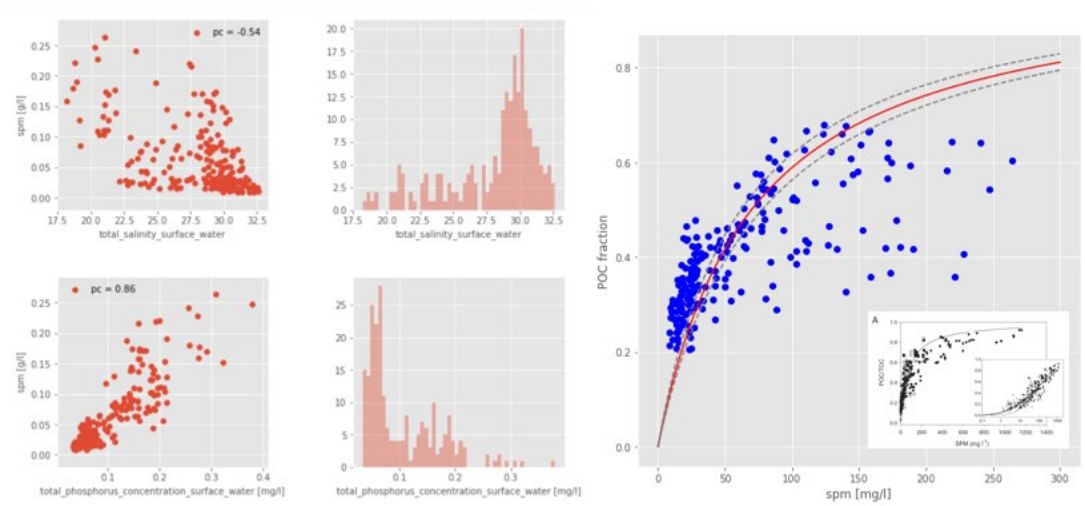


Figure 3.2. Left: Some correlation plots of spm with 2 model features. The quantity pc stands for the Pearson correlation. **Right:** Organic matter partitioning of the Wadden Sea. The inlet shows the same for other water bodies along the European coast. See Middelburg & Herman (2007) for more details.

Another validation that we performed in the post-processed dataset was to infer the organic matter partitioning. The organic matter partitioning is the relation between the spm content in a water body with the POC fraction. Middelburg & Herman (2007) found that the POC fraction depends on the spm content in the following way:

$$POC\_f = \frac{spm}{spm + K} \ (Eq. 1).$$

Where K is a constant that describes exchange processes between adsorption and desorption. They found that $K = 72 \pm 2.6$. By using the post-processed data described in Section 2.3, we obtain a value of $K = 70 \pm 7.8$. In the right panel of figure 3.2 we show the fitting of the Eq. 1 to the data, with the red lines being the 95% confidence intervals. In the inlet, we see the fitting that Middelburg & Herman (2007) made of Eq. 1 to their data.
The analysis and validation carried out in the cleaned data gives us confidence on the preprocessing steps that were carried out as well as in the model that we develop.

4 Machine learning model

Before using a machine learning model to predict the spm content in the Wadden Sea, it is necessary to split the data into train, validation and test sets. The train dataset is used to infer the relationship between spm and the predictors. The validation dataset is used to evaluate the prediction of spm during training, to optimize the model hyperparameters. The test set is used to evaluate the model and give unbiased and undeceiving error metrics[3]. The train and test datasets correspond to 80% and 20% of the original dataset respectively. The validation set corresponds to 20% of the train set. Therefore, the number of points in the datasets is the following:

Train set: 50,909 observations.
Validation set: 12,728 observations.
Test set: 15,910 observations.

Given its high predictability power, we use the eXtreme Gradient Boosting method (XGBoost, Chen, T. (2016)) to build a model that predicts spm in terms of weather, water-quality, biological and spatial-temporal predictors. XGBoost uses the so-called decision trees to make predictions on the spm content in the Wadden Sea. A decision tree consists of a set of conditions that are obtained by splitting the space defined by each of the predictors. If spm is highly correlated[4] with a certain predictor X, then the decision tree will select this feature several times to create the prediction rules. If spm is not correlated with a feature X1, this feature will not be selected by the decision tree. The branches of a decision tree thus, represent a set of conditions given by the predictors for which the spm is highly correlated. Every point in the dataset belongs to one of the tree's branches; therefore, the average spm of all the points that belong to a branch in the tree will correspond to the predicted spm for that branch. In the left panel of figure 4.1 we show a decision tree.
A single decision tree is not enough to make good predictions. One way to increase the prediction accuracy of a decision tree is by creating other decision trees for several subsamples of the original train dataset. The final prediction of spm would be given by the average of all the results in the decision trees. However, XGBoost goes a step further because it uses the decision trees of the subsamples, to correct the prediction made in the first decision tree. In the right panel of figure 4.1 we show how the XGBoost algorithm works.

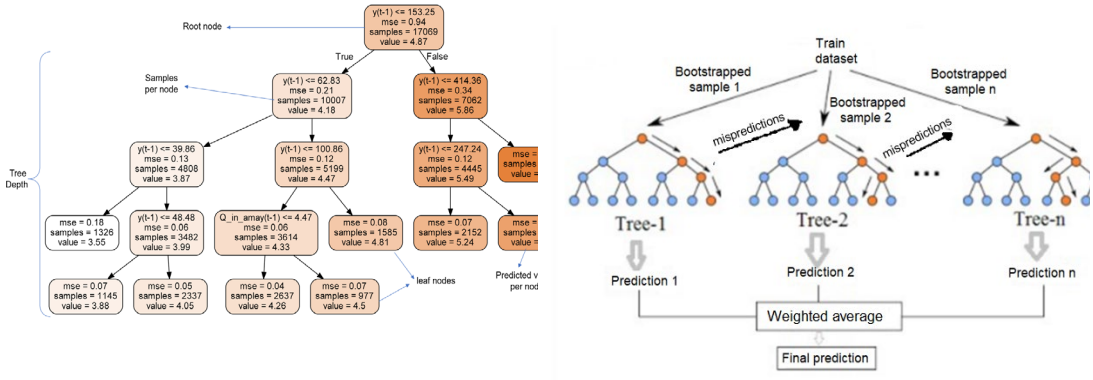


Figure 4.1. Left: Structure of a decision tree. The leaf nodes have the average value of spm of the datapoints that belong to that specific branch. **Right:** How a boosting algorithm works.

[3] It is incorrect to use the train set to give error metrics, this is a bad practice that should be avoided.

[4] This correlation does not have to be linear.

The features that are used by the XGBoost algorithm are those listed in table 2.1. Note that this algorithm does not give an analytical expression of the spm content in terms of the input variables, as a linear regression method would do; instead, the model will define a set of conditions over the predictors that will lead to the prediction of the spm content in the Wadden Sea.
We evaluate the prediction performance of the spm model by measuring the Median Absolute percentage error (MdAPE); the Root Mean Squared Error (RMSE) and the Explained variance (R^2). The error metrics are the following:
MdAPE: 14.2%
RMSE: 0.013 g/l
R^2: 0.96

In order to evaluate how good or bad is the model of spm, we compare the error metrics of above with a baseline model that always predicts on the test dataset the median spm value that is observed in the training data. The baseline model gives the following error metrics: MdAPE: 66%; RMSE: 0.07 g/l; R^2: 0.1. We can observe that the model of spm is much better than the baseline; however, a better comparison should be carried out by using a pure hydrodynamical model. This is out of the scope of this report and it is suggested to do in a future work.

In this model of spm, we do not use any mathematical transformation (e.g. logarithmic) on the target. Consequently, the model is prone to have worst predictions of small spm compared to those of large values (See Figure 4.2, Top panel). However, we see in Figure 4.2 that on average, the model can make good predictions of the spm content of the Wadden Sea.

We also explore the prediction capacity of the model on different tidal inlets and seasons (See bottom panel in Figure 4.2). We found that the highest error is for Ameland inlet, with RMSE= 0.02g/l. This is expected, since there is only one station located in this tidal inlet. The Texel and Vlie inlets have the lowest error metrics, with RMSE= 0.008g/l and RMSE= 0.007 g/l respectively. In the temporal domain, the change in the error metrics over seasons do not change significatively. The lower error is obtained in summer, where RMSE= 0.010g/l. The higher error value is obtained in autumn, where the RMSE= 0.014g/l.
One of the biggest advantages of using XGBoost is the ability to obtain the feature importance of the predictors. The feature importance is a score that indicates how useful or valuable each feature is in the construction of the decision trees within the model. Therefore, the more an attribute is used to make decisions in the trees, the higher is its relative importance. We used the Shap package (Lundberg, S. 2017) to infer the feature importance of the predictors. We show the results in Figure 4.3. According to the feature importance, we found that weather contributes with only 5% of the prediction of the spm content in the Wadden Sea. Biology, only with 4%; space and time with almost 20% and chemistry and water quality with 71%. Note that these percentages are highly dependent of the features used; which in turn depends on the availability of the data. However, these numbers give an estimate of the drivers that play an important role in the change of spm content in the Wadden Sea.

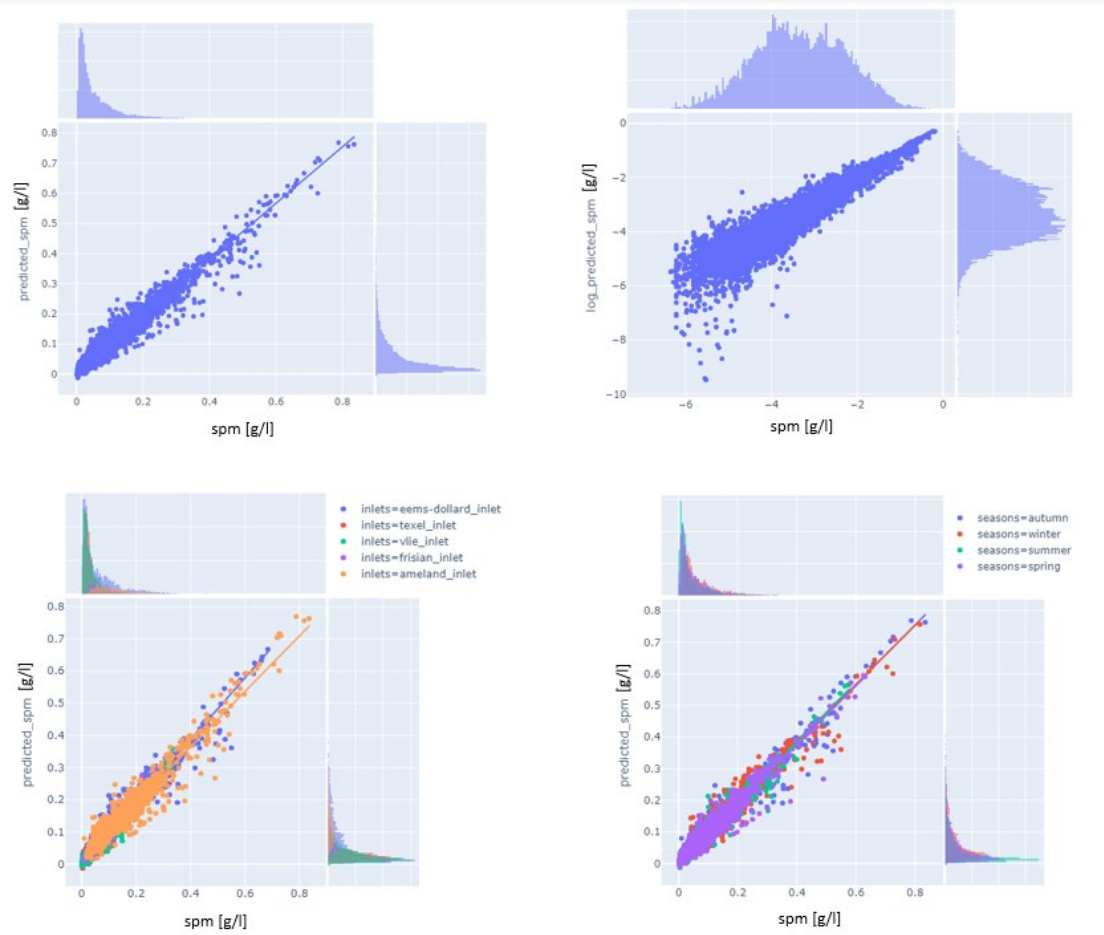


Figure 4.2. Top left: Predicted values of spm content Vs. real observation of the test set. **Top right:** Residuals plot in logarithmic scale. **Bottom left:** Residuals plot in terms of the tidal inlets. **Bottom right:** Residuals plot in terms of the seasons.

The feature importance only gives information on the features that are important in the determination of spm. In order to infer how spm change with a variation of these key drivers, we need to make a sensitivity analysis of the model or, in other terms, make use of partial dependency plots.

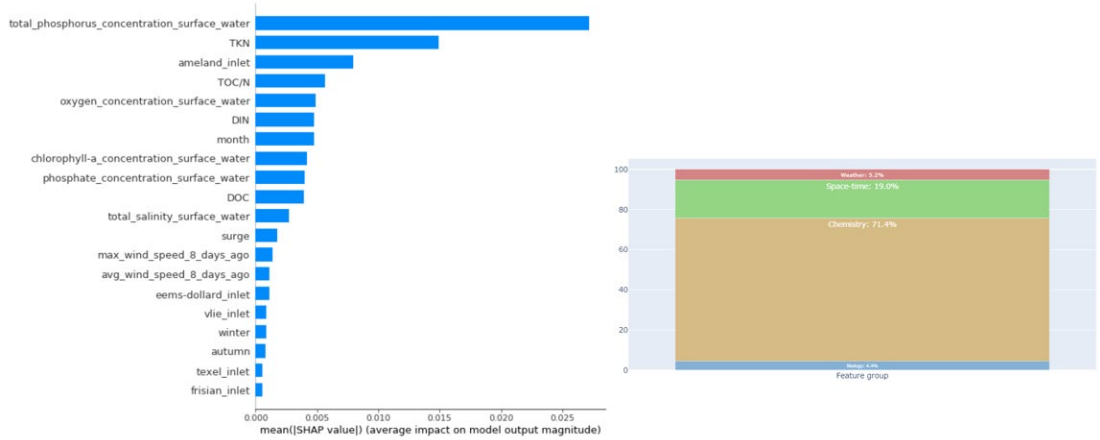


Figure 4.3: Shap values of the predictors of the model.

4.1 Partial dependency plots

The partial dependency plots (PDP) measure the change of spm over the change of a certain predictor, maintaining the others constant. This procedure is carried out for each of the points in the train dataset. Consequently, the PDP show hidden relationships that are unveiled by the machine learning model. This method also provides a way to understand the 'black box' machine learning algorithm that is being used.
In this section we present the PDP of the predictors that are important for the model; that is, those shown in the left panel of Figure 4.3.

4.1.1.1 Variation of spm under Chemical / Biological components

In Figure 4.1.1 we show the PDP of the features related to chemical and biological components. In the plots the dark line shows the change of spm under the effect of a feature, averaged over all the variations of spm measured in the train dataset. The shaded region thus, represents the 95% confidence interval of the change in spm. Note that the plots start at Δspm= 0 g/l. This means that the change in spm (for a single observation) is measured with respect to its initial value.

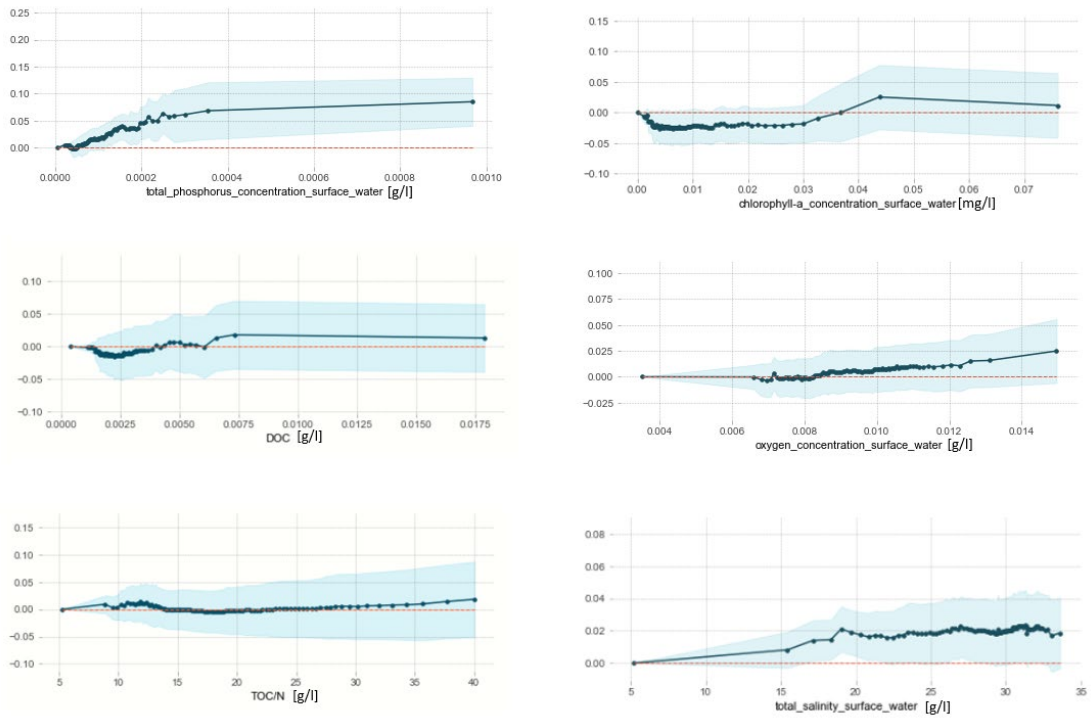


Figure 4.1.1: Change of spm in terms of different chemical/biological components.

According to Figure 4.1.1, spm has a positive linear variation with the change in the total concentration of phosphorus. When the phosphorus concentration is 0.2 mg/l, the spm varies ~0.05g/l. At higher concentrations of phosphorus, the variation rate of spm decreases, being of ~0.03 g/l. Phosphorus has a strong seasonal cycle. It reaches its minimum value around May, June or July and then starts increasing again during autumn and winter. This seasonal effect varies per tidal inlet. In Ameland, for instance, the increment of the phosphorus concentration occurs earlier in the year. The variation of the phosphorus concentration per tidal inlet depends on the ratio between tidal flats and channels. Shallow tidal inlets have different

ratio of sediment surface to water volume which means that the impact of phosphorus is different. However, the plot in figure 4.1.1 suggests that there is a relationship of spm and phosphorus that is the same across stations. This relationship can have different explanations: 1. The depth of the spm stations is small on average: this leads to both higher concentrations of phosphorus and spm. We do not have the data to test this; however, in case this hypothesis is true, the relationship shown in figure 4.1.1 is not of interest. 2. The algae highly influence the spm: during the season cycle, once the phosphorus content is low (at the end of phytoplankton bloom) the phytoplankton starts excreting mucus because it cannot maintain sugar in its cells. This means that these organisms start synthetizing amino-acids, using the phosphorus content in the water. By doing this, the phytoplankton becomes sticky, attracting particles, leading to a decrement in the spm content.
On the other hand, we see that the spm decreases with an increment of the chlorophyll-a concentration. According to Figure 4.1.1, the spm increases by 0.025 g/l when the chlorophyll-a concentration is 4x10-3mg/l. Contrary to phosphorus, the highest concentration of chlorophyll-a is in spring, around April or May. If algae grow, they are healthy and do not influence the spm content. When they are dying, they consume phosphorus and become sticky, producing a decrement in the spm content.
Oxygen concentration is an indicator of temperature and a seasonal variator. In summer, when the concentration of O_2 is the lowest (~0.006 g/l), there is no variation of spm. When the oxygen concentration gets its maximum value in winter (~0.014 g/l), the change of spm is on average ~0.025 g/l.

On the other hand, Note the variation of spm in terms of the salinity of the surface water. In fresh waters (salinity < 15), the spm can vary up to ~0.01g/l. In salty water, on the contrary, the spm content remains almost constant. This is in line with Middelburg & Herman (2007), who found that the carbon content (highly related to spm) in turbid estuaries is very uniform and low, while in river-dominated estuaries, the amount of organic matter is highly variable.

4.1.1.2 Variation of spm under weather conditions

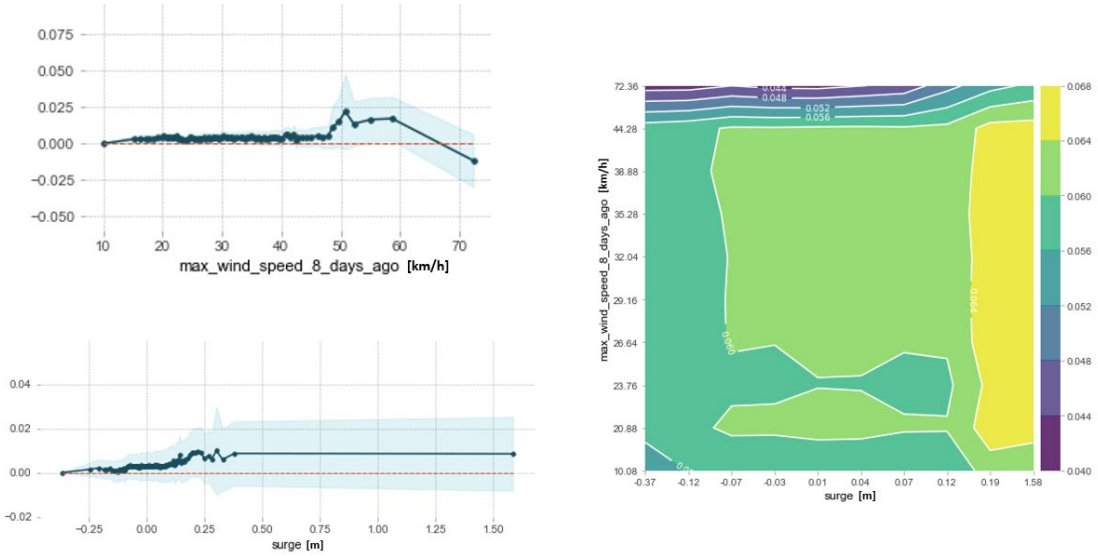


Figure 4.1.2. Left: Change of spm with maximum wind speed and surge only. **Right:** spm variation under the combined effect of surge and maximum wind speed.

In figure 4.1.2 we show the PDP of the maximum wind speed over the past 8 days and surge and the combined effect of these two features in the variation of spm. We can see that the surge has a stronger effect in the variation of spm than the max wind speed. The spm content is reduced up to ~0.04g/l when the wind has a speed higher than 50 km/h and the surge is negative (i.e. from continent to open water). These low values of spm might be related to the winter of 1996-1997, where there was ice in the water. The wind speed, even being strong, does not produce turbulence and the spm content is rather low under these conditions. If the surge reaches values of ~1.5m, the spm content in the Wadden Sea is incremented approximately by 0.06g/l.

We can see that the influence of weather conditions- particularly the surge- on the variation of spm is not as important as the chemical components. One of the reasons is the time granularity of the raw spm data. The original measurements are taken every 2 weeks, in the same occurrence period of the tidal cycle. Therefore, this effect, which is very important, cannot be observed in the model.

4.1.1.3 Variation of spm under spatial and temporal features

In Figure 4.1.3 we show the variation of spm in different months (left panel) and the values of spm in different tidal inlets and seasons. We observe that the spm content decreases during spring and summer and increases in autumn and winter. This result is in line with the observed variation of spm in terms of the oxygen content of surface water, where spm increases when the amount of oxygen is highest (which occurs in winter-autumn). These results were also obtained by Peter et al. 2018.

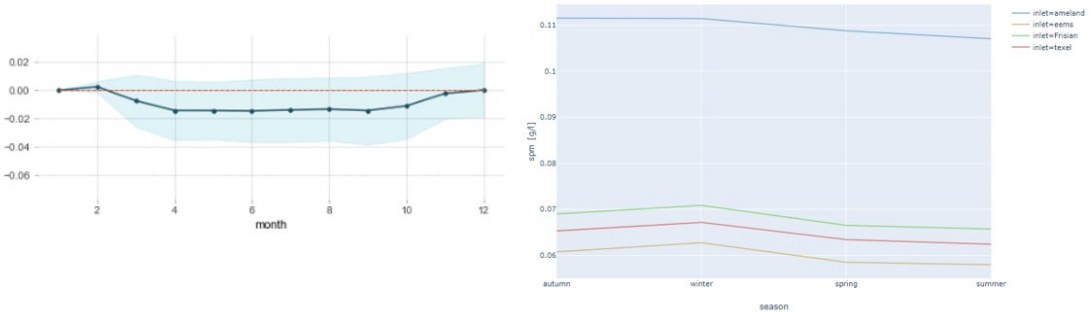


Figure 4.1.3. Left: Variation of spm over months. **Right:** spm values in terms of the tidal inlet and the season of the year. This value is averaged over all the points in the train data set.

Regarding to the change of spm due to spatial features, we observe that the spm content in the Ameland inlet is always higher on average. The reason of this effect is the fact that there are spm observations of one station only.

4.2 Capabilities and limitations of the model

The model of the spm content that we develop is purely data-driven (or empirical). This of course, has some advantages and disadvantages. First, the data-driven model can account for many physical processes that are difficult to infer analytically. Therefore, with the data-driven model we can discover many hidden relationships that govern the physics of the system under study. This, of course, depends highly on the data availability in time and space. Another advantage over a pure hydrodynamical model is the time required to obtain predictions of the spm content in the Wadden Sea. Once the machine learning model is trained, the predictions

on unseen data can be obtained instantaneously, without need of high computational power. However, it is important to note that the data-driven model described in this report is not capable of making forecasts of the spm in the future. To make a forecast, we need to account for the effect of the previous observations of the spm. In this respect, the hydrodynamical approach is a better alternative.

Despite the limitations of the data-driven model, this approach has promising applications in the areas of water management and ecology of the Wadden Sea, since it brings an instant evaluation of the spm content in different regions of this system, in terms of the chemical composition of the water, biological factors and space-temporal features. The model is flexible, since other predictors that account for the effect of sediment size, fresh water and micro-organisms can be added. Several studies within Deltares have used remote sensing to gather information on the distribution of benthic organisms and sediment in the Wadden Sea. The addition of these data would bring more insights to a deeper understanding of this complex system together with an increment of the prediction performance of the current machine learning model. A comparison between this model with a pure hydrodynamical approach is still needed; however, this study is a first attempt at building a model that can bring a more complete picture of all the factors that might influence on the dynamics of the spm content in the Wadden Sea.

5 Take home messages and future steps

We have built a predictive model of the spm content in the Wadden Sea accounting for different factors such as chemical composition of water; weather conditions, biological indicators and spatial-temporal features. This statistical model has many advantages:

- We can obtain an instantaneous value of spm in a region of the Wadden Sea.
- The model accounts for phenomena that are both hard to describe analytically and are measured during intermediate time scales. This possesses a huge advantage with respect to previous statistical analysis, as we are now able to have a quantitative description of the spm variation in the Wadden Sea in terms of these features. Therefore, the model developed in this research is a complement of the work of Herman et al. (2018), bringing more insights towards a conceptual model of mud dynamics in the Wadden Sea.

However, there are still some pieces of the puzzle that needs to be put together. In situ observations of phytoplankton and Microphytobenthos biomass are very scarce in space and time. Remote sensing observations of these biological features, found in Sentinel images, for instance, can be very beneficial in the increment of the prediction performance of the model. Additionally, in situ data of sediment characteristics (e.g. SIBES), covers all the Wadden Sea area, but it spans a time range of only 5 years. Other sources such as the Dutch archive does not offer enough data on this feature; specially in the Wadden Sea. Earth observations might also be helpful in this respect. Additionally, more frequent SPM measurements would help to gain insights at shorter timescales, at which we expect meteorological characteristics to play a bigger role.

In addition to add remote sensing data to the developed spm model, one important step is to start using this data-driven model in current projects at Deltares. We will also work towards a product that can be used by external clients for water management and ecology purposes.

6 References

Chen, T; Guestrin, C. *XGBoost: A Scalable Tree Boosting System.* 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p785-794.

Elias, E.P.L., van der Spek, A.J.F., Wang, Z.B., de Ronde, J. *Morphodynamic development and sediment budget of the Dutch Wadden Sea over the last century.* 2012. Netherlands Journal of Geosciences, 91-3, p293-310.

Hermann, P.M.J et al. *Mud dynamics in the Wadden Sea.* 2018. Deltares report.

Middelburg, J.J and Herman, P.M.J. *Organic matter processing in tidal estuaries.*2007.Marine chemistry, 106, 127-147.

Lettmann, K; Wolff, J.O; Badewien, T. *Modelling the impact of wind and waves on suspended particulate matter fluxes in the East Frisian Wadden Sea (Southern North Sea).* 2009. Ocean Dynamics, 59(2):239-262.

Lundberg, Scott M and Lee, Su-In. *A Unified Approach to Interpreting Model Predictions.* 2017. Advances in
Neural Information Processing Systems 30, p4765-4774.
Pawlowicz, R., Beardsley, B. and Lentz, S. *Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE.* 2002. Computers and Geosciences 28, p929-937.

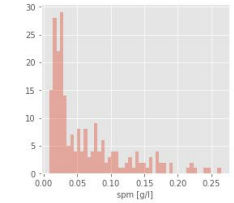
7 Appendix A

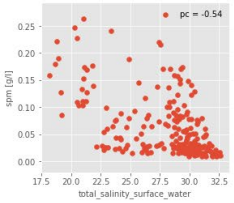
Missingness of the data provided by the Dutch archive of monitoring water data in the spm stations.

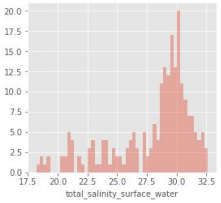
parameter	stations no missing	% stations no missing	stations missing	% stations missing
sediment_size_ diameter_smaller_than_2um	3	0.13	21	0.88
sediment_size_diameter_smaller_than_16um	3	0.13	21	0.88
sediment_size_diameter_smaller_than_20um	0	0.00	24	1.00
sediment_size_diameter_smaller_than_63um	4	0.17	20	0.83
sediment_size_diameter_bigger_than_63um	0	0.00	24	1.00
sediment_size_diameter_smaller_than_2000um	0	0.00	24	1.00
sediment_size_diameter_smaller_than_53um	0	0.00	24	1.00
sediment_size_diameter_between_16_and_2000um	2	0.08	22	0.92
salinity_surface_water	23	0.96	1	0.04
transparency_surface_water	17	0.71	7	0.29
carbon_mass_concentration_surface_water	17	0.71	7	0.29
carbon_bound_concentration_surface_water	20	0.83	4	0.17
nitrogen_concentration_surface_water	16	0.67	8	0.33
total_nitrogen_concentration_surface_water	19	0.79	5	0.21
nitrate_and_nitrite_concentration_surface_water	0	0.00	24	1.00
ammonium_concentration_surface_water	19	0.79	5	0.21
nitrite_concentration_surface_water	19	0.79	5	0.21
nitrate_concentration_surface_water	19	0.79	5	0.21
total_nitrogen_concentration_surface_water	17	0.71	7	0.29
nitrate_and_nitrite_dissolved_fraction_surface_water	19	0.79	5	0.21
total_nitrogen_bound_particulate_ surface_water	17	0.71	7	0.29
carbon_concentration_surface_water	14	0.58	10	0.42
glow_residue_concentration_surface_water	15	0.63	9	0.38
oxygen_concentration_surface_water	24	1.00	0	0.00
spm	24	1.00	0	0.00
total_phosphorus_concentration_surface_water	19	0.79	5	0.21
total_phosphorus_bound_concentration_surface_water	14	0.58	10	0.42
phosphate_concentration_surface_water	19	0.79	5	0.21
total_phosphorus_dissolved_fraction_surface_water	16	0.67	8	0.33
silica_concentration_surface_water	15	0.63	9	0.38
chlorophyll-a_concentration_surface_water	19	0.79	5	0.21
suspended_matter_diameter_smaller_than_2um	2	0.08	22	0.92
suspended_matter_diameter_smaller_than_16um	2	0.08	22	0.92
suspended_matter_diameter_smaller_than_20um	2	0.08	22	0.92
suspended_matter_diameter_greater_than_63um	0	0.00	24	1.00
wind_direction_relative_to_true_North	0	0.00	24	1.00
wind_direction_expected_to_true_North	0	0.00	24	1.00
air_pressure	0	0.00	24	1.00
air_column	0	0.00	24	1.00
wind_speed	0	0.00	24	1.00
wind_speed_expected	0	0.00	24	1.00
gust_of_wind	0	0.00	24	1.00
temperature_air	0	0.00	24	1.00
temperature_surface_water	0	0.00	24	1.00

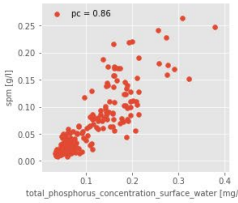
8 Appendix B

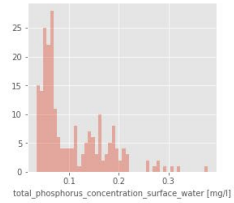
Correlations of spm with some of the model predictors.

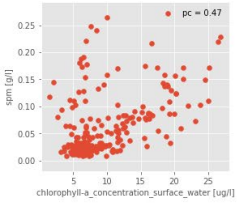


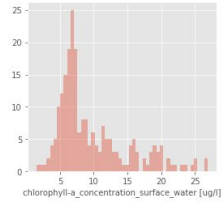


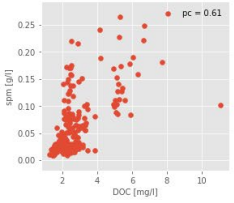


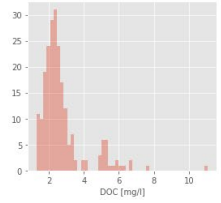


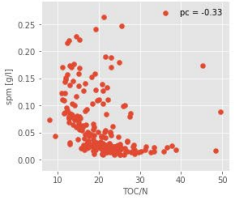


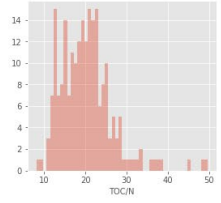


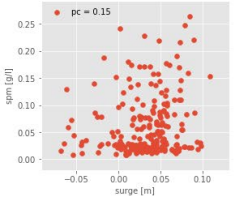


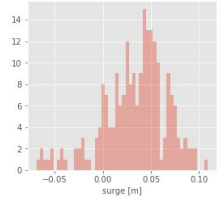












Deltares is an independent institute for applied research in the field of water and subsurface. Throughout the world, we work on smart solutions for people, environment and society.

Deltares

www.deltares.nl